Empirical Machine Translation and its Evaluation

Tesi Doctoral
per a optar al grau de
Doctor en Informàtica

per
Jesús Ángel Giménez Linares

sota la direcció del doctor
Lluís Màrquez Villodre

Programa de Doctorat en Intel·ligència Artificial
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

Barcelona, Maig de 2008

A la vida,
un gran mètode
d'aprenentatge automàtic.

Abstract

In this thesis we have exploited current Natural Language Processing technology for Empirical Machine Translation and its Evaluation.

On the one side, we have studied the problem of automatic MT evaluation. We have analyzed the main deficiencies of current evaluation methods, which arise, in our opinion, from the shallow quality principles upon which they are based. Instead of relying on the lexical dimension alone, we suggest a novel path towards heterogeneous evaluations. Our approach is based on the design of a rich set of automatic metrics devoted to capture a wide variety of translation quality aspects at different linguistic levels (lexical, syntactic and semantic). Linguistic metrics have been evaluated over different scenarios. The most notable finding is that metrics based on deeper linguistic information (syntactic/semantic) are able to produce more reliable system rankings than metrics which limit their scope to the lexical dimension, specially when the systems under evaluation are different in nature. However, at the sentence level, some of these metrics suffer a significant decrease, which is mainly attributable to parsing errors. In order to improve sentence-level evaluation, apart from backing off to lexical similarity in the absence of parsing, we have also studied the possibility of combining the scores conferred by metrics at different linguistic levels into a single measure of quality. Two valid non-parametric strategies for metric combination have been presented. These offer the important advantage of not having to adjust the relative contribution of each metric to the overall score. As a complementary issue, we show how to use the heterogeneous set of metrics to obtain automatic and detailed linguistic error analysis reports.

On the other side, we have studied the problem of lexical selection in Statistical Machine Translation. For that purpose, we have constructed a Spanish-to-English baseline phrase-based Statistical Machine Translation system and iterated across its development cycle, analyzing how to ameliorate its performance through the incorporation of linguistic knowledge. First, we have extended the system by combining shallow-syntactic translation models based on linguistic data views. A significant improvement is reported. This system is further enhanced using dedicated discriminative phrase translation models. These models allow for a better representation of the translation context in which phrases occur, effectively yielding an improved lexical choice. However, based on the proposed heterogeneous evaluation methods and manual evaluations conducted, we have found that improvements in lexical selection do not necessarily imply an improved overall syntactic or semantic structure. The incorporation of dedicated predictions into the statistical framework requires, therefore, further study.

As a side question, we have studied one of the main criticisms against empirical MT systems, i.e., their strong domain dependence, and how its negative effects may be mitigated by properly combining outer knowledge sources when porting a system into a new domain. We have successfully ported an English-to-Spanish phrase-based Statistical Machine Translation system trained on the political domain to the domain of dictionary definitions.

The two parts of this thesis are tightly connected, since the hands-on development of an actual MT system has allowed us to experience in first person the role of the evaluation methodology in the development cycle of MT systems.

Jesús Ángel Giménez Linares

TALP Research Center – Grup de Processament del Llenguatge Natural
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya
C/Jordi Girona Salgado, 1-3, E-08034 Barcelona, Spain.

e-mail: jgimenez@lsi.upc.edu
URL: http://www.lsi.upc.edu/~jgimenez

Agradecimientos

Entre los recuerdos más diáfanos de mi infancia hay uno que me gustaría compartir con vosotros. Era un día cualquiera en casa de mi tía Rosario. Mi primo David tecleaba en su Spectrum unas cuantas líneas de código copiadas de una vieja revista de Informática. Se trataba de un programa diseñado para calcular la esperanza de vida de cada persona a partir de sus respuestas a una serie de preguntas sobre sus hábitos ("¿Practica usted algún deporte?", "¿Come verdura?", "¿Fuma?", "¿Bebe Alcohol?", etc.). Al llegar mi turno, y tras responder a las preguntas, el ordenador mostró por pantalla: "Usted vivirá 92 años". ¡Increíble! Por aquel entonces la persona de más edad que yo conocía era mi abuelo José, que debía rondar los 75, así que me sentí muy afortunado. También me pareció realmente sorprendente que una máquina pudiera obtener de forma tan precisa esta información a partir de una serie de preguntas tan sencillas. Aunque, immediatamente, me asaltaba una duda: "¿Será verdad que viviré tanto?". Ciertamente, aquella era una predicción bastante poco elaborada —el programa consideraba solamente unos pocos de los factores que inciden en la longevidad y probablemente les otorgaba un peso arbitrario. Eso era algo que podría hacer uno mismo tras unos minutos de reflexión y algunas sumas y restas. No obstante, usar aquel programa era mucho más divertido[1].

Ese fue mi primer contacto con el fabuloso mundo de la Inteligencia Artificial. Sin embargo, muchos otros acontecimientos en mi vida tienen la culpa de que me inscribiese en este programa de doctorado. Tantos Sonimag's... Me viene a la memoria una calurosa mañana del verano de ¿1989? Me veo programando en Basic sobre un IBM 8086 junto a Ricardo Cucala, mi profesor de Ciencias en primaria. También recuerdo el primer ordenador que hubo en casa, un Commodore con una RAM de 128Ks y disquetera de 5 y 1/4 pulgadas, que le debió costar a mi hermano Joaquín un ojo de la cara. Luego tuvimos un 386 a 33Mhzs y con un disco duro de 80 Mbytes. Aun recuerdo a don Ricardo: "¡Qué barbaridad! Sr. Giménez Linares, ¡tiene usted disco duro para toda la vida!". Pobrecito. Tardamos sólo 6 meses en llenarlo. Fue la etapa lúdica de la Informática. Más tarde, en el instituto, el ordenador se convirtió en una herramienta de trabajo indispensable, básicamente como procesador de textos.

Ya en la universidad, algunos profesores fueron realmente una inspiración para mí. Por ejemplo, Fermín Sánchez, Elvira Pino, Javier Béjar, Jaume Sistac, Tomàs Aluja, etc. También quiero agradecerles a Climent Nadeu, y, en particular a Núria Castell, como coordinadores del MELP en 1999, por haberme dado la oportunidad de iniciar mi carrera investigadora. Además, Núria ha sido mi tutora, y directora de tesis durante el primer año. Gràcies, Núria, per la teva dedicació.

[1] Para los curiosos, hace poco encontré una versión de este programa en `http://www.klid.dk/lifetime.php`

A més, durant la realització de la tesi he tingut l'oportunitat de conèixer un munt de gent interessant. Vull destacar tot el grup de PLN, especialment els companys de despatx: Pere Comas i el seu peix blau, Montse Cuadros i el seu horripilant i triquinós optimisme, Sergi Fernández (¿iremos al jazzy?), Daniel Ferrés, Maria Fuentes, Edgar González i les seves històries surrealistes, Muntsa Padró i els seus tes (hi ha vida després de la tesi?), Emili Sapena (l'espàrring de les meves bromes sense gràcia), Meritxell González (cafè?), Roberto Asín (¿mociona ir a...?). També la resta de membres del grup: Alicia Ageno, Manuel Bertran, Bernardino Casas, Neus Català, Jordi Daudé, Gerard Escudero, Cristina España, Javier Farreres, David Farwell, Ramon Ferrer, Marta Gatius, Àngels Hernández, Samir Kanaan, Carme Martín, Lluís Padró, Jordi Poveda, Horacio Rodríguez i Jordi Turmo... i antics membres: Victoria Arranz, Jordi Atserias (¿qué se le ofrece?), Xavier Carreras, Mauro Castillo y familia (¡corresponde!), Isaac Chao, Eli Comelles, Juan Francisco Fernández (¿cuánto le costó?), Miguel García, Josep Gubianas (¡casi te engañan!), Reda Halkoum, Patrik Lambert, Eva Naqui, Francisco Javier Raya y su música, Francis Real, German Rigau, Mihai Surdeanu, Luis Villarejo (mi peloncete)... y otros compañeros y amigos: Claudia Ayala, Carlos Mérida... uff! quanta gent! Si-us-plau, disculpeu-me si em deixo algú. Vull agraïr també al personal del Departament de LSI, secretaria i laboratori de càlcul, el seu bon treball.

Quiero tener también un agradecimiento especial para German Rigau y Enrique Amigó. Enrique, lo nuestro fue amor a primera vista. Después de casi tres años desde aquel día en Ann Arbor, puedo decir orgulloso que trabajar contigo ha sido verdaderamente excitante. A veces dando palos de ciego, y otras en el clavo. ¡Menuda pareja! ¡Cuántas cosas hemos aprendido! German, te'n recordes del dia que em vas convèncer per a que traduíssim les glosses de WordNet? La teva fe en mi em va fer picar l'ham. I, la veritat és que la cosa va sortir prou bé.

I, de manera especial, vull agraïr a en Lluís Màrquez, el meu director de tesi, la seva entrega i la seva dedicació absoluta, en cos i ànima. Lluís, amb els teus consells, sempre adients, has aconseguit il·luminar el meu camí i també motivar-me per donar el màxim cada dia[2].

Por supuesto, hay muchas otras personas que me han apoyado durante estos años; especialmente, mi familia, padres, hermanos, y demás parientes cercanos, sanguíneos y políticos, presentes y no presentes... y buenos amigos que me han regalado su afecto.

Por último, gracias a ti, Sandra, mi esposa, mi incansable compañera de viaje. Gracias por tu apoyo. Gracias por tu paciencia. Gracias por renunciar a todo el tiempo que nos he robado. Gracias por mantenerme a flote. A ti van, pues, dedicadas, con todo mi cariño, todas y cada una de las horas recorridas en pos de este objetivo. A tu lado, ha sido un esfuerzo soportable. A tu lado me quedan todavía tantas cosas que aprender.

[2] Això sí, no m'ha quedat gaire temps per tocar la teva guitarra. Si Déu vol, un dia d'aquests li posarem cordes noves.

Acknowledgements

One of my most vivid remembrances from my earliest childhood was at my beloved auntie's place. My cousin David had programmed his Spectrum computer by copycatting some fifty lines of code from an old magazine. I said to him: "David, what is all this for?". He replied: "Well, I've just typed a program so my machine can tell you how many years you are going to live. You want to give it a try?". Of course, I tried. Through the screen, the computer started prompting a series of yes/no questions such as "Do you play sport?", "Do you eat vegetables?", "Do you drink alcohol?", "Do you smoke?", etc. By pressing either 'Y' or 'N' on the keyboard I answered. At the end, the computer stopped asking and prompted: "You are going to live 92 years". Wow! The oldest person I knew was my grandfather José, who was 75, so I felt pretty fortunate. But, overall, I was wondering: "How could this metal and plastic artifact have elaborated on such an accurate prediction?". Amazing, I know! After some reflection, I understood. There was nothing extraordinary in this program. On the contrary, it was something one could do by simply adding and subtracting some fixed values, associated to a list of supposed health indicators, to a heuristically-defined average expected life time. However, using that program was undoubtly much more fun[3].

That was my very first contact with the world of Artificial Intelligence. However, there are surely many other facts that lead me to enroll in this Ph.D. program. There are also a number of important people in my life, relatives, friends and colleagues, whose names I shall not repeat (see 'Agradecimientos'), without whose support I could not have get through. Let me only thank, with special affection, two important people. First, my advisor, Lluís Màrquez, for his devotion and guidance throughout this research. Second, my wife, Sandy, for her unconditional love, and for keeping me always in a healthy mental shape.

I am also grateful to the European Commission for making its proceedings available, and to the TC-STAR Consortium, NIST, WMT and IWSLT workshop organizers, for providing such valuable data sets and test beds. I must also thank a number of NLP researchers worldwide whose tools have been utilized at some point for the purpose of our research: Enrique Amigó, Johan Boss, Xavier Carreras, Eugene Charniak, Michael Collins, Brooke Cowan, Stephen Clark, James Curran, Mark Johnson, Philipp Koehn, Patrik Lambert, Lluís Padró, Andreas Stolcke, and Mihai Surdeanu.

Finally, I am thankful to the Spanish Ministries of Science and Technology (ALIADO project, TIC2002-04447-C02) and Education and Science (OpenMT project, TIN2006-15307-C03-02) for supporting this research. Our NLP group is also recognized as a Quality Research Group (2005 SGR-00130) by DURSI, the Research Department of the Catalan Government.

[3]You want to know your expected life time? You are lucky. I found a version of this program at `http://www.klid.dk/lifetime.php`.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Machine Translation (MT) is one of the earliest and most paradigmatic problems in Natural Language Processing (NLP)[1] and Artificial Intelligence (AI). Although the first writings on the use of mechanical devices for translation date back from the seventeenth century, we must situate the origins of MT as a field in the late 1940's, right after World War II, with the availability of the first electronic computers in the US. In spite of their simplicity, original MT systems, based on bilingual dictionaries and manually-defined lexicalized reordering rules, obtained very promising results (Stout, 1954). However, after an initial period of euphoria, the lack of progress attained in the following years lead the US Government to set up the Automatic Language Processing Advisory Committee (ALPAC, 1966). In their report, its committee members concluded that MT was slower, less accurate and more expensive than human translation, and, therefore, recommended replacing investment in MT by investment in basic NLP research. Hence, it was set the beginning of almost two decades of difficulties for MT. Still, some research projects were developed, but it was not until the late 1980's and early 1990's when, through the use of more powerful and faster computers, able to handle larger amounts of data, MT recovered its original vigor.

Today, turning our eyes back to the past, one may certainly tell that the ALPAC report has actually yielded very positive consequences for NLP in the long term. Many resources (e.g., tools, corpora, knowledge bases, etc.) have been developed, specially for widely-used languages, and are, thus, at our disposal for being exploited in the context of complex NLP tasks such as MT. The availability of these resources allows developers to decompose the MT problem into smaller subproblems which are easier to address. Besides, the experience accumulated in the application of empirical methods to AI in general, and to NLP in particular, provides a battery of applicable solutions for many of these problems.

This rapid development of the field together with the inherent complexity of the task, make the MT scenario very attractive and challenging for NLP researchers. At the same time, the profitability of MT as a business has motivated a number of companies, governments and institutions worldwide, to invest large amounts of money in the funding of MT related projects. Hence, these days we are living with enthusiasm wealthy times for MT research.

[1] Natural Language Processing is a subfield of Artificial Intelligence and Computational Linguistics which studies the automated understanding and generation of natural human languages.

In this thesis, following the current trend in MT research, we aim at exploiting present NLP technology for MT. Our work addresses the problem of *Empirical Machine Translation and its Evaluation*. In first place, we have studied the most notable deficiencies of current evaluation methods, which arise, in our opinion, from the shallow quality principles upon which they are based. Instead of relying on the lexical dimension alone, we suggest a novel path towards *heterogeneous* automatic MT evaluation based on a rich set of automatic similarity metrics operating at different linguistic levels (e.g., lexical, syntactic and semantic).

In parallel to our work in MT evaluation, we have studied the problem of lexical selection in Statistical Machine Translation. For that purpose, we have constructed a Spanish-English baseline phrase-based Statistical Machine Translation system and iterated across its development cycle incorporating linguistic knowledge at different points so as to improve its overall quality. As a complementary issue, we address the problem of domain dependence in empirical MT systems.

The two parts of this thesis are tightly connected, since the hands-on development of an actual MT system has allowed us to experience in first person the role of the evaluation methodology in the development cycle of MT systems.

1.1 Machine Translation

MT is formally defined as the use of a computer to translate a *message*, typically text or speech, from one natural language to another. MT is considered, quoting Martin Kay, an *NLP-complete/AI-complete* problem. The reason is that the generation of high quality translations requires a full understanding of the message under translation.

1.1.1 Natural Language Understanding

Natural Language Understanding (NLU) is difficult because of Natural Language complexity. Natural languages are expressive —they allow for many different ways to express the same message— and ambiguous —messages may have many different possible interpretations. For instance, words in a sentence may have different meanings, and even when the meaning of all words is known, still sentences may have different readings. Further, these readings may have non-compositional interpretations.

The impact of NL ambiguity on MT has been well studied since the early beginnings of the field (Kaplan, 1955; Koutsoudas & Korfhage, 1956; Harper, 1957). As an illustration, let us recall one of the most popular examples in MT literature: *"Time flies like an arrow"*[2]. This sentence has several possible interpretations: (i) time goes by very quickly just like an arrow does, (ii) you should time flies as you would time an arrow, (iii) time flies in the same manner an arrow would time them, (iv) time those flies that are like arrows, (v) time flies (as a type of insect) enjoy an arrow, etc. However, our knowledge about the use of language tells us that the most plausible interpretation is the first one; the sentence as a metaphor instead of as a literal description.

[2]We recommend Chapter 6 in (Arnold et al., 1994) for a detailed description of the linguistic problems inherent to the Translation task. The reader may find as well an excellent report on MT divergences in (Dorr, 1994). Harold Somers provides also a very nice material for discussion on this topic in his MT courses (`http://www.alta.asn.au/events/altss_w2003_proc/altss/courses/somers/somers.html`).

Moreover, even when the sentence structure is clear, still it may have different interpretations in the context of the real world. In that respect, let us reproduce another classic example provided by Yehoshua Bar-Hillel in 1960: *"Little John was looking for his toy box. Finally he found it. The box was in the pen. John was very happy."*. In order to decide whether the word *'pen'* refers to a writing instrument or to a child's play pen, real world knowledge, for instance, on the relative size of objects, is required. NLU involves, therefore, ambiguity resolution at different linguistic levels. Below, we list the most common types of ambiguity:

- **Categorial ambiguity,** i.e., words having more than one possible grammatical category.
- **Word sense ambiguity,** i.e., words having more than one possible meaning or sense.
- **Syntactic ambiguity,** i.e., sentences having more than one possible syntactic parsing, leading to multiple alternative semantic interpretations.
- **Semantic ambiguity,** i.e., sentences syntactically unambiguous having still different possible semantic interpretations.
- **Referential ambiguity,** i.e., anaphoric noun phrases having more than one possible referent.
- **Ellipsis,** i.e., incomplete sentences in which the missing constituent is not clear.
- **Pragmatic ambiguity,** i.e., when the meaning depends on the context of the current situation (e.g., discourse, real world knowledge).

The level of complexity increases in the case of spoken language. For instance, additional types of ambiguity (e.g., phonetic ambiguity, emphasis drill, etc.) and other difficulties (e.g., ungrammatical speech) appear.

1.1.2 Classification of MT systems

Approaches to MT may be classified according to several criteria. For instance, regarding the degree of human interaction, MT systems may be classified in: (i) Machine-aided Human Translation (MAHT), (ii) Human-aided Machine Translation (HAMT), and (iii) Fully Automatic Machine Translation (FAMT) systems (Yngve, 1954). Nowadays, most commercial systems implement a MAHT scheme, whereas FAMT systems are dominant in the Internet, mostly free.

According to the level of linguistic analysis that is performed, MT systems may be classified in three groups: *direct, transfer, and interlingua*. Figure 1.1 depicts an updated version of the famous Vauquois triangle. In the *direct* approach a word-by-word or phrase-by-phrase replacement is performed (Weaver, 1955; Yngve, 1955; Yngve, 1957). In the *transfer* approach the input is syntactically and/or semantically analyzed to produce a source abstract representation, which is transferred, generally through the use of linguistic rules, into an abstract target language dependent representation, from which the output is generated (Vauquois et al., 1966). The *interlingua* approach is similar to the latter but with the difference that there is a unique abstract representation (Gode, 1955; Darlington, 1962). The interlingual representation is language independent and

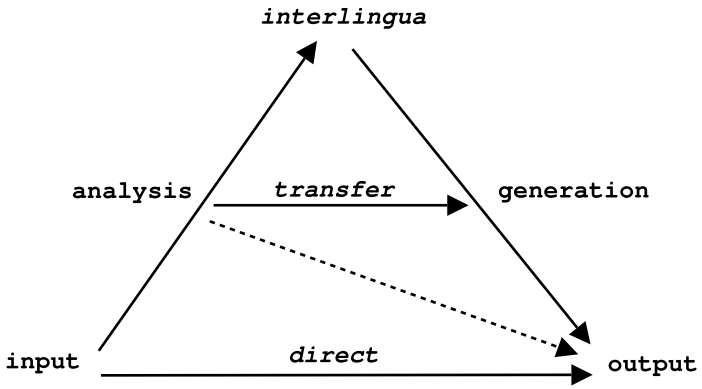


Figure 1.1: The Vauquois triangle for the classification of MT systems according to the level of linguistic analysis

deeply detailed, so all possible sentences expressing the same meaning in all languages receive the same representation. In this manner, the transfer between representations is no longer necessary.

With respect to the core technology, MT systems may be classified in two types: *rule-based* and *empirical*. In *rule-based systems*, a set of rules describing the translation process are specified by human experts. In contrast, *empirical systems* acquire this knowledge automatically from a collection of translation examples. Actually, the expression 'rule-based' is slightly inaccurate nowadays. The reason is that empirical MT systems may also use automatically induced rules. Therefore, perhaps it is more appropriate to refer to these two types of systems as *knowledge-driven* and *data-driven*. However, for historical reasons, the term 'rule-based' is still widely used.

Another distinction between rule-based and empirical systems used to be that, while rule-based systems typically performed some kind of linguistic transfer (e.g., syntactic, shallow-semantic, interlingual), empirical systems usually performed a direct translation of lexical units. This argument does not hold anymore either. Empirical systems often incorporate linguistic knowledge (e.g., syntactic parsing, see Chapter 4). In that respect, let us also note the intentional amendment of the Vauquois triangle, in Figure 1.1, with a dashed line representing the current trend in direct approaches to incorporate linguistic analysis.

Taking into account the differences and similarities between rule-based and empirical approaches, it will not be surprising that a variety of hybrid MT methods exploiting the best of both alternatives appear in the next few years. Indeed, several approaches yielding very promising results have been recently suggested (Alegría et al., 2008; Sánchez-Martínez et al., 2007; Simard et al., 2007). For instance, Simard et al. (2007) presented a valid hybridization scheme based on the statistical post-editing of the output of a rule-based MT system.

1.1.3 Current Applications

While MT technology has proved effective to aid human translation, and vice versa, it is not yet mature enough to allow for high-quality FAMT, except for literal translations in very restricted domains. This is the case, for instance, of the METEO system (Chandioux & Grimalia, 1996), which

translates Canadian weather forecasts from English into French, or the KANT system (Carbonell et al., 1992), devoted to the translation of machinery manuals from English into various languages. FAMT systems are, however, widely used in the Internet. For instance, the rule-based SYSTRAN MT system powers a number of web sites, such as Google. Also, military agencies rely on FAMT technology for the processing of languages spoken in conflict areas (e.g., Arabic, Pashto, Urdu, Dari). Moreover, the globalization of the economic system has also motivated a growing interest in the development of FAMT applications for languages in emerging markets, such as Chinese, which is also the most widely spoken language in the world with more than 1 billion speakers.

1.2 This Thesis

In this thesis, we have exploited current NLP technology for Empirical Machine Translation. Our goal is twofold. On the one side, we have studied the problem of *Automatic MT Evaluation*. We have analyzed the main deficiencies of the current methodology and suggested several complementary improvements. Our approach is based on the design of a heterogeneous set of automatic metrics devoted to capture a wide variety of translation quality aspects at different linguistic levels, from the lexical, through the syntactic, and onto the level of semantics. We also study the possibility of combining the scores conferred by different metrics into a single measure of quality.

On the other side, we have built an empirical MT system and have analyzed several of its limitations. We have incorporated linguistic knowledge into the system with the aim to improve overall translation quality. In particular, we have addressed the problem of *lexical selection*. We show that employing linguistic information allows for a better modeling of the translation context, effectively yielding an improved translation quality. As a side question, we have also studied one of the main criticisms against empirical MT systems, and empirical approaches to NLP in general, i.e., their strong domain dependence. We show how its negative effects may be mitigated by properly combining outer knowledge sources when porting a system into a new domain.

As stated in the beginning of the introduction, there is a connection between the two parts of this thesis in the sense that acting as system developers has allowed us to experience the enormous influence of evaluation methods across the different stages of the development cycle of an MT system. In the following, we outline the work deployed in each of these two research lines, as well as the circumstances that motivate it in the context of current MT research.

1.2.1 Automatic MT Evaluation

Automatic evaluation methods have notably accelerated the development cycle of MT systems in the last decade. They play a key role, allowing for fast numerical evaluations of translation quality on demand, which assist system developers in their everyday decisions. However, there are several purposes for which the behavior of current automatic evaluation methods is clearly unsatisfactory:

Evaluation of Global MT Quality. In many cases it has been argued that automatic metrics are unable to capture the quality changes which are due to the incorporation of linguistic knowledge (Yamada, 2002; Charniak et al., 2003; Och et al., 2003). The reason is that, despite

possible claims on the contrary, none of current metrics provides, in isolation, a *global* measure of quality. Indeed, all metrics focus on *partial* aspects, and, while quality dimensions are diverse, most of current metrics limit their scope to the lexical dimension.

System Optimization. The quality of a MT system depends very strongly on the metric selected to guide the development process. In other words, a system adjusted so as to maximize the score of a selected *golden* metric does not necessarily maximize the scores conferred by other metrics. We refer to this problem as *system over-tuning* (see Section 2.2.3).

Comparison of MT Systems. Current automatic evaluation metrics may not always provide reliable system evaluations. In particular, comparisons between MT systems directed towards different quality aspects have been showed to be problematic (Callison-Burch et al., 2006; Koehn & Monz, 2006). In particular, Callison-Burch et al. argue that MT researchers have possibly been overreliant on the capabilities of the BLEU measure, and, therefore, it is possible that a number of inaccurate conclusions had been drawn from past experiments. They even suggest that some of the ideas in recent literature should be revisited and reevaluated. We further discuss this issue in Section 2.2.3.

Error Analysis. Current automatic evaluation metrics fail to provide reliable evaluations at the sentence level (Blatz et al., 2003; Turian et al., 2003). Besides, they do not elaborate any interpretable information or explanation about the type of errors encountered which may help system developers to identify the strengths and weaknesses of their systems.

In order to overcome these limitations, we have deployed, in Chapter 3, a novel evaluation framework for *heterogeneous* automatic MT evaluation. Our proposal is based on a *divide and conquer* strategy. Instead of relying on individual metrics, we study how the scores conferred by different metrics can be combined into a single measure of quality. For that purpose, we have compiled a rich set of specialized automatic metrics operating at different linguistic levels (lexical, syntactic, and semantic). Our evaluation methodology has been validated over several test beds from recent well-known international evaluation campaigns. Besides, it is used, in Chapters 5 and 6, so as to assist us while iterating across the development cycle of the SMT system built for the purposes detailed in Section 1.2.2.

The main contributions of this thesis in this research line are:

- We present a heterogeneous set of similarity measures operating at different linguistic levels (Giménez & Màrquez, 2007b; Giménez & Màrquez, 2008c). Our approach provides a general framework for the definition of linguistic metrics which has been instantiated over particular similarity aspects.

- We show that linguistic metrics at more abstract levels may provide more reliable system rankings than metrics which limit their scope to the lexical dimension, specially in the case of systems belonging to different paradigms (Giménez & Màrquez, 2007b).

- We have studied the behavior of linguistic metrics in an extreme evaluation scenario corresponding to low-quality translation (Giménez & Màrquez, 2008c). We show that linguistic

metrics are robust against parsing errors committed by the automatic linguistic processors upon which they are based, particularly in the case of system-level evaluation. At the sentence level, some of these metrics (e.g., based on semantic parsing) suffer a significant decrease.

- We have exploited the possibility of combining metrics at different linguistic levels (Giménez & Màrquez, 2008b). Our approach offers the important advantage of not having to adjust the relative contribution of each metric to the overall score. A significantly improved evaluation quality at the sentence level is reported.

- We have showed how to apply linguistic metrics for the purpose of error analysis (Giménez & Màrquez, 2008d). Our proposal allows developers to rapidly obtain detailed automatic linguistic reports on their system's capabilities.

- As a by-pass product, we have developed a software package for heterogeneous MT evaluation, IQ_{MT}, which may be freely downloaded for research purposes (Giménez et al., 2005a; Giménez & Amigó, 2006; Giménez, 2007).

- We have studied the problem of meta-evaluation in the context of MT (Amigó et al., 2006). We have found that there is a tight relationship between human likeness and human acceptability.

1.2.2 Empirical MT

The second part of this thesis focuses on the study of fully automatic empirical MT of written Natural Language. By fully automatic we emphasize the fact that very light human interaction is required. By written Natural Language we distinguish text translation from speech translation.

Figure 1.2 depicts the prototypical architecture of an empirical MT system. Translation knowledge is acquired from a parallel corpus produced by human translators encoding translation examples between the languages involved. Parallel corpora are machine-readable document collections in two or more languages, such that each document is available in all languages, either as a source document or as the human translation of the associated source document. Typically, parallel corpora are automatically aligned at the paragraph or sentence level (Gale & Church, 1993). Minimal aligned units are often referred to as *segments*. Parallel corpora are also called bitexts when there are only two languages represented.

Empirical systems address MT as the problem of deciding, given an input text and acquired MT knowledge models, which is the most appropriate translation according to a given optimization criterion. Pre-processing and post-processing steps (e.g., tokenization, dedicated treatment of particular expressions such as dates, etc.) are optional.

Among empirical MT systems, the two most well-studied paradigms are Example-based Machine Translation (EBMT) and Statistical Machine Translation (SMT). Originally, these two approaches were clearly differentiable. EBMT methods used to be linguistically guided whereas SMT methods were statistically guided. Also, EBMT methods used to exploit source similarity while SMT systems exploited target similarity. These distinctions do not hold anymore. Indeed, the two approaches seem to be suavely merging into a single empirical MT paradigm (Way & Gough, 2005; Groves & Way, 2005).

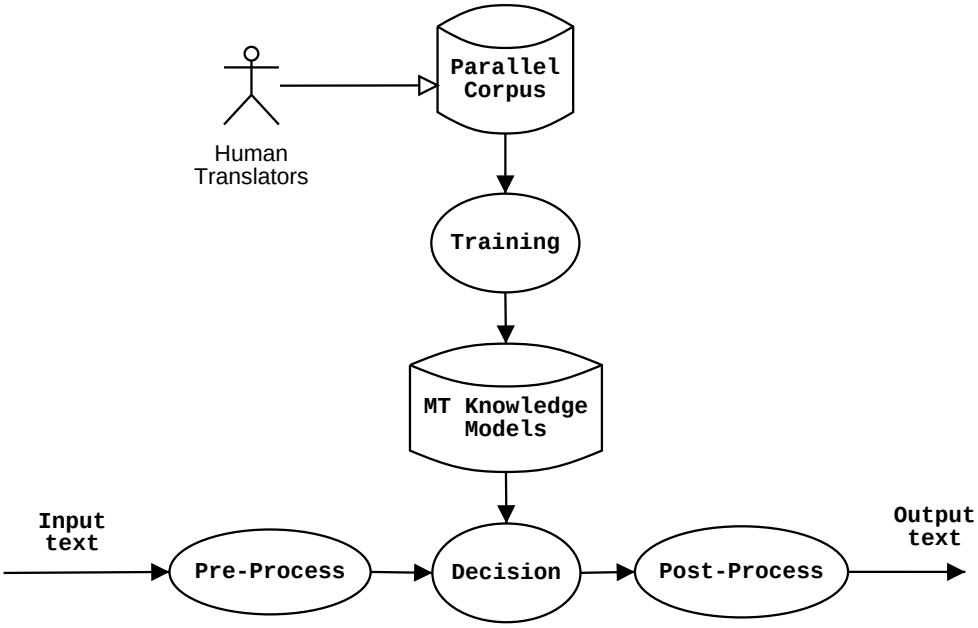


Figure 1.2: Architecture of an Empirical MT system

We have focused on SMT, which is today the most popular empirical approach to MT. SMT is also very well founded from a theoretical viewpoint. But the main reason for selecting SMT is that it allows for obtaining competitive results without using no additional linguistic information further than that implicitly encoded by lexical units. So, the room for potential improvement is in principle very large, and, at the same time, increasing the system quality is very challenging.

In our work, we suggest using current NLP technology and knowledge for improving an SMT system. Therefore, our golden assumption, and that of many other researchers (see Chapter 4), is that a system working with richer linguistic knowledge should be able to make better decisions. For that purpose, we have analyzed several points in the system architecture where improvements could take place. See Figure 1.3 as compared to Figure 1.2. Again, we would start from a parallel corpus. Linguistic processors would be used to annotate it with information at different levels. This linguistically enriched corpus would be used to train more informed knowledge models. At translation time, given a (linguistically) pre-processed input, these models would be used to provide more accurate translations. The resulting system output could be (linguistically) post-processed. Additional external knowledge sources, such as lexical ontologies or dictionaries, could be used at any stage.

In order to deploy such an architecture, first, we have adapted a number of NLP tools based on Machine Learning (ML), such as part-of-speech taggers and shallow syntactic parsers (see Appendix B). We have also collected resources such as parallel corpora, dictionaries and multilingual lexical databases. Then, we have constructed a state-of-the-art phrase-based SMT system, and studied how to incorporate these tools and resources into the system for several distinct purposes and with the final intent to improve the overall MT quality of the system (see Chapters 5, 6 and 7). In

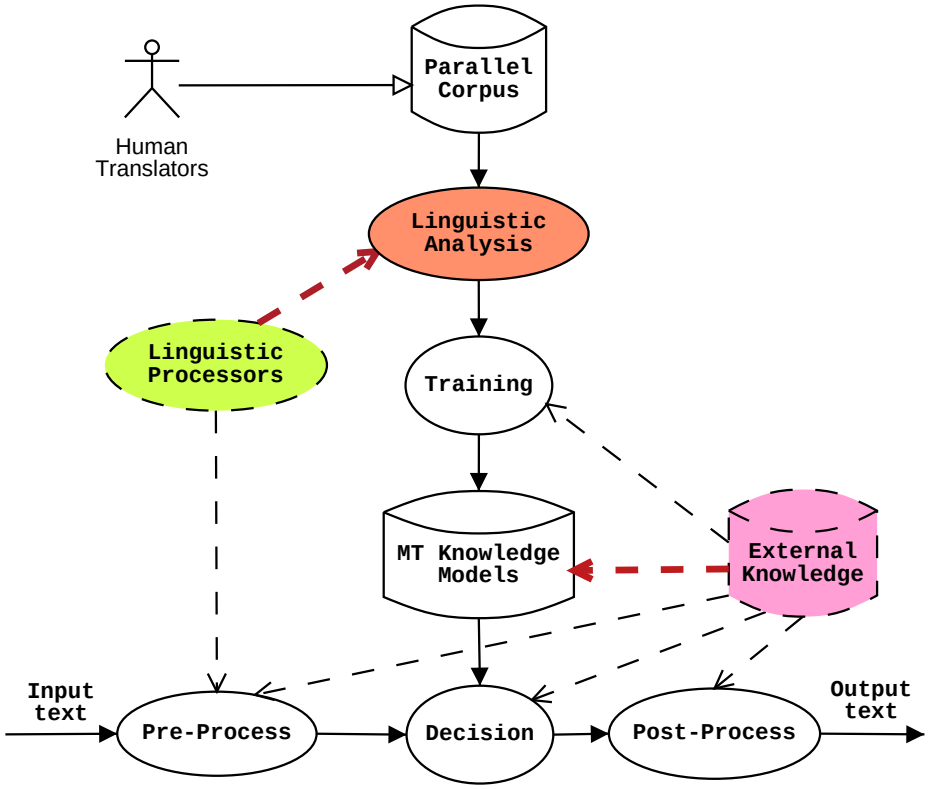


Figure 1.3: Architecture of a Linguistically-aided Empirical MT system

Chapters 5 and 6, we have studied whether it is possible to improve the modeling of translation probabilities in SMT by using automatically annotated linguistic knowledge at levels deeper than the lexical level. We have focused on the problem of lexical selection, i.e., on deciding, for a given lexical unit (word or phrase), which is the best translation among the set of possible translations candidates available (see Section 4.1.2). We have built shallow-syntactic translation models and improved their performance by building dedicated phrase translation models which are able to take into account a wider feature context. Our approach is based on supervised discriminative learning. As a side question, in Chapter 7, we have studied one of the main criticisms against empirical MT systems, i.e., *domain dependence*. We present a case study in which we apply several techniques for improving the behavior of SMT systems when used in new domains.

The main contributions of this thesis in this research line are:

- We show that linguistic information at the level of shallow-syntax may be successfully used to improve phrase-based translation models (Giménez & Màrquez, 2005; Giménez & Màrquez, 2006b). Our approach is based on building shallow-syntactic word and phrase alignments. We also present two valid phrase alignment combination schemes for translation modeling.

- We show how to build dedicated discriminative translation models (Giménez & Màrquez, 2007a; Giménez & Màrquez, 2008a). These models allow for a better representation of the

source translation context in which phrases occur, which leads to a more reliable estimation of phrase translation probabilities. Apart from exhibiting a higher local accuracy than the baseline approach based on maximum likelihood, we show that these models can be successfully integrated into a phrase-based SMT system and applied to the full translation task, yielding a significantly improved lexical selection. However, through heterogeneous automatic evaluations, we have observed that an improved lexical choice does not necessarily imply an improved overall syntactic or semantic structure. Manual evaluations have confirmed that improvements are mainly related to the adequacy dimension.

- We have studied the problem of domain dependence (Giménez et al., 2005b; Giménez & Màrquez, 2006a). First, we have used in-domain corpora to build specialized language and translation models. We show that it is possible to adapt an existing SMT system to a very different domain using only a very small amount of data belonging to the new domain. Second, we show that corpora from a similar domain may be helpful both for language and translation modeling. Third, we have built domain-independent translation models based on WordNet (Fellbaum, 1998). These models have not reported, however, much impact on translation quality, except for the case of unknown words.

- Our work in the development of an SMT system evinces the need for a heterogeneous MT evaluation methodology as the one proposed in this thesis.

1.2.3 Document Overview

The rest of this thesis is organized as follows.

- **Part I. MT Evaluation**

  - **Chapter 2. Machine Translation Evaluation**
    This chapter provides an extensive review on MT evaluation methods. We discuss both manual and automatic measures as well as the the role of evaluation methods in the context of the current development cycle of MT systems.
  - **Chapter 3. Towards Heterogeneous Automatic MT Evaluation**
    In this chapter, we present our proposal towards heterogeneous automatic MT evaluation. We describe a wide set of metrics operating at different linguistic levels and study their individual and collective application over several evaluation scenarios. We also present our approach to metric combination and to automatic error analysis.

- **Part II. Empirical MT**

  - **Chapter 4. Statistical Machine Translation**
    In this chapter, we give an overview of the fundamentals and current trends in Statistical Machine Translation. We describe the shift from word-based to phrase-based translation, as well as some of the most prominent extensions suggested in the last decade, with special focus on the incorporation of linguistic knowledge. We also discuss the problem of domain dependence in SMT.

  - **Chapter 5. Shallow Syntactic Alignments and Translation Models**
    This Chapter presents the construction of a baseline Spanish-English phrase-based SMT system based on a collection of Proceedings from the European Parliament, and its enhancement through the use of shallow-syntactic translation models. Linguistic knowledge is incorporated during the word and phrase alignment processes.
  - **Chapter 6. Discriminative Phrase Selection for SMT**
    This Chapter explores the application of discriminative learning to the problem of phrase selection in SMT. We build dedicated local phrase translation classifiers which are able to take further advantage of the source context. We also show how local predictions can be softly integrated into a phrase-based SMT system for the purpose of the global translation task.
  - **Chapter 7. Domain Adaptation of an SMT System**
    This Chapter presents a practical case study on the adaptation of the empirical MT system built on the previous chapters, from the political domain (i.e., European Parliament Proceedings) to the domain of dictionary definitions (i.e., WordNet glosses). Several complementary improvement techniques are presented.

- **Chapter 8. Conclusions**
  In this chapter, main conclusions are drawn, and future work is outlined.

- **Appendices**

  - **Appendix A. Author's Publications**
    This appendix is a full list of author's publications while enrolled in this PhD program.
  - **Appendix B. Linguistic Processors and Tag Sets**
    This appendix provides information on the linguistic processors utilized as well as a series of tables describing the associated tag sets.
  - **Appendix C. Metric Sets**
    This appendix provides a full list of metric variants in the current metric set. These are grouped in several families according to the linguistic level at which they operate.

How to read this document

As sketched across the introduction, there are two well-differentiated parts in this thesis. The first part (Chapters 2 and 3) addresses the problem of MT evaluation. Readers familiar with this subfield may skip most of the sections in Chapter 2. However, for a better understanding of the motivations of our research work, it is highly advisable to revise Sections 2.2 (specially Section 2.2.3), and 2.4 (specially Sections 2.4.2 and 2.4.4). Then, in Chapter 3, we introduce our proposal towards heterogeneous automatic MT evaluation, and validate it over several evaluation scenarios. Thus, in this part of the thesis, we have acted mainly as metric developers. However, the methods presented will also assist us in the second part of the thesis, in our complementary role as system developers.

The second part (Chapters 4 to 7) is devoted to the construction and development of an SMT system. Chapter 4 is essentially a survey on the state-of-the-art in SMT. Readers familiar with

this topic might want to proceed directly to Chapter 5, although Sections 4.2, 4.4 and 4.5 will be referenced back, since they describe a selection of the most relevant works respectively related to the contents of the following three chapters. Chapters 5 and 6 deal with the problem of lexical selection. First, Chapter 5 describes the construction of a Spanish-to-English baseline system improved with shallow-syntactic translation models. Then, in Chapter 6, this system is further improved building dedicated discriminative phrase translation models also relying on shallow-syntactic information. Chapter 7 studies the separate problem of domain dependence, and it is only related to the two previous chapters in that the baseline SMT system is the same, although in the reverse direction (i.e., English-to-Spanish).

Finally, in Chapter 8, we present a summary of results and contributions, as well as the main conclusions that can be derived. Future research work and directions are also outlined.

Part I

MT Evaluation

Chapter 2

Machine Translation Evaluation

Since its origins, research in MT has been accompanied by research in MT Evaluation (Miller & Beebe-Center, 1956; Pfafflin, 1965). In particular, there has been a wide interest in automatic evaluation methods. The reason is that these methods allow for considerably accelerating the development cycle of MT systems, and NLP applications in general (Thompson, 1991).

However, evaluating translation quality is a complex issue. This arises from the fact that MT is an *open* NLP task. Given a certain input, the set of solutions is not closed; every human subject could potentially produce a different translation, and all of them could be in principle equally valid. This is due to the expressiveness and ambiguity of Natural Language itself (see Section 1.1.1).

A number of evaluation methods have been suggested. Either manual or automatic, all share the common characteristic of operating over predefined test suites, i.e., over fixed sets of translation test cases (King and Falkedal 1990)[1]. Therefore, a first important concept to bear in mind is that test suites introduce a significant bias in the evaluation process. For instance, if the test bed does not cover a representative set of test cases, evaluation results bias accordingly. Also, if the set of manual reference translations represents only a small part of the whole space of solutions, the significance of the results is affected. Similarly, if the set of automatic translations represents only a small subset of MT systems (e.g., systems belonging to the same paradigm or different versions of the same system), or a specific language pair, or translation domain, the validity of the evaluation results will be restricted to the specific evaluation scenario.

In the following, we have elaborated a thorough review on MT evaluation. First, in Section 2.1, we talk about context-based evaluation of MT systems. Then, we focus on what relates to the research work deployed in this thesis. Section 2.2 discusses the role of the evaluation scheme in the MT development cycle. In Sections 2.3 and 2.4 we respectively describe some of the most relevant approaches to manual and automatic evaluation.

2.1 Context-based Evaluation

Although the focus of our work is in the evaluation of translation quality independently of the context of the MT system, this section is a brief note on context-based evaluation. This line of research

[1] A test case typically consists of a source sentence and a set of human reference translations.

promotes the idea that potential users of MT technology should first evaluate the suitability of this solution for their specific purpose. In that respect, Church and Hovy (1993) analyzed what requirements a good niche application for MT should meet. They suggested six desiderata: (i) it should set reasonable expectations, (ii) it should make sense economically, (iii) it should be attractive to the intended users, (iv) it should exploit the strengths of the machine and not compete with the strengths of the human, (v) it should be clear to the users what the system can and cannot do, and (vi) it should encourage the field to move forward toward a sensible long-term goal. These principles were further discussed and extended by the Evaluation Working Group of the ISLE Project (1999-2002)[2]. The main focus of this working group was the development of a classification or taxonomy of the features that are relevant to machine translation evaluation. They organized several workshops, and, overall, they developed FEMTI[3], a framework for context-based MT evaluation (Hovy et al., 2002). FEMTI provides a methodology to evaluate MT systems according to a wide range of characteristics and quality aspects such as functionality, reliability, usability, efficiency, maintainability, portability, cost, etc. FEMTI is made of two interrelated classifications or taxonomies. The first classification enables evaluators to define an intended context of use for the MT system to evaluate. The second classification links the selected relevant quality characteristics to a set of metrics associated. Once the context of the evaluation is defined, in response, FEMTI generates appropriate evaluation plans to be executed by the user.

2.2 The Role of Evaluation Methods

The current development cycle of MT systems follows the flow chart depicted in Figure 2.1. In each loop of the cycle, system developers must identify and analyze possible sources of errors. Eventually, they focus on a specific subproblem and think of possible mechanisms to address it. Then, they implement one of these mechanisms, and test it. If the system behavior improves (i.e., the number of the selected type of errors diminishes without harming the overall system performance), the mechanism is added to the system. Otherwise, it is discarded. In the context of MT system development, evaluation methods are necessary for three main purposes:

- **Error Analysis**, i.e., to detect and analyze possible cases of error. A fine knowledge of the system capabilities is essential for improving its behavior.
- **System Comparison**, i.e., to measure the effectiveness of the suggested mechanisms. This is done by comparing different versions of the same system. It is also common to compare translations by different systems, so system developers may borrow successful mechanisms from each other. This allows the research community to advance together.
- **System Optimization**, i.e, the adjustment of internal parameters. Typically, these parameters are adjusted so as to maximize overall system quality as measured according to an evaluation method at choice.

[2] http://www.issco.unige.ch/projects/isle/
[3] http://www.issco.unige.ch/femti

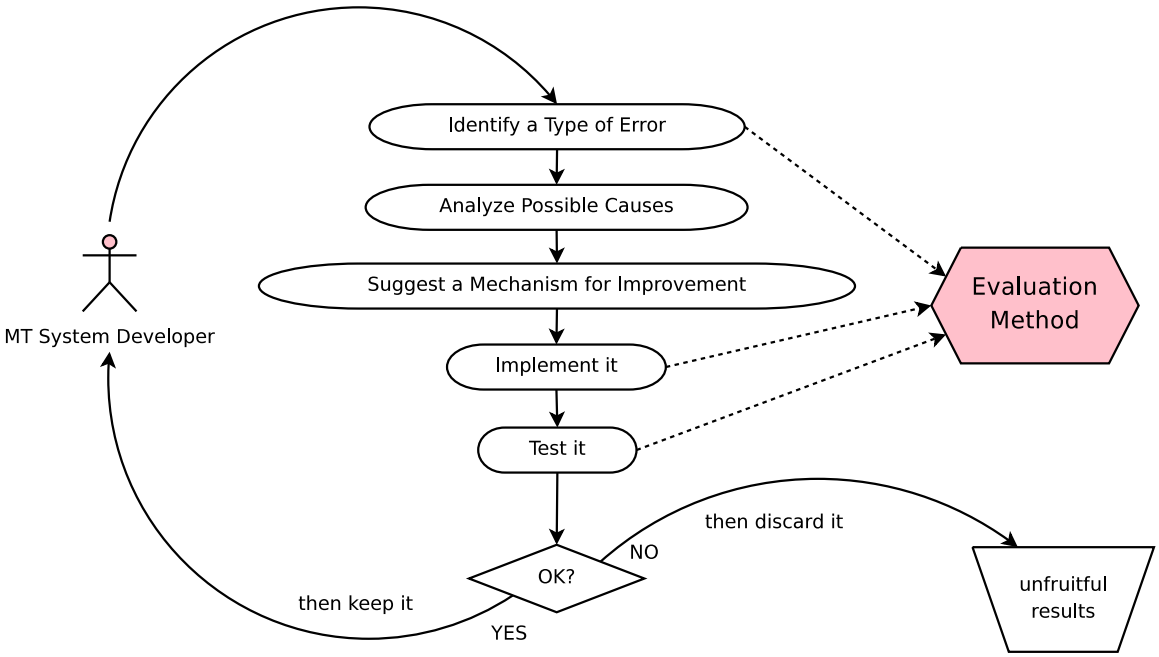


Figure 2.1: MT system development cycle

2.2.1 A Review

In the following, we provide a brief historical overview on the evolution of the evaluation scheme in the context of the MT system development. The original development scheme, prior to the availability of automatic evaluation metrics, was entirely based on human evaluations (see top-left flow chart in Figure 2.2). In this scheme, system developers iterated across the development cycle constantly introducing new changes so as to improve their prototype systems (process I). Eventually, they performed manual evaluations in order to evaluate the degree of progress attained, possibly at the time of running a competitive evaluation exercise (process II). Manual evaluations produced one or more *manual rankings* (depending on how many quality aspects were considered), which system developers could take into account for further system improvement.

The main drawback of the original scheme was that human assessments are expensive to acquire. Therefore, system developers could not monitor system improvements with enough regularity. In order to accelerate the development cycle, in the current scheme (see top-right flow chart in Figure 2.2), a process of *automatic evaluation* (process III) was added to the development cycle (Thompson, 1991). Automatic evaluation is based on *automatic metrics* which determine the quality of a *system output* according to its similarity to a predefined set of *references* generated by *human subjects*.

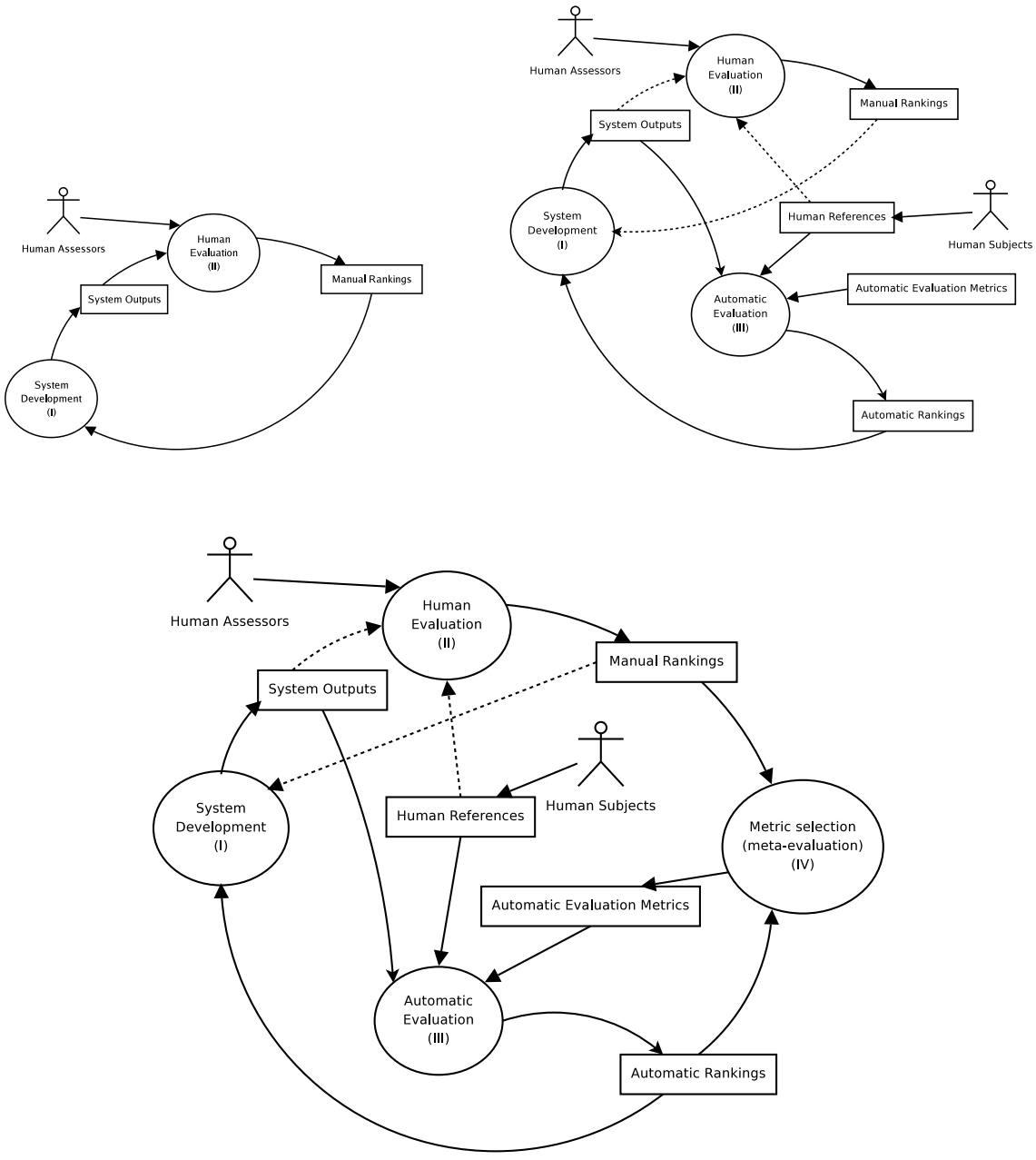


Figure 2.2: Evolution from the evaluation scheme entirely based on Human Assessors (top-left chart) to the evaluation scheme based on human assessors and automatic metrics (top-right chart). The role of meta-evaluation in this latter evaluation scheme is illustrated in the bottom chart.

2.2.2 Meta-Evaluation

Automatic evaluation metrics allow researchers to evaluate and optimize their systems without requiring the intervention of expensive human assessors. However, the usage of automatic evaluation measures generates in its turn an additional step in the development cycle: meta-evaluation, i.e., the evaluation of evaluation measures.

The bottom flow chart in Figure 2.2 illustrates the role of meta-evaluation in the current MT task development cycle (process IV). Prior to starting to iterate across the system development cycle, developers must decide which is the most suitable evaluation metric for the task at hand. This decision will have an enormous influence over the whole development cycle, since the metric selected will be responsible for guiding the developer in identifying the system weaknesses and deciding which modifications should be introduced. Moreover, the metric will also be used to judge whether the modifications are helpful or not. And, commonly, the metric will also govern any process of adjustment of parameters guiding the system towards configurations which maximize the quality aspects the metric is able to capture. In the following, we describe the two most well studied meta-evaluation criteria.

Human Acceptability

The quality of automatic MT evaluation metrics is usually estimated in terms of their ability to capture the degree of acceptability to humans of automatic translations, i.e., their ability to emulate human assessors. This is usually measured on the basis of correlation between automatic metric scores and human assessments of translation quality (Papineni et al., 2001; Callison-Burch et al., 2007). The underlying assumption is that *good* translations should be acceptable to human evaluators. For that reason, we call this type of meta-evaluation as based on *Human Acceptability*. Typically, metrics are evaluated against adequacy or fluency assessments, or a combination of the two, using either Pearson (1914, 1924, 1930), Spearman (1904) or Kendall (1938; 1955) correlation coefficients.

Most of current metrics have been developed on the basis of human acceptability. For instance, Papineni et al. (2001) say: *"We propose a method of automatic machine translation evaluation that is quick, inexpensive, and language independent, that correlates highly with human evaluation, and that has little marginal cost per run."*, Turian et al. (2003) say: *"The most important criterion for an automatic MT evaluation measure is that it ranks MT systems the same way that a human judge would rank them."*, Lin and Och (2004a) say: *"[...] the first criterion to assess the usefulness of an automatic evaluation measure is to show that it correlates highly with human judgments in different evaluation settings."*, Kulesza and Shieber (2004) say: *"The resulting metric [...] is shown to significantly improve upon current automatic metrics, increasing correlation with human judgments [...]"*, and Banerjee and Lavie (2005) say: *"We evaluate METEOR by measuring the correlation between the metric scores and human judgements of translation quality"*.

Actually, correlation with human assessments is a reasonable criterion, since automatic evaluation metrics were originally meant to replace human assessments, and therefore correlation with them seems the most direct (and interpretable) way of ensuring that such replacement is possible.

However, meta-evaluation on the basis of human acceptability presents the major drawback of relying on human evaluations, which are, expensive, not reusable, subjective, and possibly partial

(see Section 2.3). As a result, the behavior of automatic metrics is usually validated only in very few and specific evaluation scenarios, often in the context of an evaluation campaign or shared task, and over a limited number of samples. For instance, most meta-evaluation reports focus on a single language pair, a specific translation domain, and a small set of systems typically belonging to the same MT paradigm.

The problem of meta-evaluating on a very specific scenario is that results are not guaranteed to port well to other evaluation scenarios. The reason is that the quality aspects distinguishing high quality from low quality translations may vary significantly from one scenario to another, and, consequently, the performance of metrics operating on different quality dimensions may vary as well. In other words, the behavior of automatic metrics depends on a number of variables such as the language pair, the specific domain of the translation task, and the typology of systems under evaluation. Thus, it would seem reasonable to conduct a meta-evaluation process prior to any evaluation stage or campaign. However, meta-evaluation is in most cases ignored, or conducted only a posteriori. The reason is that human acceptability is a too costly solution for that purpose.

Human Likeness

A prominent alternative criterion is to evaluate metrics in terms of their ability to capture the degree of *human likeness* of automatic translations. The underlying assumption is that *good* translations should resemble human translations. Human likeness is usually measured in terms of *discriminative power*, i.e., the metric ability to capture the features which distinguish human from automatic translations (Corston-Oliver et al., 2001; Lin & Och, 2004b; Kulesza & Shieber, 2004; Amigó et al., 2005; Gamon et al., 2005). The idea is that, given that human translations are gold standard, a *good* metric should never rank automatic translations higher (in quality) than human translations. Then, when a system receives a high score according to such a metric, we can ensure that the system is able to emulate the behaviour of human translators.

The main advantage of human likeness is that it is a much more cost-effective alternative, since the need for human assessments disappears. Human likeness opens, thus, the path towards a new development scheme entirely based on automatic metrics (see Figure 2.3 as compared to the bottom flow chart in Figure 2.2). In this scheme, human subjects are only required for solving the test cases (as systems do) and, thus, to serve as models (i.e., providing human references) for the evaluation process. Avoiding human assessments eliminates also one subjective factor: the assessment evaluation guidelines. In addition, human assessments are static, while discriminative power can be updated if new human references or system outputs are incorporated to the test bed along time.

However, meta-evaluation based on human likeness presents a major shortcoming; just like automatic evaluation, it depends strongly on the heterogeneity/representativeness of the test beds employed (i.e., sets of test cases, and associated automatic system outputs and human reference translations). For instance, if the set of reference translations per test case is small it may not represent well the full set of acceptable solutions, and the meta-evaluation process may be biased. Therefore, the applicability of human likeness as meta-evaluation criterion must be further studied and validated.

In this respect, in a joint effort with Enrique Amigó and Julio Gonzalo, from the *"Universidad*

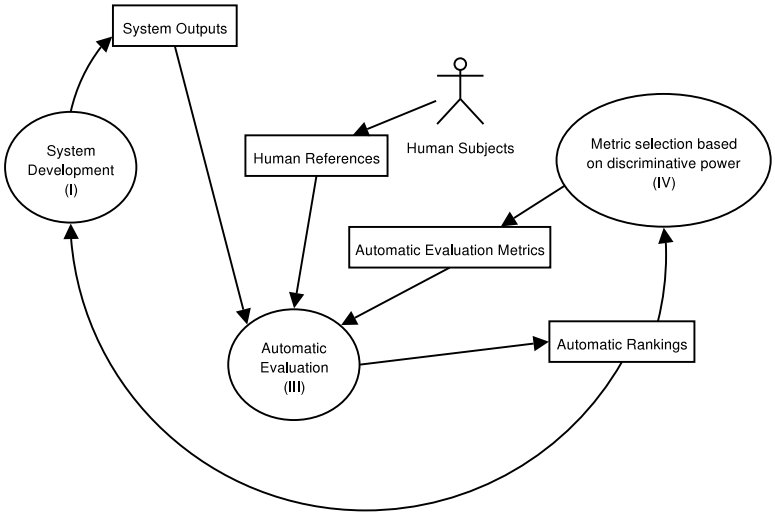


Figure 2.3: MT task development cycle entirely based on automatic metrics

de Educación a Distancia" (UNED), in Madrid, we have conducted a comparative study on the behavior of human likeness and human acceptability as meta-evaluation criteria in the context of open NLP tasks, such as Machine Translation and Automatic Summarization. Results have revealed that there is an interesting relationship between them (Amigó et al., 2006). While human likeness is a sufficient condition to attain human acceptability, human acceptability does not guarantee human likeness. In other words, human judges consider acceptable translations that are human-like, but they may also consider acceptable many other automatic translations that would be rarely generated by a human translator. Therefore, given that human likeness is a stronger condition, it seems reasonable to think that basing the development cycle on it should lead to similar results. This hypothesis is currently under study.

2.2.3 The Metric Bias Problem

Evaluation measures are all focused on partial aspects of quality (e.g., adequacy, fluency, lexical similarity, etc.). The main problem of partial measures is that they may generate strongly biased evaluations. Besides, since evaluations are required at several stages, this bias may propagate across the whole system development cycle, leading developers to derive inaccurate conclusions and, consequently, to make wrong decisions. We refer to this problem as the *metric bias* problem.

In the following, we illustrate the negative effects of metric bias through three different examples, respectively based on system evaluation, system optimization, and system development.

Unfair System Comparisons

Often, it is the case that different metrics produce different system quality rankings over the same set of test cases. The reason is that quality aspects are diverse and not necessarily interrelated. Thus, metrics based on different similarity assumptions may confer different scores.

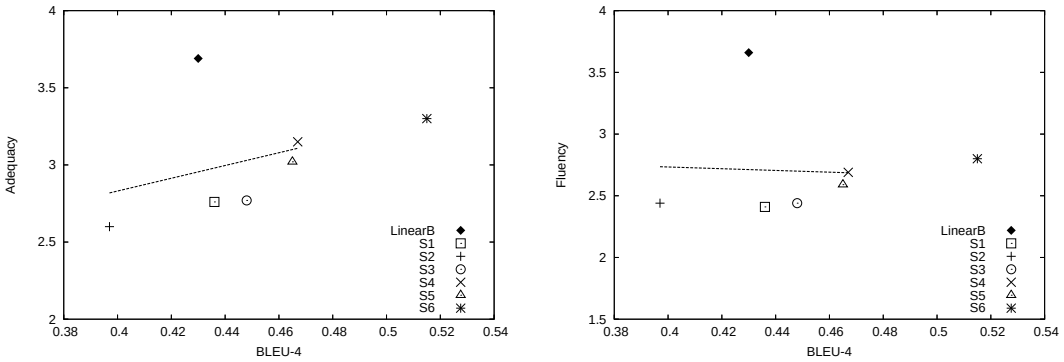


Figure 2.4: NIST 2005 Arabic-to-English. System BLEU scores vs. human assessments

For instance, Charniak et al. (2003), who worked on syntax-based language modeling for SMT, reported a significantly improved translation quality according to manual evaluations. However, the BLEU metric did not capture this improvement, but reflected, instead, a severe 30% quality drop[4].

Other similar cases have been recently reported. For instance, Callison-Burch et al. (2006) and Koehn and Monz (2006) detected several problematic cases related to the automatic evaluation and ranking of MT systems based on different paradigms (e.g., human-aided vs. statistical, rule-based vs. statistical) and, therefore, oriented towards different quality aspects. They noticed a strong disagreement between human and automatic evaluations. The problem was that they used BLEU, a metric based on lexical matching, to evaluate systems using different lexica.

As an example, Figure 2.4 illustrates the case of the Arabic-to-English 2005 NIST MT Evaluation Exercise[5] reported by Callison-Burch et al. (2006). BLEU scores are plotted against average human scores on adequacy (left) and fluency (right). It can be observed how BLEU rankings do not fully correspond to the manual evaluation. In particular, the LinearB system was ranked 1st by human judges and 6th by BLEU. The reason is that BLEU favors MT systems which share the expected reference lexicon (i.e., statistical systems), and penalizes those which use a different one.

These findings agree with those by Coughlin (2003), who presented a very rigorous study on the correlation between BLEU and NIST scores and human assessments of translation quality over a large variety of evaluation scenarios (including different MT systems, different language pairs, and varying number of reference translations available). They found out that n-gram based metrics tend to favor statistical systems vs. rule-based/hybrid systems. The reason is that statistical systems are likelier to match the sublanguage (e.g., lexical choice and order) represented by the set of reference translations, when, indeed, lexical similarity is not a sufficient neither a necessary condition so that two sentences convey the same meaning. On the contrary, as we have seen in Section 1.1.1, natural languages are expressive and ambiguous at different levels.

This problem is further analyzed in Section 3.2. We show how metrics at deeper linguistic levels provide more reliable system rankings than metrics which limit their scope to the lexical dimension.

[4]BLEU score decreased from 0.1031 to 0.0717.

[5]`http://www.nist.gov/speech/tests/summaries/2005/mt05.htm`

System Overtuning

Adjustment of parameters is a crucial step in the development of an SMT system. Particularly critical is the tuning of the parameters that govern the search. Commonly, a minimum error rate iterative strategy is followed (Och, 2003). At each iteration the MT system is run over a so-called development set under a certain parameter configuration. At the end of the process, the configuration producing the output of lowest error rate is selected to translate new text. Error rate is typically measured according to an evaluation metric at choice, typically BLEU.

Optimizing over an error measure based on a single metric presents a major drawback. The system may end strongly biased towards configurations which maximize this metric score but may not necessarily maximize the scores conferred by other metrics. We refer to this problem as *system overtuning*. Some authors have tried to overcome this problem by defining error measures over linear combinations of metrics (Hewavitharana et al., 2005; Chen et al., 2005). However, in these cases, metric combinations are selected arbitrarily, or, at the least, the criterion employed to select them is either uncertain or ad-hoc.

In Section 6.3.3, we present a practical case study on the effects of the metric selected to guide the optimization process of our SMT system enhanced with dedicated lexical selection models. Specifically, we compare the results optimizing over BLEU with respect to results optimizing over a combination of lexical metrics on the basis of human likeness. In a joint effort with Lambert et al. (2006), we conducted a similar study, in this case optimizing the TALP N-gram based SMT system (Mariño et al., 2006). Manual evaluations showed that this alternative leads to more robust system configurations than relying on the BLEU measure alone.

Blind System Development

Automatic evaluation methods play, as discussed before, a very important role in the context of MT system development. Indeed, evaluation methods are not only important but they are also an upper bound on the attainable success of the development process itself. In other words, improvements may take place as long as developers count on mechanisms to measure them. Otherwise, the development cycle is blind. A paradigmatic case of blind development occurred in the Johns Hopkins University 2003 Summer Workshop on *"Syntax for Statistical Machine Translation"* (Och et al., 2003)[6]. A team of leading researchers and motivated students devoted 6 weeks to improve a phrase-based SMT system through the incorporation of syntactic knowledge. Although they suggested a rich smorgasbord of syntax-based features, only a moderate improvement (from 31.6% to 33.2% according to BLEU) was attained, which, indeed, came almost exclusively from using the IBM 1 model word alignment probabilities to compute a lexical weighting feature function. They argued two main reasons for this result. First, they observed that syntactic parsers introduce many errors. Second, and most important, they noted that the BLEU metric, which they used for development and test, was not able to capture improvements due to a better syntactic sentence structure.

For the sake of robustness, we argue that the development cycle must be always *metricwise*, i.e., the metric (or set of metrics) guiding the development process must be able to capture the possible quality variations induced by system modifications. We further discuss this issue in Chapter 6.

[6] `http://www.clsp.jhu.edu/ws03/groups/translate/`

2.3 Human Evaluation

Manual evaluations present the main advantage of allowing system developers to measure the quality of their systems over a wide range of partial aspects of quality and over a set of potential end-users. Several approaches to human evaluation have been suggested (Lehrberger & Bourbeau, 1988; Falkedal, 1994; Arnold et al., 1994; Dabbadie et al., 2002). In the following, we give an overview on the most well known.

2.3.1 ALPAC's Approach

One of the constituent parts of the ALPAC report (1966) was a study comparing different levels of human translation with machine translation output, using human subjects as judges. Two variables were considered:

- **Fidelity** (or Accuracy) was a measure of how much information the translated sentence retained compared to the original (on a scale of 0-9).
- **Intelligibility** was a measure of how 'understandable' the automatic translation was (on a scale of 1-9).

Each point on the scale was associated with a textual description. For example, 3 on the intelligibility scale was described as *"Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence"*. Intelligibility was measured without reference to the original, while fidelity was measured indirectly. The translated sentence was presented, and after reading it and absorbing the content, the original sentence was presented. The judges were asked to rate the original sentence on informativeness. So, the more informative the original sentence, the lower the quality of the translation. The study showed that the fidelity and intelligibility were highly correlated when the human judgement was averaged per sentence. The variation among raters was small, but the researchers recommended that, at least, three or four raters should be used. The evaluation methodology managed to separate translations by humans from translations by machines with ease. The study concluded that, *"highly reliable assessments can be made of the quality of human and machine translations"*.

2.3.2 ARPA's Approach

As part of the *Human Language Technologies Program*, the *Advanced Research Projects Agency* (ARPA) created a methodology to evaluate machine translation systems (White et al., 1994; White, 1995). The evaluation program started in 1991, and continues to this day. It involved testing several systems based on different theoretical approaches (e.g., statistical, rule-based and human-assisted). A number of methods for the evaluation of the output from these systems were tested in 1992 and the most recent suitable methods were selected for inclusion in the programs for subsequent years. The evaluation measures were:

- **Comprehension Evaluation.** This method, also referred to as *informativeness*, is intended to directly compare systems based on the results from multiple choice comprehension tests,

as in (Church & Hovy, 1993). It is, therefore, an *extrinsic* evaluation measure. MT quality is indirectly evaluated by having human subjects read automatically translated texts and then answer several related questions.

- **Quality Panel Evaluation.** This method consisted in submitting translations to a panel of expert native speakers who were professional translators. The evaluations were done on the basis of a metric, modeled on a standard US government metric used to rate human translations. The principal value of this approach was that the metric was externally motivated, since it was not specifically developed for machine translation (White et al., 1994). However, setting up quality panel evaluations was very difficult in terms of logistics, since they required having a number of experts together in one place for several days. Furthermore, reaching consensus among experts was complicated. Therefore, this method was abandoned.

- **Adequacy and Fluency.** A group of human subjects is required to judge a collection of translations of one or more documents (LDC, 2005). Judges are presented with a translation segment, and asked to rate it for these two variables. Adequacy refers to the degree to which information present in the original is also communicated in the translation. It is intended to capture translation fidelity. Fluency refers to the degree to which the target is well formed according to the rules of the target language (usually Standard Written English). It is intended to capture translation intelligibility.

  These measures are very similar to the *fidelity* and *intelligibility* measures used in the ALPAC report (1966). In this case, however, scores are assessed according to a 1-5 scale. A brief interpretation of adequacy and fluency scores may be found in Table 2.1. This technique was found to cover the relevant parts of the quality panel evaluation, while, at the same time, being easier to deploy, as it did not require expert judgement. However, because these measures operate at the sentence level they may fail to capture discourse phenomena. Along with informativeness, evaluation based on adequacy and fluency is these days the standard methodology for the ARPA evaluation program[7].

Score	Adequacy	Fluency
5	All information	Flawless English
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Table 2.1: Interpretation of Adequacy and Fluency scores

- **Preferred Translation.** This measure has been proposed very recently. It consists in having human subjects to perform pairwise system comparisons at the sentence-level, i.e., deciding if output by a system 'A' is better, equal to, or worse than output by a system 'B'.

[7]The evaluation plan corresponding to the 2008 NIST Evaluation Campaign is available at `http://www.nist.gov/speech/tests/mt/doc/MT08_EvalPlan.v1.1.pdf`.

2.3.3 Other Evaluation Measures

Other evaluation measures less commonly used are:

- **Meaning Maintenance.** This measure intends to compare the meaning of the translation with the source (Eck & Hori, 2005). It is similar to adequacy, although it is more concerned with the actual meaning of a translation. There is, however, a high correlation between adequacy and meaning maintenance. A brief interpretation of meaning maintenance scores may be found in Table 2.2.

Score	Description
4	Exactly the same meaning
3	Almost the same meaning
2	Partially the same meaning and no new information
1	Partially the same meaning but misleading information is introduced
0	Totally different meaning

Table 2.2: Interpretation of Meaning Maintenance scores

- **Read Time.** Reading time relates to the amount of time a potential user needs to read a document to a *sufficient* level of understanding. It is essentially a time comprehension test (Slype, 1979).
- **Required Post-Editing.** Minimum number of key strokes required to transform the automatic translation into a valid translation.
- **Post-Edit Time.** Time required to transform the automatic translation into a valid translation.
- **Cloze Test.** A test of readability based on measuring the ability of a reader to fill in the blanks after intentionally removing single words from the automatic translations (Slype, 1979). Supposedly, it takes into account both fidelity and intelligibility.
- **Clarity.** Human raters are asked to score the clarity of each sentence on a 0-3 scale (Vanni & Miller, 2002). A brief description of clarity scores may be found in Table 2.3.

Score	Description
3	Meaning of sentence is perfectly clear on first reading
2	Meaning of sentence is clear only after some reflection
1	Some, although not all, meaning is able to be gleaned from the sentence with some effort
0	Meaning of sentence is not apparent, even after some reflection

Table 2.3: Interpretation of Clarity scores

2.3.4 Problems of Human Evaluation

Human evaluation are very informative, but they present several important limitations. Human evaluations are:

- **Expensive (and slow).** Human evaluations are labor-intensive and time-consuming. Human judges must often evaluate each automatic translation according to several quality criteria (e.g., adequacy, fluency, etc.). As a result, current evaluation campaigns produce human assessments only for a subset of systems and sentences (see, for instance, Tables 3.6, 3.10 and 3.13, in Chapter 3, for a numerical description of several standard evaluation test beds from recent MT evaluation campaigns). In addition, the number of system variants allowed for each participant to be selected for manual evaluation is typically limited to a primary submission.

- **Not Reusable.** While MT systems are dynamic components which may improve along time, human assessments are static, and therefore not reusable as systems improve.

- **Subjective.** Human assessments are subjective; not only because different judges may produce different quality assessments over the same test case, but also because they depend on evaluation guidelines involving several quality criteria which may differ between evaluation campaigns. Besides, assessors may consider additional knowledge (e.g., about language, about the world, etc.) which may be different among them.

- **Possibly Partial.** Most often, human assessments are limited to partial quality dimensions such as adequacy and fluency. Thus, it may well happen that a system 'A' is judged to produce more adequate outputs than a system 'B', while system 'B' is judged to produce more fluent outputs. In this case it is not clear which system exhibits the highest overall quality. We could either consider that adequacy is more important and thus say 'A' is best, or rely on fluency thus preferring system 'B'. Alternatively, we could combine different quality aspects into a single value. For instance, a common option is to use the sum of adequacy and fluency as a global measure of quality. However, in doing so we are implicitly considering that both dimensions are equally important which may not be always the case.

2.4 Automatic Evaluation

In contrast to manual evaluations, automatic evaluations are fast (vs. slow), inexpensive (vs. expensive), objective (vs. subjective), and updatable (vs. not reusable). Overall, automatic metrics allow for fast numerical evaluations on demand, which is a crucial aspect for their application in the system development cycle:

- **Error Analysis.** Automatic evaluation allows researchers to perform inexpensive and objective sentence-level evaluations, and, thus, identify problematic cases requiring improvement.

- **System Comparison.** Automatic evaluation allows for fast comparisons between different systems, or between different versions of the same system (system-level evaluation).

- **System Optimization.** Automatic evaluation allows system developers to adjust system parameters without having to elaborate expensive human assessments for each of the possible system configurations.

However, automatic evaluations are partial and often devoted to shallow aspects of quality (e.g., lexical similarity). In addition, as discussed in the beginning of the chapter, the significance of automatic evaluations depends very strongly on the availability of a heterogeneous set of reference translations.

A large number of metrics, based on different similarity criteria, have been suggested in the last decade. Most are based on comparisons between automatic and human reference translations. There exist, as well, several approaches to MT evaluation without human references (Quirk, 2004; Gamon et al., 2005; Albrecht & Hwa, 2007b). In the following, we provide an overview of the most well-known approaches to automatic MT evaluation. We distinguish between metrics which limit their scope to the lexical dimension and those which compute similarities at deeper linguistic levels.

2.4.1 Metrics based on Lexical Matching

Metrics based on computing lexical similarities (also called n-gram based metrics), are today the dominant approach to automatic MT evaluation. These metrics have demonstrated a notable ability to emulate the behavior of human evaluators over a variety of evaluation scenarios (Coughlin, 2003). All work by rewarding lexical similarity (n-gram matchings) among the system output and a set of reference translations. The main differences are related to the calculation of lexical similarity. Below, we briefly describe the most popular, grouped according to the type of measure computed.

Edit Distance Measures

These measures provide an estimate of translation quality based on the number of changes which must be applied to the automatic translation so as to transform it into a reference translation:

- **WER.** Word Error Rate (Nießen et al., 2000). This measure is based on the Levenshtein distance (Levenshtein, 1966) —the minimum number of substitutions, deletions and insertions that have to be performed to convert the automatic translation into a valid translation (i.e., a human reference).

- **PER.** Position-independent Word Error Rate (Tillmann et al., 1997). A shortcoming of the WER is that it does not allow reorderings of words. In order to overcome this problem, the position independent word error rate (PER) compares the words in the two sentences without taking the word order into account.

- **TER.** Translation Edit Rate (Snover et al., 2006). TER measures the amount of post-editing that a human would have to perform to change a system output so it exactly matches a reference translation. Possible edits include insertions, deletions, and substitutions of single words as well as shifts of word sequences. All edits have equal cost.

Precision-oriented Measures

These metrics compute lexical precision, i.e., the proportion of lexical units (typically n-grams of varying size) in the automatic translation covered by the reference translations:

- **BLEU.** Bilingual Evaluation Understudy (Papineni et al., 2001). This metric computes n-gram lexical precision among n-grams up to length 4.
- **NIST.** An improved version of BLEU by the National Institute of Standards and Technology (Doddington, 2002). The main difference with BLEU is in the way of averaging n-gram scores. While BLEU relies on a geometric mean, NIST performs an arithmetic mean. Also NIST takes into account n-grams up to length 5. In addition, NIST weights more heavily n-grams which occur less frequently, as an indicator of their higher informativeness.
- **WNM.** A variant of BLEU which weights n-grams according to their statistical salience estimated out from a large monolingual corpus (Babych & Hartley, 2004).

Recall-oriented Measures

These metrics compute lexical recall, i.e., the proportion of lexical units in the reference translations covered by the automatic translation:

- **ROUGE.** Recall-Oriented Understudy for Gisting Evaluation (Lin & Och, 2004a). ROUGE computes lexical recall among n-grams up to length 4. It also allows for considering stemming and discontinuous matchings (skip bigrams).
- **CDER.** Cover/Disjoint Error Rate; a recall-oriented measure modeling block reordering (Leusch et al., 2006). Based on the $\overline{CD}CD$ distance introduced by Lopresti and Tomkins (1997), CDER models movement of word blocks as an edit operation.

Measures Balancing Precision and Recall

These metrics combine lexical precision and recall:

- **GTM.** An F-measure (Melamed et al., 2003; Turian et al., 2003). The importance of the length of n-gram matchings may be adjusted.
- **METEOR.** An F-measure based on unigram alignment (Banerjee & Lavie, 2005). METEOR also includes a fragmentation score which accounts for word ordering. Besides, it allows for considering stemming and synonymy lookup based on WordNet (Fellbaum, 1998).
- **BLANC.** A family of trainable dynamic n-gram based evaluation metrics (Lita et al., 2005). Their algorithm performs an efficient full overlap search over variable-size non-contiguous word sequences, with the particularity that it can be optimized for highest agreement with human assessments.

- **SIA.** Stochastic Iterative Alignment, a metric based on loose sequence alignment but enhanced with alignment scores, stochastic word matching and an iterative alignment scheme (Liu & Gildea, 2006).

2.4.2 The Limits of Lexical Similarity

The use of N-gram based metrics in the context of system development has represented a significant advance in MT research in the last decade. Indeed, these metrics —particularly BLEU— have been widely accepted by the SMT research community as a 'de facto' standard evaluation procedure. However, they have also received many criticisms (Culy & Riehemann, 2003; Turian et al., 2003; Zhang et al., 2004; Zhang & Vogel, 2004; Callison-Burch et al., 2006; Koehn & Monz, 2006). For instance, Culy and Riehemann argue that although n-gram based metrics may correlate reliably with human rankings based on adequacy and fluency, they also present several deficiencies: (i) n-gram based metrics rely on a flawed model of translation, (ii) n-gram based metrics over-rate SMT systems, and (iii) poor reference translations tend to improve n-gram based scores.

The problem with n-gram based metrics is that, rather than translation quality measures, they are indeed document similarity measures. Their value as measures of translation goodness comes from the assumption that a good translation of a text will be similar to other good translations of the same text. Unfortunately, this assumption may not always hold. Although high n-gram scores are indicative of high translation quality, low n-gram scores are not necessarily indicative of poor translation quality.

Another weakness of n-gram metrics is that their reliability depends very strongly on the number of reference translations available. As explained in Section 1.1.1, natural languages allow for many different ways of expressing the same idea. In order to capture this flexibility a very large number of human reference translations would be required. Unfortunately, in most cases only a single reference translation is available. Besides, it is critical to control the type of translation represented by the reference translations (e.g., style, literality, etc.). Overall, Culy and Riehemann (2003) found that there is a complex relationship between acceptability perceived by lexical metrics and the suitability of the output. In their opinion, n-gram based metrics should be recalibrated for each language pair and text type.

Finally, lexical metrics are not well suited for sentence-level error analysis. For instance, Turian et al. (2003) criticize the applicability of BLEU. First, because it does not have a clear interpretation. Second, because it punishes very severely translations with a low level of n-gram matching[8]. Third, because in order to punish candidate translations that are too long/short, BLEU computes a heuristically motivated brevity penalty factor.

As an example on the limits of n-gram based metrics for sentence-level evaluation, Table 2.4 shows a particular case of Spanish-to-English translation in which incorrect translations receive higher scores than correct ones. Observe how highest scores are obtained by output *'B'*, which is wrong and nonsense. In contrast, output *'A'*, which conveys most of the meaning of the input, attains much lower scores. As to output *'C'*, which is completely valid, and in which only the first word is changed with respect to output *'A'*, it receives a dramatic null BLEU score.

[8] At least one 4-gram must be shared with a reference translation; otherwise BLEU score is 0

Source Text	la casa verde estaba situada justo delante del lago .			
Reference	the green house was right in front of the lake .			
		BLEU	**GTM**	**NIST**
Output *A*	the green house was by the lake shore .	0.30	0.70	2.29
Output *B*	the green potato right in front of the lake was right .	**0.52**	**0.87**	**2.90**
Output *C*	a green house was by the lake shore .	**0.00**	0.60	1.96

Table 2.4: An example on the deficiencies of n-gram based metrics

2.4.3 Beyond Lexical Similarity

Having reached a certain degree of maturity, current MT technology requires nowadays the usage of more sophisticated metrics. In the last few years, several approaches have been suggested. Some of them are based on extending the reference lexicon. For instance, ROUGE and METEOR allow for morphological variations by applying stemming. Additionally, METEOR may perform a lookup for synonymy in WordNet (Fellbaum, 1998). Others have suggested taking advantage of paraphrasing support (Russo-Lassner et al., 2005; Zhou et al., 2006; Kauchak & Barzilay, 2006; Owczarzak et al., 2006).

But these are still attempts at the lexical level. At a deeper linguistic level, we may find, for instance, the work by Liu and Gildea (2005) who introduced a series of syntax-based metrics. They developed the Syntactic Tree Matching (STM) metric based on constituency parsing, and the Head-Word Chain Matching (HWCM) metric based on dependency parsing. Also based on syntax, Mehay and Brew (2007) suggested flattening syntactic dependencies only in the reference translations so as to compute string-based similarities without requiring syntactic parsing of the possibly ill-formed automatic candidate translations. We may find as well the work by Owczarzak et al. (2007a; 2007b) who presented a metric which compares dependency structures according to a probabilistic Lexical-Functional Grammar. They used paraphrases as well. Their metric obtains very competitive results, specially as a fluency predictor. Other authors have designed metrics based on shallow-syntactic information. For instance, Popovic and Ney (2007) proposed a novel method for analyzing translation errors based on WER and PER measures computed over different parts-of-speech. At the semantic level, prior to this thesis, we know only about the 'NEE' metric defined by Reeder et al. (2001), which was devoted to measure MT quality over named entities[9].

The need for improving the performance of current metrics is also reflected by the recent organization of two evaluation shared tasks:

1. The evaluation shared-task at the *ACL 2008 Third Workshop On Statistical Machine Translation* (WMT'08)[10]. After the 2007 pilot experiment, this year, a separate shared task on automatic MT evaluation has been officially set up.

2. The *"NIST Metrics MATR Challenge 2008"*[11] organized by NIST in the context of the *8th Conference of the Association for Machine Translation in the Americas (AMTA)*.

[9]The 'NEE' metric is similar to the 'NE-M_e-$\star$' metric described in Section 3.1.5.

[10]`http://www.statmt.org/wmt08/`

[11]`http://www.nist.gov/speech/tests/metricsmatr/`

In Chapter 3, we present a very rich set of metrics operating at different linguistic levels, from the lexical, through the syntactic, and up to the level of and semantics. These metrics are successfully applied to the evaluation of heterogeneous systems and to the generation of detailed error analysis reports.

2.4.4 Metric Combinations

Integrating the scores conferred by different metrics into a single measure seems the most natural and direct way to improve over the individual quality of current metrics. This solution requires two important ingredients:

Combination Strategy, i.e., how to combine several metric scores into a single score. We distinguish between *parametric* and *non-parametric* approaches. In parametric approaches the contribution of each metric to the global score is individually weighted through an associated parameter. In contrast, in the non-parametric case, metric contribution is based on a global non-parameterized criterion.

Meta-Evaluation Criterion, i.e., how to evaluate the quality of a metric combination. As we have seen in Section 2.2.2, there exist at least two different meta-evaluation criteria: human likeness (i.e., the metric ability to discern between automatic and human translations) and human acceptability (i.e., correlation with human assessments).

In the following, we describe the most relevant approaches to metric combination. All implement a *'parametric'* combination strategy. The main difference between these methods can be found in the meta-evaluation criterion underlying. We distinguish between approaches relying on human likeness and approaches relying on human acceptability.

Approaches based on Human Likeness

The first approach to metric combination based on human likeness was that by Corston-Oliver et al. (2001) who used decision trees to distinguish between human-generated ('good') and machine-generated ('bad') translations. They suggested to use classifier confidence scores directly as a quality indicator. High levels of classification accuracy were obtained. However, they focused on evaluating only the well-formedness of automatic translations (i.e., subaspects of fluency). Preliminary results using Support Vector Machines were also discussed.

Kulesza and Shieber (2004) extended the approach by Corston-Oliver et al. (2001) to take into account other aspects of quality further than fluency alone. Instead of decision trees, they trained Support Vector Machines (SVM). They used features inspired by well-known metrics such as BLEU, NIST, WER, and PER. Metric quality was evaluated both in terms of classification accuracy and in terms of correlation with human assessments at the sentence level. A significant improvement with respect to standard individual metrics was reported.

Gamon et al. (2005) presented a similar approach which, in addition, had the interesting property that the set of human translations was not required to correspond, as references, to the set of automatic translations. Instead of human references, they used a language model estimated from a target-language corpus of the same domain.

Approaches based on Human Acceptability

In a different research line, Akiba et al. (2001) suggested directly predicting human scores of acceptability, approached as a multiclass classification task. They used decision tree classifiers trained on multiple edit-distance features based on combinations of lexical, morphosyntactic and lexical semantic information (e.g., word, stem, part-of-speech, and semantic classes from a thesaurus). Promising results were obtained in terms of local accuracy over an internal predefined set of overall quality assessment categories[12].

Quirk (2004) presented a similar approach, also with the aim to approximate human quality judgements, with the particularity that human references were not required. It relied only on human assessments[13]. They defined a rich collection of features, extracted by their syntax-based MT system itself (Quirk et al., 2005). These were grouped in three categories: (i) features related to the source sentence and how difficult it was to parse, (ii) features about the translation process itself, (iii) features accounting for the proportion of words and substrings covered by the training corpus. They applied a variety of supervised machine learning algorithms (e.g., Perceptron, SVM, decision trees, and linear regression). All proved very effective, attaining high levels of accuracy, with a significant advantage in favor of linear regression. However, experiments were run on automatic outputs by a single MT system, so it is not clear how well these would generalize.

Recently, Paul et al. (2007) extended these works so as to account for separate aspects of quality: adequacy, fluency and acceptability. They used SVM classifiers to combine the outcomes of different automatic metrics at the lexical level (BLEU, NIST, METEOR, GTM, WER, PER and TER). Their main contribution is on the variety of schemes they applied to binarize the multiclass classification problem (one-vs-all/all-pairs/boundary-based), and how the outcome by distinct classifiers is combined so as to decide on the final prediction.

Also very recently, Albrecht and Hwa (2007a; 2007b) re-examined the SVM-classification approach by Kulesza and Shieber (2004) and Corston-Oliver et al. (2001) and, inspired by the work of Quirk (2004), suggested a regression-based learning approach to metric combination, with and without human references. Their SVM-based regression model learns a continuous function that approximates human assessments in training examples. They used four kinds of features: (i) string-based metrics over references (BLEU, NIST, WER, and PER, ROUGE-inspired, METEOR-based), (ii) syntax-based metrics over references, (iii) string-based metrics over a large corpus, and (iv) syntax-based metrics over a large corpus. Their results outperformed those by Kulesza and Shieber (2004) in terms of correlation with human assessments. Besides, their method is shown to generalize reasonably well across different evaluation scenarios. They conducted two generalization studies: (i) on how well the trained metrics evaluate systems from other years and systems developed for a different source language, and (ii) on how variations in the set of training examples affect the metric's ability to generalize to distant systems.

In a different approach, Ye et al. (2007) suggested approaching sentence level MT evaluation as a ranking problem. They used the Ranking SVM algorithm to sort candidate translation on the

[12](A) Perfect: no problems in both information and grammar, (B) Fair: easy-to-understand, with either some unimportant information missing of flawed grammar, (C) Acceptable: broken, but understandable with effort, (D) Nonsense: important information has been translated incorrectly.

[13]Assessments were based on a 1-4 scale similar to overall quality categories used by Akiba et al. (2001). (4) Ideal, (3) Acceptable, (2) Possibly Acceptable, (1) Unacceptable.

basis of several distinct features of three different types: n-gram based, dependency-based, and translation perplexity according to a reference language model. A slight but significantly improved correlation with fluency human assessments was reported.

As an alternative to machine learning techniques, Liu and Gildea (2007) suggested a simpler approach based on linear combinations of metrics. They followed a *Maximum Correlation Training*, i.e., the weight for the contribution of each metric to the overall score was adjusted so as to maximize the level of correlation with human assessments at the sentence level. They showed this approach to significantly outperform that of Kulesza and Shieber (2004) in terms of correlation with human assessments.

All the methods described above implement a parametric combination scheme. In Section 3.4, we present a *non-parametric* alternative approach to metric combination in which metrics are combined without any a priori weighting of their relative importance.

Chapter 3

Towards Heterogeneous Automatic MT Evaluation

As discussed in Section 2.2.3, evaluation methods may introduce a bias in the development cycle of MT systems, which may cause serious problems. In order to overcome the metric bias problem, instead of relying on *partial* metrics, system developers should rely on *global* evaluations methods, i.e., methods which could take into account a wide range of quality aspects.

Doubtless, the design of a golden metric that is able to capture all the quality aspects that distinguish correct translations from incorrect ones is an ambitious and difficult goal. Instead, we suggest following a *divide and conquer* strategy. For that purpose, we have compiled a heterogeneous set of specialized metrics, devoted to capture partial aspects of MT quality at different linguistic levels: lexical, syntactic, and semantic[1]. Our goal is twofold: (i) to verify that partial metrics at different linguistic levels capture relevant and complementary pieces of information, and, are, thus, useful for the purpose of automatic MT evaluation, and (ii) to study how to combine the scores conferred by different metrics into a single measure of quality.

The rest of the chapter is organized as follows. First, in Section 3.1, we present the rich set of metrics employed in our experiments. These metrics are applied, in Section 3.2, to the evaluation of MT systems over different scenarios. We show how individual metrics based on deeper linguistic information are able to produce more reliable system rankings than metrics based on lexical matching alone, specially when the systems under evaluation are different in nature. These metrics present, however, an important shortcoming: they rely on automatic linguistic processors which are prone to error[2]. Thus, it could be argued that their performance would decrease when applied to low-quality translations. In order to clarify this issue, in Section 3.3, we study the performance of syntactic and semantic metrics in the extreme evaluation scenario of speech-to-speech translation between non-related languages. We show that these metrics exhibit a very robust behavior at the system level, whereas at the sentence level some of them suffer a significant decrease. In Section 3.4, we study the viability of working on metric combinations. We show that non-parametric schemes provide

[1] A complete list of metrics is available in Appendix C.

[2] A description of the tools utilized and related tag sets is available in Appendix B.

a robust means of combining metrics at different linguistic levels, effectively yielding a significantly improved evaluation quality at the sentence level. As a complementary issue, in Section 3.5, we show how the heterogeneous set of metrics can be also successfully applied to error analysis. Finally, in Section3.6, main conclusions are summarized and future work is outlined.

3.1 A Heterogeneous Set of Metrics

For our study, we have compiled a rich set of metric variants at 5 different linguistic levels (lexical, shallow-syntactic, syntactic, shallow-semantic and semantic). We have resorted to several existing metrics, and we have also developed new ones. Although from different viewpoints, and based on different similarity assumptions, in all cases, translation quality is measured by comparing automatic translations against a set of human reference translations. In the following subsections, we provide a description of the metrics according to the linguistic level at which they operate.

3.1.1 Lexical Similarity

We have included several variants from different standard metrics (e.g., BLEU, NIST, GTM, METEOR, ROUGE, WER PER and TER)[3]. Below we list all the variants included in our study:

- **BLEU-n | BLEUi-n:** Accumulated and individual BLEU scores for several n-gram levels ($n = 1...4$) (Papineni et al., 2001). We use version '11b' of the NIST MT evaluation kit[4] for the computation of BLEU scores. Seven variants are computed[5].
- **NIST-n | NISTi-n:** Accumulated and individual NIST scores for several n-gram levels ($n = 1...5$) (Doddington, 2002). We use version '11b' of the NIST MT evaluation kit for the computation of NIST scores. Nine variants are computed[6].
- **GTM-e:** General Text Matching F-measure (Melamed et al., 2003). We use GTM version 1.4. Three variants, corresponding to different values of the e parameter controlling the reward for longer matchings ($e \in \{1, 2, 3\}$), are computed.
- **METEOR:** We use METEOR version 0.6. (Banerjee & Lavie, 2005). Four variants are computed[7]:
  - **METEOR$_{exact}$** $\rightarrow$ running 'exact' module.
  - **METEOR$_{stem}$** $\rightarrow$ running 'exact' and 'porter_stem' modules, in that order. This variant considers morphological variations through the Porter stemmer (Porter, 2001).
  - **METEOR$_{wnstm}$** $\rightarrow$ running 'exact', 'porter_stem' and 'wn_stem' modules, in that order. This variant includes morphological variations obtained through WordNet (Fellbaum, 1998).

[3]The list of the variants selected is also available in Table C.1.
[4]The NIST MT evaluation kit is available at `http://www.nist.gov/speech/tests/mt/scoring/`.
[5]We use 'BLEU' to refer to the 'BLEU-4' variant. 'BLEU-1' and 'BLEUi-1' refer to the same metric variant.
[6]We use 'NIST' to refer to the 'NIST-5' variant. 'NIST-1' and 'NISTi-1' refer to the same metric variant.
[7]We use 'METEOR' to refer to the 'METEOR$_{wnsyn}$' variant.

  - **METEOR$_{wnsyn}$** $\longrightarrow$ running 'exact', 'porter_stem', 'wn_stem' and 'wn_synonymy' modules, in that order. This variant performs a lookup for synonyms in WordNet.

- **ROUGE:** We use ROUGE version 1.5.5 (Lin & Och, 2004a). We consider morphological variations through stemming. Options are '-z SPL -2 -1 -U -m -r 1000 -n 4 -w 1.2 -c 95 -d'. Eight variants are computed:
  - **ROUGE-n** $\longrightarrow$ for several n-gram lengths ($n = 1...4$)
  - **ROUGE$_L$** $\longrightarrow$ longest common subsequence (LCS).
  - **ROUGE$_{S\star}$** $\longrightarrow$ skip bigrams with no max-gap-length.
  - **ROUGE$_{SU\star}$** $\longrightarrow$ skip bigrams with no max-gap-length, including unigrams.
  - **ROUGE$_W$** $\longrightarrow$ weighted longest common subsequence (WLCS) with weighting factor $w = 1.2$.
- **WER:** Word Error Rate. We use $1 -$ WER (Nießen et al., 2000).
- **PER:** Position-independent Word Error Rate. We use $1 -$ PER (Tillmann et al., 1997).
- **TER:** Translation Edit Rate. We use $1 -$ TER (Snover et al., 2006).

3.1.2 Beyond Lexical Similarity

It is an evidence that MT quality aspects are diverse. However, metric families listed in Section 3.1.1 limit their scope to the lexical dimension. This may result, as discussed in Section 2.2.3, in unfair evaluations. For instance, let us show in Table 3.1, a real case extracted from the NIST 2005 Arabic-to-English translation exercise[8]. A high quality translation (by LinearB system) according to human assessments (adequacy = 4 / 5, fluency = 4 / 5) unfairly attains a low BLEU score (BLEU = 0.25). This is due to the low level of lexical matching. From all n-grams up to length four in the automatic translation only one 4-gram out of fifteen, two 3-grams out of sixteen, five 2-grams out of seventeen, and thirteen 1-grams out of eighteen can be found in at least one reference translation. Table 3.2 shows, for these n-grams in decreasing length order, the number of reference translations in which they occur.

The main problem with metrics based only on lexical similarities is that they are strongly dependent on the sublanguage represented by the set of human references available. In other words, their reliability depends on the heterogeneity (i.e., representativity) of the reference translations. These may in its turn depend not only on the number of references, but on their lexica, grammar, style, etc. Besides, while similarities between two sentences can take place at deeper linguistic levels, lexical metrics limit their scope to the surface. We believe that an explicit use of linguistic information could be very beneficial. Besides, current NLP technology allows for automatically obtaining such information.

Thus, we argue that the degree of overlapping at more abstract levels is a far more robust indicator of actual MT quality. For instance, Figure 3.1 compares automatically obtained syntactico-semantic representations for the automatic translation in the previous example (top) and reference

[8]The case corresponds to sentence 498 in the test set.

LinearB	On **Tuesday** several **missiles** and **mortar shells fell** in **southern Israel** , but there were **no casualties** .
Ref 1	Several **Qassam rockets** and **mortar shells** were **fired** on **southern Israel** today **Tuesday without victims** .
Ref 2	Several **Qassam rockets** and **mortars hit southern Israel** today **without causing any casualties** .
Ref 3	A number of **Qassam rockets** and **Howitzer missiles fell** over **southern Israel** today , **Tuesday**, **without causing any casualties** .
Ref 4	Several **Qassam rockets** and **mortar shells fell** today , **Tuesday** , on **southern Israel without causing any victim** .
Ref 5	Several **Qassam rockets** and **mortar shells fell** today , **Tuesday** , in **southern Israel without causing any casualties** .
Subject	Qassam rockets / Howitzer missiles / mortar shells
Action	fell / were fired / hit
Location	southern Israel
Time	Tuesday (today)
Result	no casualties / victims

Table 3.1: NIST 2005 Arabic-to-English. A Case of Analysis (sentence #498)

n-gram	#occ	n-gram	#occ	n-gram	#occ
and mortar shells fell	2	casualties .	3	shells	3
and mortar shells	3	on	2	fell	3
mortar shells fell	2	Tuesday	4	southern	5
and mortar	3	several	4	Israel	5
mortar shells	3	missiles	1	,	3
shells fell	2	and	4	casualties	3
southern Israel	5	mortar	3	.	5

Table 3.2: NIST 2005 Arabic-to-English. A Case of Analysis (sentence #498). Lexical matching

#5 (bottom)[9]. In first place, with respect to syntactic similarity, notice that a number of subtrees are shared (particularly, noun phrases and prepositional phrases). Also notice that the main verbal form ('fell') is shared. As to the semantic roles associated, predicates in both sentences share several arguments (A1, AM-TMP, and AM-LOC) with different degrees of lexical overlapping. All these features, that are making the difference in this case, are invisible to shallow metrics such as BLEU.

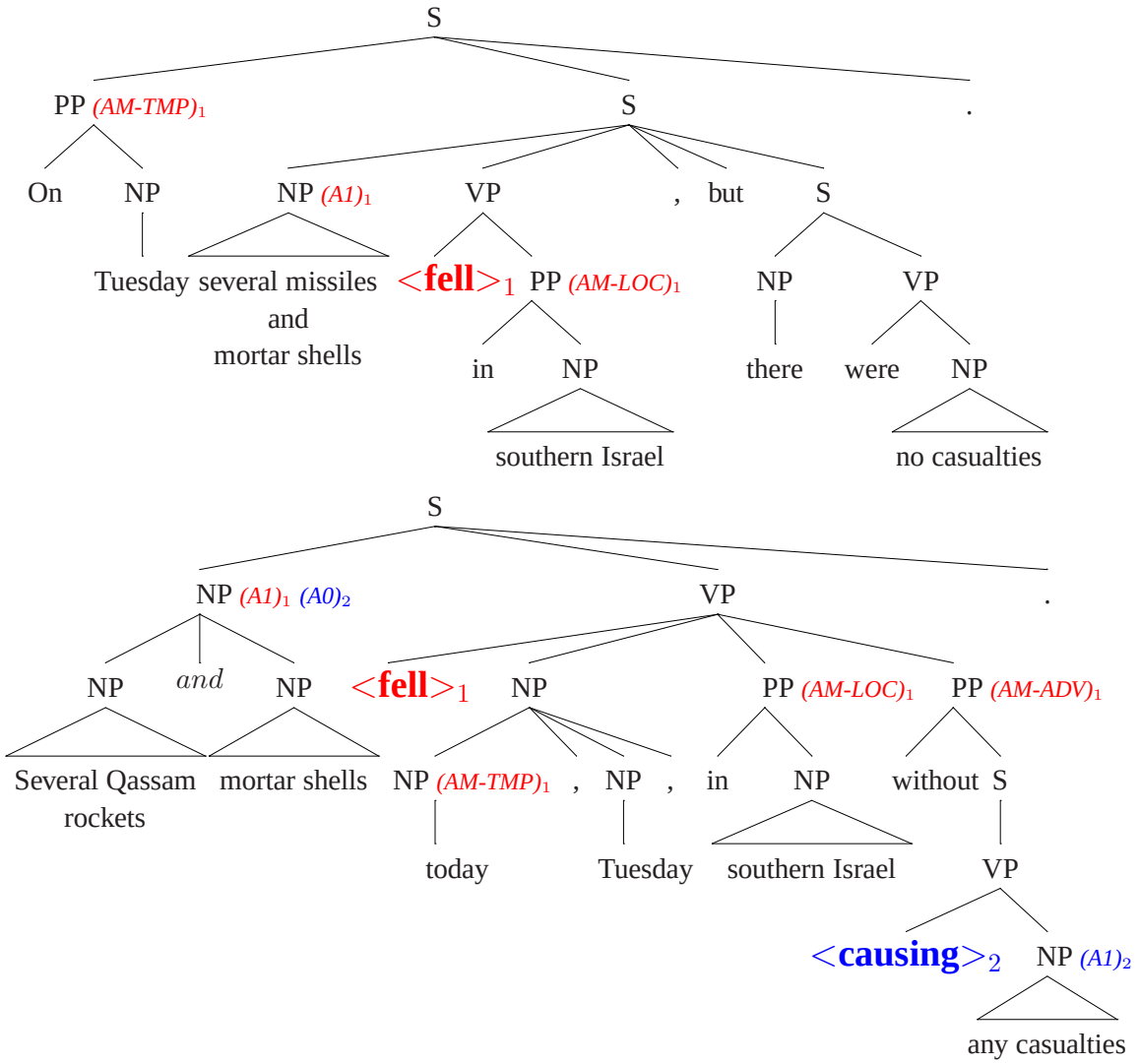


Figure 3.1: NIST 2005 Arabic-to-English. A Case of Analysis (sentence #498). Syntactico-semantic Representation

[9]Parts-of-speech and syntactic notation are based on the Penn Treebank (Marcus et al., 1993). Notation for semantic roles is based on the Proposition Bank (Palmer et al., 2005). We distinguish semantic roles associated to different verbs by indexing them with the position the related verb would occupy in a left-to-right list of verbs, starting at position 1.

Linguistic Elements

Modeling linguistic features at deeper linguistic levels requires the usage of more complex linguistic structures. We will refer to linguistic units, structures, or relationships as *linguistic elements* (LEs). Possible kinds of LEs could be, for instance, word forms, parts-of-speech, dependency relations, syntactic constituents, named entities, semantic roles, discourse representations, etc. A sentence, thus, may be seen as a bag of LEs. Each LE may consist, in its turn, of one or more LEs, which we call items inside the LE. For instance, a phrase constituent LE may consist of part-of-speech items, word form items, etc. LEs may also consist of combinations of items. For instance, a phrase constituent LE may be seen as a sequence of 'word-form:part-of-speech' items.

Hovy et al. (2006) defined a similar type of linguistic structures, so-called basic elements (BEs), for the evaluation of automated summarization systems. Their method consisted in breaking down reference sentences into sets of BEs before comparing system outputs against them. However, in contrast to LEs, they limited the information captured by BEs to the syntactic level, whereas LEs allow for representing any kind of linguistic information. Thus, BEs could be actually seen as a particular case of LEs.

Similarity Measures over Linguistic Elements

We are interested in comparing linguistic structures, and linguistic units. LEs allow for comparisons at different granularity levels, and from different viewpoints. For instance, we might compare the syntactic/semantic structure of two sentences (e.g., which verbs, semantic arguments and adjuncts exist) or we might compare lexical units according to the syntactic/semantic role they play inside the sentence. We use two very simple kinds of similarity measures over LEs: *Overlapping* and *Matching*. Below, we provide general definitions which will be instantiated over particular cases in the following subsections:

- **Overlapping** *between items inside LEs, according to their type*. Overlapping provides a rough measure of the proportion of items inside elements of a certain type that have been successfully translated. Formally:

$$\text{Overlapping}(t) = \frac{\sum_{i \in (items_t(hyp) \cap items_t(ref))} count_{hyp}(i,t)}{\sum_{i \in (items_t(hyp) \cup items_t(ref))} max(count_{hyp}(i,t), count_{ref}(i,t))}$$

  where t is the LE type, 'hyp' and 'ref' refer, respectively, to the candidate and reference translations, $items_t(s)$ refers to the set of items occurring inside LEs of type t in sentence s, and $count_s(i,t)$ denotes the number of times i appears in sentence s inside a LE of type t. LE types vary according to the specific LE class. For instance, in the case of the 'named entity' class, types may be 'PER' (i.e., person), 'LOC' (i.e., location), 'ORG' (i.e., organization),

etc. In the case of the 'semantic role' class, types may be 'A0' (i.e., prototypical subject), 'AM-TMP' (i.e., temporal adjunct), 'AM-MNR' (i.e., manner adjunct), etc.

We also introduce a coarser metric, Overlapping($\star$), which considers the averaged overlapping over all types:

$$\text{Overlapping}(\star) = \frac{\sum_{t \in T} \sum_{i \in (items_t(hyp) \cap items_t(ref))} count_{hyp}(i,t)}{\sum_{t \in T} \sum_{i \in (items_t(hyp) \cup items_t(ref))} max(count_{hyp}(i,t), count_{ref}(i,t))}$$

where T is the set of all LE types associated to the given LE class. For instance, we may define a metric which computes average lexical overlapping over all semantic roles types. This would roughly estimate to what degree translated lexical items play the expected semantic role in the context of the full candidate sentence.

- **Matching** *between items inside LEs, according to their type*. Its definition is analogous to the Overlapping definition, but in this case the relative order of the items is important. All items inside the same element are considered as a single unit (i.e., a sequence in left-to-right order). In other words, we are computing the proportion of fully translated elements, according to their type. Formally:

$$\text{Matching}(t) = \frac{\sum_{e \in (elems_t(hyp) \cap elems_t(ref))} count_{hyp}(e,t)}{\sum_{e \in (elems_t(hyp) \cup elems_t(ref))} max(count_{hyp}(e,t), count_{ref}(e,t))}$$

where t is the LE type, and $elems_t(s)$ refers to the set of LEs (as indivisible sequences of consecutive items) of type t in sentence s.

As in the case of 'Overlapping', we introduce a coarser metric, Matching($\star$), which considers the averaged matching over all types:

$$\text{Matching}(\star) = \frac{\sum_{t \in T} \sum_{e \in (elems_t(hyp) \cap elems_t(ref))} count_{hyp}(e,t)}{\sum_{t \in T} \sum_{e \in (elems_t(hyp) \cup elems_t(ref))} max(count_{hyp}(e,t), count_{ref}(e,t))}$$

Notes on Overlapping/Matching Measures

1. Overlapping and Matching operate on the assumption of a single reference translation. The reason is that, when it comes to more abstract levels, LEs inside the same sentence may be strongly interrelated, and, therefore, similarities across reference translations may not be a reliable quality indicator. The extension to the multi-reference setting is computed by assigning the maximum value attained over all human references individually.

2. Overlapping and Matching are general metrics. We may apply them to specific scenarios by defining the class of linguistic elements and items to be used. In subsections 3.1.3 to 3.1.6, these measures are instantiated over several particular cases.

3. As to abbreviated nomenclature, the first two letters of metric names identify the LE class, which indicates the level of abstraction at which they operate. In this document, we use 'SP' for shallow parsing, 'DP' for dependency parsing, 'CP' for constituency parsing, 'NE' for named entities, 'SR' for semantic roles, and 'DR' for discourse representations. Then, we find the type of similarity computed. Overlapping and Matching measures are represented by the 'O' and 'M' symbols, respectively. Additionally, these symbols may be accompanied by a subindex representing the type of LEs and items employed. For instance, 'SR-O_{rl}-$\star$' operates at the level of semantic roles (SR), and represents average Overlapping among lexical items according to their role. If the LE and item types are not specified, it is assumed that the metric computes lexical overlapping over the top-level items available. For instance, these are also valid names for the 'SR-O_{rl}-$\star$' metric: 'SR-O_r-$\star$', 'SR-O_l-$\star$', and 'SR-O-$\star$'. In the following sections and chapters, we use 'SR-O_r-$\star$' equivalent, and similarly for other metrics and LE classes.

Lexical Overlapping

We instantiate the overlapping measure at the lexical level, by defining the 'O_l' metric, which computes lexical overlapping directly over word forms. As an example, Table 3.3 shows the computation of the 'O_l' score for the case depicted in Figure 3.1, as compared to lexical precision, recall and F-measure. A and H denote, respectively, the automatic translation and the human reference. Text has been lower cased. It can be observed that lexical overlapping is, indeed, just another simple method for balancing precision and recall.

An Example Beyond the Lexical Level

Table 3.4 shows an example on how to compute average lexical overlapping among semantic roles, i.e., SR-O_r-($\star$), for the case depicted in Figure 3.1. The semantic role labeler detected one argument ('A1$_1$') and two adjuncts ('AM-TMP$_1$' and 'AM-LOC$_1$') in the automatic translation, whereas three arguments ('A1)$_1$', 'A0$_2$', and 'A1$_2$') and three adjuncts ('AM-TMP$_1$', 'AM-LOC$_1$' and 'AM-ADV$_1$') were detected for the human reference. Associated LE representations are showed for each LE type. We also provide individual lexical overlapping scores, and average overlapping.

A on **tuesday several** missiles **and mortar shells fell in southern israel ,** but there were no **casualties .**

H **several** qassam rockets **and mortar shells fell** today , **tuesday , in southern israel** without causing any **casualties .**

$A \cap H = \{$ ‘tuesday’, ‘several’, ‘and’, ‘mortar’, ‘shells’, ‘fell’, ‘in’, ‘southern’, ‘israel’, ‘,’, ‘casualties’, ‘.’ $\}$

$H \cup H = \{$ ‘on’, ‘tuesday’, ‘several’, ‘missiles’, ‘and’, ‘mortar’, ‘shells’, ‘fell’, ‘in’, ‘southern’, ‘israel’, ‘,’, ‘but’, ‘there’, ‘were’, ‘no’, ‘casualties’, ‘.’, ‘qassam’, ‘rockets’, ‘today’, ‘,’, ‘without’, ‘causing’, ‘any’ $\}$

$O_l = \frac{|A| \cap |H|}{|A| \cup |H|} = \frac{12}{25}$ $P = \frac{|A| \cap |H|}{|A|} = \frac{12}{18}$ $R = \frac{|A| \cap |H|}{|H|} = \frac{12}{19}$ $F = \frac{P+R}{2} = \frac{\frac{12}{18}+\frac{12}{19}}{2}$

Table 3.3: Lexical overlapping score for the case from Table 3.1

$\mathrm{A}_{\mathrm{A1}} = \{$ **‘several’**, ‘missiles’, **‘and’**, **‘mortar’**, **‘shells’** $\}$
$\mathrm{H}_{\mathrm{A1}} = \{$ ‘several’, ‘qassam’, ‘rockets, ‘and’, ‘mortar’, ‘shells’, ‘any’, ‘casualties’ $\}$

$\mathrm{A}_{\mathrm{A0}} = \emptyset$
$\mathrm{H}_{\mathrm{A0}} = \{$ ‘several’, ‘qassam’, ‘rockets, ‘and’, ‘mortar’, ‘shells’ $\}$
$\mathrm{A}_{\mathrm{AM\text{-}TMP}} = \{$ ‘on’, ‘tuesday’ $\}$
$\mathrm{H}_{\mathrm{AM\text{-}TMP}} = \{$ ‘today’ $\}$
$\mathrm{A}_{\mathrm{AM\text{-}LOC}} = \{$ **‘in’**, **‘southern’**, **‘israel’** $\}$
$\mathrm{H}_{\mathrm{AM\text{-}LOC}} = \{$ ‘in’, ‘southern’, ‘israel’ $\}$
$\mathrm{A}_{\mathrm{AM\text{-}ADV}} = \emptyset$
$\mathrm{H}_{\mathrm{AM\text{-}ADV}} = \{$ ‘without’, ‘causing’, ‘any’, ‘casualties’ $\}$

SR-O_r(A1) $= \frac{4}{9}$
SR-O_r(A0) $= \frac{0}{6}$
SR-O_r(AM-TMP) $= \frac{0}{3}$
SR-O_r(AM-LOC) $= \frac{3}{3}$
SR-O_r(AM-ADV) $= \frac{0}{4}$

SR-$O_r(\star) = \frac{4+0+0+3+0}{9+6+3+3+4} = \frac{7}{25}$

Table 3.4: Average semantic role (lexical) overlapping score for the case from Table 3.1

3.1.3 Shallow Syntactic Similarity

Metrics based on shallow parsing (*SP*) analyze similarities at the level of parts-of-speech (PoS), word lemmas, and base phrase chunks. Sentences are automatically annotated using the SVMTool (Giménez & Màrquez, 2004b), Freeling (Carreras et al., 2004) and Phreco (Carreras et al., 2005) linguistic processors, as described in Appendix B, Section B.1. We instantiate 'Overlapping' over parts-of-speech and chunk types. The goal is to capture the proportion of lexical items correctly translated, according to their shallow syntactic realization. Two metrics have been defined:

SP-O_p-t Lexical overlapping according to the part-of-speech 't'. For instance, 'SP-O_p-NN' roughly reflects the proportion of correctly translated singular nouns, whereas 'SP-O_p-VBN' reflects the proportion of correctly translated past participles. We also define the 'SP-O_p-$\star$' metric, which computes the average lexical overlapping over all parts-of-speech.

SP-O_c-t Lexical overlapping according to the base phrase chunk type 't'. For instance, 'SP-O_c-NP', and 'SP-O_c-VP' respectively reflect the successfully translated proportion of noun and verb phrases. We also define the 'SP-O_c-$\star$' metric, which computes the average lexical overlapping over all chunk types.

At a more abstract level, we use the NIST metric (Doddington, 2002) to compute accumulated/individual scores over sequences of:

SP-NIST(i)-n Lemmas.

SP-NIST(i)$_p$-n Parts-of-speech.

SP-NIST(i)$_c$-n Base phrase chunks.

SP-NIST(i)$_{iob}$-n Chunk IOB labels[10].

For instance, 'SP-NIST$_l$-5' corresponds to the accumulated NIST score for lemma n-grams up to length 5, whereas 'SP-NISTi$_p$-5' corresponds to the individual NIST score for PoS 5-grams. 'SP-NIST$_{iob}$-2' corresponds to the accumulated NIST score for IOB n-grams up to length 2, whereas 'SP-NISTi$_c$-4' corresponds to the individual NIST score for chunk 4-grams. A complete list of SP metric variants is available in Appendix C, Table C.2.

3.1.4 Syntactic Similarity

On Dependency Parsing (DP)

DP metrics capture similarities between dependency trees associated to automatic and reference translations. Dependency trees are obtained using the MINIPAR parser (Lin, 1998), as described in Appendix B, Section B.2. We use two types of metrics:

[10] IOB labels are used to denote the position (Inside, Outside, or Beginning of a chunk) and, if applicable, the type of chunk.

DP-$O_l|O_c|O_r$ These metrics compute lexical overlapping between dependency trees from three different viewpoints:

DP-O_l-l Overlapping between words hanging at the same level, $l \in [1..9]$, or deeper. For instance, 'DP-O_l-4' reflects lexical overlapping between nodes hanging at level 4 or deeper. Additionally, we define the 'DP-O_l-$\star$' metric, which corresponds to the averaged values over all levels.

DP-O_c-t Overlapping between words *directly hanging* from terminal nodes (i.e., grammatical categories) of type 't'. For instance, 'DP-O_c-A' reflects lexical overlapping between terminal nodes of type 'A' (Adjective/Adverbs). Additionally, we define the 'DP-O_c-$\star$' metric, which corresponds to the averaged values over all categories.

DP-O_r-t Overlapping between words ruled by non-terminal nodes (i.e., grammatical relations) of type 't'. For instance, 'DP-O_r-s' reflects lexical overlapping between subtrees of type 's' (subject). Additionally, we define the 'DP-O_r-$\star$' metric, which corresponds to the averaged values over all relation types.

DP-HWC(i)-l This metric corresponds to the Head-Word Chain Matching (HWCM) metric presented by Liu and Gildea (2005). All head-word chains are retrieved. The fraction of matching head-word chains of a given length, $l \in [1..9]$, between the candidate and the reference translation is computed. Average accumulated scores up to a given chain length may be used as well. Opposite to the formulation by Liu and Gildea, in our case reference translations are considered individually. Moreover, we define three variants of this metric according to the items head-word chains may consist of:

DP-HWC(i)$_w$-l chains consist of words.

DP-HWC(i)$_c$-l chains consist of grammatical categories, i.e., parts-of-speech.

DP-HWC(i)$_r$-l chains consist of grammatical relations.

For instance, 'DP-HWCi$_w$-4' retrieves the proportion of matching word-chains of length-4, whereas 'DP-HWC$_w$-4' retrieves average accumulated proportion of matching word-chains *up to* length-4. Analogously, 'DP-HWC$_c$-4', and 'DP-HWC$_r$-4' compute average accumulated proportion of category/relation chains up to length-4.

The extension of 'DP-HWC' metrics to the multi-reference setting is computed by assigning to each metric the maximum value attained when individually comparing to all the trees associated to the different human references.

A complete list of DP metric variants is available in Appendix C, Table C.3.

On Constituency Parsing (CP)

CP metrics analyze similarities between constituency parse trees associated to automatic and reference translations. Constituency trees are obtained using the Charniak-Johnson's Max-Ent reranking

parser (Charniak & Johnson, 2005), as described in Appendix B, Section B.2. Three types of metrics are defined:

CP-STM(i)-l This metric corresponds to the Syntactic Tree Matching (STM) metric presented by Liu and Gildea (2005). All syntactic subpaths in the candidate and the reference trees are retrieved. The fraction of matching subpaths of a given length, $l \in [1..9]$, is computed. Average accumulated scores up to a given tree depth d may be used as well. For instance, 'CP-STMi-5' retrieves the proportion of length-5 matching subpaths. Average accumulated scores may be computed as well. For instance, 'CP-STM-9' retrieves average accumulated proportion of matching subpaths up to length-9.

The extension of the 'CP-STM' metrics to the multi-reference setting is computed by assigning to each metric the maximum value attained when individually comparing to all the trees associated to the different human references.

CP-O_p-t Similarly to the 'SP-O_p' metric, this metric computes lexical overlapping according to the part-of-speech 't'.

CP-O_c-t These metrics compute lexical overlapping according to the phrase constituent type 't'. The difference between these metrics and 'SP-O_c-t' variants is in the phrase scope. In contrast to base phrase chunks, constituents allow for phrase embedding and overlapping.

We also define the 'CP-O_p-$\star$' and 'CP-O_c-$\star$' metrics, which compute the average lexical overlapping over all parts-of-speech and phrase constituents, respectively.

A complete list of CP metric variants is available in Appendix C, Table C.4.

3.1.5 Shallow Semantic Similarity

We have designed two new families of metrics, *NE* and *SR*, which are intended to capture similarities over Named Entities (NEs) and Semantic Roles (SRs), respectively.

On Named Entities (NE)

NE metrics analyze similarities between automatic and reference translations by comparing the NEs which occur in them. Sentences are automatically annotated using the BIOS package (Surdeanu et al., 2005), as described in Appendix B, Section B.3. BIOS requires at the input shallow parsed text, which is obtained as described in Section 3.1.3. At the output, BIOS returns the text enriched with NE information. We have defined two types of metrics:

NE-O_e-t Lexical overlapping between NEs according to their type t. For instance, 'NE-O_e-PER' reflects lexical overlapping between NEs of type 'PER' (i.e., person), which provides a rough estimate of the successfully translated proportion of person names. We also define the 'NE-O_e-$\star$' metric, which considers the average lexical overlapping over all NE types. Note that this metric considers only actual NEs, i.e., it excludes the NE type 'O' (Not-a-NE). Thus, this metric is useless when no NEs appear in the translation. In order to improve its recall, we

introduce the 'NE-O_e-$\star\star$' variant, which , considers overlapping among all items, including those of type 'O'.

NE-M_e-t Lexical matching between NEs according to their type t. For instance, 'NE-M_e-LOC' reflects the proportion of fully translated NEs of type 'LOC' (i.e., location). The 'NE-M_e-$\star$' metric considers the average lexical matching over all NE types, excluding type 'O'.

A complete list of NE metric variants is available in Appendix C, Table C.5.

On Semantic Roles (SR)

SR metrics analyze similarities between automatic and reference translations by comparing the SRs (i.e., arguments and adjuncts) which occur in the predicates. Sentences are automatically annotated using the SwiRL package (Surdeanu & Turmo, 2005), as described in Appendix B, Section B.3. This package requires at the input shallow parsed text enriched with NEs, which is obtained as described in Section 3.1.5. At the output, SwiRL returns the text annotated with SRs following the notation of the Proposition Bank (Palmer et al., 2005). We have defined three types of metrics:

SR-O_r-t Lexical overlapping between SRs according to their type t. For instance, 'SR-O_r-A0' reflects lexical overlapping between 'A0' arguments. We also consider 'SR-O_r-$\star$', which computes the average lexical overlapping over all SR types.

SR-M_r-t Lexical matching between SRs according to their type t. For instance, the metric 'SR-M_r-AM-MOD' reflects the proportion of fully translated modal adjuncts. Again, 'SR-M_r-$\star$' considers the average lexical matching over all SR types.

SR-O_r This metric reflects role overlapping, i.e., overlapping between semantic roles independently from their lexical realization.

Note that in the same sentence several verb predicates, with their respective argument structures, may co-occur. However, the metrics described above do not distinguish between SRs associated to different verbs. In order to account for such a distinction we introduce a more restrictive version of these metrics ('SR-M_{rv}-t', 'SR-O_{rv}-t', 'SR-M_{rv}-$\star$', 'SR-O_{rv}-$\star$', and 'SR-O_{rv}'), which require SRs to be associated to the same verb.

A complete list of SR metric variants is available in Appendix C, Table C.6.

3.1.6 Semantic Similarity

On Discourse Representations (DR)

At the properly semantic level, we have developed a novel family of metrics based on the Discourse Representation Theory (DRT) by Kamp (1981). DRT is a theoretical framework offering a representation language for the examination of contextually dependent meaning in discourse. A discourse is represented in a discourse representation structure (DRS), which is essentially a variation of first-order predicate calculus —its forms are pairs of first-order formulae and the free variables that occur

in them. *DR* metrics analyze similarities between automatic and reference translations by comparing their respective DRSs. Sentences are automatically analyzed using the C&C Tools (Clark & Curran, 2004), as described in Appendix B, Section B.4. DRS are viewed as semantic trees. As an example, Table 3.5 shows the DRS for "*Every man loves Mary.*".

```
drs([[4]:Y],
    [[4]:named(Y,mary,per,0),
     [1]:imp(drs([[1]:X],
                 [[2]:pred(X,man,n,1)]),
             drs([[3]:E],
                 [[3]:pred(E,love,v,0),
                  [3]:rel(E,X,agent,0),
                  [3]:rel(E,Y,patient,0)]))])
```

```
 ________          ____________________   ________________
| x0     |        | x1                 | | x2             |
|________|        |____________________| |________________|
| man(x0) | ==> (| named(x1,mary,per) |A| love(x2)       |)
|________|        |____________________| | event(x2)      |
                                         | agent(x2,x0)   |
                                         | patient(x2,x1) |
                                         |________________|
```

Table 3.5: An example of DRS-based semantic tree

We have defined three groups of metrics over DRSs:

DR-STM(i)-l This metric is similar to the *'STM'* metric defined by Liu and Gildea (2005), in this case applied to DRSs instead of constituent trees. All semantic subpaths in the candidate and the reference trees are retrieved. The fraction of matching subpaths of a given length, $l \in [1..9]$, is computed. Average accumulated scores up to a given tree depth d may be used as well. For instance, 'DR-STMi-5' retrieves the proportion of length-5 matching subpaths. Average accumulated scores may be computed as well. For instance, 'DR-STM-9' retrieves average accumulated proportion of matching subpaths up to length-9.

DR-O_r-t These metrics compute lexical overlapping between discourse representation structures (i.e., discourse referents and discourse conditions) according to their type 't'. For instance, 'DR-O_r-pred' roughly reflects lexical overlapping between the referents associated to predicates (i.e., one-place properties), whereas 'DR-O_r-imp' reflects lexical overlapping between referents associated to implication conditions. We also introduce a the 'DR-O_r-$\star$' metric, which computes average lexical overlapping over all DRS types.

DR-O_{rp}-t These metrics compute morphosyntactic overlapping (i.e., between grammatical categories –parts-of-speech– associated to lexical items) between discourse representation structures of the same type. We also define the 'DR-O_{rp}-$\star$' metric, which computes average morphosyntactic overlapping over all DRS types.

Note that in the case of some complex conditions, such as implication or question, the respective order of the associated referents in the tree is important. We take this aspect into account by making the order information explicit in the construction of the semantic tree. We also make explicit the type, symbol, value and date of conditions which have type, symbol, value or date, such as predicates, relations, named entities, time expressions, cardinal expressions, or anaphoric conditions.

A complete list of DR metric variants is available in Appendix C, Table C.7.

3.2 Automatic Evaluation of Heterogeneous MT Systems

Most metrics used in the context of automatic MT evaluation are based on the assumption that *acceptable* translations tend to share the lexicon (i.e., word forms) in a predefined set of manual reference translations. This assumption works well in many cases. However, as we have seen in Section 2.2.3, several results in recent MT evaluation campaigns have cast serious doubts on its general validity. For instance, Callison-Burch et al. (2006) and Koehn and Monz (2006) reported and analyzed several cases of strong disagreement between system rankings provided by human assessors and those produced by the BLEU metric (Papineni et al., 2001). In particular, they noted that when the systems under evaluation are different in nature (e.g., rule-based vs. statistical or human-aided vs. fully automatic) BLEU may be not a reliable MT quality indicator. The reason is that BLEU favors MT systems which share the expected reference lexicon (e.g., statistical systems), and penalizes those which use a different one.

Indeed, as discussed in Section 2.2.3, the underlying cause is much simpler. In general, lexical similarity is not a sufficient neither a necessary condition so that two sentences convey the same meaning. On the contrary, natural languages are expressive and ambiguous at different levels. Consequently, the similarity between two sentences may involve different linguistic dimensions.

Hence, in the cases in which lexical metrics fail to capture actual translation quality, it should still be possible to obtain reliable evaluation results by analyzing similarities at more abstract linguistic levels. In order to verify this hypothesis, we have conducted a comparative study on the behavior of the set of metrics described in Section 3.1 applied, among other scenarios, to the evaluation puzzles described in Section 2.2.3.

3.2.1 Experimental Setting

We have selected a set of coarse-grained metric variants (i.e., accumulated/average scores over linguistic units and structures of different kinds). We distinguish different evaluation contexts. First, we study the case of a single reference translation being available. In principle, this scenario should diminish the reliability of metrics based on lexical matching alone, and favor metrics based on deeper linguistic features. Second, we study the case of several reference translations available. In this scenario, the deficiencies caused by the shallowness of metrics based on lexical matching should be less apparent. We also analyze separately the case of *homogeneous* systems (i.e., all systems being of the same nature), and the case of *heterogeneous* systems (i.e., there exist systems based on different paradigms).

As to the metric meta-evaluation criterion, since we count on human assessments, metrics can be evaluated on the basis of human acceptability. Specifically, we use Pearson correlation coefficients

	In-domain			Out-of-domain			
	#Snt	Adeq.	Fluen.	#Snt	Adeq.	Fluen.	#Systems
French-to-English	2,247	3.81	3.37	1,274	3.39	3.03	11/14
German-to-English	2,401	3.20	2.96	1,535	2.93	2.70	10/12
Spanish-to-English	1,944	3.87	3.33	1,070	3.56	3.06	11/15

Table 3.6: WMT 2006 Shared Task. Test bed description

between metric scores and human assessments at the document level. The reason is that the purpose of our work is to provide more reliable system rankings. In order to avoid biasing towards either adequacy or fluency, we use the average sum of adequacy and fluency assessments as a global measure of quality, thus, assigning both dimensions equal importance.

3.2.2 Single-reference Scenario

We use some of the test beds corresponding to the *NAACL 2006 Workshop on Statistical Machine Translation (WMT 2006)* (Koehn & Monz, 2006)[11]. Since most of the linguistic features described in Section 3.1 are so far implemented only for the case of English, among the 12 translation tasks available, we studied only the 6 tasks corresponding to the Foreign-to-English direction. These correspond to three language pairs: French-to-English (fr-en), German-to-English (de-en) and Spanish-to-English (es-en); in and out of domain.

A single reference translation is available. System outputs consist of 2,000 and 1,064 sentences for the 'in-domain' and 'out-of-domain' test beds, respectively. In each case, human assessments on adequacy and fluency are available for a subset of systems and sentences. Table 3.6 shows the number of sentences assessed in each case. Each sentence was evaluated by two different human judges. System scores have been obtained by averaging over all sentence scores. In order to give the reader an idea of the translation quality exhibited by automatic systems, average adequacy and fluency scores are also provided ('Adeq.' and 'Fluen.' columns, respectively). The '#Systems' column shows the number of systems counting on human assessments with respect to the total number of systems which presented to each task.

Table 3.7 shows meta-evaluation results on the basis of human acceptability for some metric representatives at different linguistic levels. Highest correlation coefficients attained in each task appear highlighted. In four of the six translation tasks under study, all the systems are statistical except 'SYSTRAN', which is rule-based. This is the case of the German/French-to-English in-domain/out-of-domain tasks (columns 1-4). Although the four cases are different, we have identified several regularities. For instance, BLEU and, in general, all metrics based on lexical matching alone, except METEOR, obtain significantly lower levels of correlation than metrics based on deeper linguistic similarities. The problem with lexical metrics is that they are unable to capture the actual quality of the 'SYSTRAN' system. Interestingly, METEOR obtains a higher correlation, which, in the case of French-to-English, rivals the top-scoring metrics based on deeper linguistic features. The reason, however, does not seem to be related to its additional linguistic operations

[11] http://www.statmt.org/wmt06/

		Heterogeneous				Homogeneous	
		de-en		fr-en		es-en	
Level	**Metric**	**in**	**out**	**in**	**out**	**in**	**out**
Lexical	1-WER	0.35	0.37	0.75	0.76	0.89	0.84
	1-PER	0.59	0.43	0.73	0.67	0.83	0.79
	1-TER	0.52	0.47	0.76	0.78	0.89	0.86
	BLEU	0.62	0.67	0.73	0.88	0.90	**0.88**
	NIST	0.57	0.63	0.75	0.83	0.89	0.86
	GTM ($e = 1$)	0.67	0.71	0.85	0.88	0.87	0.83
	GTM ($e = 2$)	0.55	0.68	0.73	0.88	**0.91**	0.87
	O_l	0.67	0.71	0.85	0.87	0.87	0.83
	ROUGE$_W$	0.66	0.79	0.85	**0.95**	**0.91**	**0.88**
	METEOR$_{exact}$	0.72	0.67	0.90	**0.94**	0.88	0.84
	METEOR$_{wnsyn}$	0.78	0.82	**0.94**	**0.95**	0.87	0.85
Shallow Syntactic	SP-O_p-$\star$	0.66	0.72	0.82	0.90	0.89	0.82
	SP-O_c-$\star$	0.67	0.77	0.82	0.91	0.90	0.86
	SP-NIST$_l$	0.58	0.64	0.75	0.83	0.89	0.86
	SP-NIST$_p$	0.79	0.75	0.76	**0.93**	0.87	0.87
	SP-NIST$_{iob}$	0.80	0.66	0.70	0.81	0.86	0.82
	SP-NIST$_c$	0.78	0.55	0.73	0.89	0.86	0.84
Syntactic	DP-O_l-$\star$	0.83	0.84	0.89	**0.95**	0.92	0.86
	DP-O_c-$\star$	**0.91**	**0.88**	0.92	**0.95**	0.89	0.85
	DP-O_r-$\star$	**0.94**	**0.88**	0.89	**0.98**	0.90	0.86
	DP-HWC$_w$-4	0.64	0.73	0.78	0.89	**0.95**	0.84
	DP-HWC$_c$-4	**0.91**	0.73	**0.96**	**0.97**	**0.93**	**0.89**
	DP-HWC$_r$-4	**0.93**	0.75	**0.96**	**0.98**	**0.93**	**0.90**
	CP-O_p-$\star$	0.66	0.74	0.82	0.90	0.89	0.82
	CP-O_c-$\star$	0.71	0.79	0.86	**0.93**	**0.91**	0.85
	CP-STM-4	0.88	**0.86**	0.90	**0.97**	0.90	0.84
	CP-STM-5	0.90	**0.87**	0.91	**0.97**	0.90	0.85
	CP-STM-9	**0.95**	**0.87**	**0.96**	**0.96**	0.91	0.87
Shallow Semantic	NE-O_e-$\star$	**0.94**	0.64	0.79	0.77	0.72	0.71
	NE-Me-$\star$	**0.95**	0.69	0.81	0.81	0.76	0.77
	NE-O_e-$\star\star$	0.64	0.71	0.82	0.89	0.89	0.82
	SR-O_r-$\star$	**0.93**	**0.88**	0.89	**0.95**	**0.92**	**0.91**
	SR-M_r-$\star$	**0.93**	0.81	0.82	**0.96**	0.86	0.82
	SR-O_{rv}-$\star$	0.78	**0.88**	0.81	**0.94**	**0.91**	**0.90**
	SR-M_{rv}-$\star$	0.74	0.82	0.75	**0.97**	0.89	0.85
	SR-O_r	**0.93**	0.78	**0.96**	0.89	**0.92**	**0.91**
	SR-O_{rv}	0.83	**0.88**	0.84	**0.93**	**0.91**	**0.91**
Semantic	DR-O_r-$\star$	0.68	0.80	0.77	0.87	**0.91**	0.86
	DR-O_{rp}-$\star$	**0.92**	0.85	0.86	**0.92**	0.90	0.85
	DR-STM-4	0.80	**0.87**	0.87	0.87	0.90	0.84
	DR-STM-9	**0.93**	**0.89**	**0.96**	0.89	0.89	0.85

Table 3.7: WMT 2006 Shared Task. Meta-evaluation results based on human acceptability at the system level

(i.e., stemming or synonymy lookup), but rather to the METEOR matching strategy itself (unigram precision/recall).

Metrics at the shallow syntactic level are in the same range of lexical metrics. At the syntactic level, metrics obtain in most cases high correlation coefficients. However, 'DP-HWC$_w$-4', which, although from the viewpoint of dependency relationships, still considers only lexical matching, obtains a lower level of correlation. This reinforces the idea that metrics based on rewarding long n-grams matchings may not be a reliable quality indicator in these cases.

At the level of shallow semantics, while NE metrics are not equally useful in all cases, SR metrics prove very effective. For instance, correlation attained by 'SR-O_r-$\star$' reveals that it is important to translate lexical items according to the semantic role they play inside the sentence. Moreover, correlation attained by the 'SR-M_r-$\star$' metric is a clear indication that in order to achieve a high quality, it is important to 'fully' translate 'whole' semantic structures (i.e., arguments/adjuncts). The existence of all the semantic structures ('SR-O_r'), specially associated to the same verb ('SR-O_{rv}'), is also important.

At the semantic level, metrics based on discourse representations attain also high levels of correlation, although the 'DR-O_r-$\star$' metric, which computes average lexical overlapping over DR structures, exhibits only a modest performance.

In the two remaining tasks, Spanish-to-English in-domain/out-of-domain, all the systems are statistical (columns 5-6 in Table 3.7). In this case, BLEU proves very effective, both in-domain and out-of-domain. Indeed, all metrics based on lexical matching obtain high levels of correlation with human assessments. However, still metrics based on deeper linguistic analysis attain, in general, higher correlation coefficients, although the difference is not as significant as in the case of heterogeneous systems.

3.2.3 Multiple-reference Scenario

We study the case reported by Callison-Burch et al. (2006) in the context of the Arabic-to-English exercise of the *2005 NIST MT Evaluation Campaign*[12] (Le & Przybocki, 2005)[13]. All systems are statistical except *LinearB*, a human-aided MT system (Callison-Burch, 2005). Five reference translations are available. System outputs consist of 1,056 sentences. For six of the systems we counted on a subjective manual evaluation based on adequacy and fluency over a subset of 266 sentences, thus, summing up to a total of 1,596 cases assessed. Each case was evaluated by two different human judges. System scores have been obtained by averaging over all sentence scores.

Table 3.8 shows the level of correlation with human assessments for some metric representatives (see 'ALL' column). In this case, lexical metrics obtain extremely low levels of correlation. Again, the problem is that lexical metrics are unable to capture the actual quality of LinearB. At the shallow syntactic level, only metrics which do not consider any lexical information ('SP-NIST$_p$' and 'SP-NIST$_c$') attain a significantly higher quality. At the syntactic level, all metrics attain a higher correlation. In particular head-word chain matching over grammatical relations ('DP-HWC$_r$') proves very effective. At the shallow semantic level, again, while NE metrics are not specially useful, SR metrics exhibit a high degree of correlation. Finally, at the semantic level, DR metrics obtain also

[12] `http://www.nist.gov/speech/tests/summaries/2005/mt05.htm`

[13] A brief numerical description is available in Table 3.13, column 3.

Level	Metric	ALL	SMT
Lexical	1-WER	-0.50	0.69
	1-PER	-0.35	0.75
	1-TER	-0.40	0.74
	BLEU	0.06	0.83
	NIST	0.04	0.81
	GTM ($e = 1$)	0.03	**0.92**
	GTM ($e = 2$)	0.16	0.89
	O_l	-0.15	0.80
	ROUGE$_W$	0.11	0.83
	METEOR$_{exact}$	0.03	0.88
	METEOR$_{wnsyn}$	0.05	**0.90**
Shallow Syntactic	SP-O_p-$\star$	0.03	0.84
	SP-O_c-$\star$	0.04	0.88
	SP-NIST$_l$	0.04	0.82
	SP-NIST$_p$	0.42	0.89
	SP-NIST$_{iob}$	0.49	0.82
	SP-NIST$_c$	0.44	0.68
Syntactic	DP-O_l-$\star$	0.51	**0.94**
	DP-O_c-$\star$	0.53	**0.91**
	DP-O_r-$\star$	**0.72**	**0.93**
	DP-HWC$_w$-4	0.52	0.86
	DP-HWC$_c$-4	**0.80**	0.75
	DP-HWC$_r$-4	**0.88**	0.86
	CP-O_p-$\star$	-0.02	0.84
	CP-O_c-$\star$	0.11	0.80
	CP-STM-4	0.54	**0.91**
	CP-STM-5	0.61	**0.91**
	CP-STM-9	**0.72**	**0.93**
Shallow Semantic	NE-O_e-$\star$	0.24	0.83
	NE-Me-$\star$	0.33	0.79
	NE-O_e-$\star\star$	-0.06	0.80
	SR-O_r-$\star$	0.54	0.83
	SR-M_r-$\star$	**0.68**	**0.91**
	SR-O_{rv}-$\star$	0.41	0.81
	SR-M_{rv}-$\star$	0.61	**0.92**
	SR-O_r	**0.66**	0.75
	SR-O_{rv}	0.46	0.81
Semantic	DR-O_r-$\star$	0.51	0.89
	DR-O_{rp}-$\star$	**0.81**	**0.95**
	DR-STM-4	**0.72**	**0.90**
	DR-STM-9	**0.69**	**0.91**

Table 3.8: NIST 2005. Arabic-to-English. Meta-evaluation results based on human acceptability at the system level

high correlation coefficients, with a special mention for the variant dealing with morphosyntactic overlapping over discourse representations ('DR-O_{rp}-$\star$').

On the other hand, if we remove the output by the LinearB system (see 'SMT' column), lexical metrics attain a much higher correlation, in the same range of metrics based on deeper linguistic information. However, still metrics based on syntactic parsing, semantic roles, and discourse representations, exhibit, in general, a slightly higher quality.

3.2.4 The WMT 2007 Shared Task

Recently, together with other metric developers, we participated in a pilot meta-evaluation experiment in the context of the *ACL 2007 Second Workshop On Statistical Machine Translation* (WMT'07) (Callison-Burch et al., 2007)[14]. In particular, we submitted two of our metrics, 'DP-O_r-$\star$' (i.e., dependency overlap) and 'SR-O_r-$\star$' (i.e., semantic role overlap) to the evaluation of the results of the shared task on translation between several European languages. Metric quality was evaluated in terms of correlation with human assessments at the system level. Several quality criteria were used:

Adequacy and fluency on a 1-5 scale (see Section 2.3.2).

Ranking of translation sentences. Judges were asked to rank sentence translations from best to worst relative to the other choices (ties were allowed).

Ranking of translation constituents. Judges were asked to rank only the translation of highlighted parts of the sentences. These were selected based on automatic constituency parsing and word alignments, according to the following criteria:

- the constituent could not be the whole source sentence.
- the constituent had to be longer than three words, and no longer than fifteen words.
- the constituent had to have a corresponding phrase with a consistent word alignment in each of the translations.

Reproducing results in (Callison-Burch et al., 2007), Table 3.9, shows the average correlation over all the Foreign-to-English translation tasks between metric scores and human assessments according to these quality criteria for all automatic metrics presented to the evaluation. This involved four tasks: Czech-English, French-English, German-English and Spanish-English. Two test sets were available for each task but for the Czech-English tasks which was only evaluated 'out of domain'. Metrics are sorted according to their level of correlation in decreasing order. It can be observed that the the 'SR-O_r-$\star$' metric exhibited the highest overall correlation among all metrics, and the top-correlation in three quality criteria, being second in the fourth.

[14] `http://www.statmt.org/wmt07/`

Metric	Adequacy	Fluency	Rank	Constituent	Overall
Semantic Role Overlap	**.774**	**.839**	**.803**	.741	**.789**
ParaEval-Recall	.712	.742	.768	**.798**	.755
METEOR	.701	.719	.745	.669	.709
BLEU	.690	.722	.672	.602	.671
1-TER	.607	.538	.520	.514	.644
Max Adequacy Correlation	.651	.657	.659	.534	.626
Max Fluency Correlation	.644	.653	.656	.512	.616
GTM	.655	.674	.616	.495	.610
Dependency Overlap	.639	.644	.601	.512	.599
ParaEval-Precision	.639	.654	.610	.491	.598
1-WER of Verbs	.378	.422	.431	.297	.382

Table 3.9: WMT 2007 Shared Task. Official meta-evaluation results for Foreign-to-English tasks

3.3 On the Robustness of Linguistic Features

In the previous section, we have showed that linguistic features based on syntactic and semantic information are useful for the purpose of automatic MT evaluation. These metrics have proved very effective, in particular when applied to test beds with a rich system typology, i.e., test beds in which there are automatic outputs produced by systems based on different paradigms (statistical, rule-based, human-aided, etc.). The reason is that they are able to capture deep MT quality distinctions which occur beyond the shallow level of lexical similarities.

However, these metrics present the major limitation of relying on automatic linguistic processors, which are not equally available for all languages and whose performance may vary depending on the type of linguistic analysis and the application domain. Thus, it could be argued that they should suffer a significant performance decrease when applied to a very different translation domain, or to heavily ill-formed sentences (p.c., Young-Suk Lee, IBM). In this work, we have studied this issue by conducting a contrastive empirical study on the behavior of a heterogeneous set of metrics over several evaluation scenarios of decreasing translation quality. In particular, we have studied the case of Chinese-to-English translation of automatically recognized speech, which is a paradigmatic example of low quality and heavily ill-formed automatic translations.

3.3.1 Experimental Setting

We have used the test bed from the Chinese-to-English translation task at the *"2006 Evaluation Campaign on Spoken Language Translation"* (Paul, 2006)[15], extracted from the Basic Travel Expressions Corpus (BTEC) (Takezawa, 1999). The test set comprises 500 translation test cases corresponding to simple conversations (question/answer scenario) in the travel domain. Besides, there are 3 different evaluation subscenarios of increasing translation difficulty, according to the translation source:

[15] http://www.slc.atr.jp/IWSLT2006/

	CRR	ASR read	ASR spont
#human-references	7	7	7
#system-outputs	14	14	13
#sentences	500	500	500
#outputs$_{\text{assessed}}$	6	6	6
#sentences$_{\text{assessed}}$	400	400	400
Average Adequacy	1.40	1.02	0.93
Average Fluency	1.16	0.98	0.98

Table 3.10: IWSLT 2006 MT Evaluation Campaign. Chinese-to-English test bed description

CRR: Translation of correct recognition results (as produced by human transcribers).

ASR read: Translation of automatic read speech recognition results.

ASR spont: Translation of automatic spontaneous speech recognition results.

For the purpose of automatic evaluation, 7 human reference translations and automatic outputs by up to 14 different MT systems for each evaluation subscenario are available. In addition, we count on the results of a process of manual evaluation. For each subscenario, 400 test cases from 6 different system outputs were evaluated, by three human assessors each, in terms of adequacy and fluency on a 1-5 scale (LDC, 2005). A brief numerical description of these test beds is available in Table 3.10. It includes the number of human references and system outputs available, as well as the number of sentences per output, and the number of system outputs and sentences per system assessed. In order to give an idea of the translation quality exhibited by automatic systems, average adequacy and fluency scores are also provided.

In our experiments, metrics are evaluated both in terms of human acceptability and human likeness. In the case of human acceptability, metric quality is measured on the basis of correlation with human assessments both at the sentence and document (i.e., system) levels. We compute Pearson correlation coefficients. The sum of adequacy and fluency is used as a global measure of quality. Assessments from different judges have been averaged. In the case of human likeness, we use the probabilistic KING measure defined inside the QARLA Framework. Details on KING's formulation are available in Section 3.4.1. Although KING computations do not require human assessments, for the sake of comparison, we have limited to the set of test cases counting on human assessments.

3.3.2 Metric Performance

Table 3.11 presents meta-evaluation results for the three subscenarios defined ('CRR', 'ASR read' and 'ASR spont'). As before, metrics are grouped according to the linguistic level at which they operate. For the sake of readability, we have selected a small set of representatives from each level. Their respective behavior is evaluated in terms of human likeness (KING, columns 1-3), and human acceptability both at the sentence level (R_{snt}, columns 4-6) and system level (R_{sys}, columns 7-9). Highest correlation coefficients attained in each case appear highlighted, whereas italics are used to indicate low correlation coefficients.

		KING			R_{snt}			R_{sys}		
Level	**Metric**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**
	1-WER	0.63	0.69	0.71	0.47	0.50	0.48	0.50	*0.32*	0.52
	1-PER	0.71	0.79	0.79	0.44	0.48	0.45	0.67	*0.39*	0.60
	1-TER	0.69	0.75	0.77	0.49	0.52	0.50	0.66	*0.36*	0.62
	BLEU	0.69	0.72	0.73	0.54	0.53	0.52	0.79	0.74	0.62
Lexical	NIST	0.79	0.84	0.85	0.53	0.54	0.53	*0.12*	*0.26*	*-0.02*
	GTM ($e = 1$)	0.75	0.81	0.83	0.50	0.52	0.52	*0.35*	*0.10*	*-0.09*
	GTM ($e = 2$)	0.72	0.78	0.79	**0.62**	**0.64**	**0.61**	0.78	0.65	0.62
	METEOR$_{wnsyn}$	**0.81**	**0.86**	**0.86**	0.44	0.50	0.48	0.55	*0.39*	*0.08*
	ROUGE$_{W_1.2}$	0.74	0.79	0.81	0.58	0.60	0.58	0.53	0.69	0.43
	O_l	0.74	0.81	0.82	0.57	0.62	0.58	0.77	0.51	0.34
	SP-O_p-$\star$	0.75	0.80	0.82	0.54	0.59	0.56	0.77	0.54	0.48
	SP-O_c-$\star$	0.74	0.81	0.82	0.54	0.59	0.55	0.82	0.52	0.49
Shallow	SP-NIST$_l$	0.79	0.84	0.85	0.52	0.53	0.52	*0.10*	*0.25*	*-0.03*
Syntactic	SP-NIST$_p$	0.74	0.78	0.80	0.44	0.42	0.43	*-0.02*	*0.24*	*0.04*
	SP-NIST$_{iob}$	0.65	0.69	0.70	*0.33*	*0.32*	*0.35*	*-0.09*	*0.17*	*-0.09*
	SP-NIST$_c$	*0.55*	*0.59*	*0.59*	*0.24*	*0.22*	*0.25*	*-0.07*	*0.19*	*0.08*
	CP-O_p-$\star$	0.75	0.81	0.82	0.57	0.63	0.59	**0.84**	0.67	0.52
	CP-O_c-$\star$	0.74	0.80	0.82	**0.60**	**0.64**	**0.61**	0.71	0.53	0.43
	DP-O_l-$\star$	0.68	0.75	0.76	0.48	0.50	0.50	**0.84**	0.77	0.67
	DP-O_c-$\star$	0.71	0.76	0.77	0.41	0.46	0.43	0.76	0.65	0.71
Syntactic	DP-O_r-$\star$	0.75	0.80	0.81	0.51	0.53	0.51	0.81	0.75	0.62
	DP-HWC$_w$-4	*0.54*	*0.57*	*0.57*	*0.29*	*0.32*	*0.28*	0.73	0.74	0.37
	DP-HWC$_c$-4	*0.48*	*0.51*	*0.52*	*0.17*	*0.18*	*0.22*	0.73	0.64	0.67
	DP-HWC$_r$-4	*0.44*	*0.49*	*0.48*	*0.20*	*0.21*	*0.25*	0.71	0.58	0.56
	CP-STM-4	0.71	0.77	0.80	0.53	0.56	0.54	0.65	0.58	0.47
	NE-M_e-$\star$	*0.14*	*0.16*	*0.18*	*0.10*	*0.13*	*0.08*	*-0.34*	*0.24*	*-0.48*
	NE-O_e-$\star$	*0.14*	*0.17*	*0.18*	*0.10*	*0.12*	*0.07*	*-0.27*	*0.29*	*-0.31*
	NE-O_e-$\star\star$	0.74	0.80	0.82	0.56	0.61	0.58	0.76	0.55	*0.34*
	SR-M_r-$\star$	*0.40*	*0.43*	*0.45*	*0.29*	*0.28*	*0.29*	0.52	0.60	*0.20*
	SR-O_r-$\star$	*0.45*	*0.49*	*0.51*	*0.35*	*0.35*	*0.36*	0.56	0.58	*0.14*
Shallow	SR-O_r	*0.31*	*0.33*	*0.35*	*0.16*	*0.15*	*0.18*	0.68	0.73	0.53
Semantic	SR-M_{rv}-$\star$	*0.38*	*0.41*	*0.42*	*0.33*	*0.34*	*0.34*	0.79	**0.81**	0.42
	SR-O_{rv}-$\star$	*0.40*	*0.44*	*0.45*	*0.36*	*0.38*	*0.38*	0.64	0.72	0.72
	SR-O_{rv}	*0.36*	*0.40*	*0.40*	*0.27*	*0.31*	*0.29*	*0.34*	0.78	0.38
	DR-O_r-$\star$	0.67	0.73	0.75	0.48	0.53	0.50	**0.86**	0.74	0.77
Semantic	DR-O_{rp}-$\star$	0.59	0.64	0.65	0.34	0.35	0.33	**0.84**	0.78	**0.95**
	DR-STM-4	0.58	0.63	0.65	*0.23*	*0.26*	*0.26*	0.75	0.62	0.67

Table 3.11: IWSLT 2006, Chinese-to-English. Meta-evaluation results

System Level Behavior

At the system level (R_{sys}, columns 7-9), the highest quality is in general attained by metrics based on deep linguistic analysis, either syntactic or semantic We interpret the boost in performance of these metrics at the document level as an indicator that these are metrics of high precision. Parsing errors (unanalyzed or wrongly analyzed sentences) would be mainly causing a loss of recall, but for the cases in which the linguistic analysis is successful, these metrics would be able to capture fine quality distinctions.

Let us note, however, the anomalous behavior of metrics based on lexical overlapping over NEs alone, which report completely useless in this test bed. The reason is that these metrics are focused on a very partial aspect of quality, which does not seem to be important in this specific test bed. Observe how the 'NE-O_e-$\star\star$' metric, which combines lexical overlapping over NEs with lexical overlapping over the rest of words, performs similarly to lexical metrics.

As to the impact of sentence ill-formedness, while most metrics at the lexical level suffer a significant variation across the three subscenarios, the performance of metrics at deeper linguistic levels is in general quite stable. However, in the case of the translation of automatically recognized spontaneous speech (ASR spont) we have found that the 'SR-O_r-$\star$' and 'SR-M_r-$\star$' metrics, respectively based on lexical overlapping and matching over semantic roles, suffer a very significant decrease far below the performance of most lexical metrics. Although 'SR-O_r-$\star$' has performed well on other test beds (Giménez & Màrquez, 2007b), its low performance over the BTEC data suggests that it is not fully portable across different evaluation scenarios.

Finally, it is highly remarkable the degree of robustness exhibited by semantic metrics based on lexical and morphosyntactic overlapping over discourse representations ('DR-O_r-$\star$' and 'DR-O_{rp}-$\star$', respectively), which obtain a high system-level correlation with human assessments across the three subscenarios.

Sentence Level Behavior

At the sentence level (KING and R_{snt}, columns 1-6), highest quality is attained in most cases by metrics based on lexical matching. This result was expected since all MT systems are statistical and, thus, have been trained on in-domain data. Therefore, their translations have a strong tendency to share the sublanguage (i.e., word selection and word ordering) represented by the predefined set of in-domain human reference translations.

However, the most positive result is that the behavior of syntactic and semantic metrics across the three evaluation subscenarios is in general quite stable —the three values in each subrow are in a very similar range. Thus, sentence ill-formedness does not seem to be a key factor on their performance.

Metrics based on lexical overlapping and matching over shallow syntactic categories and syntactic structures ('SP-O_p-$\star$', 'SP-O_c-$\star$', 'CP-O_p-$\star$', 'CP-O_c-$\star$', 'DP-O_l-$\star$', 'DP-O_c-$\star$', and 'DP-O_r-$\star$') perform similarly to lexical metrics. However, computing NIST scores over base phrase chunk sequences ('SP-NIST$_{iob}$', 'SP-NIST$_c$') is not as effective. Metrics based on head-word chain matching ('DP-HWC$_w$', 'DP-HWC$_c$', 'DP-HWC$_r$') suffer also a significant decrease. Interestingly, the metric based on syntactic tree matching ('CP-STM-4') performed well in all scenarios.

Metrics at the shallow semantic level suffer also a severe drop in performance. Particularly dramatic is the case of metrics based on lexical overlapping over NEs which, as in the case of system level meta-evaluation, prove very ineffective. In the case of SR metrics, the performance decrement is also very significant, particularly in the case of the 'SR-O_r' metric, which does not consider any lexical information. Interestingly, the 'SR-O_{rv}' variant, which only differs in that it distinguishes between SRs associated to different verbs, performs considerably better.

At the semantic level, metrics based on lexical and morphosyntactic overlapping over discourse representations ('DR-O_r-$\star$' and 'DR-O_{rp}-$\star$') suffer only a minor decrease. The semantic tree matching metric ('DR-STM-4') is not very effective either, specially in terms of its ability to capture human acceptability (R_{snt}).

3.3.3 Improved Sentence Level Behavior

By inspecting particular cases we have found that in many cases metrics are unable to produce any evaluation result. The number of unscored sentences is particularly significant in the case of SR metrics. For instance, the 'SR-O_r-$\star$' metric is unable to confer an evaluation score in 57% of the cases. Several reasons explain this fact. The first and most important is that metrics based on deep linguistic analysis rely on automatic processors trained on out-of-domain data, which are, thus, prone to error. Second, we argue that the test bed itself does not allow for fully exploiting the capabilities of these metrics. Apart from being based on a reduced vocabulary (2,346 distinct words), test cases consist mostly of very short segments (14.64 words on average), which in their turn consist of even shorter sentences (8.55 words on average)[16].

A natural and direct solution, in order to improve their performance, could be to back off to a measure of lexical similarity in those cases in which linguistic processors are unable to produce any linguistic analysis. This should significantly increase their recall. With that purpose, we have designed two new variants for each of these metrics. Given a linguistic metric x, we define:

- $\boldsymbol{x_b} \rightarrow$ by backing off to lexical overlapping, O_l, only when the linguistic processor was not able to produce a parsing. Lexical scores are conveniently scaled so that they are in a similar range to x scores. Specifically, we multiply them by the average x score attained over all other test cases for which the parser succeeded. Formally, given a test case t belonging to a set of test cases T:

$$x_b(t) = \begin{cases} O_l(t) * \frac{\sum_{j \in ok(T)} x(j)}{|ok(T)|} & \text{if parsing}(t) \text{ failed} \\ x(t) & \text{otherwise} \end{cases}$$

  where $ok(T)$ is the subset of test cases in T which were successfully parsed.

- $\boldsymbol{x_i} \rightarrow$ by linearly interpolating x and O_l scores for all test cases, via arithmetic mean:

$$x_i(t) = \frac{x(t) + O_l(t)}{2}$$

In both cases, system-level scores are calculated by averaging over all sentence-level scores.

[16] Vocabulary size and segment/sentence average lengths have been computed over the set of reference translations.

		KING			R_{snt}			R_{sys}		
Level	**Metric**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**	**CRR**	**ASR read**	**ASR spont**
Lexical	NIST	0.79	0.84	0.85	0.53	0.54	0.53	*0.12*	*0.26*	*-0.02*
	GTM ($e = 2$)	0.72	0.78	0.79	**0.62**	**0.64**	**0.61**	0.78	0.65	0.62
	METEOR$_{wnsyn}$	0.81	**0.86**	**0.86**	0.44	0.50	0.48	0.55	*0.39*	*0.08*
	O_l	0.74	0.81	0.82	0.57	0.62	0.58	0.77	0.51	0.34
Syntactic	CP-O_p-$\star$	0.75	0.81	0.82	0.57	0.63	0.59	0.84	0.67	0.52
	CP-O_c-$\star$	0.74	0.80	0.82	**0.60**	**0.64**	**0.61**	0.71	0.53	0.43
	DP-O_l-$\star$	0.68	0.75	0.76	0.48	0.50	0.50	0.84	0.77	0.67
Shallow Semantic	SR-M_r-$\star$	*0.40*	*0.43*	*0.45*	*0.29*	*0.28*	*0.29*	0.52	0.60	*0.20*
	SR-M_r-$\star_b$	0.68	0.72	0.73	0.31	0.30	0.31	0.52	0.60	0.20
	SR-M_r-$\star_i$	**0.84**	**0.86**	**0.88**	0.34	0.34	0.34	0.56	0.63	0.25
	SR-O_r-$\star$	*0.45*	*0.49*	*0.51*	*0.35*	*0.35*	*0.36*	0.56	0.58	*0.14*
	SR-O_r-$\star_b$	0.71	0.75	0.78	0.38	0.38	0.38	0.56	0.58	0.14
	SR-O_r-$\star_i$	**0.84**	**0.88**	**0.89**	0.41	0.41	0.41	0.62	0.60	0.22
	SR-O_r	*0.31*	*0.33*	*0.35*	*0.16*	*0.15*	*0.18*	0.68	0.73	0.53
	SR-O_{rb}	0.54	0.58	0.60	0.19	0.18	0.20	0.68	0.73	0.53
	SR-O_{ri}	0.72	0.77	0.79	0.26	0.26	0.27	0.80	0.73	0.67
	SR-M_{rv}-$\star$	*0.38*	*0.41*	*0.42*	*0.33*	*0.34*	*0.34*	0.79	0.81	0.42
	SR-M_{rv}-$\star_b$	0.70	0.73	0.74	0.34	0.35	0.34	0.79	0.81	0.42
	SR-M_{rv}-$\star_i$	**0.88**	**0.90**	**0.92**	0.36	0.38	0.37	0.81	**0.82**	0.45
	SR-O_{rv}-$\star$	*0.40*	*0.44*	*0.45*	*0.36*	*0.38*	*0.38*	0.64	0.72	0.72
	SR-O_{rv}-$\star_b$	0.72	0.76	0.77	0.38	0.40	0.39	0.64	0.72	0.72
	SR-O_{rv}-$\star_i$	**0.88**	**0.90**	**0.91**	0.40	0.42	0.41	0.69	0.74	0.74
	SR-O_{rv}	*0.36*	*0.40*	*0.40*	*0.27*	*0.31*	*0.29*	*0.34*	0.78	0.38
	SR-O_{rvb}	0.66	0.70	0.71	0.29	0.32	0.30	0.34	0.78	0.38
	SR-O_{rvi}	**0.83**	**0.86**	**0.88**	0.33	0.36	0.33	0.49	**0.82**	0.56
Semantic	DR-O_r-$\star$	0.67	0.73	0.75	0.48	0.53	0.50	0.86	0.74	0.77
	DR-O_r-$\star_b$	0.69	0.75	0.77	0.50	0.53	0.50	**0.90**	0.69	0.56
	DR-O_r-$\star_i$	**0.83**	**0.87**	**0.89**	0.53	0.57	0.53	**0.88**	0.70	0.61
	DR-O_{rp}-$\star$	0.59	0.64	0.65	0.34	0.35	0.33	0.84	0.78	**0.95**
	DR-O_{rp}-$\star_b$	0.61	0.65	0.67	0.35	0.36	0.34	0.86	0.71	0.57
	DR-O_{rp}-$\star_i$	**0.80**	**0.84**	**0.85**	0.43	0.46	0.43	**0.90**	0.75	0.70
	DR-STM-4	0.58	0.63	0.65	0.23	0.26	0.26	0.75	0.62	0.67
	DR-STM-4-b	0.64	0.68	0.71	0.23	0.26	0.27	0.75	0.62	0.67
	DR-STM-4-i	**0.83**	**0.87**	**0.87**	0.33	0.36	0.36	0.84	0.63	0.66

Table 3.12: IWSLT 2006, Chinese-to-English. Improved sentence level evaluation

Table 3.12 shows meta-evaluation results on the performance of these variants for several representatives from the SR and DR families. For the sake of comparison, we also show the scores attained by the base versions, and by some of the top-scoring metrics from other linguistic levels.

The first observation is that in all cases the new variants outperform their respective base metric, being linear interpolation the best alternative. The increase is particularly significant in terms of human likeness. The new variants even outperform lexical metrics, included the O_l metrics, which suggests that, although simple and naïve, this is a valid combination scheme. However, in terms of human acceptability, the gain is only moderate, and still their performance is far from top-scoring metrics.

Sentence-level improvements are also reflected at the system level, although to a lesser extent. Interestingly, in the case of the translation of automatically recognized spontaneous speech (ASR spont, column 9), mixing with lexical overlapping improves the low-performance 'SR-O_r' and 'SR-O_{rv}' metrics, at the same time that it causes a significant drop in the high-performance 'DR-O_r' and 'DR-O_{rp}' metrics.

Still, the performance of linguistic metrics at the sentence level is under the performance of lexical metrics. This is not surprising. After all, apart from relying on automatic processors, linguistic metrics have been designed to capture very partial aspects of quality. However, since they operate at complementary quality dimensions, their scores are suitable for being combined.

3.4 Non-Parametric Metric Combinations

Approaches described in Section 2.4.4 (Corston-Oliver et al., 2001; Kulesza & Shieber, 2004; Quirk, 2004; Gamon et al., 2005; Liu & Gildea, 2007; Albrecht & Hwa, 2007a; Albrecht & Hwa, 2007b; Paul et al., 2007), although based on different assumptions, may be classified as belonging to a same family. All implement a *parametric* combination strategy. Their models involve a number of parameters which must be adjusted. The main difference between these methods can be found in the meta-evaluation criterion underlying. While Corston-Oliver et al. (2001), Kulesza and Shieber (2004), and Gamon et al. (2005) rely on human likeness (i.e., the metric ability to distinguish between human and automatic translations), Akiba et al. (2001), Quirk (2004), Liu and Gildea (2007), Albrecht and Hwa (2007a; 2007b) and Paul et al. (2007) rely on human acceptability (i.e., the metric ability to emulate human assessments).

As an alternative, in this section, we study the behavior of *non-parametric* metric combination schemes. Non-parametric approaches offer the advantage that no training or adjustment of parameters is required. Metrics are combined without having to adjust their relative importance. We describe two different non-parametric combination methods, respectively based on human likeness and human acceptability as meta-evaluation criteria. Besides, rather than limiting to the lexical dimension, we work on the rich set of linguistic metrics described in Section 3.1.

3.4.1 Approach

Our approach to non-parametric combination schemes based on human likeness relies on the QARLA Framework (Amigó et al., 2005), which is, to our knowledge, the only existing non-parametric approach to metric combination. QARLA is a probabilistic framework originally designed for the

evaluation of automatic summaries. QARLA is non-parametric because, rather than assigning a weight to the contribution of each metric, the evaluation of a given automatic output a is addressed through a set of independent probabilistic tests (one per metric) in which the goal is to falsify the hypothesis that a is a human reference. The input for QARLA is a set of test cases A (i.e., automatic translations), a set of similarity metrics X, and a set of models R (i.e., human references) for each test case. With such a testbed, QARLA provides the two essential ingredients required for metric combination:

Combination Scheme Metrics are combined through the QUEEN measure. QUEEN operates under the *unanimity* principle, i.e., the assumption that a 'good' translation must be similar to all human references according to all metrics. $\text{QUEEN}_X(a)$ is defined as the probability, over $R \times R \times R$, that, for every metric in X, the automatic translation a is more similar to a human reference r than two other references, r' and r'', to each other. Formally:

$$\text{QUEEN}_{X,R}(a) = Prob(\forall x \in X : x(a,r) \geq x(r',r''))$$

where $x(a,r)$ stands for the similarity between a and r according to the metric x. Thus, QUEEN allows us to combine different similarity metrics into a single measure, without having to adjust their relative importance. Besides, QUEEN offers two other important advantages which make it really suitable for metric combination: (i) it is *robust* against metric redundancy, i.e., metrics covering similar aspects of quality, and (ii) it is not affected by the scale properties of metrics. The main drawback of the QUEEN measure is that it requires at least three human references, when in most cases only a single reference translation is available.

Meta-evaluation Criterion Metric quality is evaluated using the KING measure. All human references are assumed to be equally optimal and, while they are likely to be different, the best similarity metric is the one that identifies and uses the features that are common to all human references, grouping them and separating them from automatic translations. Based on QUEEN, KING represents the probability that a human reference does not receive a lower score than the score attained by *any* automatic translation. Formally:

$$\text{KING}_{A,R}(X) = Prob(\forall a \in A : \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\}}(a))$$

The closest measure to KING is ORANGE (Lin & Och, 2004b). ORANGE is defined as the ratio between the average rank of the reference translations within the combined automatic and reference translations list and the size of the list. Formally:

$$\text{ORANGE}_{A,R}(X) = Prob(r \in R, a \in A : x_{R-\{r\}}(r) \geq x_{R-\{r\}}(a))$$

However, ORANGE does not allow for simultaneously considering different metrics.

Apart from being non-parametric, QARLA exhibits another important feature which differentiates it form other approaches. Besides considering the similarity between automatic translations and human references, QARLA also takes into account the distribution of similarities among human references.

However, QARLA is not well suited to port from human likeness to human acceptability. The reason is that QUEEN is, by definition, a very restrictive measure —a 'good' translation must be similar to *all* human references according to *all* metrics. Thus, as the number of metrics increases, it becomes easier to find a metric which does not satisfy the QUEEN assumption. This causes QUEEN values to get close to zero, which turns correlation with human assessments into an impractical meta-evaluation measure.

We have *simulated* a non-parametric scheme based on human acceptability by working with uniformly averaged linear combinations (ULC) of metrics. Our approach is similar to that of Liu and Gildea (2007) except that in our case all the metrics in the combination are equally important[17]. In other words, ULC is indeed a particular case of a parametric scheme, in which the contribution of each metric is not adjusted. Formally:

$$\mathrm{ULC}_X(a, R) = \frac{1}{|X|} \sum_{x \in X} x(a, R)$$

where X is the metric set, and $x(a, R)$ is the similarity between the automatic translation a and the set of references R, for the given test case, according to the metric x. We evaluate metric quality in terms of correlation with human assessments at the sentence level (R_{snt}). We use the sum of adequacy and fluency to simulate a global assessment of quality.

3.4.2 Experimental Setting

We use the test beds from the 2004 and 2005 NIST MT Evaluation Campaigns (Le & Przybocki, 2005)[18]. Both campaigns include two different translations exercises: Arabic-to-English ('AE') and Chinese-to-English ('CE'). Human assessments of adequacy and fluency are available for a subset of sentences, each evaluated by two different human judges. A brief numerical description of these test beds is available in table 3.13. It includes the number of human references and system outputs available, as well as the number of sentences per output, and the number of system outputs and sentences per system assessed. In order to give an idea of the translation quality exhibited by automatic systems, average adequacy and fluency scores are also provided.

3.4.3 Evaluating Individual Metrics

Prior to studying the effects of metric combination, we study the isolated behaviour of individual metrics. We have selected a set of metric representatives from each linguistic level. Table 3.14 shows meta-evaluation results for the test beds described in Section 3.4.2, according both to human likeness (KING) and human acceptability (R_{snt}), computed over the subsets of sentences for which

[17] That would be assuming that all metrics operate in the same range of values, which is not always the case.

[18] `http://www.nist.gov/speech/tests/summaries/2005/mt05.htm`

	$\mathbf{AE_{2004}}$	$\mathbf{CE_{2004}}$	$\mathbf{AE_{2005}}$	$\mathbf{CE_{2005}}$
#human-references	5	5	5	4
#system-outputs	5	10	7	10
#sentences	1,353	1,788	1,056	1,082
#outputs$_{\text{assessed}}$	5	10	6	5
#sentences$_{\text{assessed}}$	347	447	266	272
Average Adequacy	2.81	2.60	3.00	2.58
Average Fluency	2.56	2.41	2.70	2.47

Table 3.13: NIST 2004/2005 MT Evaluation Campaigns. Test bed description

human assessments are available. Highest correlation coefficients attained in each task appear highlighted.

The first observation is that the two meta-evaluation criteria provide very similar metric quality rankings for a same test bed. This seems to indicate that there is a relationship between the two meta-evaluation criteria employed. We have confirmed this intuition by computing the Pearson correlation coefficient between values in columns 1 to 4 and their counterparts in columns 5 to 8. There exists a high correlation (R = 0.79).

A second observation is that metric quality varies significantly from task to task. This is due to the significant differences among the test beds employed. These are related to three main aspects: language pair, translation domain, and system typology. For instance, notice that most metrics exhibit a lower quality in the case of the 'AE_{05}' test bed. The reason is that, while in the rest of test beds all systems are statistical, the 'AE_{05}' test bed presents, as we have seen in Section 3.2, the particularity of providing automatic translations produced by heterogeneous MT systems. The fact that most systems are statistical also explains why, in general, lexical metrics exhibit a higher quality. However, highest levels of quality are not in all cases attained by metrics at the lexical level (see highlighted values). In fact, there is only one metric, 'ROUGE_W' (based on lexical matching), which is consistently among the top-scoring in all test beds according to both meta-evaluation criteria. The underlying cause is simple: current metrics do not provide a global measure of quality, but account only for partial aspects of it. Apart from evincing the importance of the meta-evaluation process, these results strongly suggest the need for conducting heterogeneous MT evaluations.

3.4.4 Finding Optimal Metric Combinations

In this section, we study the applicability of the two combination strategies above presented. Optimal metric sets are determined by maximizing over the corresponding meta-evaluation measure (KING or R_{snt}). However, because exploring all possible combinations was not viable, we have used a simple algorithm which performs an approximate search:

1. Individual metrics are ranked according to their quality (KING or R_{snt}).

2. Following that order, metrics are individually added to the optimal set of metrics only if in doing so the global quality increases.

Level	Metric	KING				R_{snt}			
		AE_{04}	**CE_{04}**	**AE_{05}**	**CE_{05}**	**AE_{04}**	**CE_{04}**	**AE_{05}**	**CE_{05}**
	1-WER	0.70	0.51	0.48	0.61	0.53	0.47	0.38	0.47
	1-PER	0.64	0.43	0.45	0.58	0.50	0.51	0.29	0.40
	1-TER	0.73	0.54	0.53	0.66	0.54	0.50	0.38	0.49
	BLEU	0.70	0.49	0.52	0.59	0.50	0.46	0.36	0.39
	NIST	**0.74**	0.53	**0.55**	**0.68**	0.53	**0.55**	0.37	0.46
Lexical	GTM.e1	0.67	0.49	0.48	0.61	0.41	0.50	0.26	0.29
	GTM.e2	0.69	0.52	0.51	0.64	0.49	0.54	0.43	0.48
	$ROUGE_L$	0.73	**0.59**	0.49	0.65	**0.58**	**0.60**	0.41	**0.52**
	$ROUGE_W$	**0.75**	**0.62**	**0.54**	**0.68**	**0.59**	**0.57**	**0.48**	**0.54**
	$METEOR_{wnsyn}$	**0.75**	0.56	**0.57**	**0.69**	**0.56**	**0.56**	0.35	0.41
	SP-O_p-$\star$	0.66	0.48	0.49	0.59	0.51	**0.57**	0.38	0.41
	SP-O_c-$\star$	0.65	0.44	0.46	0.59	**0.55**	**0.58**	0.42	0.41
Shallow	$SP\text{-}NIST_l$	0.73	0.51	**0.55**	**0.66**	0.53	0.54	0.38	0.44
Syntactic	$SP\text{-}NIST_p$	**0.79**	**0.60**	**0.56**	**0.70**	0.46	0.49	0.37	0.39
	$SP\text{-}NIST_{iob}$	0.69	0.48	0.49	0.59	0.32	0.36	0.27	0.26
	$SP\text{-}NIST_c$	0.60	0.42	0.39	0.52	0.26	0.27	0.16	0.16
	$DP\text{-}HWC_w$	0.58	0.40	0.42	0.53	0.41	0.08	0.35	0.40
	$DP\text{-}HWC_c$	0.50	0.32	0.33	0.41	0.41	0.17	0.38	0.32
	$DP\text{-}HWC_r$	0.56	0.40	0.37	0.46	0.42	0.16	0.39	0.43
	DP-O_l-$\star$	0.58	0.48	0.41	0.52	0.52	0.48	0.36	0.37
Syntactic	DP-O_c-$\star$	0.65	0.45	0.44	0.55	0.49	0.51	0.43	0.41
	DP-O_r-$\star$	0.71	**0.57**	**0.54**	0.64	**0.55**	**0.55**	**0.50**	**0.50**
	CP-O_p-$\star$	0.67	0.47	0.47	0.60	0.53	**0.57**	0.38	0.46
	CP-O_c-$\star$	0.66	0.51	0.49	0.62	**0.57**	**0.59**	**0.45**	0.50
	CP-STM	0.64	0.42	0.43	0.58	0.39	0.13	0.34	0.30
	NE-O_e-$\star\star$	0.65	0.45	0.46	0.57	0.47	0.56	0.32	0.39
Shallow	SR-O_r-$\star$	0.48	0.22	0.34	0.41	0.28	0.10	0.32	0.21
Semantic	SR-O_{rv}	0.36	0.13	0.24	0.27	0.27	0.12	0.25	0.24
	DR-O_r-$\star$	0.62	0.47	0.50	0.55	0.47	0.46	0.43	0.37
Semantic	DR-O_{rp}-$\star$	0.58	0.42	0.43	0.50	0.37	0.35	0.36	0.26
Optimal Combination		**0.79**	**0.64**	**0.61**	**0.70**	**0.64**	**0.63**	**0.54**	**0.61**

Table 3.14: NIST 2004/2005 MT Evaluation Campaigns. Meta-evaluation results

$Opt.K(AE.04)$	=	$\{$*SP-NIST*$_p\}$
$Opt.K(CE.04)$	=	$\{$*ROUGE*$_W$, *SP-NIST*$_p$, *ROUGE*$_L\}$
$Opt.K(AE.05)$	=	$\{$*METEOR*$_{wnsyn}$, *SP-NIST*$_p$, *DP-O*$_r$*-**$\}$
$Opt.K(CE.05)$	=	$\{$*SP-NIST*$_p\}$
$Opt.R(AE.04)$	=	$\{$*ROUGE*$_W$, *ROUGE*$_L$, *CP-O*$_c$*-**, *METEOR*$_{wnsyn}$, *DP-O*$_r$*-**, *DP-O*$_l$*-**, *GTM.e2*, *DR-O*$_r$*-**, *CP-STM*$\}$
$Opt.R(CE.04)$	=	$\{$*ROUGE*$_L$, *CP-O*$_c$*-**, *ROUGE*$_W$, *SP-O*$_p$*-**, *METEOR*$_{wnsyn}$, *DP-O*$_r$*-**, *GTM.e2*, *1-WER*, *DR-O*$_r$*-**$\}$
$Opt.R(AE.05)$	=	$\{$*DP-O*$_r$*-**, *ROUGE*$_W\}$
$Opt.R(CE.05)$	=	$\{$*ROUGE*$_W$, *ROUGE*$_L$, *DP-O*$_r$*-**, *CP-O*$_c$*-**, *1-TER*, *GTM.e2*, *DP-HWC*$_r$, *CP-STM*$\}$

Table 3.15: NIST 2004/2005 MT Evaluation Campaigns. Optimal metric sets

Since no training is required it has not been necessary to keep a held-out portion of the data for development (see Section 3.4.5 for further discussion). Optimal metric sets are displayed in Table 3.15. Inside each set, metrics are sorted in decreasing quality order. The 'Optimal Combination' row in Table 3.14 shows the quality attained by these sets, combined under QUEEN in the case of KING optimization, and under ULC in the case of optimizing over R_{snt}. In most cases optimal sets consist of metrics operating at different linguistic levels, mostly at the lexical and syntactic levels. This is coherent with the findings in Section 3.4.3. Metrics at the semantic level are selected only in two cases, corresponding to the R_{snt} optimization in 'AE_{04}' and 'CE_{04}' test beds. Also in two cases, corresponding to the KING optimization in 'AE_{04}' and 'CE_{05}', it has not been possible to find any metric combination which outperforms the best individual metric. This is not necessarily a discouraging result. After all, in these cases, the best metric alone achieves already a very high quality (0.79 and 0.70, respectively). The fact that a single feature suffices to discern between manual and automatic translations indicates that system outputs are easily distinguishable, possibly because of their low quality and/or because systems are all based on the same translation paradigm.

3.4.5 Portability across Scenarios

It can be argued that metric set optimization is itself a training process; each metric would have an associated binary parameter controlling whether it is selected or not. For that reason, in Table 3.16, we have analyzed the portability of optimal metric sets (i) across test beds and (ii) across combination strategies. As to portability across test beds (i.e., across language pairs and years), the reader must focus on the cells for which the meta-evaluation criterion guiding the metric set optimization matches the criterion used in the evaluation, i.e., the top-left and bottom-right 16-cell quadrangles. The fact that the 4 values in each subcolumn are in a very similar range confirms that optimal metric sets port well across test beds.

The same table shows the portability of optimal metric sets across combination strategies. In other words, although QUEEN and ULC are thought to operate on metric combinations respectively optimized on the basis of human likeness and human acceptability, we have studied the effects of applying either measure over metric combinations optimized on the basis of the alternative meta-evaluation criterion. In this case, the reader must compare top-left vs. bottom-left (KING) and

Metric Set	KING				R_{snt}			
	$\mathbf{AE_{04}}$	$\mathbf{CE_{04}}$	$\mathbf{AE_{05}}$	$\mathbf{CE_{05}}$	$\mathbf{AE_{04}}$	$\mathbf{CE_{04}}$	$\mathbf{AE_{05}}$	$\mathbf{CE_{05}}$
$\boldsymbol{Opt.K(AE.04)}$	0.79	0.60	0.56	0.70	0.46	0.49	0.37	0.39
$\boldsymbol{Opt.K(CE.04)}$	0.78	0.64	0.57	0.67	0.49	0.51	0.39	0.43
$\boldsymbol{Opt.K(AE.05)}$	0.74	0.63	0.61	0.66	0.48	0.51	0.39	0.42
$\boldsymbol{Opt.K(CE.05)}$	0.79	0.60	0.56	0.70	0.46	0.49	0.37	0.39
$\boldsymbol{Opt.R(AE.04)}$	0.62	0.56	0.52	0.49	0.64	0.61	0.53	0.58
$\boldsymbol{Opt.R(CE.04)}$	0.68	0.59	0.55	0.56	0.63	0.63	0.51	0.57
$\boldsymbol{Opt.R(AE.05)}$	*0.75*	*0.64*	*0.59*	*0.69*	0.62	0.60	0.54	0.57
$\boldsymbol{Opt.R(CE.05)}$	0.64	0.56	0.51	0.52	0.63	0.57	0.53	0.61

Table 3.16: NIST 2004/2005 MT Evaluation Campaigns. Portability of combination strategies

top-right vs. bottom-right (R_{snt}) 16-cell quadrangles. It can be clearly seen that optimal metric sets, in general, do not port well across meta-evaluation criteria, particularly from human likeness to human acceptability. However, interestingly, in the case of 'AE_{05}' (i.e., heterogeneous systems), the optimal metric set ports well from human acceptability to human likeness (see numbers in italics). We speculate that system heterogeneity has contributed positively for the sake of robustness.

3.5 Heterogeneous Automatic MT Error Analysis

Error analysis is one of the crucial stages in the development cycle of an MT system. However, often not enough attention is paid to this process. The reason is that performing an accurate error analysis is a slow and delicate process which requires intensive human labor. Part of the effort is devoted to high-level analysis which involves a precise knowledge of the architecture of the system under development, but there is also a heavily time-consuming low-level part of the process related to the linguistic analysis of translation quality, which we believe that could be partially automatized.

Our proposal consists in having automatic evaluation metrics play a more active role in this part of the work. In our opinion, in the current error analysis scheme, evaluation metrics are only minimally exploited. They are used as quantitative measures, i.e., so as to identify low/high quality translations, but not as genuine qualitative measures which allow developers to automatically obtain detailed linguistic interpretations of the translation quality attained. This limited usage of automatic metrics for error analysis is a direct consequence of the shallow similarity assumptions commonly utilized for metric development. Until very recently, most metrics were based only on lexical similarity.

However, the availability of metrics at deeper linguistic levels, such as those described in Sections 2.4.3 and 3.1, opens a path towards heterogeneous automatic MT error analysis. This type of analysis would allow system developers to analyze the performance of their systems with respect to different quality dimensions (e.g., lexical, syntactic, and semantic), and at different levels of granularity —from very fine aspects of quality, related to how well certain linguistic structures are transferred, to coarser ones, related to how well the translation under evaluation complies with the expected overall lexical/syntactic/semantic reference structure. Thus, developers could have a more

precise idea of what quality aspects require improvement. Besides, in this manner, they would be allowed to concentrate on high-level decisions.

3.5.1 Types of Error Analysis

Error analyses may be classified, from the perspective of the system developer, according to two different criteria. First, according to the level of abstraction:

- **document-level analysis,** i.e., over a representative set of test cases. Such type of analysis allows developers to quantify the overall system performance. For that reason, it is often also referred to as analysis at the system level.
- **sentence-level analysis,** i.e., over individual test cases. This type of analysis allows developers to identify translation problems over particular instances.

Second, according to the evaluation referent:

- **isolated analysis,** i.e., with no referent other than human translations. This type of analysis allows developers to evaluate the individual performance of their MT system, independently from other MT systems.
- **contrastive analysis,** i.e., on the performance of MT systems in comparison to other MT systems. This type of analysis is crucial for the MT research community so as to advance together, by allowing system developers to borrow successful mechanisms from each other.

3.5.2 Experimental Setting

We have applied our approach to several evaluation test beds from different MT evaluation campaigns. In the following, we exemplify the application of heterogeneous MT error analysis through the case of the Arabic-to-English exercise from the 2005 NIST MT evaluation campaign discussed in Section 3.2.3 (Le & Przybocki, 2005). This test bed presents the particularity of providing automatic translations produced by heterogeneous MT systems. Therefore, it constitutes an excellent material in order to test the applicability of our approach. For that purpose, we have focused on the automatic outputs by LinearB and the best statistical system at hand (from now on, referred to as 'Best SMT'). Assisted by the heterogeneous metric set, we study system performance over a number of partial aspects of quality. We have performed isolated and contrastive analyses, both at the document and sentence levels.

3.5.3 Error Analysis at the Document Level

Tables 3.17 and 3.18 show evaluation results at the document level for several metric representatives. Table 3.17 reports on the lexical and syntactic dimensions, whereas Table 3.18 focuses on semantic features. It can be observed (columns 2-3) that, as we progress from the lexical level to deeper linguistic aspects, the difference in favor of the Best SMT system diminishes and, indeed, ends reversing in favor of the LinearB system when we enter the syntactic and semantic levels.

Level	Metric	KING	Linear B	Best SMT
Lexical	1-PER	0.63	0.65	**0.70**
	1-TER	**0.70**	0.53	**0.58**
	1-WER	0.67	0.49	**0.54**
	BLEU	0.65	0.47	**0.51**
	GTM (e=2)	0.66	0.31	**0.32**
	NIST	0.69	10.63	**11.27**
	ROUGE$_W$	0.68	0.31	**0.33**
	METEOR$_{wnsyn}$	0.68	0.64	**0.68**
Shallow Syntactic	SP-O_p-$\star$	0.64	0.52	**0.55**
	SP-O_p-J	0.26	0.53	**0.59**
	SP-O_p-N	0.53	0.57	**0.63**
	SP-O_p-V	0.43	0.39	**0.41**
	SP-O_c-$\star$	0.63	0.54	**0.57**
	SP-O_c-NP	0.60	0.59	**0.63**
	SP-O_c-PP	0.38	0.63	**0.66**
	SP-O_c-VP	0.41	0.49	**0.51**
	SP-NIST$_l$-5	0.69	10.78	**11.44**
	SP-NIST$_p$-5	**0.71**	8.74	**9.04**
	SP-NIST$_{iob}$-5	0.65	6.81	6.91
	SP-NIST$_c$-5	0.57	6.13	6.18
Syntactic	DP-HWC$_w$-4	0.59	0.14	0.14
	DP-HWC$_c$-4	0.48	**0.42**	0.41
	DP-HWC$_r$-4	0.52	**0.33**	0.31
	DP-O_l-$\star$	0.58	0.41	**0.43**
	DP-O_c-$\star$	0.60	0.50	**0.51**
	DP-O_c-aux	0.14	**0.56**	0.54
	DP-O_c-det	0.35	**0.75**	0.73
	DP-O_r-$\star$	**0.66**	0.36	0.36
	DP-O_r-fc	0.21	**0.26**	0.24
	DP-O_r-i	0.60	**0.44**	0.43
	DP-O_r-obj	0.43	**0.36**	0.35
	DP-O_r-s	0.47	**0.52**	0,45
	CP-O_c-$\star$	0.63	0.50	**0.53**
	CP-O_c-VP	*0.59*	0.49	**0.52**
	CP-STM-9	0.58	0.35	0.35

Table 3.17: NIST 2005 Arabic-to-English. Document level error analysis (lexical and syntactic features)

Level	Metric	KING	Linear B	Best SMT
Shallow Semantic	NE-M_e-$\star$	0.32	0.53	**0.56**
	NE-M_e-ORG	0.11	0.27	**0.29**
	NE-M_e-PER	0.13	0.34	0.34
	SR-M_r-$\star$	0.50	**0.19**	0.18
	SR-M_r-A0	0.33	**0.31**	0.30
	SR-M_r-A1	0.28	0.14	0.14
	SR-O_r	0.41	**0.64**	0.63
	SR-O_r-$\star$	**0.53**	0.36	**0.37**
	SR-O_r-AM-TMP	0.13	**0.39**	0.38
Semantic	DR-O_r-$\star$	**0.59**	**0.36**	0.34
	DR-O_r-card	0.12	**0.49**	0.45
	DR-O_r-dr	0.56	**0.43**	0.40
	DR-O_r-eq	0.12	**0.17**	0.16
	DR-O_r-named	0.38	**0.48**	0.45
	DR-O_r-pred	0.55	**0.38**	0.36
	DR-O_r-prop	0.39	**0.27**	0.24
	DR-O_r-rel	0.56	**0.38**	0.34
	DR-STM-9	0.40	0.26	0.26

Table 3.18: NIST 2005 Arabic-to-English. Document level error analysis (semantic features)

The heterogeneous set of metrics also allows us to analyze very specific aspects of quality. For instance, lexical metrics tell us that the LinearB system does not match well the expected reference lexicon. This is corroborated by analyzing shallow-syntactic similarities. For instance, observe how, while Best SMT is better than LinearB according to 'SP-O_p-J|N|V' metrics, which compute lexical overlapping respectively over adjectives, nouns and verbs, LinearB is better than Best SMT at translating determiners ('DP-O_c-det') and auxiliary verbs ('DP-O_c-aux'), closed grammatical categories which are, presumably, less prone to suffer the effects of a biased lexical selection.

At the syntactic level, differences between both systems are rather small. Metrics based on dependency parsing assign the LinearB system a higher quality, both overall ('DP-HWC$_r$-4' and 'DP-O_r-$\star$') and with respect to finer aspects such as the translation of finite complements ('DP-O_r-fc'), clause relations ('DP-O_r-i'), verb objects ('DP-O_r-obj'), and specially surface subjects ('DP-O_r-s'). In contrast, metrics based on constituent analysis tend to prefer the Best SMT system except for the 'CP-STM-9' metric which assigns both systems the same quality.

As to shallow-semantic metrics, it can be observed that LinearB has more problems than Best SMT to translate NEs, except for the case of person names. In the case of semantic arguments and adjuncts the two systems exhibit a very similar performance with a slight advantage on the side of LinearB, both overall ('SR-M_r-$\star$' and 'SR-O_r') and for fine aspects such as the translation of agent roles ('SR-M_r-A0') and temporal adjuncts ('SR-M_r-AM-TMP'). Also, it can be observed that both systems have difficulties to translate theme roles ('SR-M_r-A1').

At the semantic level, observe how there is not a single metric which ranks the Best SMT system first. LinearB is consistently better at translating basic discourse representation structures ('DR-O_r-dr'), cardinal expressions ('DR-O_r-card'), NEs ('DR-O_r-named'), equality conditions ('DR-O_r-eq'), predicates ('DR-O_r-pred'), relations ('DR-O_r-rel') and propositional attitudes ('DR-O_r-prop'), and overall ('DR-O_r-$\star$'). It can also be observed that both systems have problems to translate equality conditions. Finally, both systems are assigned the same quality according to semantic tree matching ('DR-STM-9').

Meta-evaluation in the Context of Error Analysis

In the previous experiment, metric quality has been evaluated on the basis of human likeness, i.e., in terms of the metric ability to discern between manual and automatic translations We have computed human likeness through the KING measure. As we have already seen in Section 3.4.1, KING is a measure of discriminative power. For instance, if a metric obtains a KING of 0.6, it means that in 60% of the test cases, it is able to explain by itself the difference in quality between manual and automatic translations. For KING computation we have used only the automatic outputs provided by the LinearB and Best SMT systems. However, we have not limited to segments counting on human assessments, but all segments have been used.

In the context of error analysis, KING serves as an estimate of the impact of specific quality aspects on the system performance. In that respect, it can be observed (Tables 3.17 and 3.18, column 1) that metrics at the lexical, shallow-syntactic and syntactic levels attain slightly higher KING values than metrics based on semantic similarities. Best results per family appear highlighted. We speculate that a possible explanation may be found in the performance of linguistic processors whose effectiveness suffers a significant decrease for deeper levels of analysis. Also, observe that

finer grained metrics such as 'SP-O_p-J' (i.e., lexical overlapping over adjectives), 'NE-M_e-ORG' (i.e., lexical matching over organization names) or 'DR-O_r-card' (i.e., lexical overlapping over cardinal expressions) exhibit a much lower discriminative power. The reason is that they cover very partial aspects of quality.

3.5.4 Error Analysis at the Sentence Level

The heterogeneous set of metrics allows us to analyze different dimensions of translation quality over individual test cases. In this manner, we can better search for problematic cases according to different criteria. For instance, we could seek translations lacking of subject (by looking at sentences with very low 'DP-O_r-s') and/or agent role ('SR-O_r-A0'). Or, at a more abstract level, by simultaneously relying on syntactic and semantic metrics, we could, for instance, locate a subset of possibly well-formed translations (i.e., high syntactic similarity) which do not match well the reference semantic structure (i.e., low semantic similarity).

A Case of Analysis

We have inspected particular cases. For instance, Table 3.19 presents the case of sentence 637 in which according to BLEU the translation by Best SMT is ranked first, whereas according to human assessments the translation by LinearB is judged of a superior quality both in terms of adequacy and fluency. This case is deeply analyzed in Table 3.20. In spite of its ill-formedness the translation by Best SMT deceives all lexical metrics. Particularly interesting is the case of 'METEOR$_{wnsyn}$', a metric designed to deal with differences in lexical selection, by dealing with morphological variations through stemming, and synonyms through dictionary lookup. METEOR is in this case, however, unable to deal with differences in word ordering.

In contrast, scores conferred by metrics at deeper linguistic levels reveal, in agreement with human evaluation, that LinearB produced a more fluent (syntactic similarity) and adequate (semantic similarity) translation. Overall syntactic and semantic scores (e.g., 'DP-O_r-$\star$, 'CP-STM-9', 'SR-O_r-$\star$, 'DR-O_r-$\star$ and 'DR-STM-9'), all lower than 0.6, also indicate that important pieces of information were not captured or only partially captured.

Getting into details, by analyzing fine shallow-syntactic similarities, it can be observed, for instance, that, while LinearB successfully translated a larger proportion of singular nouns, Best SMT translated more proper nouns and verb forms. Analysis at the syntactic level shows that LinearB captured more dependency relations of several types (e.g., word adjunct modifiers, verb objects, nominal complements of prepositions, and relative clauses), and translated a larger proportion of different verb phrase types (e.g., noun, prepositional, and verb phrases, and subordinated clauses). As to shallow-semantic similarity, it can be observed that the level of lexical overlapping over verb subjects and objects attained by LinearB is significantly higher. At the semantic level, the discourse representation associated to LinearB is, in general, more similar to the reference discourse representations. Only in the case of predicate conditions, both systems exhibit a similar performance.

Reference 1:	Over 1000 monks and nuns , observers and scientists from over 30 countries and the host country attended the religious summit held for the first time in Myanmar which started today , Thursday .
2:	More than 1000 monks , nuns , observers and scholars from more than 30 countries , including the host country , participated in the religious summit which Myanmar hosted for the first time and which began on Thursday .
3:	The religious summit , staged by Myanmar for the first time and began on Thursday , was attended by over 1,000 monks an nuns , observers and scholars from more than 30 countries and host Myanmar .
4:	More than 1,000 monks , nuns , observers and scholars from more than 30 countries and the host country Myanmar participated in the religious summit , which is hosted by Myanmar for the first time and which began on Thursday .
5:	The religious summit , which started on Thursday and was hosted for the first time by Myanmar , was attended by over 1,000 monks and nuns , observers and scholars from more than 30 countries and the host country Myanmar .
Information:	**(1)** → **subject:** over/more_than 1,000 monks and nuns, observers and scientists/scholars from over/more_than 30 countries , and/including the host country **action:** attended/participated_in **(2)** → **subject:** the religious summit **action:** began/started **temporal:** on Thursday **(3)** → **object:** the religious summit **action:** hosted **subject:** by Myanmar **mode:** for the first time
LinearB:	1000 monks from more than 30 States and the host State Myanmar attended the Summit , which began on Thursday , hosted by Myanmar for the first time .
Best SMT:	Religious participated in the summit , hosted by Myanmar for the first time began on Thursday , as an observer and the world of the 1000 monk nun from more than 30 countries and the host state Myanmar .

Table 3.19: NIST 2005 Arabic-to-English. Test case #637

Level	Metric	Linear B	Best SMT
Human	Adequacy	**3**	2
	Fluency	**3.5**	2
Lexical	1-TER	**0.53**	0.51
	BLEU	0.44	**0.45**
	METEOR$_{wnsyn}$	0.59	**0.64**
Shallow Syntactic	SP-O_p-$\star$	**0.52**	0.51
	SP-O_p-NN	**0.67**	0.38
	SP-O_p-NNP	0.60	**0.75**
	SP-O_p-V	0.40	**0.75**
Syntactic	DP-HWC$_w$-4	**0.17**	0.16
	DP-O_r-$\star$	**0.46**	0.44
	DP-O_r-mod	**0.62**	0.41
	DP-O_r-obj	**0.29**	0.00
	DP-O_r-pcomp-n	**0.71**	0.39
	DP-O_r-rel	**0.33**	0.00
	CP-O_c-$\star$	**0.59**	0.48
	CP-O_c-NP	**0.59**	0.55
	CP-O_c-PP	**0.57**	0.54
	CP-O_c-SB	**0.73**	0.00
	CP-O_c-VP	**0.64**	0.42
	CP-STM-9	**0.34**	0.23
Shallow Semantic	SR-O_r	**0.84**	0.25
	SR-O_r-$\star$	**0.56**	0.18
	SR-O_r-A0	**0.44**	0.10
	SR-O_r-A1	**0.57**	0.28
Semantic	DR-O_r-$\star$	**0.45**	0.34
	DR-O_r-dr	**0.57**	0.40
	DR-O_r-nam	**0.75**	0.24
	DR-O_r-pred	0.44	**0.45**
	DR-O_r-rel	**0.51**	0.32
	DR-STM-9	**0.32**	0.29

Table 3.20: NIST 2005 Arabic-to-English. Error analysis of test case #637

Difficult Cases

One of the main problems of current automatic MT evaluation methods is that their reliability depends very strongly on the representativity of the set of reference translations available. In other words, if reference translations cover only a small part of the space of valid solutions, the predictive power of automatic metrics will decrease. This may be particularly dangerous in the case of n-gram based metrics, which are not able to deal with differences in lexical selection. For instance, Table 3.21 presents a case in which the LinearB is unfairly penalized by lexical metrics for its strong divergence with respect to reference translations while the Best SMT system is wrongly favored for the opposite reason. LinearB translation receives high scores from human assessors, but a null BLEU score. In contrast, the Best SMT system attains a high BLEU score, but receives low scores from human assessors.

Metrics at deeper linguistic levels partly overcome this problem by inspecting syntactic and semantic structures. However, as it can be observed in the case selected, these structures may also exhibit a great variability. For instance, the translation by LinearB is considerably shorter than expected according to human references. Besides, while reference translations use "you must" or "you have", the LinearB translation uses "you should". Also, LinearB selected the verb form "cooperate" instead of "be more united and cooperative", etc. Table 3.22 shows the scores conferred by several metrics. It can be observed how lexical metrics completely fail to reflect the actual quality of the LinearB output. Indeed, only some dependency-based metrics are able to capture its quality (e.g., 'DP-HWC$_c$').

In the case depicted in Table 3.23, differences are mostly related to the sentence structure. Table 3.24 shows the scores conferred by several metrics. It can be observed, for instance, that several lexical metrics are able to capture the superior quality of the LinearB translation. In contrast, metrics at deeper linguistic levels do not reflect, in general, this difference in quality. Interestingly, only some syntax-based metrics confer a slightly higher score to LinearB (e.g., 'SP-O_p-$\star$' 'DP-HWC$_w$-4' 'CP-O_p-$\star$' 'CP-O_c-$\star$', etc.). All these metrics share the common property of computing lexical overlapping/matching over syntactic structures or grammatical categories.

In order to deal with divergences between system outputs and reference translations, other authors have suggested taking advantage of paraphrasing support so as to extend the reference material (Russo-Lassner et al., 2005; Zhou et al., 2006; Kauchak & Barzilay, 2006; Owczarzak et al., 2006). We believe the two approaches could be combined.

3.6 Conclusions of this Chapter

We have suggested a novel direction towards heterogeneous automatic MT evaluation based on a rich set of metrics operating at different linguistic levels (lexical, syntactic and semantic). We have shown that metrics based on deep linguistic information (syntactic/semantic) are able to produce more reliable system rankings than metrics which limit their scope to the lexical dimension, specially when the systems under evaluation are of a different nature.

Linguistic metrics present only a major shortcoming. They rely on automatic linguistic processors. This implies some important limitations on their applicability:

Tagging Errors Automatic tools are prone to error, specially for the deepest levels of analysis.

Processing Speed Linguistic analyzers are typically too slow to allow for massive evaluations, as required, for instance, in the case of system development.

Availability Linguistic analyzers are not equally available for all languages.

As to parsing accuracy, experimental results (see Sections 3.2 and 3.3) have shown that these metrics are very robust against parsing errors at the document/system level. This is very interesting, taking into account that, while reference translations are supposedly well formed, this is not the case of automatic translations. At the sentence level, however, results indicate that metrics based on deep linguistic analysis are, in general, not as reliable overall quality estimators as lexical metrics, at least when applied to low quality translations (e.g., the case discussed in Section 3.3). However, we have shown that backing off to lexical similarity is a valid an effective strategy so as to improve the performance of these metrics.

Regarding the problems related to parsing speed and lack of available tools, in the future, we plan to incorporate more accurate, and possibly faster, linguistic processors, also for languages other than English, as they become publicly available. For instance, we are currently adapting these metrics to Spanish and Catalan.

We have also exploited the possibility of combining metrics at different linguistic levels. Our approach offers the advantage of not having to adjust the relative contribution of each metric to the overall score. We have shown that non-parametric schemes are a valid alternative, yielding a significantly improved evaluation quality at the sentence level, both in terms of human likeness and human acceptability. Let us note, however, that we have not intended to provide a magic recipe, i.e., a combination of metrics which works well in all test beds. In the same manner that the quality aspects distinguishing high quality from low quality translations may vary for each test bed, optimal metric combinations must be determined in each case. The pursuit of a magic recipe for automatic MT evaluation is still a very challenging target for present NLP. For future work, we plan to perform an exhaustive comparison between parametric and non-parametric schemes to metric combination in order to clarify the pros and cons of either option.

As an complementary result, we have shown how to apply linguistic metrics for the purpose of error analysis. Our proposal allows developers to rapidly obtain detailed automatic linguistic reports on their system's capabilities. Thus, they may concentrate their efforts on high-level analysis. For future work, we plan to enhance the interface of the evaluation tool, currently in text format, so as to allow for a fast and elegant visual access from different viewpoints corresponding to the different dimensions of quality. For instance, missing or partially translated elements could appear highlighted in different colors. Besides, evaluation measures generate, as a by-pass product, syntactic and semantic analyses which could be displayed. This would allow users to separately analyze the translation of different types of linguistic elements (e.g., constituents, relationships, arguments, adjuncts, discourse representation structures, etc.).

We strongly believe that future MT evaluation campaigns should benefit from the results presented by conducting heterogeneous evaluations. For instance, the following set could be used:

{ *'ROUGE$_W$', 'METEOR$_{wnsyn}$', 'DP-HWC$_r$-4', 'DP-O$_c$-⋆', 'DP-O$_l$-⋆', 'DP-O$_r$-⋆', 'CP-STM-9', 'SR-O$_r$-⋆', 'SR-O$_{rv}$', 'DR-O$_{rp}$-⋆'* }

This set includes several metric representatives from different linguistic levels, which have been observed to be consistently among the top-scoring over a wide variety of evaluation scenarios. In that respect, we have recently applied our evaluation methodology to the evaluation shared-task at the *ACL 2008 Third Workshop On Statistical Machine Translation* (WMT'08)[19]. In the short term, we also plan to participate at the *NIST Metrics MATR Challenge 2008* on automatic MT evaluation[20].

We are also planning to extend the suggested methodology to perform statistical significance tests over heterogeneous metric sets, which serve to guarantee the statistical significance of evaluation results according to a wide range of measures simultaneously.

Finally, as an additional result of our work, we have developed the IQ$_{MT}$ Framework for MT Evaluation, which is freely and publicly available for research purposes[21].

[19] `http://www.statmt.org/wmt08/` (at the time of writing this document, results were not yet available)

[20] `http://www.nist.gov/speech/tests/metricsmatr/`

[21] The IQ$_{MT}$ (Inside Qarla Machine Translation) Evaluation Framework is released under LGPL licence of the Free Software Foundation. IQ$_{MT}$ may be freely downloaded at `http://www.lsi.upc.edu/~nlp/IQMT`.

LinearB:	You should cooperate and support one another .
Best SMT:	You that you will be more and more cooperative unit some of you and support each other .
Reference 1:	You must be more united and more cooperative and you must support each other .
2:	You must be more united and cooperative and supportive of each other .
3:	You must be more united and cooperative and supportive of each other .
4:	You have to be more united and more cooperative , and support each other .
5:	You have to be more united and more cooperative and you have to support each other .

Table 3.21: NIST 2005 Arabic-to-English. Translation Case #149.

Level	Metric	Linear B	Best SMT
Human	Adequacy	**4**	1.5
	Fluency	**5**	1.5
Lexical	1-PER	0.36	**0.62**
	1-TER	0.36	**0.49**
	BLEU	0.00	**0.37**
	NIST	1.64	**9.42**
	METEOR$_{wnsyn}$	0.32	**0.67**
Shallow Syntactic	SP-O_p-$\star$	0.25	**0.46**
	SP-O_p-V	0.17	**0.40**
	SP-O_c-$\star$	0.19	**0.43**
	SP-O_c-NP	0.43	**0.50**
	SP-O_c-VP	0.14	**0.40**
Syntactic	DP-HWC$_w$-4	0.07	**0.12**
	DP-HWC$_c$-4	**0.32**	0.19
	DP-HWC$_r$-4	**0.32**	0.25
	CP-STM-4	0.33	0.36
Shallow Semantic	SR-M_r-$\star$	0.14	**0.67**
	SR-O_r-$\star$	0.10	**0.75**
Semantic	DR-O_r-$\star$	0.17	**0.26**
	DR-O_{rp}-$\star$	0.24	**0.26**
	DR-O_{rp}-drs	0.27	**0.30**
	DR-O_{rp}-pred	0.29	**0.40**
	DR-O_{rp}-rel	**0.30**	0.24
	DR-STM-4	0.25	**0.45**

Table 3.22: NIST 2005 Arabic-to-English. Error analysis of test case #149

LinearB:	It is important to analyze and address these problems properly .
Best SMT:	It should be to analyze these problems and take them up properly .
Reference 1:	We must analyze these problems and handle them correctly .
2:	So we must analyze these problems and take them in the right way .
3:	We must correctly analyze and properly handle these problems .
4:	And so it is imperative that we analyze these problems and deal with them properly .
5:	And so we must correctly analyze and properly handle these problems .

Table 3.23: NIST 2005 Arabic-to-English. Translation Case #728.

Level	Metric	Linear B	Best SMT
Human	Adequacy	**4.5**	2.5
	Fluency	**5**	2.5
Lexical	1-PER	**0.63**	0.48
	1-TER	**0.55**	0.48
	BLEU	0.00	**0.46**
	NIST	7.82	**9.97**
	ROUGE$_W$	0.25	**0.29**
	METEOR$_{wnsyn}$	**0.54**	0.44
Shallow Syntactic	SP-O_p-$\star$	**0.44**	0.39
	SP-O_p-PRP	**0.50**	0.33
	SP-O_c-$\star$	0.28	**0.38**
Syntactic	DP-O_c-$\star$	**0.48**	0.47
	DP-HWC$_w$-4	**0.23**	0.16
	DP-HWC$_c$-4	0.31	**0.42**
	DP-HWC$_r$-4	0.21	**0.43**
	DP-O_r-$\star$	0.25	**0.36**
	DP-O_r_i	**0.44**	0.43
	DP-O_r_mod	0.11	**0.33**
	DP-O_r_s	0.50	0.50
	CP-O_p-$\star$	**0.45**	0.41
	CP-O_p-RB	0.50	0.50
	CP-O_c-$\star$	**0.43**	0.38
	CP-O_c-VP	**0.42**	0.38
	CP-STM-4	0.48	**0.59**
Shallow Semantic	SR-O_r-$\star$	0.42	**0.44**
	SR-O_r	**0.88**	0.86
Semantic	DR-O_r-$\star$	0.20	**0.36**
	DR-O_{rp}-$\star$	0.52	**0.60**
	DR-O_r-drs	0.22	**0.37**
	DR-O_r-pred	0.25	**0.33**
	DR-O_r-rel	0.20	**0.45**
	DR-STM-4	0.25	**0.33**

Table 3.24: NIST 2005 Arabic-to-English. Error analysis of test case #728

Part II

Empirical MT

Chapter 4

Statistical Machine Translation

SMT systems are characterized by generating translations using *statistical* models whose parameters are estimated from the analysis of large amounts of bilingual text corpora. SMT is today the dominant approach to Empirical MT. SMT systems can be built very quickly and fully automatically, provided the availability of a parallel corpus aligning sentences from the two languages involved. Several toolkits for the construction of most of its components, including automatic word alignment, language modeling, translation modeling and decoding, have been made available in the last years (Knight et al., 1999; Stolcke, 2002; Och & Ney, 2003; Koehn, 2004a; Crego et al., 2005a; Patry et al., 2006; Koehn et al., 2006; Koehn et al., 2007). Moreover, SMT systems achieve very competitive results, at least when applied to the training domain[1].

In the following, we give an overview of the recent yet fruitful history of SMT. In Section 4.1, we present its fundamentals. Then, in Section 4.2, we describe the extension from word-based to phrase-based translation, as well as some of the most relevant extensions suggested in the last decade, with special focus on the incorporation of linguistic knowledge. Sections 4.3 and 4.4 discuss dedicated approaches to the problems of word ordering and word selection, respectively. Finally, Section 4.5 is a brief note on one of the main problems of SMT and empirical models in general, i.e., their domain dependence.

4.1 Fundamentals

Statistical Machine Translation is based on ideas borrowed from the Information Theory field (Shannon, 1948; Shannon & Weaver, 1949). Weaver (1955) was first to suggest, in his "Translation" memorandum, that cryptographic methods were possibly applicable to MT. However, many years passed until, with the availability of faster computers and large amounts of machine-readable data, these ideas were put into practice. Brown et al. (1988; 1990; 1993) developed, at the IBM TJ Watson Research Center, the first statistical MT system, a French-to-English system called *Candide* trained on a parallel corpus of proceedings from the Canadian Parliament (Berger et al., 1994).

[1]Consult, for instance, official results from the NIST Open MT evaluation series (`http://www.nist.gov/speech/tests/mt/`) and from the shared-tasks at the ACL workshops on MT (`http://www.statmt.org/`).

4.1.1 The Noisy Channel Approach

The main assumption underlying their approach is that the translation process can be seen as a process of transmission of information through a *noisy channel*. During the transmission through the channel, for instance, between brain and mouth, the original signal which encodes a given message in a given source language is distorted into a signal encoding the same message in a different target language. Given a distorted signal, it is possible to approximate the original undistorted signal as long as we count on an accurate model of the distortion process, i.e., the noisy channel the signal went through.

Brown et al. suggested that the distortion process could be modeled using statistical methods. For that purpose, they took the view that every sentence in one language is a possible translation of any sentence in the other. Accordingly, they assigned every pair of sentences (f, e) a probability, $P(e|f)$, to be interpreted as the probability that a human translator will produce e in the target language as a valid translation when presented with f in the source language[2]. Then, based on the noisy channel assumption, the automatic translation of a given source sentence f may be reformulated as the problem of searching for the most probable target sentence e according to the probability table modeling the translation process. In other words, we must choose e so as to maximize $P(e|f)$, denoted: $\hat{e} = \text{argmax}_e\, P(e|f)$. Applying Bayes' rule, $P(e|f)$ may be decomposed as:

$$P(e|f) = \frac{P(f|e) * P(e)}{P(f)} \tag{4.1}$$

Because the denominator does not depend on e, it can be ignored for the purpose of the search:

$$\hat{e} = \underset{e}{\text{argmax}}\, P(e|f) = \underset{e}{\text{argmax}}\, \frac{P(f|e) * P(e)}{P(f)} = \underset{e}{\text{argmax}}\, P(f|e) * P(e) \tag{4.2}$$

Equation 4.2 devises two probability models:

$P(e)$, so-called *language model*, which is typically estimated from large monolingual corpora. The language modeling problem is recasted as the problem of computing the probability of a single word given all the words that precede it in a sentence. Commonly, only the last few preceding words are considered. The reason is that otherwise there would be so many histories that probabilities could not be reliably estimated (Jelinek & Mercer, 1980). However, recently, there is a growing 'brute force' trend for using as long histories as possible, estimated from large amounts of data extracted from the web (Brants et al., 2007). Clearly, the more coverage the better. On the other side, there have been also several works suggesting the construction of more accurate language models based on richer linguistic information, either syntactic (Charniak et al., 2003) or shallow-syntactic (Kirchhoff & Yang, 2005).

[2]For historical reasons, f and e are commonly used to respectively refer to source and target sentences, honoring the French-to-English Candide system.

$P(f|e)$, so-called *translation model*, which is usually estimated from parallel corpora. Originally, translation modeling was approached in a word-for-word basis. An excellent and very detailed report on the mathematics of word-based translation models may be found in (Brown et al., 1993), later extended by Och and Ney (2000). The underlying assumption behind these models is that every word in the target language is a possible translation of any word in the source language. Thus, for every possible source-target word pair (f, e) we must estimate $P(f|e)$, i.e., the probability that f was produced when translating e, usually called *alignment probability*. Word alignment is a vast research topic. Because annotating word alignments in a parallel corpus is a complex and expensive task, first alignment methods were all based on unsupervised learning. The most popular approaches used the Expectation-Maximization (EM) algorithm (Baum, 1972; Dempster et al., 1977). However, recently, there is a growing interest in applying supervised discriminative learning to the problem of word alignment (Taskar et al., 2005; Moore, 2005; Moore et al., 2006; Blunsom & Cohn, 2006; Fraser & Marcu, 2006).

Additionally, Brown et al. described two other models, namely the *distortion* and *fertility* models. The distortion model accounts for explicit word reordering. The fertility model allows for one-to-many alignments, i.e., for the cases in which a word is translated into several words. Words which do not have a translation counterpart are treated as translated into the artificial 'null' word.

Equation 4.2 devises a third component, the *decoder*, responsible for performing the 'argmax' search. MT decoding is a very active research topic. The main difficulty is that performing an optimal decoding requires an exhaustive search, which is an exponential problem in the length of the input (Knight, 1999). Thus, a naive greedy implementation of the decoder is infeasible. Efficient implementations based on dynamic programming techniques exist but for very simple models. When complex reordering models are introduced again exact search is not feasible. For that reason, most decoders perform a suboptimal search usually by introducing reordering constraints or by heuristically pruning the search space. Among the most popular recent approaches to decoding, we may find A* search (Och et al., 2001), greedy search (Germann, 2003), stack-based beam search (Koehn, 2004a), approaches based on integer programming (Germann et al., 2001), based on Graph Theory (Lin, 2004), and based on parsing (Yamada & Knight, 2002; Melamed, 2004).

4.1.2 Word Selection and Word Ordering

As we have seen, SMT systems address the translation task as a search problem. Given an input string in the source language, the goal is to find the output string in the target language which maximizes the product of a series of probability models over the search space defined by all possible phrase partitions of the source string and all possible reorderings of the translated units. This search process implicitly decomposes the translation problem into two separate but interrelated subproblems: word selection and word ordering.

Word selection, also referred to as *lexical choice*, is the problem of deciding, given a word f in the source sentence, which word e in the target sentence is the appropriate translation. This problem is mainly addressed by the translation model $P(f|e)$, which serves as a probabilistic bilingual dictionary. Translation models provide for each word in the source vocabulary a list of translation

candidates with associated translation probabilities. During the search there is another component which addresses word selection, the language model. This component helps the decoder to move towards translations which are more appropriate, in terms of grammaticality, in the context of the target sentence.

Word ordering refers to the problem of deciding which position must the word translation candidate e occupy in the target sentence. This problem is mainly addressed by the reordering model, which allows for certain word movement inside the sentence. Again, the language model may help the decoder, in this case to move towards translations which preserve a better word ordering according to the syntax of the target language.

4.2 Phrase-based Translation

Word translation models suggested by Brown et al. exhibit a main deficiency: the translation modeling of the source context in which words occur is very weak. Translation probabilities, $P(f|e)$, do not take into account, for instance, which are the words surrounding f and e. Thus, this information is ignored for the purpose of word selection. These models are, therefore, also unable to provide satisfactory translations for the case of non-compositional phrases.

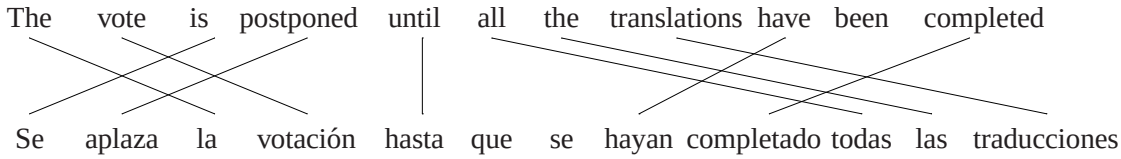
On the other hand, it is well known that the translation process does not actually occur on a word-for-word basis. On the contrary, there are many regularities in phrasal movement (Fox, 2002); words inside a phrase tend to stay together during translation. Therefore, a straightforward improvement to word-based models consists in extending the scope of the translation unit, i.e., moving from words to *phrases*[3]. Phrase-based models allow for *many-to-many* alignments, thus capturing phrasal cohesion in a very natural way. Phrase-based models take local word context into account, and allow for translation of non-compositional phrases.

4.2.1 Approaches

A number of approaches to the estimation of phrase-based models have been suggested. Wang (1998) and Wang and Waibel (1998) were the first to demonstrate the intuition shared with many other researchers that word-based alignment was a major cause of errors in MT. They proposed a new alignment model based on shallow phrase structures automatically acquired from a parallel corpus. At the same time, Alshawi et al. (1998) suggested a method for fully automatic learning of hierarchical finite state translation models in which phrases are modeled by the topology of the transducers.

But the most influential approach was that by Och et al. (1999) who presented a phrase-based translation system in which phrases are modeled as *alignment templates*. They used phrases rather than single words as the basis for the alignment models. Phrases were automatically induced from word alignments in a parallel corpus. A group of adjacent words in the source sentence could be aligned to a group of adjacent words in the target only if they were consistent with word alignments, i.e., there was no word in either phrase which was aligned to a word outside their counterpart phrase.

[3]The term *'phrase'* used hereafter in this context refers to a sequence of words not necessarily syntactically motivated.



The vote is postponed	Se aplaza la votación
The vote is postponed until	Se aplaza la votación hasta
The vote is postponed until	Se aplaza la votación hasta que
The vote is postponed until	Se aplaza la votación hasta que se
The vote is postponed until all the translations have been completed	Se aplaza la votación hasta que se hayan completado todas las traducciones
until	hasta
until	hasta que
until	hasta que se
until all the translations have been completed	hasta que se hayan completado todas las traducciones
all the translations have been completed	que se hayan completado todas las traducciones
all the translations have been completed	se hayan completado todas las traducciones

Figure 4.1: Phrase Extraction. An example

An example of phrase alignment and phrase extraction is showed in Figure 4.1. This corresponds to a case of English-Spanish translation extracted from the corpus of European Parliament Proceedings described in Section 5.1.1. At the top, the reader may find an English sentence and its Spanish translation, as well as a set of possible word alignments denoted by lines connecting words in both sentences. Below, we show all bilingual phrase pairs that can be extracted following the phrase-extract algorithm described in (Och, 2002). Each row corresponds to a phrase pair. English and Spanish counterparts appear at the right and left, respectively. This example evinces also the importance of the translation task in the sessions of the European Parliament. Let us mention, as a curiosity, that over 10% of the times the word *'translation'* was found in the corpus, it happened to occur together with the word *'problem'*.

Alignment templates associate source and target phrases and represent the correspondence between the words in each phrase by keeping the word alignment information. Koehn et al. (2003) suggested a simpler approach, in which word alignments are removed from phrase translation pairs, obtaining similar results. This approach will constitute our baseline system in the following chapters.

All the models listed above are similar in that they estimate the *conditional probability* that a target phrase e is generated as the appropriate translation of the source phrase f, i.e., $P(e|f)$. In contrast, Marcu and Wong (2002) presented a phrase-based *joint probability* model which does not

try to capture how source phrases are mapped into target phrases, but rather how source and target phrases could have been generated simultaneously out from a bag of concepts, i.e., $P(f, e)$. The main drawback of their approach was related to the computational cost of training and decoding algorithms in terms of efficiency and memory requirements.

Other approaches to joint probability translation model exist. For instance, Tillmann and Xia (2003) suggested a unigram phrase-based model based on bilingual phrase units, they called *blocks*. More recently, Mariño et al. (2005b; 2006) suggested an interesting joint probability model based on bilingual phrase units, so-called *tuples*. Their proposal presents the particularity that translation modeling is addressed as a bilingual language modeling problem. In this manner, their models can take full advantage of standard back-off n-gram smoothing techniques applied in regular language modeling.

4.2.2 The Log-linear Scheme

The phrase-based approach was extended by Och and Ney (2002) so as to allow for considering additional arbitrary *feature functions* further than the language and translation probability models. Formally:

$$\hat{e} = \underset{e}{\operatorname{argmax}} \{\log\ P(e|f)\} = \underset{e}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(f, e) \right\} \tag{4.3}$$

The weight (λ_m) of each feature function (h_m) is adjusted through discriminative training based on the maximum entropy principle (Berger et al., 1996). However, the application of this approach is limited by the feasibility of computing feature functions during the search. In other words, complex feature functions which may not be efficiently handled by the decoder are impractical. For that reason, today, feature functions are typically limited to alternative language and translation models, as well as brevity penalty functions, and lexical weightings.

A crucial aspect of this approach is the adjustment of parameters. The default optimization criterion is intended to minimize the number of wrong decisions over the training data. However, Och (2003) argued that there is a mismatch between this criterion and MT quality evaluation measures. He suggested an alternative optimization strategy in which these parameters are adjusted so as to minimize the translation error rate of the system, as measured by an automatic evaluation metric at choice, typically BLEU.

4.2.3 Other Extensions

A major shortcoming of standard phrase-based SMT models is that they do not make use of explicit linguistic knowledge. Sentences are treated as sequences of words, and words are treated as atomic units. In order to overcome this limitation, several extensions to standard phrase-based models have been proposed. Below, we briefly describe some of the most relevant, without entering yet into *syntax-based* models, which will be described in Section 4.3.

Word Classes. Och and Ney (2000) revised and improved the parameter estimation of word alignment models proposed by Brown et al. (1993) through the incorporation of word classes automatically trained from parallel corpora. These word classes grouped together words that are realized in similar contexts, and could be thought of, in some way, as unsupervised parts-of-speech.

Reranking of N-best lists. Following the ideas by Collins et al. (2000; 2005) for the reranking of syntactic parse trees, Och et al. (2003) defined a *reranking* approach to SMT. Instead of producing a single best translation, their system generates a series of n best candidates which are then reranked according to a collection of linguistic features (Shen et al., 2004). The top ranked translation is selected as the system output. The main advantage of this method is that it allows for introducing a number of global sentence features (e.g., about overall sentence grammaticality or semantic structure) without increasing decoding complexity, although at the cost of possibly discarding valid translations when compiling the n-best list.

Och et al. (2003; 2004) suggested a smorgasbord of more than 450 syntactically motivated different feature functions for the reranking of 1000-best lists of candidates applied to Chinese-to-English translation. However, only a moderate improvement, in terms of BLEU score, was reported (see Section 2.2.3). They argued that linguistic processors introduce many errors and that BLEU is not specially sensitive to the grammaticality of MT output.

Reranking techniques have been also successfully applied to other NLP tasks such as Semantic Role Labeling (Toutanova et al., 2005; Haghighi et al., 2005; Toutanova et al., 2008).

Dedicated Local Reordering. Tillmann and Zhang (2005) suggested using discriminative models based on maximum entropy to model word reordering. Their models allowed for restricted local block swapping.

Translation Based on Shallow Parsing. Several approaches to exploiting morphological and shallow syntactic information for the estimation of phrase-based translation models have been suggested. For instance, Koehn and Knight (2002) proposed, in their *ChunkMT* system, to integrate morphosyntactic analysis (part-of-speech tags) and shallow parsing (base phrase chunks). They obtained promising results. However, the applicability of their work was limited to very short sentences. They later abandoned this approach and focused on the particular case of noun phrase translation. They developed special modeling and features which integrated into a phrase-based SMT system (Koehn, 2003b).

Schafer and Yarowsky (2003) suggested a combination of models based on shallow syntactic analysis (part-of-speech tagging, lemmatization and base phrase chunking). They followed a back-off strategy in the application of their models. Decoding was based on *finite state automata*. Although no significant improvement in MT quality was reported, results were promising taking into account the short time spent in the development of the linguistic tools.

Koehn et al. (2003) published a very interesting negative result, in the view of later research. They found that limiting the phrases in a standard phrase-based translation model only to those syntactically motivated severely harmed the system performance.

In Chapter 5, we present an approach based on shallow parsing which allows to softly integrate translation models based on richer linguistic information with phrase-based models at the lexical level. We show that it is possible to robustly combine translation models based on different kinds of information, yielding a moderate improvement according to several standard automatic MT evaluation metrics. We also show that the quality loss noted by Koehn et al. (2003) when limiting to a set of syntactically motivated phrases is mainly related to a drop in recall.

Factored Models. These models are an extension of phrase-based translation models which, instead of simple words, allow for using a factored representation, i.e., a feature vector for each word derived from a variety of information sources (Koehn et al., 2006; Koehn & Hoang, 2007). These features may be the surface form, lemma, stem, part-of-speech tag, morphological information, syntactic, semantic or automatically derived categories, etc. This representation is then used to construct statistical translation models that can be combined together to maximize translation quality. An implementation of factored MT models is available inside the Moses SMT toolkit.

The work described in Chapter 5 can be also reframed as factored MT.

4.3 Syntax-based Translation

Another limitation of standard phrase-based systems is that reordering models are very simple. For instance, non-contiguous phrases are not allowed, long distance dependencies are not modeled, and syntactic transformations are not captured. Syntax-based approaches seek to remedy these deficiencies by explicitly taking into account syntactic knowledge. Approaches to *syntax-based MT* differ in several aspects: (i) side of parsing (source, target, or both sides), (ii) type of parsing (dependencies vs. constituents), (iii) modeling of probabilities (generative vs. discriminative), (iv) core (structured predictions vs. transformation rules), and (v) type of decoding (standard phrase-based, modeled by transducers, based on parsing, graph-based). Below, we list some of the most relevant approaches. We group them in three different families:

Bilingual Parsing. The translation process is approached as a case of synchronous bilingual parsing. Derivation rules are automatically learned from parallel corpora, either annotated or unannotated (Wu, 1997; Wu, 2000; Alshawi, 1996; Alshawi et al., 2000; Melamed, 2004; Melamed et al., 2005; Chiang, 2005; Chiang, 2007). Below, we briefly describe some selected works in this family:

- Wu (1997; 2000) presented a stochastic *inversion transduction grammar* formalism for bilingual language modeling of sentence pairs. They introduced the concept of *bilingual parsing* and applied it, among other tasks, to phrasal alignment.
- Dependency translation models by Alshawi (1996; 2000), based on finite state transducers may be seen as well as a case of bilingual parsing.
- Melamed (2004) and Melamed et al. (2005) suggested approaching MT as synchronous parsing, based on multitext grammars, in which the input can have fewer dimensions

than the grammar. However, their approach requires the availability of annotated multitreebanks[4] which are very expensive to build.

- Chiang (2005; 2007) proposed a phrase-based model that uses hierarchical phrases, i.e., phrases that contain subphrases. Their model is formally a *synchronous context-free grammar* but is learned from a bitext without any syntactic information.

Tree-to-String, String-to-Tree and Tree-to-Tree Models. These models exploit syntactic annotation, either in the source or target language or both, to estimate more informed translation and reordering models or translation rules (Yamada & Knight, 2001; Yamada, 2002; Gildea, 2003; Lin, 2004; Quirk et al., 2005; Cowan et al., 2006; Galley et al., 2006; Marcu et al., 2006). Below, we describe a selection of the most relevant works:

- Yamada and Knight (2001; 2002) presented a syntax based tree-to-string probability model which transforms a source language parse tree into a target string by applying stochastic operations at each node. Decoding is approached following a CKY-alike parsing algorithm. However, they did not obtain any improvement in terms of BLEU.

  Some time later, in a joint effort with Eugene Charniak, they presented a syntax-based language model based upon the language model described in (Charniak, 2001), which combined with their syntax based translation model, achieved a notable improvement in terms of grammaticality (Charniak et al., 2003). This improvement was measured following a process of manual evaluation. Interestingly, the BLEU metric was unable to reflect it.

- Gildea (2003) presented a study on tree-to-tree translation models. In spite of their degree of sophistication these models did not achieve significant improvements on standard evaluation metrics. Gildea (2004) tried also working with dependency trees instead of constituents. They found constituent trees to perform better.

  Later, Zhang and Gildea (2004) made a direct comparison between syntactically supervised and unsupervised syntax-based alignment models. Specifically, they compared the unsupervised model by Wu (1997) to the supervised model by Yamada and Knight (2001). They concluded that automatically derived trees resulted in better agreement with human-annotated word-level alignments for unseen test data.

- Cowan et al. (2006) presented a discriminative model for tree-to-tree translation based on the concept of *aligned extended projection* (AEP). AEPs are structures that contain information about the main verb of a clause and its arguments, as well as the relationship between source-language arguments and target-language arguments (i.e., their alignment to one another). These structures allow for solving translation problems such as missing or misplaced verbs, subjects, and objects. They suggested a number of features for the case of German-to-English translation, and used the perceptron algorithm to learn their weights. They reported a performance in the same range of standard phrase-based approaches (Koehn et al., 2003). Similar proposals have been suggested by other authors (Chang & Toutanova, 2007).

[4] A multitreebank is basically a multilingual parsed parallel corpus in which constituents are aligned.

- Lin (2004) proposed a path-based transfer model using dependency trees. They suggested a training algorithm that extracts a set of rules that transform a path in the source dependency tree into a fragment in the target dependency tree. Decoding was formulated as a graph-theoretic problem of finding the minimum path covering the source dependency tree. Results were under the performance of not syntactically motivated phrase-based models.
- Quirk et al. (2005) suggested a tree-based ordering model based on dependency parsing of the source side. They introduced the concept of treelet, defined as an arbitrary connected subgraph in a dependency tree. Their model had the particularity of allowing for reordering of discontinuous structures. Significant improvements were reported on small-scale domain-specific test sets.
- Galley et al. (2004; 2006) suggested approaching translation as the application of syntactically informed transformation rules. They used the framework by Graehl and Knight (2004; 2005) based on finite state tree-to-tree and tree-to-string transducers. Results presented are promising.
- Marcu et al. (2006) presented a syntactified target language translation model. Phrases were decorated with syntactic constituent information. Their models also relied on the extended tree-to-string transducers introduced by Graehl and Knight (2004; 2005). Significant improvements on a large-scale open-domain translation task were reported according to both automatic and manual evaluation.

Source Reordering. Another interesting approach consists in reordering the source text prior to translation using syntactic information so it shapes to the appropriate word ordering of the target language (Collins et al., 2005; Crego et al., 2006; Li et al., 2007). Significant improvements have been reported using this technique.

As it can be seen, many efforts are being devoted to the construction of syntactically informed SMT systems. Indeed, syntax-based models have became state-of-the-art among SMT systems, proving slightly more effective than top-quality phrase-based systems when applied to distant language pairs such as Chinese-to-English which present important differences in word ordering.

4.4 Dedicated Word Selection

Another major limitation of the standard phrase-based approach is that word (or phrase) selection is poorly modeled. In particular, the source sentence context in which phrases occur is ignored. Thus, all the occurrences of the same source phrase are assigned, no matter what the context is, the same set of translation probabilities. For instance, the phrase *'brilliant play'* in the text segment *"A brilliant play written by William Locke"* would receive the same translation probabilities when appearing in the segment *"A brilliant play by Ronaldinho that produced a wonderful goal"*. Thus, phrase selection takes place, during decoding, with the only further assistance of the language model, which involves knowledge only about the target context. Besides, in most cases translation

probabilities are estimated on the basis of relative frequency counts, i.e., Maximum Likelihood Estimates (MLE).

For these reasons, recently, there is a growing interest in the application of discriminative learning to word selection (Bangalore et al., 2007; Carpuat & Wu, 2007b; Giménez & Màrquez, 2007a; Stroppa et al., 2007; Vickrey et al., 2005). Discriminative models allow for taking into account a richer feature context, and probability estimates are more informed than simple frequency counts.

Interest in discriminative word selection has also been motivated by recent results in Word Sense Disambiguation (WSD). The reason is that SMT systems perform an implicit kind of WSD, except that instead of working with word senses, SMT systems operate directly on their potential translations. Indeed, recent semantic evaluation campaigns have treated word selection as a separate task, under the name of *multilingual lexical sample* (Chklovski et al., 2004; Jin et al., 2007). Therefore, the same discriminative approaches which have been successfully applied to WSD, should be also applicable to SMT. In that spirit, instead of relying on MLE for the construction of the translation models, approaches to discriminative word selection suggest building dedicated translation models which are able to take into account a wider feature context. Lexical selection is addressed as a classification task. For each possible source word (or phrase) according to a given bilingual lexical inventory (e.g., the translation model), a distinct classifier is trained to predict lexical correspondences based on local context. Thus, during decoding, for every distinct instance of every source phrase a distinct context-aware translation probability distribution is potentially available.

Brown et al. (1991a; 1991b) were the first to suggest using dedicated WSD models in SMT. In a pilot experiment, they integrated a WSD system based on mutual information into their French-to-English word-based SMT system. Results were limited to the case of binary disambiguation, i.e., deciding between only two possible translation candidates, and to a reduced set of very common words. A significantly improved translation quality was reported according to a process of manual evaluation. However, apparently, they abandoned this line of research.

Some years passed until these ideas were recovered by Carpuat and Wu (2005b), who suggested integrating WSD predictions into a phrase-based SMT system. In a first approach, they did so in a *hard* manner, either for decoding, by constraining the set of acceptable word translation candidates, or for post-processing the SMT system output, by directly replacing the translation of each selected word with the WSD system prediction. However, they did not manage to improve MT quality. They encountered several problems inherent to the SMT architecture. In particular, they described what they called the *language model effect* in SMT: *"The lexical choices are made in a way that heavily prefers phrasal cohesion in the output target sentence, as scored by the language model"*. This problem is a direct consequence of the hard interaction between their WSD and SMT systems. WSD predictions cannot adapt to the surrounding target context. In a later work, Carpuat and Wu (2005a) analyzed the converse question, i.e., they measured the WSD performance of SMT systems. They showed that dedicated WSD models significantly outperform the WSD ability of current state-of-the-art SMT models. Consequently, SMT should benefit from WSD predictions.

Simultaneously, Vickrey et al. (2005) studied the application of *context-aware* discriminative word selection models based on WSD to SMT. Similarly to Brown et al. (1991b), they worked with translation candidates instead of word senses, although their models were based on maximum entropy and dealt with a larger set of source words and higher levels of ambiguity. However, they did not approach the full translation task but limited to the *blank-filling* task, a simplified version of

the translation task, in which the target context surrounding the word translation is available. They did not encounter the language model effect because: (i) the target context was fixed a priori, and (ii) they approached the task in a soft way, i.e., allowing WSD-based probabilities to interact with other models during decoding.

Following similar approaches to that of Vickrey et al. (2005), Cabezas and Resnik (2005) and Carpuat et al. (2006) used WSD-based models in the context of the full translation task to aid a phrase-based SMT system. They reported a small improvement in terms of BLEU score, possibly because they did not work with phrases but limited to single words. Besides, they did not allow WSD-based predictions to interact with other translation probabilities.

More recently, other of authors, including ourselves, have extended these works by moving from words to phrases and allowing discriminative models to cooperate with other phrase translation models as an additional feature. Moderate improvements have been reported (Bangalore et al., 2007; Carpuat & Wu, 2007b; Carpuat & Wu, 2007a; Giménez & Màrquez, 2007a; Giménez & Màrquez, 2008a; Stroppa et al., 2007; Venkatapathy & Bangalore, 2007). All these works were being elaborated at the same time, and were presented in very near dates with very similar conclusions. We further discuss the differences between them in Chapter 6.

Other integration strategies have been tried. For instance, Specia et al. (2008) used dedicated predictions for the reranking of n-best translations. Their models were based on Inductive Logic Programming (ILP) techniques (Specia et al., 2007). They limited to a small set of words from different grammatical categories. A very significant BLEU improvement was reported.

In a different approach, Chan et al. (2007) used a WSD system to provide additional features for the hierarchical phrase-based SMT system based on bilingual parsing developed by Chiang (2005; 2007). These features were intended to give a bigger weight to the application of rules that are consistent with WSD predictions. A moderate but significant BLEU improvement was reported.

Finally, Sánchez-Martínez et al. (2007) integrated a simple lexical selector, based on source lemma co-occurrences in a very local scope, into their hybrid corpus-based/rule-based MT system.

Overall, apart from showing that this is a very active research topic, most of the works listed in this section evince that dedicated word selection models might be useful for the purpose of MT. Our approach to discriminative phrase selection will be deeply described in Chapter 6. Further details on the comparison among other approaches and ours will be also discussed in Section 6.4.

4.5 Domain Dependence

One of the main criticisms against empirical methods in NLP is their strong domain dependence. Since parameters are estimated from a corpus belonging to a specific domain, the performance of the system on a different domain is often much worse. This flaw of statistical and machine learning approaches is well known and has been largely described in recent literature for a variety of tasks such as parsing (Sekine, 1997), word sense disambiguation (Escudero et al., 2000), and semantic role labeling (He & Gildea, 2006).

In the case of SMT, domain dependence has very negative effects in translation quality. For instance, in the 2007 edition of the ACL MT workshop (WMT07), an extensive comparative study

between in-domain and out-of-domain performance of MT systems built for several European languages was conducted (Callison-Burch et al., 2007). Results showed a significant drop in MT quality consistently according to a number of automatic evaluation metrics for all statistical systems. In contrast, the decrease reported in the case of rule-based or hybrid MT systems was less significant or inexistent. Even, in some cases their out-of-domain performance was higher than in-domain. The reason is that, while these systems are often built on the assumption of an open or general domain, SMT systems are heavily specialized on the training corpora. A change in domain implies a significant shift in the sublanguage (i.e., lexical choice and lexical order) employed, and, consequently, statistical models suffer a significant lack both of recall —due to unseen events— and precision —because event probability distributions differ substantially. Notice that we intentionally talk about *events* instead of words or phrases. In this manner, we have intended to emphasize that the decrease is not only due to unknown vocabulary, but also to other types of linguistic phenomena, such as syntactic or semantic structures, either unseen or seen in different contexts. In other words, domain dependence is not only a problem related to lexical selection, but also to other aspects such as syntactic ordering and semantic interpretations.

Domain adaptability is, thus, a need for empirical MT systems. Interest in domain adaptation lies in the fact that while there are large amounts of data electronically available (e.g., in the web), most often, these belong to a specific domain which is not always the target application domain. Typically, none or very few in-domain data are available. For that reason, domain adaptation is a very active research topic. For instance, the special challenge of the WMT07 shared-task was on domain adaptation. Several interesting approaches were suggested (Civera & Juan, 2007; Koehn & Schroeder, 2007).

Other authors have looked at the same problem the other way around. For instance, Vogel and Tribble (2002) studied whether an speech-to-speech SMT system built on a small in-domain parallel corpus could be improved by adding out-of-domain knowledge sources.

In Chapter 7, we discuss the problem of domain dependence in the context of SMT and present several techniques which can be applied so as to mitigate its negative effects.

Chapter 5

Shallow Syntactic Alignments and Translation Models

As we have seen in Sections 4.2 and 4.3, in the last years, there is a growing interest in the incorporation of linguistic knowledge into SMT systems. For instance, the use of syntactic information has lead to notable improvements, particularly in terms of word ordering, e.g., approaches based on bilingual parsing (Chiang, 2005), source reordering (Collins et al., 2005; Li et al., 2007), and syntactified target language models (Charniak et al., 2003; Kirchhoff & Yang, 2005; Marcu et al., 2006). However, dedicated reordering models (Quirk et al., 2005; Cowan et al., 2006; Chang & Toutanova, 2007), syntax-based translation models (Yamada & Knight, 2001; Gildea, 2003), or approaches based on using syntactic information for the reranking of n-best translations (Och et al. 2003; 2004), have only reported moderate improvements. As possible reasons for these results researchers have argued that (i) current metrics, such as BLEU, are not able to capture syntactic improvements, and that (ii) linguistic processors, often trained on out-of-domain data, introduce many errors. In addition, we argue that a third possible cause is data sparsity. While the translation between two languages may involve a wide range of possible syntactic movements, their observation in training data is often very sparse, thus leading to poor parameter estimations.

In this chapter, we present a simple approach for the incorporation of linguistic knowledge into translation models. Instead of modeling syntactic reordering, we suggest exploiting shallow syntactic information for the purpose of lexical selection. Our approach is similar to the so-called *factored machine translation models* which have emerged very recently (Koehn et al., 2006; Koehn et al., 2007; Koehn & Hoang, 2007). First, we redefine the translation unit so it may contain additional linguistic information beyond the lexical level. Then, following the standard approach, we build word alignments over these enriched translation units and perform phrase extraction over these alignments. Resulting translation models, based on different types of information, are then suitable for being combined as additional features in the log-linear scheme (Och & Ney, 2002), yielding a significantly improved translation quality.

The rest of the chapter is organized as follows. First, in Section 5.1, we describe the construction of a phrase-based baseline system. Then, in Section 5.2, we give the details of our proposal. Experimental results are presented in Section 5.2.2. Main conclusions are summarized in Section5.3.

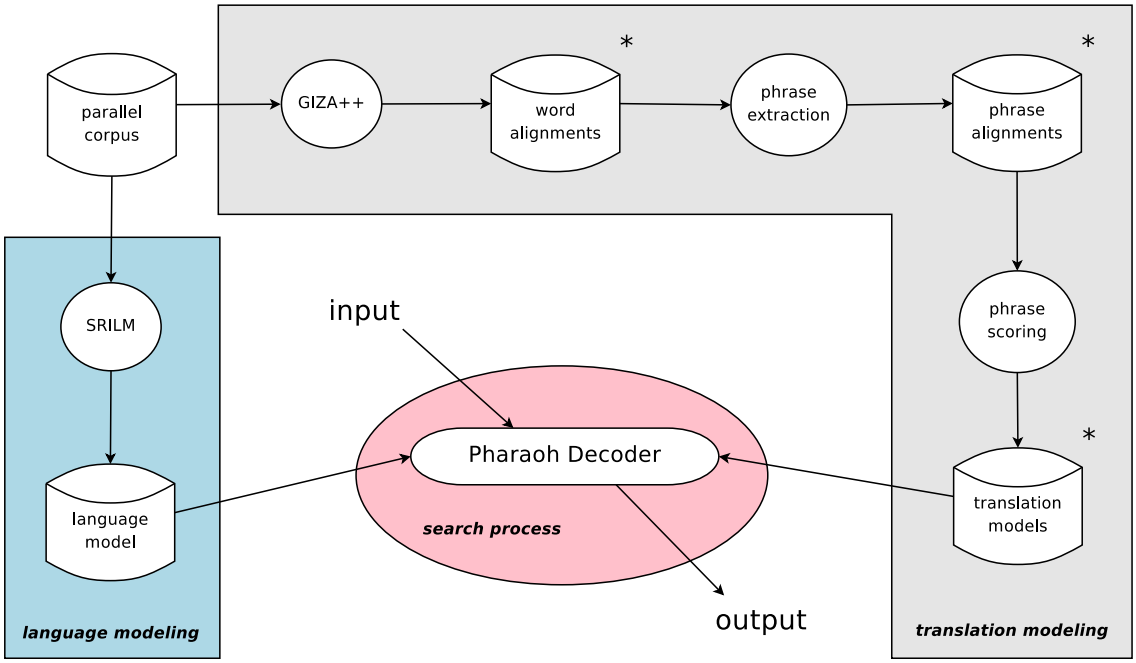


Figure 5.1: Architecture of the baseline phrase-based SMT system

5.1 Building a Baseline System

Our baseline system implements a standard phrase-based SMT architecture (see Figure 5.1). This involves, as seen in Section 4.1, three main components: (i) translation model(s), (ii) language model(s), and (iii) the decoder.

For translation modeling, we follow the approach by Koehn et al. (2003) in which phrase pairs are automatically induced from word alignments. These are generated using the *GIZA++ SMT Toolkit* in its default configuration, i.e., 5 iterations for IBM model 1, 4 iterations for IBM model 3, 3 iterations for IBM model 4, and 5 iterations for HMM model (Och & Ney, 2003)[1]. Phrase extraction is performed following the *phrase-extract* algorithm described by Och (2002). This algorithm takes as input a word aligned parallel corpus and returns, for each sentence, a set of phrase pairs that are *consistent* with word alignments. A phrase pair is said to be consistent with the word alignment if all the words within the source phrase are only aligned to words within the target phrase, and vice versa. We work with the union of source-to-target and target-to-source word alignments, with no heuristic refinement. Only phrases up to length five are considered. Also, phrase pairs appearing only once are discarded, and phrase pairs in which the source/target phrase is more than three times longer than the target/source phrase are ignored. Phrase pairs are scored on the basis of relative frequency (i.e., Maximum Likelihood Estimates). Formally, let ph_f be a phrase in the source language (f) and ph_e a phrase in the target language (e), we define a function $count(ph_f, ph_e)$ which counts the number of times the phrase ph_f has been seen aligned to phrase ph_e in the training data. The

[1] `http://www.fjoch.com/GIZA++.html`

conditional probability that ph_f maps into ph_e is estimated as:

$$P(ph_f|ph_e) = \frac{count(ph_f, ph_e)}{\sum_{ph_f} count(ph_f, ph_e)} \tag{5.1}$$

For language modeling, we use the *SRI Language Modeling Toolkit*[2] (SRILM) (Stolcke, 2002). SRILM supports creation and evaluation of a variety of language model types based on N-gram statistics, as well as several related tasks, such as statistical tagging and manipulation of N-best lists and word lattices. We build trigram language models applying linear interpolation and Kneser-Ney discounting for smoothing.

Regarding the 'argmax' search, we use the *Pharaoh*[3] beam search decoder (Koehn, 2004a), which naturally fits with the previous tools. *Pharaoh* is an implementation of an efficient dynamic programming stack-based search algorithm with lattice generation and XML markup for external components. In order to speed up the translation process, we have fixed several of the decoder parameters. In particular, we have limited the number of candidate translations to 30, the maximum beam size (i.e., stack size) to 300, and used a beam threshold of 10^{-5} for pruning the search space. We have also set a distortion limit of 6 positions.

We extend the baseline by combining generative and discriminative translation models, $P(e|f)$ and $P(f|e)$, following the log-linear formulation suggested by Och and Ney (2002). See Section 4.2.2. We have used the Pharaoh's default heuristic distortion model and word penalty feature.

Let us also note that, keeping with usual practice, prior to building translation and language models, the parallel corpus is case lowered. However, for the purpose of evaluation, word case is automatically recovered using the *Moses* package (Koehn et al. 2006; 2007). We did not use Moses for decoding because most of the experimental work is previous to its public release.

5.1.1 Data Sets

We have constructed our system using the *'EuroParl'* parallel corpus of *European Parliament Proceedings* (Koehn, 2003a)[4]. Specifically, we have used the Europarl release from the Openlab 2006 Initiative[5] promoted by the TC-STAR Consortium[6]. This test suite is entirely based on European Parliament Proceedings covering April 1996 to May 2005.

We have focused on the Spanish-to-English translation task. Figure 5.2 shows a short fragment extracted from the Spanish-English EuroParl parallel corpus. The training set consists of 1,272,046 parallel sentences. Besides, for evaluation purposes we count on a separate set of 1,008 sentences. Three human references per sentence are available. We have randomly split this set in two halves, which are respectively used for development and test. A brief numerical description of the data sets is available in Table 5.1. We show the number of sentences and words after tokenization. As to the vocabulary size, we give the number of distinct words after case lowering.

[2] `http://www.speech.sri.com/projects/srilm/download.html`.
[3] `http://www.isi.edu/licensed-sw/pharaoh/`.
[4] `http://www.statmt.org/europarl/`
[5] `http://tc-star.itc.it/openlab2006/`
[6] `http://www.tc-star.org/`

```
[English]

1. President .
2. - I declare resumed the session of the European Parliament adjourned
   on Thursday , 28 March 1996 .
3. President .
4. - Ladies and gentlemen , on behalf of the House let me welcome a
   delegation from the Grand Committee of the Finnish Parliament , i.e.
   , the European Affairs Committee of the Finnish Parliament , led
   by its chairman , Mr Erkki Tuomioja . I bid you a warm welcome !
5. ( Applause )
6. We are pleased at this visit , which reflects the increasingly close
   cooperation between us and the national parliaments in the Union ,
   and I wish our Finnish colleagues a pleasant stay in Strasbourg and
   , of course , useful and interesting discussions in this House !
7. President .
8. - The Minutes of the sitting of Thursday , 28 March 1996 have been
   distributed .
9. Are there any comments ?


[Spanish]

1. El Presidente .
2. - Declaro reanudado el período de sesiones del Parlamento Europeo
   , interrumpido el 28 de marzo de 1996 .
3. El Presidente .
4. - Deseo dar la bienvenida a los miembros de una delegación de
   la  '' Gran Comisión '' , es decir , la Comisión de Asuntos
   Europeos , del Parlamento finlandés , dirigida por su Presidente
   , el Sr. Erkki Tuomioja , delegación que acaba de llegar a la
   tribuna de invitados .
5. ( Aplausos )
6. Nos alegramos de esta visita , que se enmarca en la cooperación cada
   vez más estrecha entre nosotros y los Parlamentos nacionales de la
   Unión . Deseo que nuestros colegas finlandeses tengan una agradable
   estancia en Estrasburgo y también , naturalmente , que tengamos
   ocasión de hablar en esta Asamblea de manera provechosa e interesante .
7. El Presidente .
8. - El Acta de la sesión del jueves 28 de marzo de 1996 ha sido
   distribuida .
9. ¿ Hay alguna observación ?
```

Figure 5.2: A short fragment of the Spanish-English Europarl parallel corpus

	Set	#sentences	#tokens	#distinct tokens
Spanish	Train	1,272,046	36,072,505	138,056
	Test	504	13,002	2,471
	Dev	504	12,731	2,386
English	Train	1,272,046	34,590,575	94,604
	Test	504	13,219	2,103
	Dev	504	12,851	2,010

Table 5.1: Description of the Spanish-English corpus of European Parliament Proceedings

5.1.2 Adjustment of Parameters

The adjustment of the parameters that control the contribution of each log-linear feature during the search is of critical importance in SMT systems. Most commonly, a minimum error rate training (MERT) strategy is followed (Och, 2003). A certain number of parameter configurations are tried for the translation of a held-out development data set. At the end of the process, the configuration yielding the highest score, according to a given automatic evaluation measure at choice, typically BLEU, is selected to translate the test set.

In our case, a greedy iterative optimization strategy is followed. In the first iteration only two values $\min$ and $\max$, taken as preliminary minimum and maximum values, are tried for each parameter. For translation, language, and distortion models, first values are $\{0.1, 1\}$. For word penalty, values are $\{-3, 3\}$. In each following iteration, n values in the interval centered at the top scoring value from the previous iteration are explored at a resolution of $\frac{1}{n}$ the resolution of the previous iteration. The resolution of the first iteration is $\max - \min$. The process is repeated until a maximum number of iterations I is reached. In our experiments we have set $n = 2$ and $I = 5$. In that manner, the number of configurations visited, with possible repetitions, is: $2^{t+3} + (I-1) * 3^{t+3}$, where t is the number of translation models utilized. Thus, $t+3$ considers a single language model, word penalty and distortion model. For instance, in the default setting, in which two translation models are used (i.e., $P(e|f)$ and $P(f|e)$), the optimization algorithm inspects 1,004 parameter configurations ($1004 = 2^5 + 4 * 3^5$).

Unless stated otherwise, system optimization is guided by the BLEU measure.

5.1.3 Performance

Prior to improving the baseline system, we analyze its performance. Experimental results are showed in Table 5.2. Based on the meta-evaluation results from Section 3.2.2, we have selected several metrics at the lexical level obtaining high levels of correlation with human assessments at the evaluation of the Spanish-to-English translation of European Parliament Proceedings.

First, we study the impact of the symmetrization heuristics used during phrase extraction. Four different methods are compared:

System	METEOR (wnsyn)	ROUGE (w_1.2)	GTM (e=2)	BLEU
∅	0.7422	0.4327	0.4158	0.6135
∩	0.7380	0.4291	0.4106	0.6010
∩/∪	0.7467	0.4360	0.4196	0.6166
∪	**0.7528**	0.4370	0.4217	0.6217
∪+	0.7512	**0.4375**	**0.4230**	**0.6234**
SYSTRAN	0.7175	0.3971	0.3621	*0.4910*

Table 5.2: Baseline system. Automatic evaluation of MT results

- ∅ ⟶ no symmetrization (only source-to-target word alignments)
- ∩ ⟶ intersection of source-to-target an target-to-source word alignments.
- ∪ ⟶ union of source-to-target an target-to-source word alignments.
- ∩/∪ ⟶ exploring the space between the intersection and the union of word alignments, as described by Och and Ney (2003).

It can be observed that, over this test bed, best results are obtained when using the union of word alignments, consistently according to all metrics, with a slight but significant advantage over exploring the space between the union and the intersection. Interestingly, working on the intersection is worse than skipping symmetrization. Unless stated otherwise, statistical significance of evaluation results is verified using the bootstrap resampling test described by Koehn (2004b), applied over the BLEU metric and based on 1,000 test samples.

Second, we study the influence of the phrase length. We compare the default setting, applying the union heuristic limited to length-5 phrases to a setting in which phrases up to length 10 are allowed. It can be observed ('∪+' row) that incorporating longer phrases reports a minimal improvement. Therefore, for the sake of efficiency, in the rest of the chapter, we will use the '∪' system as our baseline, and will apply this same heuristic in the construction of all phrase tables.

As a complementary issue, we compare the baseline system to a general-purpose commercial MT system, SYSTRAN[7], based on manually-defined lexical and syntactic transfer rules. As expected, the performance of the out-of-domain rule-based system is significantly lower (see last row in Table 5.2), specially in terms of BLEU. We have applied the methodology for heterogeneous MT evaluation described in Chapter 3 to further analyze the differences between SYSTRAN and the baseline system based on the '∪' heuristic (see Table 5.3). Interestingly, although all lexical metrics consider that the SMT system is significantly better than SYSTRAN, according to several syntactic and semantic metrics, the difference between both systems is much smaller (see highlighted values). For instance, metrics based on head-word chain matching over dependency relationships

[7]We use the on-line version 5.0 of SYSTRAN, available at `http://www.systransoft.com/`.

Level	Metric	SYSTRAN	SMT baseline
Lexical	1-PER	0.7131	0.7657
	1-TER	0.6408	0.6969
	1-WER	0.6134	0.6750
	BLEU	0.4910	0.6217
	NIST	9.6494	11.0780
	GTM ($e = 2$)	0.3621	0.4217
	ROUGE$_W$	0.3971	0.4370
	METEOR$_{wnsyn}$	0.7175	0.7528
Syntactic	SP-NIST$_p$	**9.3279**	**9.9268**
	SP-NIST$_c$	**6.7231**	**6.9483**
	DP-HWC$_w$-4	0.2048	0.2612
	DP-HWC$_c$-4	**0.4825**	**0.4823**
	DP-HWC$_r$-4	**0.4279**	**0.4270**
	CP-STM-5	0.5979	0.6358
Semantic	SR-M_r-$\star$	**0.2126**	**0.2165**
	SR-O_r-$\star$	**0.3470**	**0.3533**
	SR-O_r	**0.5546**	**0.5564**
	DR-O_r-$\star$	0.4613	0.5355
	DR-O_{rp}-$\star$	0.6248	0.6504
	DR-STM-5	0.4759	0.5148

Table 5.3: Baseline system vs. SYSTRAN. Heterogeneous evaluation

(small 'DP-HWC$_r$-4') and grammatical categories (small 'DP-HWC$_c$-4') even assign SYSTRAN a higher quality, although the difference is not significant.

This fact reveals that in-domain statistical and out-of-domain rule-based systems operate on different quality dimensions. Therefore, it reinforces the belief that hybrid statistical/rule-based approaches must be investigated. Moreover, this result corroborates the need for heterogeneous evaluation methodologies as the one proposed in Chapter 3.

5.2 Linguistic Data Views

Far from full syntactic complexity, we suggest to go back to the simpler alignment methods first described by Brown et al. (1993), but applied over redefinable alignment units beyond the shallow level of lexical units. Our approach explores the possibility of using additional linguistic annotation up to the level of shallow parsing. For that purpose, we introduce the general concept of *linguistic data view* (LDV), which is defined as any possible linguistic representation of the information contained in a bitext. Data views are enriched with linguistic features, such as the *part-of-speech (PoS)*, *lemma*, and *base phrase chunk IOB label*.

Let us illustrate the applicability through an example. Figure 5.2 shows two sentence pairs, for the case of Spanish-English translation, in which, the English word form *'play'* is translated into

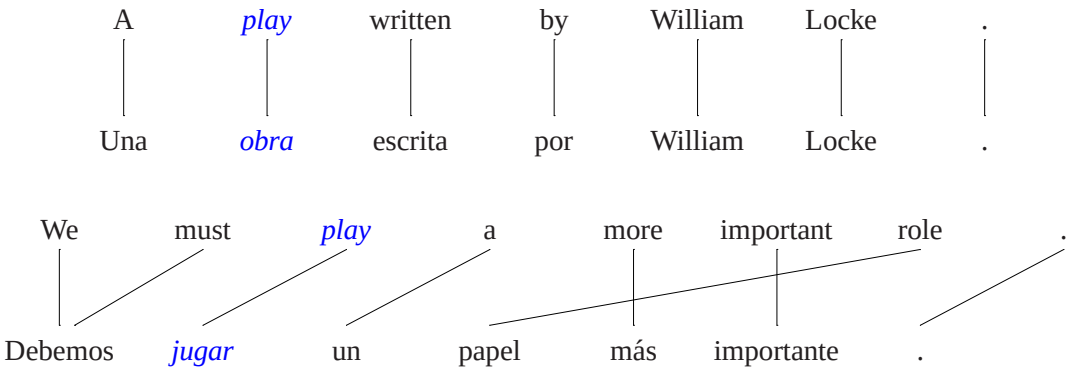


Figure 5.3: Linguistic Data Views. A motivating example

Spanish as *'obra'* and *'jugar'*, respectively. Conventional word alignment methods (Brown et al., 1993; Och & Ney, 2000), based on the EM algorithm, will update frequency counts for *'(play, obra)'* and *'(play, jugar)'*. However, in that manner, they are ignoring the fact that these two realizations of the word form 'play' correspond indeed to different words that happen to be homographs. In the first case, *'play'* is acting as a noun, and as the head of the noun phrase *'A brilliant play'*, whereas in the second case *'play'* acts as a verb, and as the head of the verb phrase *'must play'*. In the same way, phrase alignments will consider *'obra'* and *'jugar'* as valid translations for *'play'*. However, representing the two realizations of the word *'play'*, for instance, as *'play*$_{NN}$*'* and *'play*$_{VB}$*'* would allow us to distinguish between them. This would have a direct implication in the estimation of translation probabilities, since, proceeding in this manner, they will be considered as distinct events. This should lead, therefore, to more accurate word and phrase alignments. In addition, we hypothesize that translation models built over these alignments should yield an improved translation quality.

The use of shallow syntactic information for translation modeling is, as we have seen in Section 4.2.3, not a new idea. For instance, Schafer and Yarowsky (2003) suggested combining lexical, PoS and lemma translation models, following a back-off strategy. Our approach is also very similar, although previous, to the recently suggested factored machine translation models (Koehn et al., 2006; Koehn & Hoang, 2007). However, in our case, apart from enriching the alignment unit, we also allow for redefining its scope, by working with alignments at two different levels of granularity, lexical (i.e., words) and shallow syntactic (i.e., chunks).

5.2.1 Construction

Using linguistic data views requires data to be automatically annotated for the two languages involved. Thus, prior to case lowering, parallel segments are automatically PoS-tagged, lemmatized, and base phrase chunked, using the SVMTool (Giménez & Màrquez, 2004b), Freeling (Carreras et al., 2004) and Phreco (Carreras et al., 2005) linguistic processors, as described in Appendix B.1.

	Spanish			English		
Data View	**Train**	**Dev**	**Test**	**Train**	**Dev**	**Test**
W	138,056	2,471	2,386	94,604	2,103	2,010
L	103,112	1,839	1,753	84,189	1,767	1,661
WP	181,420	2,643	2,562	117,365	2,333	2,227
WC	239,680	2,972	2,911	175,142	2,754	2,668
WPC	275,428	3,085	3,012	201,368	2,882	2,785
Cw	1,125,605	3,168	3,093	1,384,739	2,017	2,932

Table 5.4: Linguistic Data Views. Vocabulary sizes

Notice that it is not necessary that the two parallel counterparts of a bitext share the same data view, as long as they share the same granularity.

In order to simplify the experiments, we have worked with 6 different LDV types: word (W), lemma (L), word and PoS (WP), word and chunk label (WC), word, PoS and chunk label (WPC), and chunk of words (Cw). By chunk label we refer to the IOB label[8] associated to every word inside a chunk, e.g., *'I_{B-NP} $declare_{B-VP}$ $resumed_{I-VP}$ the_{B-NP} $session_{I-NP}$ of_{B-PP} the_{B-NP} $European_{I-NP}$ $Parliament_{I-NP}$ $._{O}$'*. We build chunk tokens by explicitly connecting words in the same chunk, e.g. *'$(I)_{NP}$ $(declare\_resumed)_{VP}$ $(the\_session)_{NP}$ $(of)_{PP}$ $(the\_European\_Parliament)_{NP}$'*. Table 5.4 shows vocabulary sizes (i.e., number of distinct tokens) for each data view over training, development and test sets, both for English and Spanish. It can be seen how vocabulary size increases as more linguistic information is added. Only in the case of replacing words for lemmas the vocabulary size diminishes. An example of data view annotation is available in Table 5.5.

Following the process described in Section 5.1, first, we build word alignments on each of these data views. Then phrase alignments are extracted ($\cup$ heuristic) and scored on the basis of relative frequency. We also build language models for each data view. Moreover, prior to evaluation, automatic translations must be post-processed in order to remove the additional linguistic annotation and split chunks back into words. Finally, translations are recased so they can be compared to reference translations.

5.2.2 Experimental Results

Table 5.6 presents evaluation results. For the sake of readability, we have limited to a representative subset of source-target LDV combinations. The main observation is that no individual data view improves over the 'W-W' baseline. As a main explanation for this result, we argue on the level of data sparsity motivated by the incorporation of linguistic knowledge. Data sparsity may lead to biased parameter estimations, thus, causing a possible decrease in precision, but, in addition, there is also an important decrease in recall, which varies considerably among data views. We have measured this latter argument by observing the size of phrase-based translation models built for each data view, after being filtered for the test set by selecting only the source phrases applicable. For instance, translation models built over word-based data-views suffer a decrement in size, with respect to the 'W' data view, sorted in increasing order, of 3% ('W-WPC'), 6.5% ('WP-WP'), 22%

[8] Inside-Outside-Begin.

W	It would appear that a speech made at the weekend by Mr Fischler indicates a change of his position . Fischler pronunció un discurso este fin de semana en el que parecía haber cambiado de actitud .

L	It would appear that a speech *make* at the weekend by Mr Fischler *indicate* a change of his position . Fischler *pronunciar uno* discurso este fin de semana en el que *parecer haber cambiar* de actitud .

WP	It$_{PRP}$ would$_{MD}$ appear$_{VB}$ that$_{IN}$ a$_{DT}$ speech$_{NN}$ made$_{VBN}$ at$_{IN}$ the$_{DT}$ weekend$_{NN}$ by$_{IN}$ Mr$_{NNP}$ Fischler$_{NNP}$ indicates$_{VBZ}$ a$_{DT}$ change$_{NN}$ of$_{IN}$ his$_{PRP\$}$ position$_{NN}$.$_{.}$ Fischler$_{VMN}$ pronunció$_{VMI}$ un$_{DI}$ discurso$_{NC}$ este$_{DD}$ fin$_{NC}$ de$_{SP}$ semana$_{NC}$ en$_{SP}$ el$_{DA}$ que$_{PR0}$ parecía$_{VMI}$ haber$_{VAN}$ cambiado$_{VMP}$ de$_{SP}$ actitud$_{NC}$.$_{Fp}$

WC	It$_{B-NP}$ would$_{B-VP}$ appear$_{I-VP}$ that$_{B-SBAR}$ a$_{B-NP}$ speech$_{I-NP}$ made$_{B-VP}$ at$_{B-PP}$ the$_{B-NP}$ weekend$_{I-NP}$ by$_{B-PP}$ Mr$_{B-NP}$ Fischler$_{I-NP}$ indicates$_{B-VP}$ a$_{B-NP}$ change$_{I-NP}$ of$_{B-PP}$ his$_{B-NP}$ position$_{I-NP}$.$_{O}$ Fischler$_{B-VP}$ pronunció$_{B-VP}$ un$_{B-NP}$ discurso$_{I-NP}$ este$_{B-NP}$ fin$_{I-NP}$ de$_{B-PP}$ semana$_{B-NP}$ en$_{B-PP}$ el$_{B-SBAR}$ que$_{I-SBAR}$ parecía$_{B-VP}$ haber$_{I-VP}$ cambiado$_{I-VP}$ de$_{B-PP}$ actitud$_{B-NP}$.$_{O}$

WPC	It$_{[PRP:B-NP]}$ would$_{[MD:B-VP]}$ appear$_{[VB:I-VP]}$ that$_{[IN:B-SBAR]}$ a$_{[DT:B-NP]}$ speech$_{[NN:I-NP]}$ made$_{[VBN:B-VP]}$ at$_{[IN:B-PP]}$ the$_{[DT:B-NP]}$ weekend$_{[NN:I-NP]}$ by$_{[IN:B-PP]}$ Mr$_{[NNP:B-NP]}$ Fischler$_{[NNP:I-NP]}$ indicates$_{[VBZ:B-VP]}$ a$_{[DT:B-NP]}$ change$_{[NN:I-NP]}$ of$_{[IN:B-PP]}$ his$_{[PRP\$:B-NP]}$ position$_{[NN:I-NP]}$.$_{[.:O]}$ Fischler$_{[VMN:B-VP]}$ pronunció$_{[VMI:B-VP]}$ un$_{[DI:B-NP]}$ discurso$_{[NC:I-NP]}$ este$_{[DD:B-NP]}$ fin$_{[NC:I-NP]}$ de$_{[SP:B-PP]}$ semana$_{[NC:B-NP]}$ en$_{[SP:B-PP]}$ el$_{[DA:B-SBAR]}$ que$_{[PR0:I-SBAR]}$ parecía$_{[VMI:B-VP]}$ haber$_{[VAN:I-VP]}$ cambiado$_{[VMP:I-VP]}$ de$_{[SP:B-PP]}$ actitud$_{[NC:B-NP]}$.$_{[Fp:O]}$

Cw	(It) (would_appear) (that) (a_speech) (made) (at) (the_weekend) (by) (Mr_Fischler) (indicates) (a_change) (of) (his_position) (.) (Fischler) (pronunció) (un_discurso) (este_fin) (de) (semana) (en) (el_que) (parecía_haber_cambiado) (de) (actitud) (.)

Table 5.5: Linguistic data views. An example

Data View		METEOR	ROUGE	GTM	
Source	Target	(wnsyn)	(w_1.2)	(e=2)	BLEU
W	**W**	**0.7528**	**0.4370**	0.4217	**0.6217**
WP	**WP**	0.7444	0.4304	0.4180	0.6172
WC	**WC**	0.7420	0.4290	0.4134	0.6090
WPC	**WPC**	0.7474	0.4279	0.4167	0.6008
W	**WPC**	0.7477	0.4350	0.4203	0.6185
WPC	**W**	0.7517	0.4351	**0.4235**	0.6157
L	**W**	0.7356	0.4231	0.3960	0.5732
L	**WPC**	0.7328	0.4200	0.3963	0.5708
W	**Cw**	0.6621	0.3838	0.3428	0.4587
Cw	**W**	0.5304	0.2410	0.2800	0.3327
Cw	**Cw**	0.5401	0.2518	0.2902	0.3475

Table 5.6: Linguistic data views. Individual performance (A)

Word Alignment Data View		Phrase Alignment Data View		METEOR	ROUGE	GTM	
Source	Target	Source	Target	(wnsyn)	(w_1.2)	(e=2)	BLEU
W	**W**	**W**	**W**	0.7528	0.4370	0.4217	0.6217
WP	**WP**	**WP**	**WP**	0.7444	0.4304	0.4180	0.6172
WP	**WP**	**W**	**W**	**0.7556**	**0.4390**	**0.4279**	**0.6230**
WP	**WP**	**W**	**WP**	0.7489	0.4364	**0.4246**	**0.6253**
WP	**WP**	**WP**	**W**	0.7505	0.4350	0.4212	0.6187
WC	**WC**	**WC**	**WC**	0.7420	0.4290	0.4134	0.6090
WC	**WC**	**W**	**W**	0.7447	0.4323	0.4189	0.6206
WC	**WC**	**W**	**WC**	0.7491	0.4344	0.4168	0.6116
WC	**WC**	**WC**	**W**	0.7518	0.4334	**0.4247**	0.6151
WPC	**WPC**	**WPC**	**WPC**	0.7474	0.4279	0.4167	0.6008
WPC	**WPC**	**W**	**W**	0.7479	0.4345	0.4200	0.6220
WPC	**WPC**	**W**	**WPC**	0.7487	0.4326	0.4179	0.6105
WPC	**WPC**	**WPC**	**W**	0.7491	0.4323	0.4204	0.6143
Cw	**Cw**	**Cw**	**Cw**	0.5401	0.2518	0.2902	0.3475
Cw	**Cw**	**W**	**W**	0.7173	0.4120	0.3902	0.5524

Table 5.7: Linguistic data views. Individual performance (B)

('WC-WC'), 25% ('WPC-W'), and 26% ('WPC-WPC'). And, translation models for chunk-based data views, which attain the lowest translation quality, exhibit by large also the highest translation model size decrement (75% for 'Cw-W' and 87.5% for 'Cw-Cw'). Logically, the more information is added to a data view, specially to the source side, the larger the recall decrement. The case of 'L-W' and 'L-WPC' data views is different, since filtered translation models are indeed larger (25-30% size increment). Thus, the drop in translation quality for lemma-based data views may only be attributable to a lack of precision. This result is not surprising. After all, it only confirms the common intuition that ignoring morphology in the translation of rich morphological languages such as Spanish is not a good idea.

In order to mitigate the effects of data sparsity we post-process word alignments prior to phrase extraction, removing linguistic information, so only lexical units and alignment information remain. Evaluation results are presented in Table 5.7. We distinguish between word-alignment data views and phrase-alignment data views. It can be observed that translation quality improves substantially, although only in the case of 'WP-WP' alignment data views there is a slight increase over the baseline. With the intent to further improve these results we study the possibility of combining alignments from data views based on different linguistic information. We consider two different combination schemes: *local* and *global* phrase extraction.

Local Phrase Extraction (L-phex)

A separate phrase extraction process is performed for each source-target LDV word alignment. Resulting translation models are then combined as additional log-linear features. Note that there is a limitation in this approach. Although word alignments may be based on different data views, phrase alignments must all be based on the same source-target data view.

Experimental results are showed in Table 5.8. Because our approach to parameter adjustment based on MERT does not scale well when the number of features increases[9], we have only optimized individual weights in the case of combining a maximum of three translation models. The most positive result is that all pair combinations significantly outperform the baseline system consistently according to all metrics. In the case of combining two models, best results are obtained by the 'W+WPC' combination, according to all metrics. LDV triplets exhibit a similar performance, although BLEU confers a significant advantage to the 'W+WPC+Cw' triplet, whereas METEOR prefers the 'W+WP+Cw' one.

When more than three translation models are combined adjusting their weights becomes impractical. Thus, we decided to set their respective contribution uniform so that all models receive the same weight and weights sum up to one. The global contribution of translation models is adjusted only with respect to language, word penalty and distortion models. Interestingly, using uniform weights leads also to an improved translation quality, specially in terms of BLEU and GTM. Best results are attained by the 'W+WP+Cw' and 'W+WPC+Cw' triplets. However, when more than three features are combined, improvements are minimal. Indeed, according to ROUGE and METEOR this option underperforms the baseline. Therefore, uniform weighting is not a practical solution when the number of features in the log-linear combination increases.

[9]For instance, when 3, 4 and 5 translation models are combined, the number of parameter configurations to visit increases up to 2,980, 8,876 and 26,500, respectively.

Data View	METEOR (wnsyn)	ROUGE (w_1.2)	GTM (e=2)	BLEU
	Baseline			
W	0.7528	0.4370	0.4217	0.6217
	L-phex — Adjusted Contribution			
W+Cw	0.7567	0.4383	0.4229	0.6293
W+WP	0.7566	0.4397	0.4280	0.6268
W+WC	0.7565	0.4395	0.4253	0.6312
W+WPC	0.7594	**0.4409**	**0.4310**	0.6350
WP+WC	0.7561	0.4394	0.4252	0.6304
W+WP+WC	0.7540	0.4400	0.4259	0.6349
W+WP+Cw	**0.7609**	0.4402	0.4273	0.6316
W+WPC+Cw	0.7530	0.4372	0.4272	**0.6390**
	L-phex — Uniform Contribution			
W+WP+WC	0.7529	0.4369	0.4261	0.6324
W+WP+Cw	0.7543	0.4361	0.4237	0.6291
W+WC+Cw	0.7561	**0.4380**	0.4274	**0.6372**
W+WPC+Cw	**0.7589**	**0.4380**	**0.4292**	0.6306
W+WP+WC+Cw	*0.7494*	*0.4350*	0.4232	0.6317
W+WP+WC+WPC+Cw	*0.7503*	*0.4340*	0.4244	0.6333
	G-phex — Adjusted Contribution			
W+L+WP+WC+WPC+Cw	0.7566	0.4376	4268	0.6316

Table 5.8: Linguistic data views. Local vs. global phrase extraction

Global Phrase Extraction (G-phex)

A unique phrase extraction is performed over the *union* of word alignments corresponding to different source-target data views, thus, resulting in a unique translation model. The main advantage of this alternative is that the complexity of the parameter tuning process does not vary. Besides, phrase alignments do not have to implement all the same source-target data view.

Experimental results on the combination of the 6 data views (last row in Table 5.8) show that global phrase extraction outperforms the baseline system consistently according to all metrics. Their performance is under the performance of L-phex. However, this approach has the main advantage of making the process of requiring a much lighter parameter optimization process.

5.2.3 Heterogeneous Evaluation

We have applied the methodology for heterogeneous MT evaluation described in Chapter 3 to perform a contrastive error analysis between the baseline system and the several LDV combinations based on local and global phrase extraction. Prior to analyzing individual cases, Table 5.9 reports on system-level evaluation. Several metric representatives from each linguistic level have been selected. Since we do not count on human assessments, metrics are evaluated only in terms of their ability to capture human likeness, using the KING measure. We have also computed to variants of the QUEEN measure, namely QUEEN(X^+) and QUEEN(X^+_{LF}). The first value corresponds to the application of QUEEN to the optimal metric combination based on lexical features only ($X^+ =$ { METEOR$_{wnsyn}$ }), whereas the second value corresponds to QUEEN applied to the optimal metric combination considering linguistic features at different levels ($X^+_{LF} =$ { SP-NIST$_p$, SP-NIST$_c$ }). Interestingly, this set consists only of shallow-syntactic metrics. Optimal metric combinations have been obtained following the procedure described in Section 6.3.4.

The most important observation, is that the difference in quality between the baseline system and combined data views is significantly and consistently reflected by metrics at all linguistic levels. Only in the case of global phrase extraction there are a few exceptions ('ROUGE$_W$', 'DP-O_c-$\star$', 'SR-O_r-$\star_b$' and 'SR-O_{rb}'). In all cases, the highest scores are attained by the local phrase extraction method, although there is no clear consensus on which combination is best.

5.2.4 Error Analysis

We inspect particular cases at the sentence level. For instance, Table 5.10 presents a negative case on the behavior of the global phrase extraction system. Observe how *'todo depende'* is wrongly translated into *'all a matter'* instead of *'everything depends'*. This case also reveals that global phrase extraction suffers also a slight decrease in recall. For instance, no translation is found for *'asumirse'* whereas the baseline system successfully translates it into *'taken'*. The reason is that the union of word alignments produces fewer phrase alignments. In other words, as word links are added to the alignment matrix it becomes more difficult to find phrase pairs consistent with word alignment (see Section 5.1). Only phrase pairs supported by all data views are extracted. Thus, phrase pairs occurring few times in the training data may easily disappear from the translation table. Summing up, global phrase extraction is a method oriented towards precision. However, the increase in precision of phrase alignments is attained at the cost of recall. In contrast, the local

Metric	KING	W	L-phex				G-phex
		baseline	uniform W WC+Cw	tuned W+WPC	tuned W+ WP+Cw	tuned W+ WPC+Cw	
1-WER	0.1462	0.6750	0.6855	0.6866	0.6851	**0.6888**	0.6794
1-PER	0.1250	0.7657	0.7718	0.7709	**0.7730**	0.7708	0.7671
1-TER	0.1442	0.6969	0.7072	0.7071	0.7076	**0.7090**	0.7022
BLEU	0.1131	0.6217	0.6372	0.6350	0.6316	**0.6390**	0.6316
NIST	0.1435	11.0780	**11.2840**	11.2260	11.2070	11.2296	11.2039
GTM ($e = 1$)	0.1071	0.8833	0.8830	**0.8881**	0.8870	0.8811	0.8854
GTM ($e = 2$)	0.1336	0.4217	0.4274	**0.4310**	0.4273	0.4272	0.4268
O_l	0.0985	0.7012	0.7096	**0.7109**	**0.7113**	0.7071	0.7062
ROUGE$_W$	0.1574	0.4370	0.4380	**0.4409**	0.4402	0.4372	0.4376
METEOR$_{exact}$	0.1475	0.7140	0.7172	**0.7204**	**0.7205**	0.7137	0.7195
METEOR$_{wnsyn}$	**0.1667**	0.7528	0.7561	**0.7594**	0.7609	0.7530	0.7566
QUEEN(X^+)	0.1667	0.5647	0.5705	0.5773	**0.5810**	0.5558	0.5692
SP-O_p-$\star$	0.1157	0.6799	0.6878	0.6880	**0.6894**	0.6848	0.6827
SP-O_c-$\star$	0.1157	0.6824	0.6910	**0.6916**	**0.6915**	0.6870	0.6849
SP-NIST$_l$	0.1422	11.1838	**11.3946**	11.3378	11.3282	11.3383	11.2970
SP-NIST$_p$	**0.2097**	9.9268	**10.1114**	10.0305	10.0350	10.0900	10.0274
SP-NIST$_c$	0.1779	6.9483	**7.0498**	7.0018	7.0043	**7.0494**	6.9125
SP-NIST$_{iob}$	0.1825	7.5888	**7.6958**	7.6600	7.6782	**7.6868**	7.5950
QUEEN(X^+_{LF})	**0.2149**	0.3659	0.3736	**0.3763**	0.3689	0.3690	0.3678
DP-O_l-$\star$	0.1409	0.4975	0.5038	**0.5138**	0.5088	0.5024	0.5031
DP-O_c-$\star$	0.1587	0.5993	**0.6080**	0.6037	0.6066	0.6016	0.6003
DP-O_r-$\star$	**0.1700**	0.4637	**0.4735**	0.4667	0.4705	0.4685	0.4665
DP-HWC$_w$-4	0.1078	0.2612	0.2683	**0.2806**	0.2772	0.2696	0.2671
DP-HWC$_c$-4	**0.1766**	0.4823	0.4967	**0.5008**	**0.5006**	0.4916	0.4936
DP-HWC$_r$-4	**0.1687**	0.4270	0.4428	**0.4439**	0.4429	0.4326	0.4385
CP-O_p-$\star$	0.1138	0.6768	0.6853	0.6864	**0.6888**	0.6815	0.6807
CP-O_c-$\star$	0.1111	0.6481	0.6585	**0.6606**	0.6597	0.6553	0.6560
CP-STM-4	0.1462	0.6763	0.6862	**0.6877**	**0.6873**	0.6821	0.6827
NE-M_e-$\star$	0.0443	0.5315	0.5337	0.5348	**0.5356**	0.5286	0.5234
NE-O_e-$\star$	0.0562	0.5513	0.5538	0.5509	**0.5546**	0.5471	0.5398
NE-O_e-$\star\star$	0.1151	0.6842	0.6933	0.6921	**0.6946**	0.6889	0.6880
SR-M_r-$\star_b$	0.0926	0.2165	0.2242	0.2252	**0.2290**	0.2194	0.2192
SR-O_r-$\star_b$	0.1071	0.3533	0.3559	0.3623	**0.3652**	0.3478	0.3533
SR-O_{rb}	0.1204	0.5564	0.5588	0.5662	**0.5710**	0.5518	0.5559
DR-O_r-$\star$	0.1382	0.5355	0.5416	**0.5475**	0.5439	0.5370	0.5415
DR-O_{rp}-$\star$	**0.1508**	0.6504	0.6585	**0.6611**	0.6546	0.6498	0.6545
DR-STM-4	0.1362	0.5670	0.5737	**0.5767**	**0.5772**	0.5687	0.5729

Table 5.9: Baseline system vs combined data views. Heterogeneous evaluation

Source	Por supuesto , todo depende de lo que se haya calculado y de cuánto deba asumirse .
Ref 1	It does , of course , depend on what is being estimated , and on how much is to be taken up .
Ref 2	Of course , everything depends on what has been calculated and on how much must be assumed .
Ref 3	Of course , everything depends on what se was calculated and on how much should be assumed .
Baseline	Of course , everything depends on what has been calculated and of how much should be taken .
G-phex	Of course , **all a matter** of what has been calculated and of how much should **asumirse** .
L-phex	Of course , everything depends on what has been calculated and of how much should be taken .

Table 5.10: Linguistic data views. G-phex method fails

Level	Metric	Baseline	G-phex
Lexical	BLEU	0.7526	0.3381
	GTM ($e = 2$)	0.6283	0.3562
	ROUGE$_W$	0.4442	0.2831
	METEOR$_{wnsyn}$	0.8354	0.5321
	QUEEN	0.5556	0.0000
Syntactic	DP-HWC$_r$-4	1.0000	0.0000
	DP-O_r-$\star$	0.7085	0.2512
	CP-O_c-$\star$	0.6857	0.3537
	CP-STM-9	0.8734	0.3173
Semantic	SR-O_r-$\star$	0.3636	0.0789
	SR-M_r-$\star$	0.4444	0.2222
	SR-O_r	0.5000	0.3226
	DR-O_r-$\star$	0.6419	0.1917
	DR-O_{rp}-$\star$	1.0000	0.2431
	DR-STM-4	0.8476	0.1783

Table 5.11: Linguistic data views. G-phex method fails (heterogeneous evaluation of case from Table 5.10)

Source	Los miles de decisiones y leyes aprobadas por la Comisión rara vez se hacen en la propia Comisión ; casi siempre las hacen los grupos de trabajo en los que hay participantes de los que no sabemos nada .
Ref 1	The thousands of decisions and laws adopted by the Commission are rarely made in the Commission itself but , more often than not , in working groups involving participants of whom we have no knowledge .
Ref 2	The thousands of decisions and laws approved by the Commission are seldom taken by the Commission itself ; they are almost always taken by the working groups that include participants about whom we know nothing .
Ref 3	The thousands of decisions and laws approved by the Commission are rarely made in the actual Commission ; they are nearly always made by the work groups in which there are participants of whom we know nothing about .
Baseline	The thousands of decisions and laws passed by the Commission rarely being made in the Commission itself ; almost always the do the working groups where there participants from those who we know nothing .
G-phex	The thousands of decisions and laws **adopted by** the Commission rarely **are made** in the Commission itself ; almost always the **made by** the working groups **in which** there participants from *which* we know nothing .
L-phex	The thousands of decisions and laws **adopted by** the Commission rarely **are made** in the Commission itself ; almost always the *make* the working groups **in which** there is participants of those **who** we know nothing .

Table 5.12: Linguistic data views. LDV models help

Level	Metric	Baseline	G-phex	L-phex
Lexical	1-PER	0.6486	0.7297	0.6757
	1-TER	0.6216	0.6486	0.6216
	BLEU	0.4126	0.6162	0.5669
	GTM ($e = 1$)	0.7778	0.9315	0.8889
	ROUGE$_W$	0.3016	0.3443	0.3336
	METEOR$_{wnsyn}$	0.6199	0.7114	0.6763
	QUEEN	0.2222	0.5556	0.5556
Syntactic	SP-O_p-$\star$	0.4792	0.5957	0.5417
	SP-O_c-$\star$	0.5102	0.5625	0.5417
	SP-NIST$_p$	9.0035	10.1480	9.7283
	SP-NIST$_c$	6.4328	5.9954	5.8059
	DP-O_l-$\star$	0.7005	0.7851	0.7725
	DP-O_c-$\star$	0.3464	0.5081	0.4933
	DP-O_r-$\star$	0.3866	0.5273	0.5240
	DP-HWC$_w$-4	0.4410	0.7647	0.7734
	CP-O_p-$\star$	0.4800	0.6304	0.5745
	CP-O_c-$\star$	0.4244	0.5848	0.5059
	CP-STM-4	0.5223	0.6197	0.5920
Semantic	SR-M_r-$\star$	0.1250	0.2353	0.2500
	SR-O_r-$\star$	0.4043	0.6327	0.4259
	SR-O_r	0.6098	0.9512	0.5750
	DR-O_r-$\star_b$	0.2305	0.3055	0.2797
	DR-O_{rp}-$\star_b$	0.2800	0.3692	0.3385
	DR-STM-4$_b$	0.2722	0.3612	0.3267

Table 5.13: Linguistic data views. LDV models help (heterogeneous evaluation of case from Table 5.12)

phrase extraction technique exhibits a more robust behavior ('L-phex' corresponds to the output by the 'W+WPC+Cw' system). Heterogeneous evaluation results are shown in Table 5.11.

Table 5.12 shows a positive case in which the use of linguistic data views leads to an improved translation. For instance, *'aprobadas'* is better translated into *'adopted'* instead of *'passed'*, *'se hacen'* into *'are made'* instead of *'being made'*, and *'en los que'* into *'in which'* instead of *'where'*. Heterogeneous evaluation results, reported in Table 5.13, show that improvements take place in several quality dimensions. Most metrics prefer the output by the 'G-phex' system, with a slight advantage over the 'L-phex' system.

5.3 Conclusions of this Chapter

This chapter deals with the construction and development of a Spanish-to-English phrase-based SMT system trained on European Parliament Proceedings. First, we have analyzed its performance as compared to an open-domain rule-based MT system. Interestingly, while lexical metrics give a significant advantage to the SMT system, several metrics at deeper linguistic levels confer both systems a similar score.

In order to improve the baseline SMT system, we introduce the concept of linguistic data view. Six different data views at the shallow-syntactic level have been used to build alternative word and phrase alignment models. The first observation is that individual translation models based on enriched data views underperform the baseline system. This result is mainly attributable data sparsity, which leads to biased parameter estimations, causing a loss of precision and recall. Thus, we have shown that data sparsity is a major cause for the lack of success in the incorporation of linguistic knowledge to translation modeling in SMT.

As a solution, we study the possibility of combining translation models based on different data views. We have presented and discussed the pros and cons of two different combination schemes. Interestingly, combined models yield a significantly improved translation quality. This confirms that they actually carry complementary kinds of information about the translation process. Besides, error analyses show that improvements take place at deeper quality dimensions beyond the lexical level.

We leave for further work the experimentation with new data views using deeper linguistic information, such as full syntactic constituents, grammatical dependencies, and semantic roles. We also speculate that linguistic information could be used to compute alternative translation probabilities and also to prune translation tables according to linguistic criteria and/or constraints.

Finally, across this chapter we have observed that system optimization is a crucial and complex issue. Specifically, we are concerned about its scalability, and about the effects of system overtuning. These two problems require further study.

Chapter 6

Discriminative Phrase Selection for SMT

As we have seen in Section 4.4, a major limitation of the standard phrase-based approach to SMT is that lexical selection is poorly modeled. For instance, the source sentence context in which phrases occur is completely ignored. Thus, all occurrences of the same phrase are assigned, no matter what the context is, the same translation probabilities. Besides, the estimation of translation probabilities is often very simple. Typically, they are estimated on the basis of relative frequency (i.e., maximum likelihood, see Section 4.2) (Koehn et al., 2003).

In order to overcome this limitation, this chapter explores the application of discriminative learning to the problem of phrase selection in SMT. Instead of relying on MLE for the construction of translation models, we suggest using local classifiers which are able to take further advantage of contextual information. We present experimental results on the application of DPT models to the Spanish-to-English translation of European Parliament Proceedings.

The chapter is organized as follows. First, in Section 6.1, our approach to Discriminative Phrase Translation (DPT) is fully described. Then, In Section 6.2, prior to considering the full translation task, we measure the local accuracy of DPT classifiers at the isolated *phrase translation* task. In this task, the goal is not to translate the whole sentence but only individual phrases without having to integrate their translations in the context of the target sentence. We present a comparative study on the performance of four different classification settings based on two different learning paradigms, namely Support Vector Machines and Maximum Entropy models.

In Section 6.3, we tackle the full translation task. We have built a state-of-the-art factored phrase-based SMT system based on linguistic data views at the level of shallow parsing as described in Chapter 5. We compare the performance of DPT and MLE-based translation models built on the same parallel corpus and phrase alignments. DPT predictions are integrated into the SMT system in a *soft* manner, by making them available to the decoder as an additional log-linear feature so they can fully interact with other models (e.g., language, distortion, word penalty and additional translation models) during the search. We separately study the effects of using DPT predictions for all phrases as compared to focusing on a small set of very frequent phrases.

This chapter has also served us to experience in first person, through a practical case study, the role of automatic evaluation metrics in the context of system development. In particular, we have studied the influence of the metric guiding the adjustment of the internal parameters of an SMT system. We have applied the methodology for heterogeneous automatic MT evaluation described in

Chapter 3, which allows for separately analyzing quality aspects at different linguistic levels, e.g., lexical, syntactic, and semantic. As we have seen, this methodology also offers a robust mechanism to combine different similarity metrics into a single measure of quality based on *human likeness*. We have complemented automatic evaluation results through error analysis and by conducting a number of manual evaluations. Main conclusions are summarized in Section 6.5.

6.1 Discriminative Phrase Translation

Instead of relying on MLE estimation to score the phrase pairs (f_i, e_j) in the translation table, DPT models deal with the translation of every source phrase f_i as a multiclass classification problem, in which every possible translation of f_i is a class. As an illustration, in Figure 6.1, we show a real example of Spanish-to-English phrase translation, in which the source phrase *"creo que"*, in this case translated as *"I believe that"*, has several possible candidate translations.

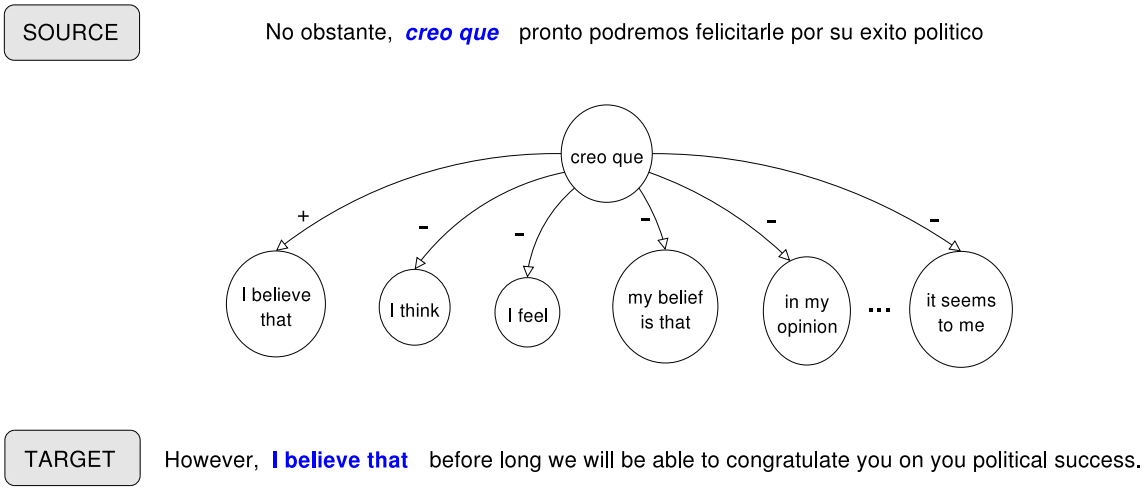


Figure 6.1: Discriminative phrase translation. An example

6.1.1 Problem Setting

Training examples are extracted from the same training data as in the case of conventional MLE-based models, i.e., a phrase-aligned parallel corpus (see Section 6.3.1). We use each occurrence of each source phrase f_i to generate a positive training example for the class corresponding to the actual translation e_j of f_i in the given sentence, according to the automatic phrase alignment. Let us note that phrase translation is indeed a multilabel problem. Since word alignments allow words both in the source and the target sentence to remain unaligned (see Figure 4.1, in Section 4.2), the phrase extraction algorithm employed allows each source phrase to be aligned with more than one target phrase, and vice versa, with the particularity that all possible phrase translations are embedded or

overlap. However, since the final goal of DPT classifiers is not to perform local classification but to provide a larger system with more accurate translation probabilities, in our current approach no special treatment of multilabel cases has been performed.

6.1.2 Learning

There exist a wide variety of learning algorithms which can be applied to the multiclass classification scenario defined. In this work we have focused on two families, namely Support Vector Machines, SVM, (Vapnik, 1995; Cristianini & Shawe-Taylor, 2000), and Maximum Entropy, ME, (Jaynes, 1957). Both methods have been widely and successfully applied to WSD and other NLP problems (Berger et al., 1996; Ratnaparkhi, 1998; Joachims, 1999; Màrquez et al., 2006). We have tried four different learning settings:

1. Linear Binary SVMs (SVMlinear)
2. Degree-2 Polynomial Binary SVMs (SVMpoly2)
3. Linear Multiclass SVMs (SVMmc)
4. Multiclass ME models (MaxEnt)

In all cases, classifiers have been constructed using publicly available software. SVMs have been learned using the SVM^{light} and SVM^{struct} packages by Thorsten Joachims (Joachims, 1999)[1]. ME models have been been learned using the MEGA package by Daumé III (2004)[2], and the MaxEnt package, by Zhang Le[3]. MEGA follows the Limited Memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization method for parameter estimation, whereas MaxEnt additionally allows for using the Generative Iterative Scaling (GIS) optimization method.

Binary vs. Multiclass Classification

While approaches 3 and 4 implement by definition a multiclass classification scheme, approaches 1 and 2 are based on binary classifiers, and, therefore, the multiclass problem must be binarized. We have applied *one-vs-all* binarization, i.e., a binary classifier is learned for every possible translation candidate e_j in order to distinguish between examples of this class and all the rest. Each occurrence of each source phrase f_i is used to generate a positive example for the actual class (or classes) corresponding to the aligned target phrase (or phrases), and a negative example for the classes corresponding to the other possible translations of f_i. At classification time, given a source phrase f_i, SVMs associated to each possible candidate translation e_j of f_i will be applied, and the most confident candidate translation will be selected as the phrase translation.

[1] `http://svmlight.joachims.org`
[2] `http://www.cs.utah.edu/~hal/megam/`
[3] `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`

Support Vector Machines vs. Maximum Entropy

The SVM and ME algorithms are based on different principles. While the SVM algorithm is a linear separator which relies on margin maximization, i.e. on finding the hyperplane which is more distant to the closest positive and negative examples, ME is a probabilistic method aiming at finding the least biased probability distribution that encodes certain given information by maximizing its entropy. An additional interest of comparing the behavior of SVM and ME classifiers is motivated by the nature of the global MT system architecture. While the outcomes of ME classifiers are probabilities which can be easily integrated into the SMT framework, SVM predictions are unbounded real numbers. This issue will be further discussed in Section 6.3.2.

Linear vs Polynomial Kernels

Although SVMs allow for a great variety of kernel functions (e.g., polynomial, gaussian, sigmoid, etc.), in this work, based on results published in recent WSD literature (Lee & Ng, 2002; Màrquez et al., 2006), we have focused on linear and polynomial kernels of degree-2 (see Section 6.2). The main advantage of using linear kernels, over other kernel types, is that this allows for working in the primal formulation of the SVM algorithm and, thus, to take advantage of the extreme sparsity of example feature vectors. This is a key factor, in terms of efficiency, since it permits to considerably speed up both the training and classification processes (Giménez & Màrquez, 2004a). The usage of linear kernels requires, however, the definition of a rich feature set.

6.1.3 Feature Engineering

We have built a feature set which considers different kinds of information, always from the source sentence. Each example has been encoded on the basis of the *local context* of the phrase to be disambiguated and the *global context* represented by the whole source sentence.

As for the local context, we use n-grams ($n \in \{1, 2, 3\}$) of: word forms, parts-of-speech, lemmas, and base phrase chunking IOB labels, in a window of 5 tokens to the left and to the right of the phrase to disambiguate. We also exploit part-of-speech, lemmas and chunk information inside the source phrase, because, in contrast to word forms, these may vary and thus report very useful information. Text has been automatically annotated using the following tools: SVMTool for PoS tagging (Giménez & Màrquez, 2004b), Freeling for lemmatization (Carreras et al., 2004), and Phreco for base phrase chunking (Carreras et al., 2005), as described in Section B.1. These tools have been trained on the WSJ Penn Treebank (Marcus et al., 1993), for the case of English, and on the 3LB Treebank (Navarro et al., 2003) for Spanish, and, therefore, rely on their tag sets. However, for the case of parts-of-speech, because tag sets take into account fine morphological distinctions, we have additionally defined several coarser classes grouping morphological variations of nouns, verbs, adjectives, adverbs, pronouns, prepositions, conjunctions, determiners and punctuation marks.

As for the global context, we collect topical information by considering content words (i.e., nouns, verbs, adjectives and adverbs) in the source sentence as a bag of lemmas. We distinguish between lemmas at the left and right of the source phrase being disambiguated.

As an illustration, Table 6.1 shows the feature representation for the example depicted in Figure 6.1. At the top, the source sentence appears annotated at the level of shallow syntax (following a

Source Sentence creo$_{[\text{creer:VMI:B-VP}]}$ que$_{[\text{que:CS:B-CONJP}]}$ pronto$_{[\text{pronto,AQ,O}]}$ podremos$_{[\text{podremos,VMS,B-VP}]}$ felicitarle$_{[\text{felicitarle,VMN,I-VP}]}$ por$_{[\text{por,SP,B-PP}]}$ su$_{[\text{su,DP,B-NP}]}$ éxito$_{[\text{exito,NC,I-NP}]}$ político$_{[\text{politico,AQ,I-NP}]}$ ·$_{[.,\text{Fp,O}]}$

Source phrase features

Lemma n-grams	(creer)$_1$, (que)$_2$, (creer,que)$_1$
PoS n-grams	(VMI)$_1$, (CS)$_2$, (VMI,CS)$_1$
Coarse PoS n-grams	(V)$_1$, (C)$_2$, (V,C)$_1$
Chunk n-grams	(B-VP)$_1$, (B-CONJP)$_2$, (B-VP,B-CONJP)$_1$

Source sentence features

Word n-grams	(pronto)$_1$, (podremos)$_2$, (felicitarle)$_3$, (por)$_4$, (su)$_5$, (_,pronto)$_{-1}$, (pronto,podremos)$_1$, (podremos,felicitarle)$_2$, (felicitarle,por)$_3$, (por,su)$_4$, (_,_,pronto)$_{-2}$, (_,pronto,podremos)$_{-1}$, (pronto,podremos,felicitarle)$_1$, (podremos,felicitarle,por)$_2$, (felicitarle,por,su)$_3$
Lemma n-grams	(pronto)$_1$, (poder)$_2$, (felicitar)$_3$, (por)$_4$, (su)$_5$, (_,pronto)$_{-1}$, (pronto,poder)$_1$, (poder,felicitar)$_2$, (felicitar,por)$_3$, (por,su)$_4$, (_,_,pronto)$_{-2}$, (_,pronto,poder)$_{-1}$,(pronto,poder,felicitar)$_1$, (poder,felicitar,por)$_2$, (felicitar,por,su)$_3$
PoS n-grams	(AQ)$_1$, (VMS)$_2$, (VMN)$_3$, (SP)$_4$, (DP)$_5$, (_,AQ)$_{-1}$, (AQ,VMS)$_1$, (VMS,VMN)$_2$, (VMN,SP)$_3$, (SP,DP)$_4$, (_,_,AQ)$_{-2}$, (_,AQ,VMS)$_{-1}$, (AQ,VMS,VMN)$_1$, (VMS,VMN,SP)$_2$, (VMN,SP,DP)$_3$
Coarse PoS n-grams	(A)$_1$, (V)$_2$, (V)$_3$, (S)$_4$, (D)$_5$ (_,A)$_{-1}$, (A,V)$_1$, (V,V)$_2$, (V,S)$_3$, (S,D)$_4$ (_,A,V)-1, (_,_,A)$_{-2}$, (A,V,V)$_1$, (V,V,S)$_2$, (V,S,D)$_3$
Chunk n-grams	(O)$_1$, (B-VP)$_2$, (I-VP)$_3$, (B-PP)$_4$, (B-NP)$_5$, (_,O)$_{-1}$, (O,B-VP)$_1$, (B-VP,I-VP)$_2$, (I-VP,B-PP)$_3$, (B-PP,B-NP)$_4$, (_,_,O)$_{-2}$, (_,O,B-VP)$_{-1}$, (O,B-VP,I-VP)$_1$, (B-VP,I-VP,B-PP)$_2$, (I-VP,B-PP,B-NP)$_3$
Bag-of-lemmas	left = $\emptyset$ right = { pronto, poder, felicitar, éxito, político }

Table 6.1: Discriminative phrase translation. An example of feature representation

'word$_{[\text{lemma:PoS:IOB}]}$' format). Below, the corresponding source phrase and source sentence features are shown. We have not extracted any feature from the target phrase, nor the target sentence, neither the correspondence between source and target phrases (i.e., word alignments). The reason is that our purpose was to use DPT models to aid an existing SMT system to make better lexical choices. However, using these type of features would have forced us to build a new and more complex decoder.

6.2 Local Performance

Analogously to the *word translation* task definition by Vickrey et al. (2005), rather than predicting the sense of a word according to a given sense inventory, in *phrase translation* the goal is to predict the correct translation of a *phrase*, for a given target language, in the context of a sentence. This task is simpler than the full translation task in that phrase translations of different source phrases do not have to interact in the context of the target sentence. However, it provides an insight to the gain prospectives.

6.2.1 Data Sets and Settings

We have used the same data sets corresponding to the Spanish-English translation of European Parliament Proceedings used in Chapter 5 (see Section 5.1.1). After performing phrase extraction over the training data (see details in Section 6.3.1), also discarding source phrases occurring only once (around 90%), translation candidates for 1,729,191 source phrases were obtained. In principle, we could have built classifiers for all these source phrases. However, in many cases learning could be either unfruitful or not necessary at all. For instance, 27% of these phrases are not ambiguous (i.e., have only one associated possible translation), and most phrases count on few training examples. Based on these facts, we decided to build classifiers only for those source phrases with more than one possible translation and 100 or more occurrences. Besides, due to the fact that phrase alignments have been obtained automatically and, therefore, include many errors, source phrases may have a large number of associated phrase translations. Most are wrong and occur very few times. We have discarded many of them by considering only as possible phrase translations those which are selected more than 0.5% of the times as the actual translation[4]. The resulting training set consists of 30,649 Spanish source phrases. Table 6.2 presents a brief numerical description of the phrase set. For instance, it can be observed that most phrases are trained on less than 5,000 examples. Most of them are length-2 phrases and most have an entropy lower than 3.

As to feature selection, we discarded features occurring only once in the training data, and constrained the maximum number of dimensions of the feature space to 100,000, by discarding the less frequent features.

[4]This value was empirically selected so as to maximize the local accuracy of classifiers on a small set of phrases of varying number of examples.

#Occurrences	#Phrases	Phrase length	#Phrases	Phrase entropy	#Phrases
(100, 500]	23,578	**1**	7,004	**[0, 1)**	6,154
(500, 1,000]	3,340	**2**	12,976	**[1, 2)**	11,648
(1,000, 5,000]	2,997	**3**	7,314	**[2, 3)**	8,615
(5,000, 10,000]	417	**4**	2,556	**[3, 4)**	3,557
(10,000, 100,000]	295	**5**	799	**[4, 5)**	657
$>$ **100,000**	22			**[5, 6)**	18

Table 6.2: Discriminative phrase translation. Numerical description of the set of 'all' phrases

	Evaluation scheme	
#Examples	**Development and test**	**Test only**
2-9	leave-one-out	
10..99	10-fold cross validation	
100..499	5-fold cross validation	
500..999	3-fold cross validation	
1,000..4,999	train(80%)–dev(10%)–test(10%)	train(90%)–test(10%)
5,000..9,999	train(70%)–dev(15%)–test(15%)	train(80%)–test(20%)
$>$ **10,000**	train(60%)–dev(20%)–test(20%)	train(75%)–test(25%)

Table 6.3: Discriminative phrase translation. Evaluation scheme for the local phrase translation task

6.2.2 Evaluation

Local DPT classifiers are evaluated in terms of accuracy against automatic phrase alignments, which are used as gold standard. Let us note that, in the case of multilabel examples, we count the prediction by the classifier as a hit if it matches any of the classes in the solution. Moreover, in order to maintain the evaluation feasible, a heterogeneous evaluation scheme has been applied (see Table 6.3). Basically, when there are few examples available we apply cross-validation, and the more examples available the fewer folds are used. Besides, because cross-validation is costly, when there are more than 1,000 examples available we simply split them into training, development and test sets, keeping most of the examples for training and a similar proportion of examples for development and test. Also, as the number of examples increases, the smaller proportion is used for training and the bigger proportion is held out for development and test. In all cases, we have preserved, when possible, the proportion of samples of each phrase translation so folders do not get biased.

6.2.3 Adjustment of Parameters

Supervised learning algorithms are potentially prone to overfit training data. There are, however, several alternatives in order to fight this problem. In the case of the SVM algorithm, the contribution of training errors to the objective function of margin maximization is balanced through the C regularization parameter of the soft margin approach (Cortes & Vapnik, 1995). In the case of the ME algorithm, the most popular method is based on the use of a gaussian prior on the parameters of

#Occurrences	#Phrases	Phrase length	#Phrases	Phrase entropy	#Phrases
(100, 500]	790	**1**	213	**[1, 2)**	467
(500, 1,000]	100	**2**	447	**[2, 3)**	362
(1,000, 5,000]	92	**3**	240	**[3, 4)**	139
(5,000, 10,000]	11	**4**	78	**[4, 5)**	31
(10,000, 50,000]	7	**5**	22	**[5, 6)**	1

Table 6.4: Discriminative phrase translation. Numerical description of the representative set of 1,000 phrases selected

the model, whose variance, σ^2, may be balanced (Chen & Rosenfeld, 1999). Learning parameters are typically adjusted so as to maximize the accuracy of local classifiers over held-out data. In our case, a greedy iterative strategy, similar to the optimization strategy described in Section 5.1.2, has been followed. In the first iteration several values are tried. In each following iteration, n values around the top scoring value of the previous iteration are explored at a resolution of $\frac{1}{n}$ the resolution of the previous iteration, and so on, until a maximum number of iterations I is reached[5].

6.2.4 Comparative Performance

We present a comparative study of the four learning schemes described in Section 6.1.2. For the case of ME models, we show the results distinctly applying the LM-BFGS and GIS optimization methods. In order to avoid overfitting, the C and σ^2 parameters have been adjusted. However, because parameter optimization is costly, taking into account the large number of classifiers involved, we have focused on a randomly selected set of 1,000 representative source phrases with a number of examples in the [100, 50,000] interval. Phrases with a translation entropy lower than 1 have not been considered. A brief numerical description of this set is available in Table 6.4.

Table 6.5 shows comparative results, in terms of accuracy. The local accuracy for each source phrase is evaluated according to the number of examples available, as described in Table 6.3. DPT classifiers are also compared to the *most frequent translation* baseline (MFT), which is equivalent to selecting the translation candidate with highest probability according to MLE. The 'macro' column shows macro-averaged results over all phrases, i.e., the accuracy for each phrase counts equally towards the average. The 'micro' column shows micro-averaged accuracy, where each test example counts equally[6]. The 'optimal' columns correspond to the accuracy computed on optimal parameter values, whereas the 'default' columns correspond to the accuracy computed on default C and σ^2 parameter values. In the case of SVMs, we have used the SVM^{light} default value for the C parameter[7]. In the case of ME, we have set σ^2 to 1 for all classifiers. The reason is that this was the most

[5] In our case, $n = 2$ and $I = 3$. In the case of the C parameter of SVMs first iteration values are set to 10^i (for $i \in [-4, +4]$), while for the σ^2 of ME prior gaussians, values are $\{0, 1, 2, 3, 4, 5\}$.

[6] The contribution of each phrase to micro-averaged accuracy has been conveniently weighted so as to avoid the extra weight conferred to phrases evaluated via cross-validation.

[7] The C parameter for each binary classifier is set to $\frac{\sum (\vec{x_i}\vec{x_i})^{-1}}{N}$, where $\vec{x_i}$ is a sample vector and N corresponds to the number of samples. In the case of multiclass SVMs, the default value is 0.01.

Model	Optimal		Default	
	Macro (%)	**Micro (%)**	**Macro (%)**	**Micro (%)**
MFT	64.66	67.60	64.66	67.60
SVMlinear	70.81	74.24	69.50	73.56
SVMpoly2	71.31	**74.75**	**69.94**	**73.91**
SVMmc	70.11	73.48	*57.68*	*63.49*
MaxEnt$_{bfgs}$	**71.38**	74.44	69.87	73.41
MaxEnt$_{gis}$	71.08	74.34	*67.38*	*70.69*

Table 6.5: Discriminative phrase translation. Local accuracy over a selected set of 1,000 phrases based on different learning types vs. the MFT baseline

common return value, with a frequency over 50% of the cases, of the parameter tuning process on the selected set of 1,000 phrases.

When the C and σ^2 are properly optimized, all learning schemes, except linear multiclass SVMs, exhibit a similar performance, with a slight advantage in favor of polynomial SVMs. The increase with respect to the MFT baseline is comparable to that described by Vickrey et al. (2005). These results are, taking into account the differences between both tasks, also coherent with results attained in WSD (Agirre et al., 2007). However, when default values are used, all models suffer a significant decrease. For instance, it can be observed that using GIS for parameter estimation causes a severe drop in the performance of ME models. More dramatic is the case of multiclass SVMs, which fall even below the MFT baseline. These two approaches, thus, require an exhaustive process of adjustment of parameters.

6.2.5 Overall Performance

The aim of this subsection is to analyze which factors have a bigger impact on the performance of DPT classifiers applied to the set of *all* phrases. In this scenario, no matter how greedy the process is, the adjustment of the C and σ^2 becomes impractical. For that reason we have used fixed default values. In the case of SVMs, for the sake of efficiency, we have limited to the use of linear kernels. In the case of ME, we encountered problems when running the MEGA software over all phrases[8]. These seemed to be related to parameter estimation —MEGA follows the LM-BFGS optimization method. In order to solve these problems, we shifted to the GIS optimization method, using the MaxEnt software. An excellent comparison on the performance of these two algorithms was published by Malouf (2002).

Phrase translation results are shown in Table 6.6. Again, phrases are evaluated according to the number of examples available, as described in Table 6.3. We distinguish between the case of using *all* the 30,649 phrases counting on 100 or more examples (columns 1 and 2), and the case of considering only a small subset of 317 very *frequent* phrases occurring more than 10,000 times (columns 3 and 4).

The first observation is that both DPT learning schemes outperform the MFT baseline when default learning parameters are used. However, as expected, ME models based on the GIS method

[8]MEGA exited abruptly and unexpectedly before termination.

Model	All		Frequent	
	Macro (%)	Micro (%)	Macro (%)	Micro (%)
MFT	70.51	80.49	79.77	86.12
SVMlinear	74.52	**85.48**	86.32	**91.33**
MaxEnt$_{gis}$	72.73	82.53	82.31	87.94

Table 6.6: Discriminative phrase translation. Overall local accuracy

Phrase	#Occurrences	Entropy	Δ_{acc}
que	605,329	1.6803	0,1642
en	557,745	1.4494	0,0744
no	239,287	2.3292	0,0670
una	211,665	1.3297	0,0586
un	221,998	1.2292	0,0481
del	244,132	1.9712	0,0375
a	206,437	0.9429	0,0230
los	269,679	0.7635	0,0057
de	633,120	0.4936	0,0032
.	1,078,835	0.1019	0,0012
y	714,353	0.6503	-0,0000
el	473,770	0.4623	-0,0000
,	1,232,833	0.4705	-0,0000
la	801,318	0.6494	-0,0001
es	271,696	1.5350	-0,0023

Table 6.7: Discriminative phrase translation. Local performance of most frequent phrases

for parameter estimation are much less effective than linear SVMs. A second observation is that the difference, in terms of micro-averaged accuracy gain with respect to the MFT baseline, between using all phrases and focusing on a set of very frequent ones is very small. The reason is that the set of frequent phrases dominates indeed the evaluation with 51.65% of the total number of test cases. In contrast, macro-averaged results confer a significantly wider advantage to DPT models applied to the set of frequent phrases, specially in the case of linear SVMs. This result is significant taking account the high results of the MFT baseline on this set. A third, marginal, observation is that frequent phrases are easier to disambiguate, presumably because of their lower entropy (see MFT performance).

In Figure 6.2 we analyze several factors which have a direct influence on the behavior of DPT classifiers. All plots correspond to the case of linear SVMs. For instance, the top-left plot shows the relationship between the local accuracy gain and the number of training examples, for all source phrases. As expected, DPT classifiers trained on fewer examples exhibit the most unstable behavior, yielding a maximum accuracy gain of 0.65 and a maximum decrease of 0.30. However, in general, with a sufficient number of examples (over 10,000), DPT classifiers outperform the MFT baseline. It can also be observed that for most of the phrases trained on more than around 200,000 examples

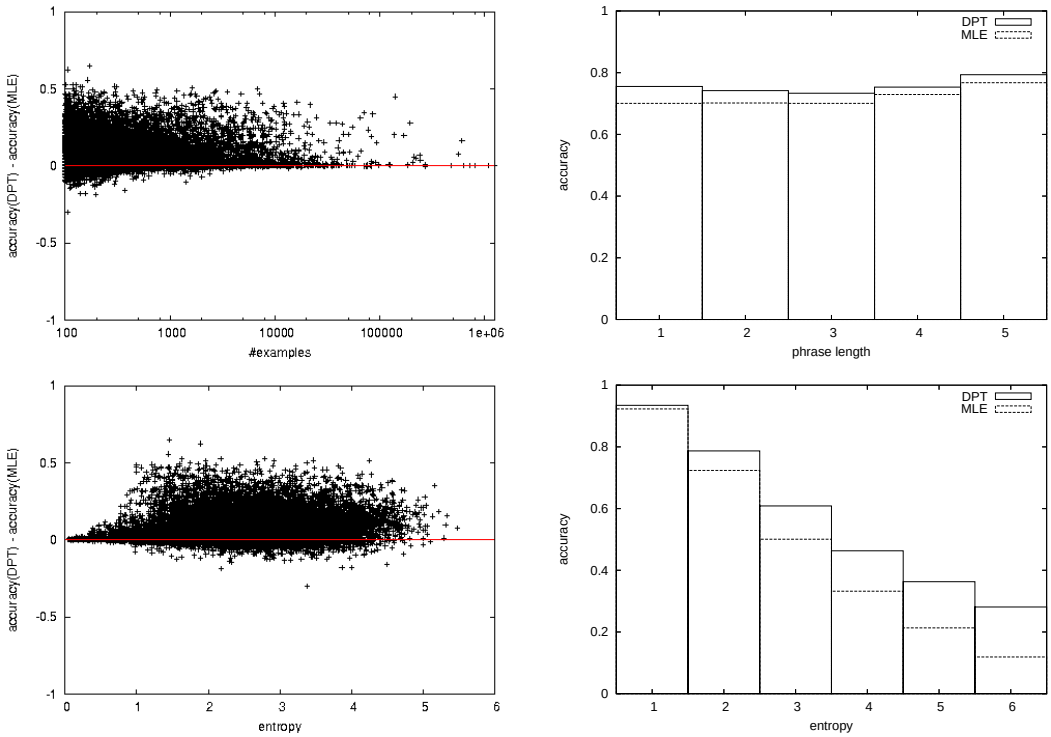


Figure 6.2: Discriminative phrase translation. Analysis of phrase translation results

the accuracy gain is very low. The reason, however, is in the fact that these are phrases with very low translation entropy, mostly stop words, such as punctuation marks (*“.”*, *“,”*), determiners (*“el”*, *“la”*, *“los”*, *“las”*, *“un”*, *“una”*), or conjunctions and prepositions (*“y”*, *“de”*, *“en”*, *“a”*). There is a very interesting positive case, that of phrase *“que”*, which acts mostly as a conjunction or relative pronoun, and that most often gets translated into *“that”* or *“which”*. This phrase, which appears more than 600,000 times in the data with a translation entropy of 1.68, attains an accuracy gain of 0.16. Let us show, in Table 6.7, some illustrative cases of the translation of very frequent phrases, sorted in decreasing order according to the accuracy gain.

The top-right plot in Figure 6.2 shows the relationship between micro-averaged accuracy and source phrase length. There is improvement across all phrase lengths, but, in general, the shorter the phrase the larger the improvement. This plot also indicates that phrases up to length-3 are on average harder to disambiguate than longer phrases. Thus, there seems to be a trade-off between phrase length, level of ambiguity (i.e., translation entropy), and number of examples. Shorter phrases are harder because they exhibit higher ambiguity. DPT is a better model for these phrases because it is able to properly take advantage of the large number of training examples. Longer phrases phrases are easier to model because they present a lower ambiguity. Middle length phrases are hardest because they present a high ambiguity and not many examples.

We further investigate this issue in the two bottom plots. The bottom-left plot shows the relationship between the local accuracy gain and translation entropy, for all source phrases. It can be

observed that for phrases with entropy lower than 1 the gain is very small, while for higher entropy levels the behavior varies. In order to clarify this scenario, we analyze the relationship between micro-averaged accuracy and phrase translation entropy at different intervals (bottom-right plot). As expected, the lower the entropy the higher the accuracy. Interestingly, it can also be observed that as the entropy increases the accuracy gain in favor of DPT models increases as well.

6.3 Exploiting Local Models for the Global Task

In this section, we analyze the impact of DPT models when the goal is to translate the whole sentence. First, we describe our phrase-based SMT baseline system and how DPT models are integrated into the system. Then, some aspects of evaluation are discussed, with special focus on the adjustment of the parameters governing the search process. Finally, MT results are evaluated and analyzed, and several concrete cases are commented.

6.3.1 Baseline System

Our system follows the phrase-based SMT architecture described in Chapter 5, enhanced with *linguistic data views* up to the level of shallow syntax. Phrase alignments are extracted from a word-aligned parallel corpus linguistically enriched with part-of-speech information, lemmas, and base phrase chunk labels. We have followed the *global phrase extraction* strategy described in Section 5.2, i.e., a single translation table is built on the union of alignments corresponding to different linguistic data views. We have not used the *local phrase extraction* strategy because it introduces more complexity into the process of adjustment of parameters.

The integration of DPT predictions into the log-linear scheme is straightforward:

$$\begin{aligned}\log P(e|f) \approx\ & \lambda_{lm} \log P(e) + \lambda_g \log P_{\text{MLE}}(f|e) + \lambda_d \log P_{\text{MLE}}(e|f) \\ & + \lambda_{\text{DPT}} \log P_{\text{DPT}}(e|f) + \lambda_d \log P_d(e,f) + \lambda_w \log w(e)\end{aligned}$$

DPT predictions are integrated as an additional feature. $P(e)$ stands for the language model probability. $P_{\text{MLE}}(f|e)$ corresponds to the MLE-based generative translation model, whereas $P_{\text{MLE}}(e|f)$ corresponds to the analogous discriminative model. $P_{\text{DPT}}(e|f)$ corresponds to the DPT model which uses DPT predictions in a wider feature context. Finally, $P_d(e,f)$ and $w(e)$, correspond to the distortion and word penalty models[9]. The λ parameters controlling the relative importance of each model during the search must be adjusted. We further discuss this issue in subsection 6.3.4.

[9]We have used default *Pharaoh*'s word penalty and distortion models.

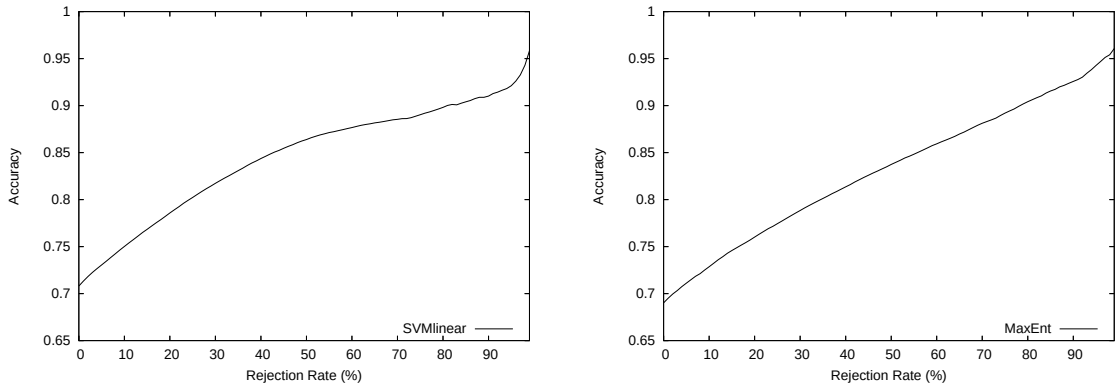


Figure 6.3: Discriminative phrase translation. Rejection curves. Linear SVMs + softmax (left) vs. ME (right)

6.3.2 Soft Integration of Dedicated Predictions

We consider every instance of f_i as a separate classification problem. In each case, we collect the classifier outcome for all possible phrase translations e_j of f_i. In the case of ME classifiers, outcomes are directly probabilities. However, in the case of SVMs, outcomes are unbounded real numbers. We transform them into probabilities by applying the *softmax function* described by Bishop (1995):

$$P(e_j|f_i) = \frac{e^{\gamma\ score_{ij}}}{\sum_{k=1}^{K} e^{\gamma\ score_{ik}}}$$

where K denotes the number of possible target phrase translations for a given source phrase f_i, and score$_{ij}$ denotes the outcome for target phrase e_j according to the SVM classifier trained for f_i. Other transformation techniques can be found in recent literature. For instance, Platt (2000) suggested using a sigmoid function.

In order to verify the suitability of the softmax function, we computed rejection curves for the estimated output probabilities with respect to classification accuracy. For that purpose, we have used the representative set of 1,000 phrases from subsection 6.2.4. This set offers almost 300,000 predictions. In order to calculate rejection curves, the probability estimates for these predictions are sorted in decreasing order. At a certain level of rejection (n%), the curve plots the classifier accuracy when the lowest scoring n% subset is rejected. We have collected values for 100 rejection levels at a resolution of 1%. We tested different values for the γ parameter of the softmax function. The selected final value is $\gamma = 1$. In Figure 6.3 (left) we plot the rejection curve for linear SVMs. For the sake of comparison, the rejection curve for ME classifiers is also provided (right plot). It can be observed that both rejection curves are increasing and smooth, indicating a good correlation between probability estimates and classification accuracy.

At translation time, we do not constrain the decoder to use the translation e_j with highest probability. Instead, we make all predictions available and let the decoder choose. We have precomputed

f_i	e_j	$P_{MLE}(e\|f)$	$P_{DPT}(e\|f)$
...			
creo$_{14}$ que$_{441}$	i believe that	0.3624	0.2405
creo$_{14}$ que$_{441}$	i think that	0.1975	0.0506
creo$_{14}$ que$_{441}$	i think	0.1540	0.0475
creo$_{14}$ que$_{441}$	i feel that	0.0336	0.0511
creo$_{14}$ que$_{441}$	i think it	0.0287	0.0584
creo$_{14}$ que$_{441}$	i believe that it	0.0191	0.0487
creo$_{14}$ que$_{441}$	i think that it	0.0114	0.0498
creo$_{14}$ que$_{441}$	believe that	0.0108	0.0438
creo$_{14}$ que$_{441}$	i believe that this	0.0077	0.0482
creo$_{14}$ que$_{441}$	i believe it	0.0060	0.0439
...			

Table 6.8: Discriminative phrase translation. An example of translation table

all DPT predictions for all possible translations of all source phrases appearing in the test set. The input text is conveniently transformed into a sequence of identifiers[10], which allows us to uniquely refer to every distinct instance of every distinct word and phrase in the test set. Translation tables are accordingly modified so that each distinct occurrence of every single source phrase has a distinct list of phrase translation candidates with their corresponding DPT predictions. Let us note that, as described in Section 6.2.1, for each source phrase, not all associated target translations which have a MLE-based prediction have also a DPT prediction, but only those with a sufficient number of training examples. In order to provide equal opportunities to both models, we have incorporated translation probabilities for these phrases into the DPT model by applying linear discounting.

As an illustration, Table 6.8 shows a fragment of the translation table corresponding to the phrase *"creo que"* in the running example. Notice how this concrete instance has been properly identified by indexing the words inside the phrase ("creo que" $\rightarrow$ "creo$_{14}$ que$_{441}$"). We show MLE-based and DPT predictions (columns 3 and 4, respectively) for several phrase candidate translations sorted in decreasing MLE probability order. The first observation is that both methods agree on the top-scoring candidate translation, "I believe that". However, the distribution of the probability mass is significantly different. While, in the case of the MLE-based model, there are three candidate translations clearly outscoring the rest, concentrating more than 70% of the probability mass, in the case of the DPT model predictions give a clear advantage to the top-scoring candidate although with less probability, and the rest of candidate translations obtain a very similar score.

By integrating DPT predictions in this manner, we have avoided having to implement a new decoder. However, because translation tables may become very large, this technique involves an extra cost in terms of memory and disk consumption. Besides, it imposes a limitation on the kind of features the DPT system may use. In particular, features from the target sentence under construction and from the correspondence between source and target (i.e., alignments) can not be used.

[10]In our case a sequence of w_i tokens, where w is a word and i corresponds to the number of occurrences of word w seen in the test set before the current occurrence number. For instance, the source sentence in the example depicted in Figure 6.1 is transformed into *"creo$_{14}$ que$_{441}$ pronto$_0$ podremos$_0$ felicitarle$_0$ por$_{109}$ su$_0$ éxito$_3$ político$_4$.$_{366}$"*.

6.3.3 Evaluation

Evaluating the effects of using DPT predictions in the full translation task presents two serious difficulties. In first place, the actual room for improvement caused by a better translation modeling is smaller than estimated in Section 6.2. This is mainly due to the SMT architecture itself which relies on a search over a probability space in which several models cooperate. For instance, in many cases errors caused by a poor translation modeling may be corrected by the language model. In a recent study over the same data set (Spanish-to-English translation of the Openlab 2006 corpus), Vilar et al. (2006) found that only around 28% of the errors committed by their SMT system were related to word selection. In half of these cases errors are caused by a wrong word sense disambiguation, and in the other half the word sense is correct but the lexical choice is wrong. In second place, most conventional automatic evaluation metrics have not been designed for this purpose and may, therefore, not be able to reflect possible improvements attained due to a better word selection. For instance, n-gram based metrics such as BLEU (Papineni et al., 2001) tend to favor longer string matchings, and are, thus, biased towards word ordering. In order to cope with evaluation difficulties, we have applied several complementary actions, which are described below.

Heterogeneous Automatic MT Evaluation

We follow the evaluation methodology described in Chapter 3 for heterogeneous automatic MT evaluation. For our experiments, we have selected a representative set of around 50 metrics at different linguistic levels: lexical (i.e., on word forms), shallow-syntactic (e.g., on word lemmas, part-of-speech tags, and base phrase chunks), syntactic (e.g., on dependency and constituency trees), shallow-semantic (on named entities and semantic roles), and semantic (e.g., on discourse representations).

MT Evaluation based on Human Likeness

Heterogeneous MT evaluations might be very informative. However, a new question arises. Since metrics are based on different similarity criteria, and, therefore, biased towards different aspects of quality, scores conferred by different metrics may be controversial. Thus, as system developers we require an additional tool, a meta-evaluation criterion, which allows us to select the most appropriate metric or set of metrics for the task at hand.

As seen in Section 2.2.2, the two most prominent meta-evaluation criteria are human acceptability and human likeness. In this chapter, partly because we do not count on human assessments, we have relied on human likeness. We follow the approach *QARLA*'s approach (Amigó et al., 2005), applied in two complementary steps. First, we determine the set of metrics with highest discriminative power by maximizing over the KING measure. Second, we use QUEEN to measure overall MT quality according to the optimal metric set[11]. QUEEN exhibits several properties which make it really practical for the purpose of our task. First, since QUEEN focus on unanimously supported quality distinctions, it is is a measure of high precision. Second, QUEEN provides a robust means

[11]The KING and QUEEN measures are available inside IQMT.

of combining several metrics into a single measure of quality; it is robust against metric redundancy, i.e., metrics devoted to very similar quality aspects, and with respect to metric scale properties.

A Measure of Phrase Translation Accuracy

For the purpose of evaluating the changes related only to a specific set of phrases (e.g., 'all' vs. 'frequent' sets), we introduce a new measure, $\mathbf{A_{pt}}$, which computes *phrase translation accuracy* for a given list of source phrases. For every test case, A_{pt} counts the proportion of phrases from the list appearing in the source sentence which have a valid[12] translation both in the target sentence and in at least one reference translation. Cases in which no valid translation is available in any reference translation are not taken into account. Moreover, in order to avoid using the same target phrase more than once for the same translation case, when a phrase translation is used, source and target phrases are discarded. In fact, because in general source-to-target alignments are either unknown or automatically acquired, A_{pt} calculates an approximate solution. Current A_{pt} implementation inspects phrases from left to right in decreasing length order.

Manual Evaluation

Along this research, we have contrasted automatic evaluation results by conducting a number of manual evaluations. This type of evaluation offers the advantage of being directly interpretable. However, it is expensive to produce, not reusable, possibly partial, and subjective. In order to reduce the degree of subjectivity we have simplified the manual evaluation process to the case of pairwise system comparisons. Human assessors are presented a collection of translation test cases with associated source and reference translations, and automatic outputs by two different systems, 'A' and 'B'. For each case, assessors must judge whether the output by system 'A' is better, equal to or worse than the output by system 'B', with respect to adequacy (i.e., preservation of the meaning), fluency (i.e., sentence well-formedness), and overall quality. In order to prevent judges from biasing towards either system during the evaluation, the respective position in the display of the sentences corresponding to each system is randomized. In all cases, statistical significance is determined using the sign-test (Siegel, 1956). Agreement between judges has been estimated based on the Kappa measure (Cohen, 1960).

6.3.4 Adjustment of Parameters

As we have seen in Section 6.2, DPT models provide translation candidates only for specific subsets of phrases. Therefore, in order to translate the whole test set, alternative translation probabilities for all the source phrases in the vocabulary which do not have a DPT prediction must be provided. We have used MLE-based predictions to complete DPT tables. However, interaction between DPT and MLE models is problematic. Problems arise when, for a given source phrase, f_i, DPT predictions must compete with MLE predictions for larger source phrases f_j overlapping with or containing f_i (See Section 6.3.6). We have mitigated these problems by splitting DPT tables in 3 subtables: (i)

[12] Valid translations are provided by the translation table.

phrases with DPT prediction, (ii) phrases with DPT prediction only for subphrases of it, and (iii) phrases with no DPT prediction for any subphrase. Formally:

$$P_{\text{DPT}'}(e|f) = \begin{cases} \lambda'_d P_{\text{DPT}}(e|f) & \text{if } \exists\ P_{\text{DPT}}(e|f) \\ \lambda_o P_{\text{MLE}}(e|f) & \text{if } (\neg\exists\ P_{\text{DPT}}(e|f)) \wedge\ (\exists\ P_{\text{DPT}}(e'|f') \wedge\ (f' \cap f \neq \emptyset)) \\ \lambda_\neg P_{\text{MLE}}(e|f) & otherwise \end{cases}$$

As discussed in Section 5.1.2, in order to perform fair comparisons, all λ parameters governing the search must be adjusted. We have followed the greedy algorithm described in Section 5.1.2, although performing only three iterations instead of five. In that manner, between 500 and 1,000 different parameter configurations are tried. For that reason, results in Chapter 5 are not directly comparable. The parameter configuration yielding the highest score, according to a given automatic evaluation measure x, over the translation of the development set will be used to translate the test set. Let us remark that, since metrics are based on different similarity assumptions, optimal parameter configurations may vary very significantly depending on the metric used to guide the optimization process. Most commonly, the BLEU metric is selected. However, in this work, we additionally study the system behavior when λ parameters are optimized on the basis of human likeness, i.e, by maximizing translation quality according to the QUEEN measure over the metric combination of highest discriminative power according to KING.

For the sake of efficiency, we have limited to the set of lexical metrics provided by IQ$_{\text{MT}}$. Metrics at deeper linguistic levels have not been used because their computation is currently too slow to allow for massive evaluation processes as it is the case of parameter adjustment. KING optimization has been carried out following the algorithm described in Section 3.4.4. The KING measure has been computed over a representative set of baseline systems based on different non-optimized parameter configurations. The resulting optimal set is: $X^+ = \{$ METEOR$_{wnsyn}$, ROUGE$_{w_1.2}$ $\}$, which includes variants of METEOR and ROUGE, metrics which, interestingly, share a common ability to capture lexical and morphological variations (use of stemming, and dictionary lookup).

6.3.5 Results

We compare the performance of DPT and MLE-based models in the full translation task. For that purpose, we use the development and test sets described in Section 5.1.1, each consisting of 504 test cases. We have used a system which relies on MLE for the estimation of translation models *('MLE')* as a baseline. We separately study the case of (i) using DPT for the set of *'all'* phrases and that of (ii) using DPT predictions for the reduced set of *'frequent'* phrases. This latter set exhibits a higher local accuracy. However, most phrases in this set are single words. Specifically, this set consists of 240 length-1 phrases, 64 length-2 phrases, 12 length-3 phrases and 1 length-4 phrase. Thus, it constitutes an excellent material to analyze the interaction between DPT and MLE-based probabilities in the context of the global task. Besides, this set covers 67% of the words in the test, whereas the 'all' set covers up to 95% of the words. In both cases, DPT predictions for uncovered words are provided by the MLE model.

Moreover, since the adjustment of internal parameters (C and σ^2) is impractical when using all phrases, based on the results from the Section 6.2, we have limited to test the behavior of binary SVMs. Also, for the sake of efficiency, we have limited to linear kernels.

System Config.	QUEEN (lexical)	METEOR (wnsyn)	ROUGE (w_1.2)	A_{pt} (all)	A_{pt} (frq)	BLEU
	BLEU-based optimization					
MLE	0.4826	0.7894	0.4385	0.7099	0.7915	0.6331
DPT$_{all}$	0.4717	0.7841	0.4383	0.7055	0.7823	**0.6429**
DPT$_{frq}$	0.4809	0.7863	0.4386	0.7102	0.7941	0.6338
	QUEEN-based optimization					
MLE	0.4872	0.7924	0.4384	0.7158	0.8097	0.6149
DPT$_{all}$	0.4907	**0.7949**	0.4391	0.7229	0.8115	0.6048
DPT$_{frq}$	**0.4913**	0.7934	**0.4404**	**0.7245**	**0.8251**	0.6038

Table 6.9: Discriminative phrase translation. Evaluation of MT results based on lexical similarity

Table 6.9 shows automatic evaluation results, before case restoration, according to different metrics, including BLEU and QUEEN. For the sake of informativeness, METEOR$_{wnsyn}$ and ROUGE$_{w_1.2}$ scores used in QUEEN computations are provided as well. Phrase translation accuracy is evaluated by means of the A_{pt} measure, both over the set of 'all' and 'frequent' phrases. We have separately studied the cases of parameter optimizations based on BLEU (rows 1 to 3) and QUEEN (rows 4 to 6). The first observation is that in the two cases DPT models yield an improved lexical choice according to the respective evaluation metric guiding the adjustment of parameters. However, for the rest of metrics there is not necessarily improvement. Interestingly, in the case of BLEU-based optimizations, DPT predictions as an additional feature report a significant BLEU improvement over the MLE baseline only when all phrases are used (see rows 2 and 3). In contrast, in the case of QUEEN-based optimizations, improvements take place in both cases, although with less significance. It is also interesting to note that the significant increase in phrase translation accuracy (A_{pt}) only reports a very modest improvement in the rest of metrics (see rows 5 and 6). This could be actually revealing a problem of interaction between DPT predictions and other models.

BLEU vs QUEEN

Table 6.9 illustrates the enormous influence of the metric selected to guide the optimization process. A system adjusted so as to maximize the score of a specific metric does not necessarily maximize the scores conferred by other metrics. In that respect, BLEU and QUEEN exhibit completely opposite behaviors. Improvements in BLEU do not necessarily imply improvements in QUEEN, and vice versa. We have further analyzed this controversial relationship by comparing optimal parameter configurations, and observed that λ's are in a very similar range, except for the weight of the word penalty model (λ_w), close to 0 in the case of BLEU, whereas in the case of QUEEN, it takes negative values around -1, thus, favoring longer translations. This seems to indicate that the heuristically motivated brevity penalty factor of BLEU could be responsible for the 'BLEU vs QUEEN' puzzle observed. We have verified this hypothesis by inspecting BLEU values before applying the penalty factor. These are on average 0.02 BLEU points higher (0.605 $\rightarrow$ 0.625), which

explains part of the puzzle. The other part must be found in the fact that, while BLEU is based on n-gram precision, QUEEN is a meta-metric which combines different quality aspects, in this case borrowed from ROUGE and METEOR.

Heterogeneous Evaluation

In order to analyze other quality aspects beyond the lexical dimension, in Table 6.10 we provide automatic evaluation results according to several metric representatives from different linguistic levels. In order to favor performance of linguistic processors, the case of automatic translations has been automatically recovered using Moses (Koehn et al., 2006). Metrics are grouped according to the level at which they operate (i.e, lexical, shallow-syntactic, syntactic, shallow-semantic and semantic). We have also computed two different QUEEN values, namely QUEEN(X^+) and QUEEN(X^+_{LF}). The first value corresponds to the application of QUEEN to the optimal metric combination based on lexical features only, whereas the second value corresponds to QUEEN applied to the optimal metric combination considering linguistic features at different levels. In this latter case, the optimal metric combination, obtained following the procedure described in subsection 6.3.4, is: $X^+_{LF} = \{$ SP-NIST$_p$, DR-O_{rp}-$\star_i$ $\}$, which includes two metrics respectively based on part-of-speech n-gram matching, and average part-of-speech overlapping over discourse representations.

First of all, metrics are evaluated according to their ability to distinguish between manual and automatic translations, as computed by KING over the six systems under evaluation. It can be observed that all metrics exhibit little ability to capture human likeness, close or even under the KING value a random metric would obtain ($\frac{1}{6}$). The highest KING value is obtained by a metric based on shallow-syntactic similarity, 'SP-NIST$_p$', which computes the NIST score over sequences of parts-of-speech. The lowest KING values are obtained by metrics at the shallow-semantic level (NE and SR families). Metric combinations show only a modest improvement over individual metrics in terms of KING.

As to system evaluation, quality aspects are diverse, and as such, it is not always the case that all aspects improve together. However, the most positive result is in the fact that all metrics based on lexical similarity consistently prefer DPT systems over MLE baselines. This confirms that DPT predictions yield an improved lexical choice. By observing scores at the lexical level, it can be observed that most metrics prefer the 'DPT$_{all}$' system optimized over BLEU. Only some ROUGE and METEOR variants prefer the DPT systems optimized over QUEEN. After all, the X^+ set, used in the QUEEN computation, consists precisely of ROUGE and METEOR variants, so this result was expected. Let us also note, that QUEEN(X^+) values in Tables 6.10 and 6.9 do not match. The reason is that, while BLEU and ROUGE scores do not vary significantly, the METEOR family is quite sensitive to case distinctions. Observe, for instance, the difference in METEOR$_{wnsyn}$ values.

At the shallow-syntactic level, all metrics prefer again the 'DPT$_{all}$' system optimized over BLEU, except the 'SP-O_p-$\star$' metric, which does not have any clear preference. At the syntactic level, however, most metrics prefer the 'MLE' systems. Only the shallowest metrics, e.g., DP-HWC$_w$-4 (i.e., lexical head-word matching over dependency trees), CP-O_p-$\star$ and CP-O_c-$\star$ (i.e., lexical overlapping over parts-of-speech and phrase constituents) seem to prefer DPT systems, always optimized over BLEU. This is a very interesting result since it reveals that an improved lexical similarity does not necessarily lead to an improved syntactic structure.

Metric	KING	BLEU-based optim.			QUEEN-based optim.		
		MLE	DPT$_{all}$	DPT$_{frq}$	MLE	DPT$_{all}$	DPT$_{frq}$
1-WER	0.1581	0.6651	**0.6907**	0.6841	0.6797	0.6504	0.6541
1-PER	0.1448	0.7571	**0.7763**	0.7700	0.7679	0.7397	0.7481
1-TER	0.1567	0.6882	**0.7134**	0.7061	0.7030	0.6736	0.6776
BLEU	0.1177	0.6149	**0.6430**	0.6339	0.6331	0.6048	0.6039
NIST	0.1534	10.9529	**11.3406**	11.2403	11.2210	10.7512	10.8017
GTM ($e = 2$)	0.1415	0.4226	**0.4283**	0.4248	0.4265	0.4204	0.4170
O_l	0.1210	0.7056	**0.7108**	0.7087	0.7076	0.6970	0.7027
ROUGE$_L$	0.1349	0.6914	**0.6984**	0.6962	0.6958	0.6888	0.6908
ROUGE$_W$	**0.1653**	0.4384	0.4383	0.4386	0.4385	0.4391	**0.4404**
METEOR$_{exact}$	0.1495	0.7217	0.7197	0.7209	0.7214	**0.7232**	**0.7234**
METEOR$_{wnsyn}$	**0.1706**	0.7602	0.7569	0.7580	0.7589	**0.7610**	0.7599
QUEEN(X^+)	0.1753	0.5165	0.5058	0.5171	0.5152	0.5149	**0.5224**
SP-O_p-$\star$	0.1376	0.6842	0.6843	0.6842	0.6844	0.6760	0.6811
SP-O_c-$\star$	0.1349	0.6854	**0.6871**	0.6855	0.6853	0.6818	0.6814
SP-NIST$_l$	0.1534	11.0467	**11.4334**	11.3304	11.3139	10.8396	10.8970
SP-NIST$_p$	**0.2156**	9.7950	**10.0853**	9.9871	10.0258	9.6265	9.6064
SP-NIST$_{iob}$	0.1812	7.5005	**7.6598**	7.6124	7.6124	7.3858	7.3497
SP-NIST$_c$	0.1806	6.8328	**7.0240**	6.9644	6.9297	6.7105	6.6964
DP-HWC$_w$-4	0.1171	0.2704	**0.2763**	0.2661	0.2694	0.2711	0.2691
DP-HWC$_c$-4	**0.1720**	0.4920	0.4887	0.4899	**0.4951**	0.4771	0.4929
DP-HWC$_r$-4	0.1653	0.4354	0.4332	0.4324	**0.4377**	0.4202	0.4344
DP-O_l-$\star$	0.1515	**0.5055**	**0.5060**	0.5032	0.5045	0.4978	0.4992
DP-O_c-$\star$	0.1594	0.6003	0.5995	0.5999	**0.6038**	0.6006	0.6000
DP-O_r-$\star$	**0.1739**	0.4656	0.4651	0.4633	**0.4674**	0.4612	0.4624
CP-O_p-$\star$	0.1343	0.6819	**0.6836**	0.6832	0.6824	0.6747	0.6773
CP-O_c-$\star$	0.1389	0.6561	**0.6595**	0.6582	0.6570	0.6470	0.6508
CP-STM-4	0.1521	0.**6836**	0.6821	0.6821	**0.6843**	0.6792	0.6782
NE-O_e-$\star\star$	0.1356	0.6870	**0.6927**	0.6905	0.6897	0.6785	0.6834
NE-O_e-$\star$	0.0714	0.5346	**0.5479**	0.5422	0.5444	0.5368	0.5274
NE-M_e-$\star$	0.0582	0.5165	**0.5314**	0.5267	0.5279	0.5202	0.5136
SR-O_r-$\star_b$	0.1190	0.3527	0.3584	**0.3618**	0.3519	0.3353	0.3440
SR-M_r-$\star_b$	0.1012	0.2220	0.2149	**0.2227**	0.2192	0.2179	0.2185
SR-O_{rb}	0.1310	0.5526	0.5579	**0.5657**	0.5556	0.5357	0.5424
DR-O_r-$\star_b$	0.1534	0.5403	0.5379	0.5360	**0.5436**	0.5380	0.5335
DR-O_{rp}-$\star_b$	0.1640	0.6540	0.6471	0.6470	**0.6576**	0.6542	0.6437
DR-O_{rp}-$\star_i$	**0.1799**	0.5358	0.5325	0.5311	**0.5386**	0.5338	0.5253
DR-STM-4$_b$	0.1680	0.5241	0.5201	0.5203	**0.5254**	0.5194	0.5137
DR-STM-4$_i$	0.1640	0.4657	0.4640	0.4634	**0.4678**	0.4590	0.4555
QUEEN(X^+_{LF})	0.2262	0.3325	**0.3474**	0.3447	**0.3485**	0.3302	0.3062

Table 6.10: Discriminative phrase translation. Heterogeneous evaluation of MT results

	Adequacy	Fluency	Overall
DPT > MLE	**89**	68	**99**
DPT = MLE	100	76	46
DPT < MLE	39	**84**	83

Table 6.11: Discriminative phrase translation. Manual evaluation of MT results

At the shallow-semantic level, while NE-based similarities although not very informative[13], tend to prefer the 'DPT$_{all}$' system optimized over BLEU, SR metrics seem to prefer the 'DPT$_{frq}$' system optimized over BLEU, whereas at the properly semantic level, metrics based on discourse representations prefer the 'MLE' system optimized over QUEEN. Therefore, no clear conclusions can be made on which model or optimization strategy leads to a better semantic structure.

Finally, according to QUEEN(X^{+}_{LF}), i.e., combining the 'SP-NIST$_p$' and 'DR-O_{rp}-$\star_i$' metrics on the basis of human likeness, the best system is the 'MLE' baseline optimized over QUEEN, with a slight advantage over the two DPT variants optimized over BLEU.

Manual Evaluation

Several conclusions must be drawn from these results. First, the lack of consensus between metrics based on different similarity criteria reinforces the need for evaluation methodologies which allow system developers to take into account a heterogeneous set of quality aspects. Second, the fact that an improved lexical similarity does not necessarily lead to an improved syntactic or semantic structure might be revealing problems of interaction between DPT predictions and the other models in the SMT system. We have verified this hypothesis through a number of manual evaluations. These have revealed that gains are mainly related to the adequacy dimension, whereas for fluency there is no significant improvement. For instance, Table 6.11 presents manual evaluation results corresponding to the pairwise comparison of the DPT$_{frq}$ system and the MLE baseline, both optimized over QUEEN. The set of test cases was selected based on the following criteria:

- sentence length between 10 and 30 words.
- at least 5 words have a DPT prediction.
- DPT and MLE outputs differ.

A total of 114 sentences fulfilled these requirements. The manual evaluation was conducted following the procedure described in Section6.3.3. Four judges participated in the evaluation. Each judge evaluated only half of the cases. Each case was evaluated by two different judges. Thus, we obtained 228 human assessments. According to human assessors, the DPT system outperforms the MLE-based system very significantly with respect to adequacy, whereas for fluency there is a slight advantage in favor of the MLE baseline. Overall, there is a slight but significant advantage in favor of the 'DPT' system.

[13]Observe the low KING values attained, except for the case of the 'NE-O_e-$\star\star$' metric, which also considers overlapping among tokens which are not named entities.

6.3.6 Error Analysis

Tables 6.12, 6.13 and 6.14 show three sentence fragments illustrating the different behavior of the system configurations evaluated. We start, in Table 6.12, by showing a positive case in which the DPT predictions help the system to find a better translation for *'fuera sancionado'*. Observe how baseline SMT systems, whose translation models are based on MLE, all wrongfully translate *'fuera'* as *'outside'* instead of as an auxiliary verb form (e.g., *'was'* or *'were'*) or past form of the accompanying verb *'sancionado'* (e.g., *'sanctioned'* or *'penalised'*) . In contrast, 'DPT$_{all}$' systems are able to provide more appropriate translations for this phrase, regardless of the metric guiding the parameter optimization process. Observe also, how 'DPT$_{frq}$' systems, which, unfortunately, do not count on DPT predictions for this not frequent enough phrase, commit all the same mistake than MLE-based systems.

Tables 6.13 and 6.14 present two cases in which the metric guiding the optimizations has a stronger influence. In Table 6.13, all MLE baseline systems wrongfully translate *'cuyo nombre'* into *'whose behalf'*. Only the 'DPT$_{all}$' system optimized over BLEU is able to find a correct translation (*'whose name'*). In Table 6.14, while MLE-based systems provide all fairly correct translations of *'van a parar a'* into *'go to'*, DPT predictions may cause the system to wrongfully translate *'van a parar a'* into *'are going to stop to'*. Only the 'DPT$_{frq}$' system optimized over BLEU is able to find a correct translation. The underlying cause behind these two cases is that there is no DPT prediction for *'cuyo nombre'* and *'van a parar a'*, two phrases of very high cohesion, but only for subphrases of it (e.g., *'cuyo'*, *'nombre'*, *'van'*, *'a'*, *'parar'*, *'van a'* , *'a parar'*). DPT predictions for these subphrases must compete with MLE-based predictions for larger phrases, which may cause problems of interaction.

Source	Yo quisiera que el incumplimiento institucional del Consejo **fuera sancionado** [...]
Ref 1	I would like the Council 's institutional infringement **to be penalised** [...]
Ref 2	I would like the Council 's institutional non-fulfilment of its obligations **to be sanctioned** [...]
Ref 3	I would like to see the institutional non-compliance of the Council **punished** [...]
	BLEU-based optimizations
MLE	I would like to see the failure to comply with institutional **outside** of the Council **sanctioned** [...]
DPT$_{all}$	I would like to see the institutional breach of the Council **was sanctioned** [...]
DPT$_{frq}$	I would like to see the institutional breach of the Council **outside sanctioned** [...]
	QUEEN-based optimizations
MLE	I would like to see the failure to comply with the institutional Councils **outside sanctioned** [...]
DPT$_{all}$	I would like to see the failure to comply with the institutions of the Council **were to be sanctioned** [...]
DPT$_{frq}$	I would like to see the failure to comply with the institutional Councils **outside sanctioned** [...]

Table 6.12: Discriminative phrase translation. Case of Analysis #1. DPT models help

Source	[...] aquel diputado **cuyo nombre** no conozco [...]
Ref 1	[...] the Member **whose name** I do not know [...]
Ref 2	[...] the Honourable Member , **whose name** I can not recall [...]
Ref 3	[...] that Member **whose name** I ignore [...]
	BLEU-based optimizations
MLE	[...] that Member **whose behalf** I do not know [...]
DPT$_{all}$	[...] that Member **whose name** I do not know [...]
DPT$_{frq}$	[...] that Member **whose behalf** I do not know [...]
	QUEEN-based optimizations
MLE	[...] that Member **on whose behalf** I am not familiar with [...]
DPT$_{all}$	[...] that Member **on whose behalf** I am not familiar with [...]
DPT$_{frq}$	[...] that MEP **whose behalf** I am not familiar with [...]

Table 6.13: Discriminative phrase translation. Case of Analysis #2. DPT models may help

Source	[...] poco más del 40 % de los fondos **van a parar a** esos países .
Ref 1	[...] only slightly more than 40 % of the money **ends up in** those countries .
Ref 2	[...] little more than 40 % of these funds **end up in** these countries .
Ref 3	[...] little more than 40 % of the funds **are going to** those countries .
	BLEU-based optimizations
MLE	[...] little more than 40 % of the funds **go to them** .
DPT$_{all}$	[...] little more than 40 % of the funds **will stop to these countries** .
DPT$_{frq}$	[...] little more than 40 % of the funds **go to these countries** .
	QUEEN-based optimizations
MLE	[...] just a little more than 40 % of the money **goes to those countries** .
DPT$_{all}$	[...] little more than 40 % of the funds **are going to stop to these countries** .
DPT$_{frq}$	[...] little more than 40 % of the funds **are going to stop to these countries** .

Table 6.14: Discriminative phrase translation. Case of Analysis #3. DPT models may not help

6.4 Related Work

As we have seen in Section 4.4, other authors have recently conducted similar experiments. Although tightly related, there exist several important differences between the works by Carpuat and Wu (2007b), Bangalore et al. (2007), Stroppa et al. (2007), Specia et al. (2008), and ours. These differences are discussed below. We have divided them in three main categories: (i) task, (ii) system and (iii) evaluation differences.

6.4.1 Task Differences

Several translation scenarios are approached. The most important differences are related to:

- Language pair (Spanish-to-English, Chinese-to-English, Arabic-to-English, French-to-English, and English-to-Portuguese).
- Task domain, as determined by the corpus (Europarl, NIST, BTEC, Hansards, United Nations, or heterogeneous compilations).

For instance, while we work in the Spanish-to-English translation of European Parliament proceedings, Carpuat and Wu (2007b) and Bangalore et al. (2007) work on the Chinese-to-English translation of basic travel expressions and newswire articles, and Stroppa et al. (2007) work on the Chinese-to-English and Italian-to-English translation of basic travel expressions. Additionally, Bangalore et al. (2007) present results on Arabic-to-English translation of proceedings of the United Nations and on French-to-English translation of proceedings of the Canadian Parliament.

All these works focus on a single translation domain. In contrast, Specia et al. (2008), worked on the English-to-Portuguese translation of a heterogeneous data set of different domains and genres, compiled from various sources, including the Bible, literary fiction, European Parliament proceedings and a mixture of smaller sources. The significant improvements reported evince that dedicated lexical selection models are a valid solution to tackle domain shifts.

6.4.2 System Differences

Other differences are related to the translation system itself:

- System Architecture (Log-linear models vs. Finite-state transducers vs. Reranking)
- Learning scheme (Support Vector Machines, Maximum Entropy, Naïve Bayes, Boosting, Kernel PCA-based models, Memory-based learning, Inductive Logic Programming, and combined schemes).

For instance, while we rely on SVM predictions, Carpuat and Wu (2007b) use an ensemble of four combined models (naïve Bayes, maximum entropy, boosting, and Kernel PCA-based models), Stroppa et al. (2007) rely on memory-based learning, Bangalore et al. (2007) use maximum entropy, and Specia et al. (2008) use Inductive Logic Programming (ILP) and SVMs.

Besides, Bangalore et al. (2007) employ a slightly different SMT architecture based on stochastic finite-state transducers which addresses the translation task as two separate processes: (i) global lexical selection, i.e., dedicated word selection, and (ii) sentence reconstruction. Moreover, their translation models are indeed bilingual language models. They also deal with reordering in a different manner. Prior to translation, the source sentence is reordered so as to approximate the right order of the target language. This allows them to perform a monotonic decoding.

Apart from the corpus heterogeneity, the approach by Specia et al. (2008) has two other very interesting particularities. First, their dedicated models are exploited in the context of a syntax-based dependency treelet SMT system (Quirk et al., 2005). Second, their integration strategy is based on using ILP predictions as an additional feature for the reranking of n-best lists (Och et al., 2004). As explained in Section 4.2.3, reranking has the the cost of possibly discarding valid translations when compiling the n-best list. In order to overcome this limitation, they expanded the n-best list by generating new translations which include the most probable candidate translations according to dedicated predictions.

6.4.3 Evaluation Differences

There are also significant differences in the evaluation process. Bangalore et al. (2007) and Specia et al. (2008) rely on BLEU as the only measure of evaluation, Stroppa et al. (2007) additionally rely on NIST, and Carpuat and Wu (2007b) show results according to eight different standard evaluation metrics based on lexical similarity including BLEU and NIST. In contrast, we have used a set of evaluation metrics operating at deeper linguistic levels. We have also relied on the QUEEN measure, which allows for non-parametric combinations of different metrics into a single measure of quality. Besides, we have conducted several processes of manual evaluation.

6.5 Conclusions of this Chapter

In this chapter, we have shown that discriminative phrase translation may be successfully applied to SMT. Despite the fact that measuring improvements in word selection is a very delicate issue, experimental results, according to several well-known metrics based on lexical similarity, show that dedicated DPT models yield a significantly improved lexical choice over traditional MLE-based ones. However, by evaluating linguistic aspects of quality beyond the lexical level (e.g., syntactic, and semantic), we have found that an improved lexical choice and semantic structure does not necessarily lead to an improved grammaticality. This result has been verified through a number of manual evaluations, which have revealed that gains are mainly related to the adequacy dimension, whereas for fluency there is no significant improvement.

Besides, this work has also served us to study the role of automatic metrics in the development cycle of MT systems, and the importance of meta-evaluation. We have shown that basing evaluations and parameter optimizations on different metrics may lead to very different system behaviors. For system comparison, this may be solved through manual evaluations. However, this is impractical for the adjustment of parameters, where hundreds of different configurations are tried. Thus, we argue that more attention should be paid to the meta-evaluation process. In our case, metrics have been evaluated on the basis of human likeness. Other solutions exist. The main point, in our

opinion, is that system development is metricwise (see Section 2.2.3). This is a crucial issue, since, most often, system improvements focus on partial aspects of quality, such as word selection or word ordering, which can not be always expected to improve together.

Finally, the fact that improvements in adequacy do not lead to an improved fluency evinces that the integration of local DPT probabilities into the statistical framework requires further study. We believe that if DPT models considered features from the target side under generation and from the correspondence between source and target, phrase translation accuracy would improve and cooperation with the decoder would be even softer. Although, still, predictions based on local training may not always be well suited for being integrated in the target translation. Thus, we also argue that if phrase translation classifiers were trained in the context of the global task their integration would be more robust and translation quality could further improve. The possibility of moving towards a new global DPT architecture in the fashion, for instance, of those suggested by Tillmann and Zhang (2006) or Liang et al. (2006) should be considered.

Chapter 7

Domain Adaptation of an SMT System

As discussed in Chapter 4, Section 4.5, empirical MT systems suffer a significant quality drop when applied to a different domain. In this chapter, we analyze different alternatives so as to adapt an existing SMT system to a new domain when few or none domain-specific data are available.

We present a practical case study on the automatic translation into Spanish of the glosses in the English WordNet (Fellbaum, 1998). Glosses are short dictionary definitions that accompany WordNet synsets. We have selected this scenario for several reasons. First, WordNet glosses are a useful resource which has been successfully applied to many NLP tasks. For instance, Mihalcea and Moldovan (1999) suggested an automatic method for generating sense tagged corpora which uses WordNet glosses. Hovy et al. (2001) used WordNet glosses as external knowledge to improve their Webclopedia Question Answering (QA) system. Second, there exist wordnets for several languages[1], but they contain, in general, very few glosses. For instance, in the current version of the Spanish WordNet fewer than 10% of its synsets have a gloss. Conversely, since version 1.6, every synset in the English WordNet has a gloss. We believe that a method to rapidly obtain glosses for all wordnets may be helpful, and an opportunity for current empirical MT techniques to show their applicability. These glosses could serve as a starting point for a further stage of revision. Moreover, from a conceptual point of view, the idea of enriching wordnets using other wordnets results very attractive.

We start by building an out-of-domain SMT system based on a parallel corpus of European Parliament Proceedings. We have analyzed its domain dependence by applying it directly to the domain of dictionary definitions. As expected, this system fails to properly translate WordNet glosses. After inspecting particular cases, we found out that most errors are related to unseen events (e.g., unknown words, expressions and syntactic constructions). In order to adapt the system to the new domain, we suggest several techniques, all based on the incorporation of outer knowledge:

- **Use of In-domain Corpora.** We count on a small set of Spanish hand-developed glosses generated, however, without considering its English counterpart. This in-domain corpus is used to construct specialized statistical models, which are well suited for being combined with out-of-domain models.

[1] A list of wordnets currently under development is available at `http://www.globalwordnet.org/gwa/wordnet_table.htm`.

- **Use of Close-to-domain Corpora.** We count on two large monolingual Spanish electronic dictionaries. We use these dictionaries to build additional language models.
- **Use of Domain-independent Knowledge Sources.** We exploit the information contained in WordNet itself to construct domain-independent translation models.

We show that these simple techniques may yield a very significant improvement. Results are also accompanied by a detailed process of qualitative error analysis.

7.1 Corroborating Domain Dependence

Prior to elaborating on adaptation techniques, we verify the problem of domain dependence of SMT systems in the specific scenario.

7.1.1 Settings

We have built two individual baseline systems. Both are phrase-based SMT systems constructed following the procedure described in Section 5.1. The differences between them are related only to the training data utilized:

- **Out-of-domain Baseline System.** The first baseline system ('EU') is entirely based on a collection of 730,740 out-of-domain parallel sentences extracted from the Europarl corpus (Koehn, 2003a)[2], which corresponds exactly to the training data provided by the organizers of the Shared Task 2: *"Exploiting Parallel Texts for Statistical Machine Translation"* of the ACL 2005 Workshop on *"Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond"* (Koehn & Monz, 2005)[3]. A brief numerical description of this data set is available in Table 7.1.
- **In-domain Baseline System.** The second baseline system ('WNG') is entirely based on a small in-domain corpus of English–Spanish parallel glosses. This has been collected using the *Multilingual Central Repository* (MCR), a multilingual lexical-semantic database which connects several WordNets at the synset level (Atserias et al., 2004). The MCR has been developed following the EuroWordNet design (Vossen et al., 1997), in the context of the MEANING project[4]. Currently, the MCR includes linked versions of the English, Spanish, Italian, Basque and Catalan wordnets. Overall, we count on a set of 6,519 parallel glosses, which correspond to 5,698 nouns, 87 verbs, and 734 adjectives. We have removed examples and parenthesized texts. Glosses have also been tokenized and case lowered. In addition, we have discarded some of these parallel glosses based on the difference in length between the source and the target. The gloss average length for the resulting 5,843 glosses was 8.25 words for English and 8.13 for Spanish. Finally, gloss pairs have been randomly split into training

[2]The Europarl Corpus is available at `http://people.csail.mit.edu/people/koehn/publications/europarl`

[3]`http://www.statmt.org/wpt05/`.

[4]`http://www.lsi.upc.edu/~nlp/meaning/`

	Set	#sentences	#tokens	#distinct tokens
Spanish	Train	730,740	15,676,710	102,885
	Test	2,000	60,276	7,782
English	Train	730,740	15,222,105	64,122
	Test	2,000	57,945	6,054

Table 7.1: WMT 2005 Shared Task. Description of the Spanish-English data sets

	Set	#sentences	#tokens	#distinct tokens
Spanish	Train	4,843	39,311	6,358
	Test	500	3,981	1,454
	Dev	500	4,193	1,509
English	Train	4,843	40,029	6,495
	Test	500	4,036	1,544
	Dev	500	4,167	1,560

Table 7.2: Description of the small Spanish-English corpus of parallel glosses

(4,843), development (500) and test (500) sets. A brief numerical description is available in Table 7.2.

7.1.2 Results

Table 7.3 shows automatic evaluation results of the two baseline systems, both over development and test sets, according to several standard metrics based on lexical similarity. We also compare the performance of the 'EU' baseline on these data sets with respect to its performance on the Europarl test set from the ACL 2005 MT workshop ('acl05-test'). The first observation is that, as expected, there is a very significant decrease in performance (from 0.24 to 0.08 according to BLEU) when the 'EU' baseline system is applied to the new domain. Some of this decrement is also due to a certain degree of free translation exhibited by the set of available *quasi-parallel* glosses. We further discuss this issue in Section 7.2.3.

Results obtained by the 'WNG' system are also very low, though significantly higher than those attained by the 'EU' system. This is a very interesting fact. Although the amount of data utilized to construct the 'WNG' baseline is 150 times smaller than the amount utilized to construct the 'EU' baseline, its performance is better consistently according to all metrics. We interpret this result as a corroboration that models estimated from in-domain data provide higher precision.

We also compare these results to those attained by a commercial system. We use the on-line

System	GTM ($e = 1$)	GTM ($e = 2$)	METEOR (wnsyn)	NIST	BLEU
Development					
EU$_{baseline}$	0.3131	0.2216	0.2881	2.8832	0.0737
WNG$_{baseline}$	0.3604	0.2605	0.3288	3.3492	0.1149
SYSTRAN	0.4257	0.2971	0.4394	3.9467	0.1625
Test					
EU$_{baseline}$	0.3131	0.2262	0.2920	2.8896	*0.0790*
WNG$_{baseline}$	0.3471	0.2510	0.3219	3.1307	0.0951
SYSTRAN	0.4085	0.2921	0.4295	3.7873	0.1463
acl05-test					
EU$_{baseline}$	0.5699	0.2429	0.5153	6.5848	0.2381

Table 7.3: Translation of WordNet glosses. Baseline performance

version 5.0 of SYSTRAN[5], a general-purpose MT system based on manually-defined lexical and syntactic transfer rules. The performance of the baseline systems is significantly worse than SYSTRAN's on both development and test sets. This confirms the widespread assumption that rule-based systems are more robust than SMT systems against changes in domain. The difference with respect to the specialized 'WNG' also suggests that the amount of data used to train the 'WNG' baseline is clearly insufficient.

7.1.3 Error Analysis

Tables 7.4 and 7.5 show, respectively, several positive and negative cases on the performance of the 'EU' out-of-domain system, based on the GTM F-measure ($e = 2$). We use this measure because, in contrast to BLEU, it has an intuitive interpretation. It represents the fraction of the automatic-reference translation grid covered by aligned blocks. Automatic translations are accompanied by the scores attained. Interestingly, most of the positive cases are somehow related to the domains of politics and/or economy (e.g., cases 1, 2, 4, 5, 6 and 7 in Table 7.4), which are close to the domain of the corpus of parliament proceedings utilized. As to low quality translations, in many cases these are due to unknown vocabulary (e.g., cases 2, 3, 5 and 6 in Table 7.5). However, we also found many translations unfairly scoring too low due to *quasi-parallelism*, i.e., divergences between source and reference glosses (e.g., cases 1, 4, and 7 in Table 7.5).

7.2 Combining Knowledge Sources

In order to improve the baseline results, first, we use monolingual electronic dictionaries so as to build close-to-domain language models. Then, we study different alternatives for combining in-domain and out-of-domain translation models. Thus, we move to a working scenario in which the

[5] http://www.systransoft.com/

	Synset		Content
1	00392749#n	**Source**	the office and function of *president*
		Reference	cargo y función de presidente
	F_1=0.9090	**Target**	el cargo y función de presidente
2	00630513#n	**Source**	the action of *attacking the enemy*
		Reference	acción y efecto de atacar al enemigo
	F_1=0.6871	**Target**	acción de atacar al enemigo
3	00785108#n	**Source**	the act of giving hope or support to someone
		Reference	acción de dar esperanza o apoyo a alguien
	F_1=0.9412	**Target**	la acción de dar esperanza o apoyo a alguien
4	00804210#n	**Source**	the combination of two or more *commercial companies*
		Reference	combinación de dos o más empresas
	F_1=0.7142	**Target**	la combinación de dos o más comerciales compañías
5	05359169#n	**Source**	the act of *presenting a proposal*
		Reference	acto de presentar una propuesta
	F_1=0.9090	**Target**	el acto de presentar una propuesta
6	06089036#n	**Source**	a *military unit* that is part of an *army*
		Reference	unidad militar que forma parte de un ejército
	F_1=0.0609	**Target**	unidad militar que forma parte de un ejército
7	06213619#n	**Source**	a group of *representatives* or *delegates*
		Reference	grupo de representantes o delegados
	F_1=1	**Target**	grupo de representantes o delegados
8	06365607#n	**Source**	a safe place
		Reference	lugar seguro
	F_1=1	**Target**	lugar seguro
9	01612822#v	**Source**	perform an action
		Reference	realizar una acción
	F_1=1	**Target**	realizar una acción

Table 7.4: Translation of WordNet glosses. Error analysis #1 (good translations)

	Synset		Content
1	00012865#n	**Source**	a feature of the mental life of a living organism
		Reference	rasgo psicológico
	F_1=0	**Target**	una característica de la vida mental de un organismo vivo
2	00029442#n	**Source**	the act of departing politely
		Reference	acción de marcharse de forma educada
	F_1=0	**Target**	el acto de *departing politely*
3	02581431#n	**Source**	a kitchen appliance for disposing of garbage
		Reference	cubo donde se depositan los residuos
	F_1=0	**Target**	*kitchen* una *appliance* para *disposing* de *garbage*
4	05961082#n	**Source**	people in general
		Reference	grupo de gente que constituye la mayoría de la población y que define y mantiene la cultura popular y las tradiciones
	F_1=0.0833	**Target**	gente en general
5	07548871#n	**Source**	a painter of theatrical scenery
		Reference	persona especializada en escenografía
	F_1=0	**Target**	una *painter* de *theatrical scenery*
6	10069279#n	**Source**	rowdy behavior
		Reference	comportamiento escandaloso
	F_1=0	**Target**	*rowdy behavior*
7	00490201#a	**Source**	without reservation
		Reference	sin reservas
	F_1=0	**Target**	movido por una devoción o un compromiso entusiasta y decidido

Table 7.5: Translation of WordNet glosses. Error analysis #2 (bad translations)

Dictionary	#definitions	#tokens	distinct tokens
D1	142,892	2,111,713	79,063
D2	168,779	1,553,311	72,435

Table 7.6: Description of two Spanish electronic dictionaries

large out-of-domain corpus contributes with recall by covering unseen events, while the in-domain corpus contributes mainly with precision, by providing more accurate translations.

7.2.1 Adding Close-to-domain Language Models

In first place we turned our eyes to language modeling. In addition to the language model built from the Europarl corpus ('EU') and the specialized language model based on the small training set of parallel glosses ('WNG'), we have built two specialized language models, 'D1' and 'D2', based on two large monolingual Spanish electronic dictionaries:

- **D1** Gran diccionario de la Lengua Española (Martí, 1996).
- **D2** Diccionario Actual de la Lengua Española (Vox, 1990).

A brief numerical description of these dictionaries, after case lowering, is available in Table 7.6.

Automatic evaluation results, over the development set, are shown in Table 7.7. We tried several configurations. In all cases, language models were combined with equal probability. As expected, the closer the language model is to the target domain, the better results. The first observation is that using language models 'D1' and 'D2' outperforms the results using the out-of-domain 'EU' language model. A second observation is that best results are in all cases consistently attained when using the 'WNG' language model, either alone or combined with close-to-domain language models. This means that language models estimated from small sets of in-domain data are helpful. A third observation is that a significant gain is obtained by incrementally adding in-domain or close-to-domain specialized language models to the baseline systems, according to all metrics but BLEU for which no combination seems to significantly outperform the 'WNG' baseline alone. Observe that the best results are obtained, except in the case of BLEU, by the system using the out-of-domain 'EU' translation model combined with in-domain and, optionally, close-to-domain language models. We interpret this result as an indicator that translation models estimated from out-of-domain data are helpful because they provide recall. Another interesting point is that adding an out-of-domain language model ('EU') does not seem to help, at least combined with equal probability to in-domain models. Same conclusions hold for the test set, too.

Tuning the System

Adjusting the Pharaoh parameters that control the importance of the different probabilities governing the search may yield significant improvements. In our base SMT system, there are 4 important

Translation Model	Language Model	GTM ($e = 1$)	GTM ($e = 2$)	METEOR (wnsyn)	NIST	BLEU
EU	**EU**	0.3131	0.2216	0.2881	2.8832	0.0737
EU	**WNG**	**0.3714**	0.2631	**0.3377**	**3.4831**	0.1062
EU	**D1**	0.3461	0.2503	0.3158	3.2570	0.0959
EU	**D2**	0.3497	0.2482	0.3163	3.2518	0.0896
EU	**D1 + D2**	0.3585	0.2579	0.3244	3.3773	0.0993
EU	**EU + D1 + D2**	0.3472	0.2499	0.3160	3.2851	0.0960
EU	**D1 + D2 + WNG**	0.3690	0.**2662**	**0.3372**	**3.4954**	**0.1094**
EU	**EU + D1 + D2 + WNG**	0.3638	0.2614	0.3321	3.4248	0.1080
WNG	**EU**	0.3128	0.2202	0.2689	2.8864	0.0743
WNG	**WNG**	0.3604	0.2605	0.3288	3.3492	**0.1149**
WNG	**D1**	0.3404	0.2418	0.3050	3.1544	0.0926
WNG	**D2**	0.3256	0.2326	0.2883	3.0295	0.0845
WNG	**D1 + D2**	0.3331	0.2394	0.2995	3.1185	0.0917
WNG	**EU + D1 + D2**	0.3221	0.2312	0.2847	3.0361	0.0856
WNG	**D1 + D2 + WNG**	0.3462	0.2479	0.3117	3.2238	0.0980
WNG	**EU + D1 + D2 + WNG**	0.3309	0.2373	0.2941	3.0974	0.0890

Table 7.7: Translation of WordNet glosses. Combined language models

kinds of parameters to adjust: the language model probabilities (λ_{lm}), the translation model probability (λ_ϕ), the distortion model probability (λ_d) and the word penalty factor (λ_w). In our case, it is specially important to properly adjust the contribution of the language models. We adjusted parameters by means of a software based on the *Downhill Simplex Method in Multidimensions* (William H. Press & Flannery, 2002) implemented by our fellow Patrik Lambert. The tuning was based on maximizing the BLEU score attained over the development set. We tuned 6 parameters: 4 language models (λ_{lmEU}, λ_{lmD1}, λ_{lmD2}, λ_{lmWNG}), the translation model (λ_ϕ), and the word penalty (λ_w)[6]. Evaluation results, in Table 7.8, report a substantial improvement. Highest scores are attained using the 'EU' translation model. Interestingly, the weight of language models is concentrated on the small but precise in-domain 'WNG' language model ($\lambda_{lmWNG} = 0.95$).

7.2.2 Integrating In-domain and Out-of-domain Translation Models

We also study the possibility of combining out-of-domain and in-domain translation models aiming at achieving a good balance between precision and recall that yields better MT results. Two different strategies have been tried. In a first strategy we simply concatenate the out-of-domain corpus ('EU') and the in-domain corpus ('WNG'). Then, we construct the translation model ('EUWNG') as described in Section 5.1. A second manner to proceed is to linearly combine the two different

[6]Final values when using the 'EU' translation model are $\lambda_{lmEU} = 0.22$, $\lambda_{lmD1} = 0$, $\lambda_{lmD2} = 0.01$, $\lambda_{lmWNG} = 0.95$, $\lambda_\phi = 1$, and $\lambda_w = -2.97$, while when using the 'WNG' translation model final values are $\lambda_{lmEU} = 0.17$, $\lambda_{lmD1} = 0.07$, $\lambda_{lmD2} = 0.13$, $\lambda_{lmWNG} = 1$, $\lambda_\phi = 0.95$, and $\lambda_w = -2.64$.

Translation Model	Language Model	GTM ($e = 1$)	GTM ($e = 2$)	METEOR (wnsyn)	NIST	BLEU
development						
EU	**EU + D1 + D2 + WNG**	0.3856	0.2727	0.3695	3.6094	**0.1272**
WNG	**EU + D1 + D2 + WNG**	0.3688	0.2676	0.3452	3.3740	0.1269
test						
EU	**EU + D1 + D2 + WNG**	0.3720	0.2650	0.3644	3.4180	**0.1133**
WNG	**EU + D1 + D2 + WNG**	0.3525	0.2552	0.3343	3.1084	0.1015

Table 7.8: Translation of WordNet glosses. Effects of tuning the contribution of language models

Translation Model	Language Model	GTM ($e = 1$)	GTM ($e = 2$)	METEOR (wnsyn)	NIST	BLEU
development						
EUWNG	**WNG**	0.3949	0.2832	0.3711	3.7677	0.1288
EUWNG	**EU + D1 + D2 + WNG**	0.4081	0.2944	0.3998	3.8925	0.1554
EU+WNG	**WNG**	0.4096	0.2936	0.3804	3.9743	0.1384
EU+WNG	**EU + D1 + D2 + WNG**	0.4234	0.3029	0.4130	4.1415	**0.1618**
test						
EUWNG	**WNG**	0.3829	0.2771	0.3595	3.6777	0.1123
EUWNG	**EU + D1 + D2 + WNG**	0.3920	0.2810	0.3885	3.6478	0.1290
EU+WNG	**WNG**	0.3997	0.2872	0.3723	3.8970	0.1227
EU+WNG	**EU + D1 + D2 + WNG**	0.4084	0.2907	0.3963	3.8930	**0.1400**

Table 7.9: Translation of WordNet glosses. Combined translation models

translation models into a single translation model ('EU+WNG'). In this case, we can assign different weights (ω) to the contribution of the different models to the search. We can also determine a certain threshold θ which allows us to discard phrase pairs under a certain probability. These weights and thresholds have been adjusted as detailed in Section 7.2.1. Optimal values are: $\omega_{tmEU} = 0.1$, $\omega_{tmWNG} = 0.9$, $\theta_{tmEU} = 0.1$, and $\theta_{tmWNG} = 0.01$. Interestingly, at combination time the importance of the 'WNG' translation model ($\omega_{tmWNG} = 0.9$) is much higher than the importance of the 'EU' translation model ($\omega_{tmEU} = 0.1$).

Table 7.9 shows results for the two strategies after tuning, both over development and test sets. As expected, the 'EU+WNG' strategy consistently obtains the best results according to all metrics both on the development and test sets, since it allows to better adjust the relative importance of each translation model. However, both techniques achieve a very competitive performance. For instance, according to BLEU, results improve from 0.13 to 0.16, and from 0.11 to 0.14, for the development and test sets, respectively. Improvement is also captured by all other metrics except NIST, which improves only for the development set.

We measured statistical significance of the results using the bootstrap resampling technique, over the BLEU measure, as described by Koehn (2004b). The 95% confidence intervals extracted

from the test set after 10,000 samples are the following:

$$I_{\mathrm{EU}_{\mathrm{baseline}}} = [0.0642, 0.0939]$$
$$I_{\mathrm{WNG}_{\mathrm{baseline}}} = [0.0788, 0.1112]$$
$$I_{\mathrm{EU+WNG}_{\mathrm{best}}} = [0.1221, 0.1572]$$

Since intervals do not overlap, we can conclude that differences are statistically significant at the 95% level of confidence.

How much in-domain data is needed?

In principle, the more in-domain data the better, but these may be difficult and expensive to collect. Thus, a very interesting issue in the context of our work is how much in-domain data is needed in order to improve results attained using out-of-domain data alone. To answer this question we focus on the 'EU+WNG' strategy and analyze the impact on performance of specialized models extracted from an incrementally larger number of examples. We compute three variants separately, by considering the use of the in-domain data: (i) only for the translation model (TM), (ii) only for the language model (LM), and (iii) simultaneously in both models (TM+LM). In order to avoid the possible effect of over-fitting we focus on the behavior of the test set. Note that the optimization of parameters is performed at each point in the x-axis using only the development set.

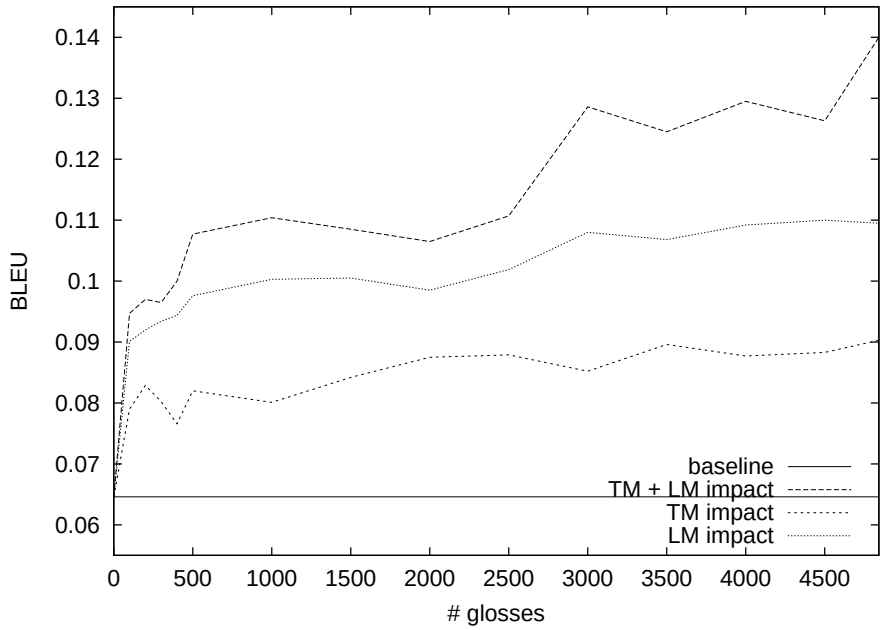


Figure 7.1: Translation of WordNet glosses. Impact of the amount of in-domain data

Results are presented in Figure 7.1. A significant initial gain of around 0.3 BLEU points is observed when adding as few as 100 glosses. In all cases, it is not until around 1,000 glosses are added that the 'EU+WNG' system stabilizes. After that, results continue improving as more in-domain data are added. We observe a very significant increase by just adding around 3,000 glosses. Another interesting observation is the boosting effect of the combination of TM and LM specialized models. While individual curves for TM and LM tend to be more stable with more than 4,000 added examples, the TM+LM curve still shows a steep increase in this last part.

7.2.3 Error Analysis

We inspected results at the sentence level for the best configuration of the 'EU+WNG' system, based on the GTM F-measure ($e = 1$). The first observation is that 196 sentences out from the 500 obtain an F-measure equal to or higher than 0.5 on the development set (181 sentences in the test set), whereas only 54 sentences obtain a score lower than 0.1. These numbers give a first idea of the relative usefulness of our system.

Table 7.10 shows some translation cases selected for discussion. We compare translations by the 'EU', 'WNG' and 'EU+WNG' systems. 'Source' and reference correspond respectively to the input gloss, and the human reference translation (i.e., expected output). Automatic translations are accompanied by the scores attained.

Case 1 is a clear example of unfair low score. The problem is that source and reference are not parallel but 'quasi-parallel'. Both glosses define the same concept but in a different way. Thus, metrics based on rewarding lexical similarities are not well suited for these cases. Cases 2, 3, 4 and 5 are examples of proper cooperation between 'EU' and 'WNG' models. 'EU' models provides recall, by suggesting translation candidates for 'bombs' or 'price below', while 'WNG' models provide precision, by choosing the right translation for 'an attack' or 'the act of'.

We also compared the 'EU+WNG' system to SYSTRAN. In the case of SYSTRAN, 167 sentences obtain a score equal to or higher than 0.5 whereas 79 sentences obtain a score lower than 0.1. These numbers are slightly under the performance of the 'EU+WNG' system. Table 7.11 shows some translation cases selected for discussion. Case 1 is again an example of both systems obtaining very low scores because of 'quasi-parallelism'. Cases 2 and 3 are examples of SYSTRAN outperforming our system. In case 2, SYSTRAN exhibits higher precision in the translation of 'accompanying' and 'illustration', whereas in case 3 it shows higher recall by suggesting appropriate translation candidates for 'fibers', 'silkworm', 'cocoon', 'threads', and 'knitting'. Cases 4, 5 and 6 are examples in which our system outperforms SYSTRAN. In case 4, our system provides higher recall by suggesting an adequate translation for 'top of something'. In case 5, our system shows higher precision by selecting a better translation for 'rate'. In case 6, our system generates a better gloss introduction, by translating 'relating to or characteristic' as 'relativo o perteneciente'. However, we observed that SYSTRAN tends in most cases to construct sentences exhibiting a higher degree of grammaticality.

Synset			Content
1	**Source**	→	of the younger of two boys with the same family name
	Reference	→	*que tiene menos edad*
	System	**GTM**	**Output**
	EU	0.0000	de acuerdo con el más joven de dos boys con la misma familia fama
	WNG	0.1333	de la younger de dos boys tiene el mismo nombre familia
	EU+WNG	0.1111	de acuerdo con el más joven de dos muchachos tiene el mismo nombre familia
2	**Source**	→	an attack by dropping bombs
	Reference	→	ataque con bombas
	System	**GTM**	**Output**
	EU	0.2857	atacar por cayendo **bombas**
	WNG	0.2500	**ataque** realizado por dropping bombs
	EU+WNG	0.5000	**ataque** realizado por cayendo **bombas**
3	**Source**	→	the temporal property of two things happening at the same time
	Reference	→	acción de pasar a la vez
	System	**GTM**	**Output**
	EU	0.1250	el temporal propiedad de dos cosas ocurren en el tiempo
	WNG	0.2222	temporal la propiedad de dos cosas sucede at la misma el tiempo
	EU+WNG	0.5882	temporal la propiedad de dos cosas , **pasar a la vez**
4	**Source**	→	the act of informing by verbal report
	Reference	→	acción y efecto de informar con una explicación verbal
	System	**GTM**	**Output**
	EU	0.1250	acto de la información por verbales ponencia
	WNG	0.7059	**acción y efecto** de informing por verbal **explicación**
	EU+WNG	0.5882	**acción y efecto** de informaba por verbales **explicación**
5	**Source**	→	a price below the standard price
	Reference	→	precio que está por debajo de lo normal
	System	**GTM**	**Output**
	EU	0.5000	un **precio por debajo de la** norma precio
	WNG	0.0000	una price below número estándar price
	EU+WNG	0.5000	un **precio por debajo de la** estándar precio

Table 7.10: Translation of WordNet glosses. Error analysis #3 (Combined knowledge sources)

Synset			Content
1	**Source**	→	a newspaper that is published every day
	Reference	→	*publicación periódica monotemática*
	System	**GTM**	**Output**
	EU+WNG	0.0000	periódico que se publica diario
	SYSTRAN	0.0000	un periódico que se publica cada día
2	**Source**	→	brief description accompanying an illustration
	Reference	→	pequeña descripción que acompaña una ilustración
	System	**GTM**	**Output**
	EU+WNG	0.1818	breve descripción adjuntas un aclaración
	SYSTRAN	0.8333	breve descripción **que acompaña una ilustración**
3	**Source**	→	fibers from silkworm cocoons provide threads for knitting
	Reference	→	fibras de los capullos de gusano de seda que proporcionan hilos para tejer
	System	**GTM**	**Output**
	EU+WNG	0.1905	fibers desde silkworm cocoons proporcionan threads para knitting
	SYSTRAN	0.7333	las **fibras** de los **capullos** del **gusano de seda** proporcionan los **hilos** de rosca para **hacer punto**
4	**Source**	→	the top of something
	Reference	→	parte superior de una cosa
	System	**GTM**	**Output**
	EU+WNG	1.0000	parte superior de una cosa
	SYSTRAN	0.0000	la tapa algo
5	**Source**	→	a rate at which something happens
	Reference	→	ritmo al que sucede una cosa
	System	**GTM**	**Output**
	EU+WNG	0.6667	un **ritmo** al que sucede algo
	SYSTRAN	0.3077	una tarifa en la cual algo sucede
6	**Source**	→	of or relating to or characteristic of peru or its people
	Reference	→	relativo o perteneciente a perú
	System	**GTM**	**Output**
	EU+WNG	0.7692	**relativo o perteneciente** a perú , su gente
	SYSTRAN	0.1053	de o en lo que concierne o característica de Perú o de su gente

Table 7.11: Translation of WordNet glosses. Comparison with SYSTRAN

Data Set	GTM ($e = 1$)	GTM ($e = 2$)	NIST	BLEU
dev	0.3091	0.2196	3.0953	0.0730
dev⋆	0.3428	0.2456	3.5655	0.0949
test	0.3028	0.2155	3.0274	0.0657
EU-test	0.5885	0.3567	7.2477	0.2725

Table 7.12: Translation of WordNet glosses. Baseline performance

7.3 Domain Independent Translation Models

As an additional question, we study the possibility of exploiting the information contained in the MCR for the purpose of MT. Other authors have previously applied information extracted from aligned wordnets. Tufis et al. (2004b) presented a method for Word Sense Disambiguation (WSD) based on parallel corpora. They utilized the aligned wordnets in BalkaNet (Tufis et al., 2004a). In our case, we suggest using the MCR to build a domain-independent word-based translation model.

7.3.1 Baseline Performance

We define a new baseline system entirely based on out-of-domain data. In that manner, we aim at favoring the occurrence of unseen events in the test data. Like in the previous sections, our baseline system follows the architecture described in Chapter 4, Section 5.1. Again, it is based on the Europarl corpus (Koehn, 2003a). However, in order to speed up the construction process, we selected a subset of 327,368 parallel segments, of length from five to twenty, for training. The Spanish side contained 4,243,610 tokens, whereas the English side consisted of 4,197,836 tokens.

For in-domain test, we counted on a preliminary version of the set of parallel glosses described in Section 7.1.1, containing 6,503 gloss pairs extracted from an earlier version of the MCR. These corresponded to 5,684 nouns, 87 verbs, and 732 adjectives. Examples and parenthesized texts were also removed. Gloss average length is 8,03 words for English and 7,83 for Spanish. Parallel glosses were also tokenized and case lowered, and randomly split into development (3,295 gloss pairs) and test (3,208 gloss pairs).

Automatic evaluation results are shown in Table 7.12. The performance of the system on the new domain is very low in comparison to the performance on a set of a held-out portion of 8490 sentences from the Europarl corpus ('EU-test'). The 'dev⋆' row refers to the results over the development set by an enhanced version of the baseline system combining the in-domain language model with the two close-to-domain language models extracted form the 'D1' and 'D2' Spanish dictionaries, as described in Section 7.2.1. Consistently with previous results, a significant improvement is obtained.

7.3.2 Exploiting the MCR

Outer knowledge may be supplied to the Pharaoh decoder by annotating the input with alternative translation options via XML-markup. In the default setting we enrich all nouns, verbs, and adjectives by looking up all possible translations for all their meanings according to the MCR. For the

3,295 glosses in the development set, a total of 13,335 words, corresponding to 8,089 nouns, 2,667 verbs and 2,579 adjectives respectively, were enriched. We have not worked on adverbs yet because of some problems with our lemmatizer. While in WordNet the lemma for adverbs is an adjective our lemmatizer returns an adverb.

Translation pairs are heuristically scored according to the number of senses which may lexicalize in the same manner. For every unknown word we create a list of candidate translations by looking up, for every sense associated to its lemma in the source language, every possible lexicalization of the corresponding senses, if any, in the target language, as provided by the MCR. Candidate translations are then scored by relative frequency. Formally, let w_f, p_f be the source word and PoS, and w_e be the target word, we define a function $Scount(w_f, p_f, w_e)$ which counts the number of senses for (w_f, p_f) which can lexicalize as w_e. Then, translation probabilities are computed according to following formula:

$$P(w_f, p_f|w_e) = \frac{Scount(w_f, p_f, w_e)}{\sum_{(w_f,p_f)} Scount(w_f, p_f, w_e)} \tag{7.1}$$

For instance, the English word *'bank'* as a noun is assigned nine different senses in WordNet. Four of these senses may lexicalize as the Spanish word *'banco'* (financial institution) whereas only one sense lexicalizes as *'orilla'* (the bank of a river). The scoring heuristic would assign these pairs a respective score of $\frac{4}{9}$ and $\frac{1}{9}$.

Let us note that in WordNet all word forms related to the same concept are grouped and represented by their lemma and part-of-speech (PoS). Therefore, input word forms must be lemmatized and PoS-tagged. WordNet takes care of the lemmatization step. For PoS-tagging we used the SVMTool package (Giménez & Màrquez, 2004b) (see SectionB.1 in Appendix B). Similarly, at the output, the MCR provides us with lemmas instead of word forms as translation candidates. A lemma extension must be performed. We utilized components from the Freeling package (Carreras et al., 2004) for this step. See, in Table 7.13, an example of XML input in which six glosses have been enriched. Source tokens appear highlighted.

Experimental Results

Several strategies have been tried. In all cases we allowed the decoder to bypass the MCR-based model when a better (i.e., likelier) solution could be found using the phrase-based model alone. Results are presented in Table 7.14.

We defined as new baseline the system which combines the three language models ('no-MCR'). In a first attempt, we enriched all content words in the validation set with all possible translation candidates ('ALL'). No improvement was achieved. By inspecting input data, apart from PoS-tagging errors, we found that the number of translation options generated via MCR was growing too fast for words with too many senses, particularly verbs. In order to reduce the degree of polysemy we tried limiting to words with 1, 2, 3, 4 and 5 different senses at most ('S1', 'S2', 'S3', 'S4' and 'S5'). Results improved slightly.

Ideally, one would wish to work with accurately word sense disambiguated input. We tried restricting translation candidates to those generated by the most frequent sense only ('ALL_{MFS}'). There was no significant variation in results. We also studied the behavior of the model applied

<NN english="consecuciones|consecución|logro|logros|
realizaciones|realización" prob="0.1666|0.1666|0.1666|
0.1666|0.1666|0.1666"> ***accomplishment*** </NN> **of an objective**

an organism such as an<NN english="insecto|insectos"
prob="0.5|0.5">***insect***</NN>**that habitually shares the**
<NN english="madriguera|madrigueras|nido|nidos"
prob="0.25|0.25|0.25|0.25"> ***nest*** </NN> **of a species of**
<NN english="hormiga|hormigas" prob="0.5|0.5"> **ant** </NN>

the part of the human<NN english="pierna|piernas"
prob="0.5|0.5"> **leg** </NN> **between the** <NN english=
"rodilla|rodillas" prob="0.5|0.5"> **knee** </NN> **and the**
<NN english="tobillo|tobillos" prob="0.5|0.5"> **ankle** </NN>

a<JJ english="casada|casadas|casado|casados"
prob="0.25|0.25|0.25|0.25"> **married** </JJ>**man**

an<NN english="abstracciones|abstracción|extracciones|
extracción|generalizaciones|generalización|pintura abstracta"
prob="0.3333|0.3333|0.0666|0.0666|0.0666|0.0666|0.0666">
abstraction </NN> **belonging to or** <JJ english="característica|
características|característico|característicos|típica|
típicas|típico|típicos" prob="0.125|0.125|0.125|0.125|0.125|
0.125|0.125|0.125">characteristic</JJ> **of two** <NNS english=
"entidad|entidades" prob="0.5|0.5"> **entities** </NNS> **or**
<NNS english="partes" prob="1"> **parts** </NNS> **together**

strengthening the concentration by removing<JJ english="irrelevante|
irrelevantes" prob="0.5|0.5">**extraneous**</JJ>**material**

Table 7.13: Domain-independent translation modeling. A sample input

Strategy	GTM ($e = 1$)	GTM ($e = 2$)	NIST	BLEU
no-MCR	0.3428	0.2456	3.5655	0.0949
ALL	0.3382	0.2439	3.4980	0.0949
ALL$_{MFS}$	0.3367	0.2434	3.4720	0.0951
S1	0.3432	0.2469	3.5774	0.0961
S2	0.3424	0.2464	3.5686	0.0963
S3	0.3414	0.2459	3.5512	0.0963
S4	0.3412	0.2458	3.5441	0.0966
S5	0.3403	0.2451	3.5286	0.0962
N$_{MFS}$	0.3361	0.2428	3.4588	0.0944
V$_{MFS}$	0.3428	0.2456	3.5649	0.0945
A$_{MFS}$	0.3433	0.2462	3.5776	0.0959
UNK$_{MFS}$	0.3538	0.2535	3.7580	**0.1035**
UNK-and-S1	0.3463	0.2484	3.6313	0.0977
UNK-or-S1	0.3507	0.2523	3.7104	**0.1026**

Table 7.14: Domain-independent translation modeling. Results on the development set

separately to nouns ('N$_{MFS}$'), verbs ('V$_{MFS}$'), and adjectives ('A$_{MFS}$'). The system worked worst for nouns, and seemed to work a little better for adjectives than for verbs.

All in all, we did not find an adequate manner to have the two translation models to cooperate properly. Therefore, we decided to use the MCR-based model only for those words unknown to the phrase-based model ('UNK$_{MFS}$'). 7.87% of the words in the development set are unknown. A significant relative improvement of 9% in BLEU score was achieved. We also tried translating only those words that were both unknown and monosemous ('UNK-and-S1'), and those that were either unknown or monosemous ('UNK-or-S1'). Results did not further improve.

System Tuning

We also performed an exhaustive process of adjustment of system parameters for the 'no-MCR' and 'UNK$_{MFS}$' strategies on the development set. As in the previous section, parameter tuning was carried out by maximizing the BLEU score over the development set. Evaluation results for the test set, in Table 7.15, report an overall 64% relative improvement with respect to the baseline.

Strategy	GTM ($e = 1$)	GTM ($e = 2$)	NIST	BLEU
baseline	0.3028	0.2155	3.0274	0.0657
noMCR	0.3431	0.2450	3.4628	0.0965
UNK$_{MFS}$	0.3554	0.2546	3.7079	0.1075

Table 7.15: Domain-independent translation modeling. Results on the test set

7.3.3 Error Analysis

Tables 7.16 and 7.17 show several cases for the 'UNK$_{MFS}$' and 'ALL$_{MFS}$' strategies, respectively. Automatic translations are accompanied by the scores attained. MCR-based models prove their usefulness (e.g., all cases in Table 7.16 and cases 1 and 2 in Table 7.17), but sometimes they may also cause the system to make a mistake (e.g., cases 3, 4, and 5 in Table 7.17).

	Synset		Content
1	00025788#n	**Source**	accomplishment of an objective
		Reference	consecución de un objetivo
	F_1=0.75	$\mathbf{T}_{baseline}$	accomplishment de un objetivo
	F_1=1	$\mathbf{T}_{MCR}$	*consecución* de un objetivo
2	00393890#n	**Source**	the position of secretary
		Reference	posición de secretario
	F_1=0.3333	$\mathbf{T}_{baseline}$	situación de secretary
	F_1=0.5714	$\mathbf{T}_{MCR}$	el cargo de *secretario*
3	00579072#n	**Source**	the activity of making portraits
		Reference	actividad de hacer retratos
	F_1=0.75	$\mathbf{T}_{baseline}$	actividad de hacer portraits
	F_1=1	$\mathbf{T}_{MCR}$	actividad de hacer *retratos*
4	00913742#n	**Source**	an organism such as an insect that habitually shares the nest of a species of ant
		Reference	organismo que comparte el nido de una especie de hormigas
		$\mathbf{T}_{baseline}$	un organismo como un insect que habitually comparte el nest de una especie de ant
	F_1=0.3752	$\mathbf{T}_{MCR}$	un organismo como un insecto que habitually comparte el *nido*
	F_1=0.5713		de una especie de *hormiga*
5	04309478#n	**Source**	the part of the human leg between the knee and the ankle
		Reference	parte de la pierna humana comprendida entre la rodilla y el tobillo
	F_1=0.3254	$\mathbf{T}_{baseline}$	parte de la persona leg entre los knee y el ankle
	F_1=0.5898	$\mathbf{T}_{MCR}$	parte de la persona *pierna* entre la *rodilla* y el *tobillo*

Table 7.16: Translation of WordNet glosses. Error analysis #4 (domain-independent translation probabilities, 'UNK$_{MFS}$' strategy)

	Synset		Content
1	00029961#n	**Source**	the act of withdrawing
		Reference	acción de retirarse
	F_1=0.2857	$\mathbf{T}_{baseline}$	el acto de retirar
	F_1=0.5714	$\mathbf{T}_{MCR}$	el acto de *retirarse*
2	00790504#n	**Source**	a favorable judgment
		Reference	opinión favorable
	F_1=0.4	$\mathbf{T}_{baseline}$	una sentencia favorable
	F_1=0.8	$\mathbf{T}_{MCR}$	una *opinión* favorable
3	04395081#n	**Source**	source of difficulty
		Reference	fuente de dificultad
	F_1=1	$\mathbf{T}_{baseline}$	fuente de dificultad
	F_1=0.6667	$\mathbf{T}_{MCR}$	fuente de *problemas*
4	04634158#n	**Source**	the branch of biology that studies plants
		Reference	rama de la biología que estudia las plantas
	F_1=0.6799	$\mathbf{T}_{baseline}$	rama de la biología que estudios plantas
	F_1=0.8	$\mathbf{T}_{MCR}$	rama de la biología que estudia *factoría*
5	10015334#n	**Source**	balance among the parts of something
		Reference	equilibrio entre las partes de algo
	F_1=1	$\mathbf{T}_{baseline}$	equilibrio entre las partes de algo
	F_1=0.8334	$\mathbf{T}_{MCR}$	equilibrio entre las partes de *entidades*

Table 7.17: Translation of WordNet glosses. Error analysis #5 (domain-independent translation probabilities, 'ALL$_{MFS}$' strategy)

7.3.4 Discussion

Overall, by working with specialized language models and MCR-based translation models we achieved a relative gain of 63.62% in BLEU score (0.0657 vs 0.1075) when porting the system to a new domain. However, there is a limitation in our approach. When we markup the input to Pharaoh we do not allow MCR-based predictions to interact with phrase-based predictions. We are somehow forcing the decoder to choose between a word-to-word translation and a phrase-to-phrase translation. Better ways to integrate MCR-based models during decoding are required. A possible solution would be to apply the method described for the integration of discriminative phrase translation probabilities described in Section 6.3.

On the other hand, more sophisticated heuristics should be considered for selecting and scoring MCR-based translation candidates. The WordNet topology could be exploited so as to build better domain-independent translation models. Relations such as hypernymy/hyponymy, holonymy/meronymy, or information such as the associated WordNet domain, or the conceptual distance between synsets, could be useful for the purpose of discriminative phrase selection.

7.4 Conclusions of this Chapter

We have studied the problem of domain-dependence in the context of SMT systems through a practical case study on the porting of an English-to-Spanish phrase-based SMT system from the domain of parliament proceedings to the domain of dictionary definitions.

The first observation is that an SMT system trained on out-of-domain data fails to properly translate in-domain data. This is mainly due to the large language variations between both domains (vocabulary, style, grammar, etc.). We have suggested several simple techniques in order to improve the performance of SMT systems when ported to new domains. Specifically, we have exploited the possibility of combining: (i) in-domain corpora, (ii) close-to-domain corpora, and (iii) domain-independent knowledge sources.

We have shown that it is possible to adapt an existing SMT system to a very different domain using only a very small amount of in-domain or close-to-domain data. We have built specialized language and translation models, and close-to-domain language models. These proposals together with a good tuning of the system parameters have lead to a notable improvement of results, according to several standard automatic MT evaluation metrics. This boost in performance is statistically significant according to the bootstrap resampling test described by Koehn (2004b) applied over the BLEU metric. The main reason behind this improvement is that the large out-of-domain corpus provides recall, while the in-domain corpus provides precision. We have presented a qualitative error analysis supporting these claims. We have also addressed the important question of how much in-domain data is needed so as to adapt an out-of-domain system. Our results show that a significant improvement may be obtained using only a minimal amount of in-domain data.

As a complementary issue, we have exploited WordNet topology to build domain-independent translation models directly extracted from the aligned wordnets in the MCR. We present a rigorous study grouping words according to several criteria (part-of-speech, ambiguity, etc.). All in all, we did not find an adequate manner to have the domain-independent translation models to properly cooperate with other translation models. Better integration techniques should be studied.

Finally, apart from experimental findings, this study has generated, as an end product, a valuable resource. We have completed the Spanish WordNet by adding automatically generated glosses for all synsets lacking of gloss. Although far from being perfect, this material has served as an excellent starting point for the process of manual revision and post-editing, which is currently ongoing.

Moreover, all the methods used are language independent, assumed the availability of the required in-domain or close-to-domain additional resources. Thus, other wordnets and similar resources could be enriched using the presented techniques.

Chapter 8

Conclusions

This chapter discusses the main contributions of this thesis as well as future research work. In Section 8.1, we present a summary of the main results and the conclusions that can be derived. Section 8.2 is a brief note on two software packages developed along this thesis, which have been made freely and publicly available to the NLP community for research purposes. Finally, Section 8.3 outlines future work and research directions.

8.1 Summary

We have exploited current NLP technology for empirical MT and its evaluation. Problems addressed in this thesis fall into two main research lines: (i) automatic MT evaluation, and (ii) development of an Empirical MT system.

8.1.1 MT Evaluation

Our main contribution in this part is the proposal of a novel direction towards heterogeneous automatic MT evaluation. In first place, we have compiled a rich set of automatic measures devoted to capture MT quality aspects at different linguistic levels (e.g., lexical, syntactic, and semantic). We have shown that metrics based on deeper linguistic information (syntactic/semantic) are able to produce more reliable system rankings than metrics which limit their scope to the lexical dimension, specially when the systems under evaluation are of a different nature. We have also presented two simple strategies for metric combination. Our approach offers the important advantage of not having to adjust the relative contribution of each metric to the overall score.

Linguistic metrics present, however, a major shortcoming. They rely on automatic linguistic processors which may be prone to error. At the document/system level, experimental results have shown that these metrics are very robust against parsing errors. The reason is that these are metrics of high precision. When they capture a similarity they are highly confident. However, at the sentence level, results indicate that these metrics are, in general, not as reliable overall quality estimators as lexical metrics, at least when applied to low quality translations. The problem is related to the lack of recall due to parsing errors or to the absence of parsing. In the latter case, we have shown that backing off to lexical similarity is an effective strategy so as to improve their performance.

We strongly believe that future MT evaluation campaigns should benefit from the results presented in this thesis, for instance, by including metrics at different linguistic levels, and metric combinations. The following set could be used:

{ *'ROUGE$_W$', 'METEOR$_{wnsyn}$', 'DP-HWC$_r$-4', 'DP-O$_c$-⋆', 'DP-O$_l$-⋆', 'DP-O$_r$-⋆', 'CP-STM-9', 'SR-O$_r$-⋆', 'SR-O$_{rv}$', 'DR-O$_{rp}$-⋆'* }

This set includes several metric representatives from different linguistic levels, which have been observed to be consistently among the top-scoring over a wide variety of evaluation scenarios.

Besides, as we have discussed, currently there is a growing interest in metric combination schemes. Thus, other researchers could exploit the rich set of metrics presented in this work to feed their combination method with linguistic features.

As an additional result, we have shown how to perform heterogeneous processes of error analysis using linguistic metrics on the basis of human likeness. Our proposal allows developers to rapidly obtain detailed automatic linguistic reports on their system's capabilities.

8.1.2 Empirical MT

The second part of the thesis is devoted to the incorporation of linguistic knowledge in the development of an empirical MT system. We have built a state-of-the-art phrase-based SMT system and completed several steps of its development cycle assisted by our evaluation methodology for heterogeneous automatic MT evaluation.

First, we have used linguistic processors to build shallow-syntactic word and phrase alignments. We have shown that data sparsity is a major cause for the lack of success in the incorporation of linguistic knowledge to translation modeling in SMT. Individual translation models based on enriched data views underperform the baseline system, mainly due to a severe decrease in recall. However, combined models yield a significantly improved translation quality. We have presented and discussed the pros and cons of two different combination schemes. Besides, the report on heterogeneous evaluation shows that improvements take place at other quality dimensions beyond the lexical level.

Our main contribution is, however, our approach to dedicated discriminative lexical selection (Giménez & Màrquez, 2008a). Despite the fact that measuring improvements in lexical selection is a very delicate issue, experimental results, according to several well-known metrics based on lexical similarity, show that dedicated DPT models yield a significantly improved lexical choice over traditional MLE-based ones. However, by evaluating linguistic aspects of quality beyond the lexical level (e.g., syntactic, and semantic), we have found that an improved lexical choice does not necessarily lead to an improved syntactic or semantic structure. This result has been verified through a number of manual evaluations, which have revealed that gains are mainly related to the adequacy dimension, whereas for fluency there is no significant improvement.

Besides, this work has also served us to study the role of automatic metrics in the development cycle of MT systems, and the importance of meta-evaluation. We have shown that basing evaluations and parameter optimizations on different metrics may lead to very different system behaviors.

For system comparison, this may be solved through manual evaluations. However, this is impractical for the adjustment of parameters, where hundreds of different configurations are tried. Thus, we argue that more attention should be paid to the meta-evaluation process. In our case, metrics have been evaluated on the basis of human likeness. Other solutions exist. The main point, in our opinion, is that system development is metricwise. That is, the metric (or set of metrics) guiding the development process must be able to capture the possible quality variations induced by system modifications. This is a crucial issue, since, most often, system improvements focus on partial aspects of quality, such as word selection or word ordering, which can not be always expected to improve together.

As a side question, we have studied the problem of domain-dependence in the context of SMT systems through a practical case study. The first observation is that, as expected, an SMT system trained on out-of-domain data fails to properly translate in-domain data. This is mainly due to the large language variations between both domains (vocabulary, style, grammar, etc.). We have suggested several simple techniques in order to improve the performance of SMT systems when ported to new domains. Our approach exploits the possibility of combining: (i) in-domain corpora, (ii) close-to-domain corpora, and (iii) domain-independent knowledge sources. We have built specialized language and translation models from in-domain a small parallel corpus, and nearly specialized language models from medium-size monolingual corpora of a similar domain. The main reason behind the obtained improvement is that the large out-of-domain corpus provides recall, while in-domain and close-to-domain corpora provide precision. A qualitative error analysis supporting these claims has been presented. In addition, we have also addressed the important question of how much in-domain data is needed so as to adapt an out-of-domain system. Our results show that a significant improvement may be obtained using only a minimal amount of in-domain data. As a complementary issue, we have also studied the possibility of exploiting WordNet topology to build domain-independent translation models directly extracted from the aligned wordnets in the MCR. A rigorous study grouping words according to several criteria (part-of-speech, ambiguity, etc.) has been presented. However, we did not find an adequate manner to have the domain-independent translation models to properly cooperate with other translation models. Better integration techniques should be applied. Finally, our study has served to enrich the MCR by providing an automatically generated gloss for all synsets in the Spanish WordNet. This material is currently under manual revision.

8.2 Software

This thesis has contributed as well with the development of two software packages which are publicly available for research purposes released under the GNU Lesser General Public License[1] (LGPL) of the Free Software Foundation[2]. These packages are:

IQ$_{MT}$: The IQ$_{MT}$ Framework for MT Evaluation is the adaptation of the QARLA Framework (Amigó et al., 2005), originally designed for the evaluation of the Automatic Summarization task, to

[1] http://www.fsf.org/licenses/lgpl.html
[2] http://www.fsf.org/

the case of MT (Giménez and Màrquez 2007b; 2006; 2005a). IQ$_{MT}$ offers a common workbench on which automatic MT evaluation metrics can be meta-evaluated, utilized and combined on the basis of human likeness. It provides (i) a measure to evaluate the quality of any set of similarity metrics (KING), (ii) a measure to evaluate the quality of a translation using a set of similarity metrics (QUEEN), and (iii) a measure to evaluate the reliability of a test bed (JACK). IQ$_{MT}$ allows also for evaluation and meta-evaluation on the basis of human acceptability. All metrics described in this work have been incorporated into the IQ$_{MT}$ package which is released under LGPL license[3].

SVMTool: The SVMTool is a simple and effective generator of sequential taggers based on Support Vector Machines (Giménez & Màrquez, 2003; Giménez & Màrquez, 2004b; Giménez & Màrquez, 2004a). We have applied the SVMTool to the problem of part-of-speech tagging. By means of a rigorous experimental evaluation, we conclude that the proposed SVM-based tagger is robust and flexible for feature modeling (including lexicalization), trains efficiently with almost no parameters to tune, and is able to tag thousands of words per second, which makes it really practical for real NLP applications. Regarding accuracy, the SVM-based tagger significantly outperforms the TnT tagger (Brants, 2000) exactly under the same conditions, and achieves a very competitive accuracy of 97.2% for English on the Wall Street Journal corpus, which is comparable to the best taggers reported up to date. It has been also successfully applied to Spanish and Catalan exhibiting a similar performance, and to other tagging problems such as chunking. Perl and C++ versions are publicly available under LGPL license[4].

8.3 Future Work

This section describes future research work.

8.3.1 Extending the Evaluation Methodology

We have in mind several extensions to the methodology for heterogeneous automatic MT evaluation deployed in Chapter 3.

Metric Improvement

The set of metrics presented in Section 3.1 covers a wide range of quality aspects. However, the fact of relying on automatic linguistic processors implies, as we have discussed, several limitations. Some are derived from their performance —linguistic processors are prone to error and often very slow. In order to improve the effectiveness and efficiency of current metrics, we plan to use newer versions of current linguistic processors as they become available as well as to study the possibility of shifting to alternative tools.

[3]The IQ$_{MT}$ software may be freely downloaded at `http://www.lsi.upc.es/~nlp/IQMT/`.

[4]The SVMTool software may be freely downloaded at `http://www.lsi.upc.es/~nlp/SVMTool/`.

Another limitation of current metrics is that they are language dependent. In the short term, we plan to adapt some of the metrics to languages other than English counting on the required linguistic processors (e.g., Arabic, Chinese). For instance, we are currently developing shallow-syntactic, syntactic and shallow-semantic metrics for Spanish and Catalan.

We also plan to incorporate new metrics. This may involve using new linguistic processors which are able to acquire new types of information, and also designing new types of similarity measures over currently available linguistic representations.

Metric Combinations

We plan to perform a thorough comparison between parametric and non-parametric metric combination schemes. The idea is to reproduce the parametric approaches by Kulesza and Shieber (2004), Albrecht and Hwa (2007a) and Liu and Gildea (2007) and to compare them to the combination strategies described in Section 3.4, over different evaluation scenarios (i.e., language-pairs, task domains, and system paradigms).

Heterogeneous Statistical Significance Tests

Statistical significance tests allow researchers to determine whether the quality attained by a system A over a fixed set of translations is higher, equal to, or lower than the quality attained by another system B over the same set of translations (Koehn, 2004b; Collins et al., 2005). Translation quality is typically measured according to an automatic metric at choice (e.g., BLEU), which causes the test to be metric-biased. A more robust alternative, in our opinion, would consist in performing heterogeneous tests that would guarantee statistical significance of the results simultaneously according to a heterogeneous set of metrics operating at different linguistic levels. For that purpose we count on the QUEEN measure. As we have seen, QUEEN is a probabilistic measure which, based on the unanimity principle, provides an estimate of the level of agreement among different metrics on the quality of automatic outputs. Thus, its application to the problem of assessing the statistical significance of translation results should be, in principle, straightforward. This hypothesis must be theoretically and empirically validated.

Adjustment of Parameters

The computational cost of some linguistic metrics turns them into impractical for the system optimization process, in which hundreds of different system configurations are tried. We plan to study the applicability of these metrics in the near future.

Automatic Error Analysis

Error analysis is a crucial stage in the development of an SMT system. In order to accelerate this process, we plan to refine the IQ_{MT} interface, currently in text format, so that it allows for a fast and elegant visual access from different viewpoints corresponding to the different dimensions of quality. For instance, missing or partially translated elements could appear highlighted in different colors. Besides, evaluation measures generate, as a by-pass product, syntactic and semantic analyses which

could be displayed. This would allow users to separately analyze the translation of different types of linguistic elements (e.g., constituents, relationships, arguments, adjuncts, discourse representation structures, etc.).

Towards a Development Cycle without Human Intervention

Meta-evaluation on the basis of human likeness eliminates the need for human assessments from the development cycle. Human labor is only required for the construction of reference translations. Moreover, as we have seen in Section 2.4, several approaches to automatic MT evaluation without using human references have been suggested (Quirk, 2004; Gamon et al., 2005; Albrecht & Hwa, 2007b). We plan to study their applicability with the intent to definitely remove all human intervention from the evaluation task. This could originate a new development cycle in which neither human assessments nor human references would be required.

8.3.2 Improving the Empirical MT System

There are several natural improvements that should be addressed.

Linguistic Knowledge

In the development of our SMT system, we have limited to using information up to the level of shallow syntax. In future experiments we plan to use information at deeper linguistic levels (e.g., based on semantic roles).

Moreover, our current implementation of WordNet-based domain-independent translation models does not fully exploit the WordNet topology. It uses the MCR merely as a multilingual dictionary, i.e., exploiting only the synonymy relationship. However, WordNet offers several other types of relationships. The possibility of incorporating features based on the WordNet topology (e.g., about domains, hyponymy/hypernymy and meronymy/holonymy relationships, and conceptual distance) should be considered. This information could be also exploited during the construction of discriminative phrase translation models.

English-to-Spanish Lexical Selection

So far, we have only exploited our dedicated shallow syntactic discriminative translation models for the case of Spanish-to-English translation. However, Spanish is known to have a richer morphology than English. Thus, the room for improvement for our models should be larger when applied in the reverse direction, i.e., English-to-Spanish. This hypothesis must be verified.

Domain Adaptation

One of the strengths of our approach to lexical selection is that it is able to model the source context, and, thus, mitigate the effects of a biased lexical selection. This property makes it specially suitable for being applied to the problem of domain adaptation. A comparative study on the performance of DPT models in a restricted domain vs. an open domain should be conducted.

8.3.3 Towards a New System Architecture

As we have seen in Chapter 6, the integration of context-aware discriminative translation probabilities into the SMT framework is problematic. First, the type of features under consideration has a direct influence in the complexity of the decoder. For instance, we have not been able to incorporate additional features from the target side (sentence under construction) and from the correspondences between source and target sides (i.e., alignments). Second, in spite of achieving a higher classification accuracy, discriminative predictions based on local training may be not necessarily well suited for being integrated in the target translation. We argue that if phrase translation classifiers were trained in the context of the global task their integration would be more robust and translation quality could further improve. Third, the cooperation between discriminative models and other models (e.g., language model and additional translation models) in the standard log-linear architecture is poorly modeled. The relative importance of the features represented by each model is determined through a simple process of global parameter adjustment, when, indeed, feature importance may vary at more local levels (e.g., sentence, constituent, semantic role, etc.). Undoubtly, the possibility of moving towards a new global empirical MT architecture in the fashion, for instance, of those suggested by Tillmann and Zhang (2006) or Liang et al. (2006) should be studied.

8.3.4 Other Directions

Different Languages

In this thesis we have focused on the translation between English and Spanish, two Indo-European languages which present a similar word order. In the future, we plan to move to new language pairs. Apart from English and Spanish, we are interested as well in Catalan, Basque, French, German, Chinese and Arabic. For instance, we have presented our DPT system to the Arabic-to-English Exercise of the 2008 NIST MT Evaluation Campaign[5] (España-Bonet, 2008). Results corroborate the findings in Chapter 6.

Entity Translation

We participated in the the Automatic Content Extraction (ACE) Entity Translation 2007 Evaluation Campaign[6], with an Arabic-to-English entity translation system (Farwell et al., 2007). Our approach to entity translation was fairly simple. We divided the task into three separate subtasks: Named Entity Recognition and Classification, Coreference Resolution, and Machine Translation, which were approached independently. In the next editions, we should study the possibility of performing a joint training of the three subsystems. This would allow the entity translation system to eliminate the noise which appears due to the interaction between modules. Ideally, in order to train such a system a parallel corpus with annotated named entities, coreferences, and correspondences between them would be required. However, bootstrapping techniques could be also applicable.

Moreover, we could use the DPT models described in Chapter 6 so as to better exploit the sentence context in which entities occur.

[5] http://www.nist.gov/speech/tests/mt/2008/

[6] http://www.nist.gov/speech/tests/ace/ace07/et/

Hybrid MT

Recently, several approaches to hybridization of MT systems have been suggested (Simard et al., 2007; Alegría et al., 2008). We plan to study the possibility of building hybrid approaches (e.g., statistical and rule-based). For instance, we have successfully reproduced the experiments by Simard et al. (2007) for the case of English-to-Spanish translation by using our SMT system to post-edit the output of the *Translendium* rule-based MT system[7]. In the short term, we plan to use DPT models to improve the lexical selection of a Basque-Spanish rule-based MT system (Alegria et al., 2005).

[7] http://www.translendium.com/

Bibliography

Agirre, E., Màrquez, L., & Wicentowski, R. (Eds.). (2007). *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Prague, Czech Republic: Association for Computational Linguistics.

Akiba, Y., Imamura, K., & Sumita, E. (2001). Using Multiple Edit Distances to Automatically Rank Machine Translation Output. *Proceedings of Machine Translation Summit VIII* (pp. 15–20).

Albrecht, J., & Hwa, R. (2007a). A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 880–887).

Albrecht, J., & Hwa, R. (2007b). Regression for Sentence-Level MT Evaluation with Pseudo References. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 296–303).

Alegria, I., de Ilarraza, A. D., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Forcada, M. L., Ortiz-Rojas, S., & Padró, L. (2005). An open architecture for transfer-based machine translation between Spanish and Basque. *Proceedings of OSMaTran: Open-Source Machine Translation, A workshop at Machine Translation Summit X (Phuket, Thailand, September 12–16)*.

Alegría, I., Màrquez, L., & Sarasola, K. (Eds.). (2008). *Mixing Approaches To Machine Translation (MATMT)*. University of the Basque Country.

ALPAC (1966). *Languages and Machines: Computers in Translation and Linguistics* (Technical Report). Automatic Language Processing Advisory Committee (ALPAC), Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

Alshawi, H. (1996). Head automata and bilingual tiling: Translation with minimal representations (invited talk). *Proceedings of the Joint 16th International Conference on Computational Linguistics and the 34th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 167–176).

Alshawi, H., Bangalore, S., & Douglas, S. (1998). Automatic Acquisition of Hierarchical Transduction Models for Machine Translation. *Proceedings of the Joint 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 41–47).

Alshawi, H., Bangalore, S., & Douglas, S. (2000). Learning Dependency Translation as Collection of Finite State Head Transducers. *Computational Linguistics, 26*(1), 45–60.

Amigó, E., Giménez, J., Gonzalo, J., & Màrquez, L. (2006). MT Evaluation: Human-Like vs. Human Acceptable. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 17–24).

Amigó, E., Gonzalo, J., nas, A. P., & Verdejo, F. (2005). QARLA: a Framework for the Evaluation of Automatic Summarization. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 280–289).

Arnold, D., Balkan, L., Meijer, S., Humphreys, R., & Sadler, L. (Eds.). (1994). *Machine Translation: an Introductory Guide*. Blackwells-NCC, London. ISBN 1855542-17x.

Atserias, J., Villarejo, L., Rigau, G., Agirre, E., Carroll, J., Magnini, B., & Vossen, P. (2004). The MEANING Multilingual Central Repository. *Proceedings of the 2nd Global WordNet Conference (GWC)*. Brno, Czech Republic. ISBN 80-210-3302-9.

Babych, B., & Hartley, T. (2004). Extending the BLEU MT Evaluation Method with Frequency Weightings. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Banerjee, S., & Lavie, A. (2005). METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*.

Bangalore, S., Haffner, P., & Kanthak, S. (2007). Statistical Machine Translation through Global Lexical Selection and Sentence Reconstruction. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 152–159).

Baum, L. E. (1972). An Inequality and Associated Maximization Technique in Statistical Estimation for Probabilistic Functions of Markov Processes. *Inequalities, 3*, 1–8.

Berger, A. L., Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., Gillet, J. R., Lafferty, J. D., Mercer, R. L., Printz, H., & Ureš, L. (1994). The Candide System for Machine Translation. *Proceedings of the ARPA Workshop on Human Language Technology*.

Berger, A. L., Pietra, S. A. D., & Pietra, V. J. D. (1996). A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics, 22*(1), 39–72.

Bishop, C. M. (1995). 6.4: Modeling Conditional Distributions. *Neural Networks for Pattern Recognition* (p. 215). Oxford University Press.

Blatz, J., Fitzgerald, E., Foster, G., Gandrabur, S., Goutte, C., Kulesza, A., Sanchis, A., & Ueffing, N. (2003). *Confidence estimation for machine translation. Final Report of Johns Hopkins 2003 Summer Workshop on Speech and Language Engineering* (Technical Report). Johns Hopkins University.

Blunsom, P., & Cohn, T. (2006). Discriminative Word Alignment with Conditional Random Fields. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 65–72).

Bos, J. (2005). Towards Wide-Coverage Semantic Interpretation. *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS-6)* (pp. 42–53).

Bos, J., Clark, S., Steedman, M., Curran, J. R., & Hockenmaier, J. (2004). Wide-Coverage Semantic Representations from a CCG Parser. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (pp. 1240–1246).

Brants, T. (2000). TnT - A Statistical Part-of-Speech Tagger. *Proceedings of the Sixth ANLP*.

Brants, T., Popat, A. C., Xu, P., Och, F. J., & Dean, J. (2007). Large Language Models in Machine Translation. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (pp. 858–867).

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Lafferty, J. D., Mercer, R. L., & Roossin, P. S. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics, 16*(2), 76–85.

Brown, P. F., Cocke, J., Pietra, S. A. D., Pietra, V. J. D., Jelinek, F., Mercer, R. L., & Roossin, P. S. (1988). A Statistical Approach to Language Translation. *Proceedings of the 12th International Conference on Computational Linguistics (COLING)*.

Brown, P. F., Pietra, S. A. D., Mercer, R. L., & Pietra, V. J. D. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics, 19*(2), 263–311.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1991a). A statistical approach to sense disambiguation in machine translation. *HLT '91: Proceedings of the workshop on Speech and Natural Language* (pp. 146–151). Morristown, NJ, USA: Association for Computational Linguistics.

Brown, P. F., Pietra, S. A. D., Pietra, V. J. D., & Mercer, R. L. (1991b). Word-sense Disambiguation Using Statistical Methods. *Proceedings of the 29th annual meeting on Association for Computational Linguistics* (pp. 264–270). Morristown, NJ, USA: Association for Computational Linguistics.

Cabezas, C., & Resnik, P. (2005). *Using WSD Techniques for Lexical Selection in Statistical Machine Translation (CS-TR-4736/LAMP-TR-124/UMIACS-TR-2005-42)* (Technical Report). University of Maryland, College Park. `http://lampsrv01.umiacs.umd.edu/pubs/TechReports/LAMP\_124/LAMP\_124.pdf`.

Callison-Burch, C. (2005). Linear B system description for the 2005 NIST MT evaluation exercise. *Proceedings of the NIST 2005 Machine Translation Evaluation Workshop*.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., & Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. *Proceedings of the ACL Workshop on Statistical Machine Translation* (pp. 136–158).

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*.

Carbonell, J. G., Mitamura, T., & III, E. H. N. (1992). The KANT Perspective: A Critique of Pure Transfer (and Pure Interlingua, Pure Statistics, ...). *Proceedings of the 4th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)* (pp. 225–235).

Carpuat, M., Shen, Y., Xiaofeng, Y., & Wu, D. (2006). Toward Integrating Semantic Processing in Statistical Machine Translation. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)* (pp. 37–44).

Carpuat, M., & Wu, D. (2005a). Evaluating the Word Sense Disambiguation Performance of Statistical Machine Translation. *Proceedings of the Second International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 122–127).

Carpuat, M., & Wu, D. (2005b). Word Sense Disambiguation vs. Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Carpuat, M., & Wu, D. (2007a). How Phrase Sense Disambiguation outperforms Word Sense Disambiguation for Statistical Machine Translation. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Carpuat, M., & Wu, D. (2007b). Improving Statistical Machine Translation Using Word Sense Disambiguation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 61–72).

Carreras, X., Chao, I., Padró, L., & Padró, M. (2004). FreeLing: An Open-Source Suite of Language Analyzers. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 239–242).

Carreras, X., Màrquez, L., & Castro, J. (2005). Filtering-Ranking Perceptron Learning for Partial Parsing. *Machine Learning*, *59*, 1–31.

Chan, Y. S., Ng, H. T., & Chiang, D. (2007). Word Sense Disambiguation Improves Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 33–40).

Chandioux, J., & Grimalia, A. (1996). Specialized Machine Translation. *Proceedings of the 2nd Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 206–212).

Chang, P.-C., & Toutanova, K. (2007). A Discriminative Syntactic Word Order Model for Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 9–16).

Charniak, E. (2001). Immediate-Head Parsing for Language Models. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Charniak, E., & Johnson, M. (2005). Coarse-to-fine n-best parsing and MaxEnt discriminative reranking. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Charniak, E., Knight, K., & Yamada, K. (2003). Syntax-based Language Models for Machine Translation. *Proceedings of MT SUMMIT IX*.

Chen, B., Cattoni, R., Bertoldi, N., Cettolo, M., & Federico, M. (2005). The ITC-irst SMT System for IWSLT-2005. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Chen, S. F., & Rosenfeld, R. (1999). *A Gaussian Prior for Smoothing Maximum Entropy Models* (Technical Report). Technical Report CMUCS -99-108, Carnegie Mellon University.

Chiang, D. (2005). A Hierarchical Phrase-based Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 263–270).

Chiang, D. (2007). Hierarchical Phrase-based Translation. *Computational Linguistics*, *33*(2), 201–228.

Chklovski, T., Mihalcea, R., Pedersen, T., & Purandare, A. (2004). The Senseval-3 Multilingual EnglishÂ–Hindi lexical sample task. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text* (pp. 5–8). Barcelona, Spain: Association for Computational Linguistics.

Church, K. W., & Hovy, E. H. (1993). Good Applications for Crummy Machine Translation. *Machine Translation*, *8*(4), 239–258.

Civera, J., & Juan, A. (2007). Domain Adaptation in Statistical Machine Translation with Mixture Modelling. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 177–180). Prague, Czech Republic: Association for Computational Linguistics.

Clark, S., & Curran, J. R. (2004). Parsing the WSJ using CCG and Log-Linear Models. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 104–111).

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*, 37–46.

Collins, M. (2000). Discriminative Reranking for Natural Language Parsing. *Proceedings of the 17th International Conference on Machine Learning* (pp. 175–182). Morgan Kaufmann, San Francisco, CA.

Collins, M. (2002). Discriminative Training Methods for Hidden Markov Models: Theory and Experiments with Perceptron Algorithms. *Proceedings of the 7th Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1–8).

Collins, M., Koehn, P., & Kucerova, I. (2005). Clause Restructuring for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Collins, M., & Koo, T. (2005). Discriminative Reranking for Natural Language Parsing. *Computational Linguistics*, *31*(1), 25–69.

Corston-Oliver, S., Gamon, M., & Brockett, C. (2001). A Machine Learning Approach to the Automatic Evaluation of Machine Translation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 140–147).

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, *20*(3), 273–297.

Coughlin, D. (2003). Correlating Automated and Human Assessments of Machine Translation Quality. *Proceedings of Machine Translation Summit IX* (pp. 23–27).

Cowan, B., Kucerova, I., & Collins, M. (2006). A Discriminative Model for Tree-to-Tree Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Crego, J. M., de Gispert, A., Lambert, P., Costa-jussà, M. R., Khalilov, M., Banchs, R., no, J. B. M., & Fonollosa, J. A. R. (2006). N-gram-based SMT System Enhanced with Reordering Patterns. *Proceedings of the HLT-NAACL Workshop on Statistical Machine Translation* (pp. 162–165).

Crego, J. M., de Gispert, A., & no, J. B. M. (2005a). An Ngram-based Statistical Machine Translation Decoder. *Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech)*.

Crego, J. M., de Gispert, A., & no, J. B. M. (2005b). The TALP Ngram-based SMT System for IWSLT'05. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Cristianini, N., & Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press.

Culy, C., & Riehemann, S. Z. (2003). The Limits of N-gram Translation Evaluation Metrics. *Proceedings of MT-SUMMIT IX* (pp. 1–8).

Dabbadie, M., Hartley, A., King, M., Miller, K., Hadi, W. M. E., Popescu-Belis, A., Reeder, F., & Vanni, M. (2002). A Hands-On Study of the Reliability and Coherence of Evaluation Metrics. *Handbook of the LREC 2002 Workshop "Machine Translation Evaluation: Human Evaluators Meet Automated Metrics"* (pp. 8–16).

Darlington, J. (1962). Interlingua and MT. *Mechanical Translation*, *7*(1), 2–7.

Daumé III, H. (2004). Notes on CG and LM-BFGS Optimization of Logistic Regression. Paper available at `http://pub.hal3.name#daume04cg-bfgs`, implementation available at `http://hal3.name/megam/`.

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*(1), 1–38.

Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Proceedings of the 2nd International Conference on Human Language Technology* (pp. 138–145).

Dorr, B. J. (1994). Machine Translation Divergences: A Formal Description and Proposed Solution. *Computational Linguistics, 20*(4), 597–633.

Eck, M., & Hori, C. (2005). Overview of the IWSLT 2005 Evaluation Campaign. *Proceedings of the International Workshop on Spoken Language Technology (IWSLT)* (pp. 11–32). Carnegie Mellon University.

Escudero, G., Màrquez, L., & Rigau, G. (2000). An Empirical Study of the Domain Dependence of Supervised Word Disambiguation Systems. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 172–180).

Falkedal, K. (1994). *Evaluation Methods for Machine Translation Systems: An Historical Survey and A Critical Account* (Technical Report). ISSCO: Interim Report to Suissetra.

Farwell, D., Giménez, J., González, E., Halkoum, R., Rodríguez, H., & Surdeanu, M. (2007). The UPC System for Arabic-to-English Entity Translation. *Proceedings of the Automatic Content Extraction (ACE) Evaluation Program*.

Fellbaum, C. (Ed.). (1998). *WordNet. An Electronic Lexical Database*. The MIT Press.

Fox, H. J. (2002). Phrasal Cohesion and Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 304–311).

Fraser, A., & Marcu, D. (2006). Semi-Supervised Training for Statistical Word Alignment. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 769–776).

Gale, W. A., & Church, K. W. (1993). A Program for Aligning Sentences in Bilingual Corpora. *Computational Linguistics, 19*(1), 75–102.

Galley, M., Graehl, J., Knight, K., Marcu, D., DeNeefe, S., Wang, W., & Thayer, I. (2006). Scalable Inference and Training of Context-Rich Syntactic Translation Models. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 961–968).

Galley, M., Hopkins, M., Knight, K., & Marcu, D. (2004). What's in a translation rule? *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 273–280).

Gamon, M., Aue, A., & Smets, M. (2005). Sentence-Level MT evaluation without reference translations: beyond language modeling. *Proceedings of EAMT* (pp. 103–111).

Germann, U. (2003). Greedy Decoding for Machine Translation in Almost Linear Time. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 72–79).

Germann, U., Jahr, M., Knight, K., Marcu, D., & Yamada, K. (2001). Fast decoding and optimal decoding for machine translation. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 228–235).

Gildea, D. (2003). Loosely Tree-Based Alignment for Machine Translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Gildea, D. (2004). Dependencies vs. Constituents for Tree-Based Alignment. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Giménez, J. (2007). *IQMT v 2.0. Technical Manual (LSI-07-29-R)* (Technical Report). TALP Research Center. LSI Department. `http://www.lsi.upc.edu/~nlp/IQMT/IQMT.v2.1.pdf`.

Giménez, J., & Amigó, E. (2006). IQMT: A Framework for Automatic Machine Translation Evaluation. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)* (pp. 685–690).

Giménez, J., Amigó, E., & Hori, C. (2005a). Machine Translation Evaluation Inside QARLA. *Proceedings of the International Workshop on Spoken Language Technology (IWSLT)* (pp. 199–206).

Giménez, J., & Màrquez, L. (2003). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *Proceedings of the 4th International Conference on Recent Advances in Natural Language Processing (RANLP)* (pp. 158–165).

Giménez, J., & Màrquez, L. (2004a). Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited. *Recent Advances in Natural Language Processing III* (pp. 153–162). Amsterdam: John Benjamin Publishers. ISBN 90-272-4774-9.

Giménez, J., & Màrquez, L. (2004b). SVMTool: A general POS tagger generator based on Support Vector Machines. *Proceedings of 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 43–46).

Giménez, J., & Màrquez, L. (2005). Combining Linguistic Data Views for Phrase-based SMT. *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 145–148).

Giménez, J., & Màrquez, L. (2006a). Low-cost Enrichment of Spanish WordNet with Automatically Translated Glosses: Combining General and Specialized Models. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 287–294).

Giménez, J., & Màrquez, L. (2006b). The LDV-COMBO system for SMT. *Proceedings of the NAACL Workshop on Statistical Machine Translation (WMT'06)* (pp. 166–169).

Giménez, J., & Màrquez, L. (2007a). Context-aware Discriminative Phrase Selection for Statistical Machine Translation. *Proceedings of the ACL Workshop on Statistical Machine Translation* (pp. 159–166).

Giménez, J., & Màrquez, L. (2007b). Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. *Proceedings of the ACL Workshop on Statistical Machine Translation* (pp. 256–264).

Giménez, J., & Màrquez, L. (2008a). Discriminative Phrase Selection for Statistical Machine Translation. In *Learning Machine Translation*, NIPS Workshop series. MIT Press.

Giménez, J., & Màrquez, L. (2008b). Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations. *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)* (pp. 319–326).

Giménez, J., & Màrquez, L. (2008c). On the Robustness of Linguistic Features for Automatic MT Evaluation. (Under submission).

Giménez, J., & Màrquez, L. (2008d). Towards Heterogeneous Automatic MT Error Analysis. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*.

Giménez, J., Màrquez, L., & Rigau, G. (2005b). Automatic Translation of WordNet Glosses. *Proceedings of Cross-Language Knowledge Induction Workshop, EUROLAN Summer School* (pp. 1–8).

Gode, A. (1955). The Signal System in Interlingua — A Factor in Mechanical Translation. *Mechanical Translation*, *2*(3), 55–60.

Graehl, J., & Knight, K. (2004). Training Tree Transducers. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 105–112).

Groves, D., & Way, A. (2005). Hybrid Example-Based SMT: the Best of Both Worlds? *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 183–190).

Haghighi, A., Toutanova, K., & Manning, C. (2005). A Joint Model for Semantic Role Labeling. *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL)* (pp. 173–176). Ann Arbor, Michigan: Association for Computational Linguistics.

Harper, K. E. (1957). Contextual Analysis. *Mechanical Translation*, *4*(3), 70–75.

He, S., & Gildea, D. (2006). *Self-training and Co-training for Semantic Role Labeling: Primary Report* (Technical Report). TR 891, Department of Computer Science, University of Rochester.

Hewavitharana, S., Zhao, B., Hildebrand, A., Eck, M., Hori, C., Vogel, S., & Waibel, A. (2005). The CMU Statistical Machine Translation System for IWSLT 2005. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*.

Hovy, E., Hermjakob, U., & Lin, C.-Y. (2001). The Use of External Knowledge of Factoid QA. *Proceedings of TREC*.

Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation, 17*(1), 43–75.

Hovy, E., Lin, C.-Y., Zhou, L., & Fukumoto, J. (2006). Automated Summarization Evaluation with Basic Elements. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)* (pp. 899–902).

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review, 106*(4), 620–630.

Jelinek, F., & Mercer, R. L. (1980). Interpolated Estimation of Markov Source Parameters from Sparse Data. *Proceedings of the Workshop on Pattern Recognition in Practice* (pp. 381–397).

Jin, P., Wu, Y., & Yu, S. (2007). SemEval-2007 Task 05: Multilingual Chinese-English Lexical Sample. *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 19–23). Prague, Czech Republic: Association for Computational Linguistics.

Joachims, T. (1999). Making large-Scale SVM Learning Practical. *Advances in Kernel Methods - Support Vector Learning*. The MIT Press.

Kamp, H. (1981). A Theory of Truth and Semantic Representation. *Formal Methods in the Study of Language* (pp. 277–322). Amsterdam: Mathematisch Centrum.

Kaplan, A. (1955). An Experimental Study of Ambiguity and Context. *Mechanical Translation, 2*(2), 39–46.

Kauchak, D., & Barzilay, R. (2006). Paraphrasing for Automatic Evaluation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 455–462).

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika, 30*, 81–89.

Kendall, M. (1955). *Rank Correlation Methods*. Hafner Publishing Co.

King, M., & Falkedal, K. (1990). Using Test Suites in Evaluation of MT Systems. *Proceedings of the 13th International Conference on Computational Linguistics (COLING)* (pp. 211–216).

Kirchhoff, K., & Yang, M. (2005). Improved Language Modeling for Statistical Machine Translation. *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 125–128).

Knight, K. (1999). Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, *25*(4).

Knight, K., Al-Onaizan, Y., Purdy, D., Curin, J., Jahr, M., Lafferty, J., Melamed, D., Smith, N., Och, F. J., & Yarowsky, D. (1999). *Final Report of Johns Hopkins 1999 Summer Workshop on Statistical Machine Translation* (Technical Report). Johns Hopkins University.

Knight, K., & Graehl, J. (2005). An Overview of Probabilistic Tree Transducers for Natural Language Processing. *Sixth International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)* (pp. 1–25).

Koehn, P. (2003a). *Europarl: A Multilingual Corpus for Evaluation of Machine Translation* (Technical Report). `http://people.csail.mit.edu/people/koehn/publications/europarl/`.

Koehn, P. (2003b). *Noun Phrase Translation*. Doctoral dissertation, University of Southern California.

Koehn, P. (2004a). Pharaoh: a Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. *Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA)*.

Koehn, P. (2004b). Statistical Significance Tests for Machine Translation Evaluation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 388–395).

Koehn, P., & Hoang, H. (2007). Factored Translation Models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 868–876).

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., & Herbst, E. (2007). *Moses: Open Source Toolkit for Statistical Machine Translation* (Technical Report). (ACL 2007) demonstration session.

Koehn, P., & Knight, K. (2002). ChunkMT: Statistical Machine Translation with Richer Linguistic Knowledge. Draft.

Koehn, P., & Monz, C. (2005). Shared Task: Statistical Machine Translation between European Languages. *Proceedings of the ACL Workshop on Building and Using Parallel Texts* (pp. 119–124). Ann Arbor, Michigan: Association for Computational Linguistics.

Koehn, P., & Monz, C. (2006). Manual and Automatic Evaluation of Machine Translation between European Languages. *Proceedings of the NAACL Workshop on Statistical Machine Translation* (pp. 102–121).

Koehn, P., Och, F. J., & Marcu, D. (2003). Statistical Phrase-Based Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Koehn, P., & Schroeder, J. (2007). Experiments in Domain Adaptation for Statistical Machine Translation. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 224–227). Prague, Czech Republic: Association for Computational Linguistics.

Koehn, P., Shen, W., Federico, M., Bertoldi, N., Callison-Burch, C., Cowan, B., Dyer, C., Hoang, H., Bojar, O., Zens, R., Constantin, A., Herbst, E., & Moran, C. (2006). *Open Source Toolkit for Statistical Machine Translation* (Technical Report). Johns Hopkins University Summer Workshop. `http://www.statmt.org/jhuws/`.

Koutsoudas, A., & Korfhage, R. (1956). Mechanical Translation and the Problem of Multiple Meaning. *Mechanical Translation*, *3*(2), 46–51.

Kulesza, A., & Shieber, S. M. (2004). A learning approach to improving sentence-level MT evaluation. *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)* (pp. 75–84).

Lambert, P., Giménez, J., Costa-jussá, M. R., Amigó, E., Banchs, R. E., Màrquez, L., & Fonollosa, J. A. R. (2006). Machine Translation System Development based on Human Likeness. *Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology*.

LDC (2005). *Linguistic Data Annotation Specification: Assessment of Adequacy and Fluency in Translations. Revision 1.5* (Technical Report). Linguistic Data Consortium. `http://www.ldc.upenn.edu/Projects/TIDES/Translation/TransAssess04.pdf`.

Le, A., & Przybocki, M. (2005). NIST 2005 machine translation evaluation official results. *Official release of automatic evaluation scores for all submissions, August*.

Lee, Y. K., & Ng, H. T. (2002). An Empirical Evaluation of Knowledge Sources and Learning Algorithms for Word Sense Disambiguation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 41–48).

Lehrberger, J., & Bourbeau, L. (1988). *Machine Translation: Linguistic Characteristics of MT Systems and General Methodology of Evaluation*. John Benjamin Publishers.

Leusch, G., Ueffing, N., & Ney, H. (2006). CDER: Efficient MT Evaluation Using Block Movements. *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)* (pp. 241–248).

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, *8*(10), 707–710.

Li, C.-H., Li, M., Zhang, D., Li, M., Zhou, M., & Guan, Y. (2007). A Probabilistic Approach to Syntax-based Reordering for Statistical Machine Translation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 720–727).

Liang, P., Bouchard-CÃ´té, A., Klein, D., & Taskar, B. (2006). An End-to-End Discriminative Approach to Machine Translation. *Proceedings of the Joint 21st International Conference on*

Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL) (pp. 761–768).

Lin, C.-Y., & Och, F. J. (2004a). Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Lin, C.-Y., & Och, F. J. (2004b). ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Lin, D. (1998). Dependency-based Evaluation of MINIPAR. *Proceedings of the Workshop on the Evaluation of Parsing Systems*.

Lin, D. (2004). A Path-Based Transfer Model for Machine Translation. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Lita, L. V., Rogati, M., & Lavie, A. (2005). BLANC: Learning Evaluation Metrics for MT. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)* (pp. 740–747).

Liu, D., & Gildea, D. (2005). Syntactic Features for Evaluation of Machine Translation. *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization* (pp. 25–32).

Liu, D., & Gildea, D. (2006). Stochastic Iterative Alignment for Machine Translation Evaluation. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 539–546).

Liu, D., & Gildea, D. (2007). Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation. *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)* (pp. 41–48).

Lopresti, D., & Tomkins, A. (1997). Block Edit Models for Approximate String Matching. *Theoretical Computer Science, 181*(1), 159–179.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. *Proceedings of the Sixth Conference on Natural Language Learning (CONLL)* (pp. 49–55).

Marcu, D., Wang, W., Echihabi, A., & Knight, K. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marcu, D., & Wong, W. (2002). A Phrase-Based, Joint Probability Model for Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Màrquez, L., Escudero, G., Martínez, D., & Rigau, G. (2006). Supervised Corpus-Based Methods for WSD. In Phil Edmonds and Eneko Agirre (Ed.), *Word Sense Disambiguation: Algorithms, Applications, and Trends*, NIPS, chapter 7. Kluwer.

Màrquez, L., Surdeanu, M., Comas, P., & Turmo, J. (2005). Robust Combination Strategy for Semantic Role Labeling. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*.

Martí, M. A. (Ed.). (1996). *Gran diccionario de la Lengua Española*. Larousse Planeta, Barcelona.

Mehay, D., & Brew, C. (2007). BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Melamed, D. (2004). Statistical Machine Translation by Parsing. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Melamed, D., Clark, S., Way, A., Wu, D., Hall, K., Hearne, M., Carpuat, M., Dreyer, M., Groves, D., Shen, Y., Wellington, B., Burbank, A., & Fox, P. (2005). *Final Report of Johns Hopkins 2003 Summer Workshop on Statistical Machine Translation by Parsing* (Technical Report). Johns Hopkins University.

Melamed, I. D., Green, R., & Turian, J. P. (2003). Precision and Recall of Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Mihalcea, R., & Moldovan, D. (1999). An Automatic Method for Generating Sense Tagged Corpora. *Proceedings of AAAI*.

Miller, G. A., & Beebe-Center, J. G. (1956). Some Psychological Methods for Evaluating the Quality of Translation. *Mechanical Translation*, *3*(3), 73–80.

Moore, R. C. (2005). A Discriminative Framework for Bilingual Word Alignment. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)* (pp. 81–88).

Moore, R. C., tau Yih, W., & Bode, A. (2006). Improved Discriminative Bilingual Word Alignment. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 513–520).

Navarro, B., Civit, M., Martí, M. A., Marcos, R., & Fernández, B. (2003). Syntactic, Semantic and Pragmatic Annotation in Cast3LB. *Proceedings of SProLaC* (pp. 59–68).

Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*.

no, J. B. M., Banchs, R., Crego, J. M., de Gispert, A., Lambert, P., Fonollosa, J. A. R., & Costa-jussà, M. R. (2006). N-gram-based Machine Translation. *Computational Linguistics*, *32*(4), 527–549.

Och, F. J. (2002). *Statistical Machine Translation: From Single-Word Models to Alignment Templates*. Doctoral dissertation, RWTH Aachen, Germany.

Och, F. J. (2003). Minimum error rate training in statistical machine translation. *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sapporo, Japan.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., & Radev, D. (2003). *Final Report of Johns Hopkins 2003 Summer Workshop on Syntax for Statistical Machine Translation* (Technical Report). Johns Hopkins University.

Och, F. J., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., & Radev, D. (2004). A Smorgasbord of Features for Statistical Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Och, F. J., & Ney, H. (2000). Improved Statistical Alignment Models. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Och, F. J., & Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 295–302).

Och, F. J., & Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, *29*(1), 19–51.

Och, F. J., Tillmann, C., & Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Och, F. J., Ueffing, N., & Ney, H. (2001). An Efficient A* Search Algorithm for Statistical Machine Translation. *Proceedings of Data-Driven Machine Translation Workshop* (pp. 55–62).

Owczarzak, K., Groves, D., Genabith, J. V., & Way, A. (2006). Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 148–155).

Owczarzak, K., van Genabith, J., & Way, A. (2007a). Dependency-Based Automatic Evaluation for Machine Translation. *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 80–87).

Owczarzak, K., van Genabith, J., & Way, A. (2007b). Labelled Dependencies in Machine Translation Evaluation. *Proceedings of the ACL Workshop on Statistical Machine Translation* (pp. 104–111).

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71–106.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). *Bleu: a method for automatic evaluation of machine translation, RC22176* (Technical Report). IBM T.J. Watson Research Center.

Patry, A., Gotti, F., & Langlais, P. (2006). Mood at work: Ramses versus Pharaoh. *Proceedings on the Workshop on Statistical Machine Translation* (pp. 126–129).

Paul, M. (2006). Overview of the IWSLT 2006 Evaluation Campaign. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)* (pp. 1–15). Kyoto, Japan.

Paul, M., Finch, A., & Sumita, E. (2007). Reducing Human Assessments of Machine Translation Quality to Binary Classifiers. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*.

Pearson, K. (1914, 1924, 1930). *The life, letters and labours of Francis Galton*. (3 volumes).

Pfafflin, S. M. (1965). Evaluation of Machine Translations by Reading Comprehension Tests and Subjective Judgments. *Mechanical Translation and Computational Linguistics*, *8*(2), 2–8.

Platt, J. C. (2000). Probabilities for SV Machines. In Alexander J. Smola and Peter L. Bartlett and Bernhard SchÃ¶lkopf and Dale Schuurmans (Ed.), *Advances in Large Margin Classifiers*, NIPS, chapter 5, 61–74. The MIT Press.

Popovic, M., & Ney, H. (2007). Word Error Rates: Decomposition over POS classes and Applications for Error Analysis. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 48–55). Prague, Czech Republic: Association for Computational Linguistics.

Porter, M. (2001). The Porter Stemming Algorithm.

Quirk, C. (2004). Training a Sentence-Level Machine Translation Confidence Metric. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 825–828).

Quirk, C., Menezes, A., & Cherry, C. (2005). Dependency Treelet Translation: Syntactically Informed Phrasal SMT. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 271–279).

Ratnaparkhi, A. (1998). *Maximum Entropy Models for Natural Language Ambiguity Resolution*. Doctoral dissertation, University of Pennsylvania.

Reeder, F., Miller, K., Doyon, J., & White, J. (2001). The Naming of Things and the Confusion of Tongues: an MT Metric. *Proceedings of the Workshop on MT Evaluation "Who did what to whom?" at Machine Translation Summit VIII* (pp. 55–59).

Russo-Lassner, G., Lin, J., & Resnik, P. (2005). *A Paraphrase-Based Approach to Machine Translation Evaluation (LAMP-TR-125/CS-TR-4754/UMIACS-TR-2005-57)* (Technical Report). University of Maryland, College Park. `http://lampsrv01.umiacs.umd.edu/pubs/TechReports/LAMP\_125/LAMP\_125.pdf`.

Sánchez-Martínez, F., Pérez-Ortiz, J. A., & Forcada, M. L. (2007). Integrating Corpus-based and Rule-based Approaches in an Open-source Machine Translation System. *Proceedings of METIS-II Workshop: New Approaches to Machine Translation* (pp. 73–82). Leuven, Belgium.

Schafer, C., & Yarowsky, D. (2003). Statistical Machine Translation Using Coercive Two-Level Syntactic Transduction. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Sekine, S. (1997). The Domain Dependence of Parsing. *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 96–102).

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal, 27*, 379–423, 623–656.

Shannon, C. E., & Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press.

Shen, L., Sarkar, A., & Och, F. J. (2004). Discriminative Reranking for Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Siegel, S. (1956). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Simard, M., Ueffing, N., Isabelle, P., & Kuhn, R. (2007). Rule-Based Translation with Statistical Phrase-Based Post-Editing. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 203–206).

Slype, G. V. (1979). *Critical Study of Methods for Evaluating the Quality of MT* (Technical Report). Technical Report BR19142, European Commission / Directorate for General Scientific and Technical Information Management (DG XIII).

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A Study of Translation Edit Rate with Targeted Human Annotation. *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 223–231).

Spearman, C. (1904). The Proof and Measurement of Association Between Two Rings. *American Journal of Psychology, 15*, 72–101.

Specia, L., Sankaran, B., & das GraÃ§as Volpe Nunes, M. (2008). n-Best Reranking for the Efficient Integration of Word Sense Disambiguation and Statistical Machine Translation. In *Computational Linguistics and Intelligent Text Processing*, vol. 4919/2008 of *Lecture Notes in Computer Science*, 399–410. Springer Berlin / Heidelberg.

Specia, L., Stevenson, M., & das GraÃ§as Volpe Nunes, M. (2007). Learning Expressive Models for Word Sense Disambiguation. *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 41–48).

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. *Proceedings of ICSLP*.

Stout, T. M. (1954). Computing Machines for Language Translation. *Mechanical Translation*, *1*(3), 41–46.

Stroppa, N., van den Bosch, A., & Way, A. (2007). Exploiting Source Similarity for SMT using Context-Informed Features. *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)* (pp. 231–240).

Surdeanu, M., & Turmo, J. (2005). Semantic Role Labeling Using Complete Syntactic Analysis. *Proceedings of CoNLL Shared Task*.

Surdeanu, M., Turmo, J., & Comelles, E. (2005). Named Entity Recognition from Spontaneous Open-Domain Speech. *Proceedings of the 9th International Conference on Speech Communication and Technology (Interspeech)*.

Takezawa, T. (1999). Building a Bilingual Travel Conversation Database for Speech Translation Research. *Proceedings of the 2nd International Workshop on East-Assian Language Resources and Evaluation —Oriental COCOSDA Workshop* (pp. 17–20).

Taskar, B., Lacoste-Julien, S., & Klein, D. (2005). A Discriminative Matching Approach to Word Alignment. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)* (pp. 73–80).

Thompson, H. S. (1991). Automatic Evaluation of Translation Quality: Outline of Methodology and Report on Pilot Experiment. *Proceedings of the Evaluators' Forum* (pp. 215–223). Les Rasses, Vaud, Switzerland: Geneva: ISSCO.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). Accelerated DP based Search for Statistical Translation. *Proceedings of European Conference on Speech Communication and Technology*.

Tillmann, C., & Xia, F. (2003). A Phrase-based Unigram Model for Statistical Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*.

Tillmann, C., & Zhang, T. (2005). A Localized Prediction Model for Statistical Machine Translation. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 557–564). Ann Arbor, Michigan: Association for Computational Linguistics.

Tillmann, C., & Zhang, T. (2006). A Discriminative Global Training Algorithm for Statistical MT. *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)* (pp. 721–728).

Toutanova, K., Haghighi, A., & Manning, C. (2005). Joint Learning Improves Semantic Role Labeling. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 589–596). Ann Arbor, Michigan: Association for Computational Linguistics.

Toutanova, K., Haghighi, A., & Manning, C. D. (2008). A Global Joint Model for Semantic Role Labeling. *To appear in Computational Linguistics. Special Issue on Semantic Role Labeling*.

Toutanova, K., Klein, D., & Manning, C. D. (2003). Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. *Proceedings of Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)* (pp. 173–180).

Tufis, D., Cristea, D., & Stamou, S. (2004a). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. *Romanian Journal on Science Technology of Information. Special Issue on Balkanet*, *7*(3–4), 9–44.

Tufis, D., Ion, R., & Ide, N. (2004b). Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. *Proceedings of the 20th International Conference on Computational Linguistics (COLING)* (pp. 1312–1318).

Turian, J. P., Shen, L., & Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation. *Proceedings of MT SUMMIT IX*.

Vanni, M., & Miller, K. (2002). Scaling the ISLE Framework: Use of Existing Corpus Resources for Validation of MT Evaluation Metrics across Languages. *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC)* (pp. 1254–1262).

Vapnik, V. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag. ISBN 0-387-98780-0.

Vauquois, B., Veillon, G., & Veyrunes, J. (1966). Syntax and Interpretation. *Mechanical Translation and Computational Linguistics*, *9*(2), 44–54.

Venkatapathy, S., & Bangalore, S. (2007). Three models for discriminative machine translation using Global Lexical Selection and Sentence Reconstruction. *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation* (pp. 152–159).

Vickrey, D., Biewald, L., Teyssier, M., & Collen, D. (2005). Word-Sense Disambiguation for Machine Translation. *Proceedings of the Joint Conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT-EMNLP)*.

Vilar, D., Xu, J., D'Haro, L. F., & Ney, H. (2006). Error Analysis of Machine Translation Output. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)* (pp. 697–702).

Vogel, S., & Tribble, A. (2002). Improving Statistical Machine Translation for a Speech-to-Speech Translation Task. *Proceedings of ICSLP-2002 Workshop on Speech-to-Speech Translation* (pp. 1901–1904).

Vossen, P., Diez-Orzas, P., & Peters, W. (1997). Multilingual Design of EuroWordNet. *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications* (pp. 1–8). New Brunswick, New Jersey: Association for Computational Linguistics.

Vox (Ed.). (1990). *Diccionario Actual de la Lengua Española*. Bibliograf, Barcelona.

Wang, Y.-Y. (1998). *Grammar Inference and Statistical Machine Translation*. Doctoral dissertation, Carnegie Mellon University.

Wang, Y.-Y., & Waibel, A. (1998). Modeling with Structures in Statistical Machine Translation. *Proceedings of the Joint 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*.

Way, A., & Gough, N. (2005). Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, *11*(3), 295–309.

Weaver, W. (1955). *Translation (1949)*. Machine Translation of Languages. Cambridge, MA: The MIT Press.

White, J. S. (1995). Approaches to Black Box MT Evaluation. *Proceedings of Machine Translation Summit V* (p. 10).

White, J. S., O'Connell, T., & O'Mara, F. (1994). The ARPA MT Evaluation Methodologies: Evolution, Lessons, and Future Approaches. *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)* (pp. 193–205).

William H. Press, Saul A. Teukolsky, W. T. V., & Flannery, B. P. (2002). *Numerical Recipes in C++: the Art of Scientific Computing*. Cambridge University Press.

Wu, D. (1997). Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, *23*(3), 377–404.

Wu, D. (2000). Bracketing and aligning words and constituents in parallel text using Stochastic Inversion Transduction Grammars. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer.

Yamada, K. (2002). *A Syntax-based Translation Model*. Doctoral dissertation, University of Southern California.

Yamada, K., & Knight, K. (2001). A Syntax-based Statistical Translation Model. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Yamada, K., & Knight, K. (2002). A Decoder for Syntax-based Statistical MT. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Ye, Y., Zhou, M., & Lin, C.-Y. (2007). Sentence Level Machine Translation Evaluation as a Ranking. *Proceedings of the Second Workshop on Statistical Machine Translation* (pp. 240–247).

Yngve, V. H. (1954). The Machine and the Man. *Mechanical Translation, 1*(2), 41–46.

Yngve, V. H. (1955). Sentence-for-Sentence Translation. *Mechanical Translation, 2*(2), 29–37.

Yngve, V. H. (1957). A Framework for Syntactic Translation. *Mechanical Translation, 4*(3), 59–65.

Zhang, H., & Gildea, D. (2004). Syntax-Based Alignment: Supervised or Unsupervised? *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*.

Zhang, Y., & Vogel, S. (2004). Measuring Confidence Intervals for the Machine Translation Evaluation Metrics. *Proceedings of The 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)* (pp. 85–94).

Zhang, Y., Vogel, S., & Waibel, A. (2004). Interpreting Bleu/NIST scores: How much improvement do we need to have a better system? *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 2051–2054).

Zhou, L., Lin, C.-Y., & Hovy, E. (2006). Re-evaluating Machine Translation Results with Paraphrase Support. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 77–84).

Appendix A

Author's Publications

In the following, we provide a full list of the publications this thesis has contributed with. Each reference is accompanied by a short abstract. Publications are grouped according to the line of research in reverse chronological order:

- **MT Evaluation:**

  - **[E10] On the Robustness of Linguistic Features for Automatic MT Evaluation**. Jesús Giménez and Lluís Màrquez. Currently under submission.
    Abstract: we present an empirical study on the behavior of a heterogeneous set of metrics based on linguistic analysis in the paradigmatic case of speech translation between non-related languages (see Section 3.3).
  - **[E9] A Smorgasbord of Features for Automatic MT Evaluation**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the Third ACL Workshop On Statistical Machine Translation*
    Abstract: This document describes the approach by the Empirical MT Group at the Technical University of Catalonia (UPC-LSI), for the shared task on Automatic Evaluation of Machine Translation at the ACL 2008 Third Workshop on Statistical Machine Translation.
  - **[E8] Towards Heterogeneous Automatic MT Error Analysis**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the The Sixth International Conference on Language Resources and Evaluation (LREC'08) Marrakech, Morocco, 2008.*
    Abstract: This work studies the viability of performing heterogeneous automatic MT error analyses. Through a practical case study, we show how these features provide an effective means of elaborating interpretable and detailed automatic reports of translation quality (see Section 3.5).
  - **[E7] Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP'08)*. Hyderabad, India, 2008.

Abstract: We suggest a novel path towards heterogeneous automatic MT evaluation which involves considering metrics at different linguistic dimensions without having to adjust their relative importance (see Section 3.4).

- **[E6] Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the ACL Workshop on Statistical Machine Translation (WMT'07)*. Prague, Czech Republic, 2007.
  Abstract: We show that automatic metrics at deep linguistic levels (syntactic and shallow-semantic) are able to produce more reliable rankings of heterogeneous systems than metrics based on lexical similarities (see Sections 3.1) and 3.2).

- **[E5] IQMT v 2.0. Technical Manual**. Jesús Giménez. *Research Report LSI-07-29-R. TALP Research Center. LSI Department*. `http://www.lsi.upc.edu/~nlp/IQMT/IQMT.v2.0.pdf`.
  Abstract: This report presents a description and tutorial on the IQ$_{MT}$ package for automatic MT evaluation based on human likeness.

- **[E4] Machine Translation System Development based on Human Likeness**. Patrik Lambert, Jesús Giménez, Marta R. Costa-jussá, Enrique Amigó, Rafael E. Banchs, Lluís Màrquez and J.A. R. Fonollosa. In *Proceedings of IEEE/ACL 2006 Workshop on Spoken Language Technology*. Palm Beach, Aruba, 2007.
  Abstract: We present a novel approach for parameter adjustment in SMT systems by working in metric combinations optimized on the basis of human likeness.

- **[E3] MT Evaluation: Human-Like vs. Human Acceptable**. Enrique Amigó, Jesús Giménez, Julio Gonzalo and Lluís Màrquez. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL'2006)*. Sydney, Australia, 2006.
  Abstract: We present a comparative study on the behaviour of human likeness and human acceptability as meta-evaluation criteria in the context of MT evaluation.

- **[E2] IQMT: A Framework for Automatic Machine Translation Evaluation**. Jesús Giménez and Enrique Amigó. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, 2006.
  Abstract: We present the IQ$_{MT}$ Framework for Machine Translation Evaluation Inside QARLA.

- **[E1] Machine Translation Evaluation Inside QARLA**. Jesús Giménez and Enrique Amigó and Chiori Hori. In *Proceedings of the International Workshop on Spoken Language Technology (IWSLT'05)*. Pittsburgh, PA, 2005.
  Abstract: Preliminary results on the application of the QARLA Framework to MT evaluation are presented.

- **Lexical Selection in SMT:**

  - **[L4] Discriminative Phrase Selection for Statistical Machine Translation** . Jesús Giménez and Lluís Màrquez. To appear in *Learning Machine Translation*. NIPS Workshop series. MIT Press, 2008.

Abstract: This work explores the application of discriminative learning to the problem of phrase selection in Statistical Machine Translation (see Section Chapter 6).

- **[L3] Context-aware Discriminative Phrase Selection for Statistical Machine Translation**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the ACL Workshop on Statistical Machine Translation (WMT'07)*. Prague, Czech Republic, 2007.
  Abstract: In this work we revise the application of discriminative learning to the problem of phrase selection in Statistical Machine Translation (see Chapter 6).
- **[L2] The LDV-COMBO system for SMT**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the NAACL Workshop on Statistical Machine Translation (WMT'06)*. New York City, 2006.
  Abstract: We describe the LDV-COMBO system presented at the Shared Task of the NAACL'06 MT Workshop.
- **[L1] Combining Linguistic Data Views for Phrase-based SMT**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Ann Arbor, MI, 2005.
  Abstract: We explore the possibility of working with alignments at different levels of abstraction, using different degrees of linguistic annotation at the level of shallow parsing. We also investigate alternative methods so as to combine different translation models built out from different linguistic data views (see Chapter 5).

- **Domain-Dependence in SMT:**

  - **[D3] The UPC System for Arabic-to-English Entity Translation**. David Farwell, Jesús Giménez, Edgar González, Reda Halkoum, Horacio Rodríguez and Mihai Surdeanu. In *Proceedings of the Automatic Content Extraction Evaluation Program (ACE 2007)*. University of Maryland, MD, 2007.
    Abstract: We describe the UPC Arabic-to-English Entity Translation System presented at the ACE/ET 2007 Evaluation Campaign, and its application to the Arabic-to-English Entity Translation task.
  - **[D2] Low-cost Enrichment of Spanish WordNet with Automatically Translated Glosses: Combining General and Specialized Models**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the joint conference of the International Committee on Computational Linguistics and the Association for Computational Linguistics (COLING/ACL'2006)*. Sydney, Australia, 2006.
    Abstract: We study the enrichment of Spanish WordNet with synset glosses automatically obtained from the English WordNet glosses using a phrase-based Statistical Machine Translation system trained on data sources from different domains (see Chapter 7).
  - **[D1] Automatic Translation of WordNet Glosses**. Jesús Giménez and Lluís Màrquez and German Rigau. In *Proceedings of Cross-Language Knowledge Induction Workshop, EUROLAN Summer School, 2005*. Cluj-Napoca, Romania, 2005.
    Abstract: We present preliminary results on the automatic translation of the glosses

in the English WordNet. We intend to generate a preliminary material which could be utilized to enrich other wordnets lacking of glosses (see Chapter 7).

We also provide a list of other publications by the author which are not directly related to the work presented in this thesis:

- **Development of NLP Tools:**

  - **[T4] Semantic Role Labeling as Sequential Tagging**. Lluís Màrquez, Pere Comas, Jesús Giménez and Neus Català. In *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL), 2005*. Ann Arbor, MI, 2005.
    Abstract: We describe the Semantic Role Labeling system presented to the CoNLL 2005 shared task.
  - **[T3] SVMTool: A general POS tagger generator based on Support Vector Machines**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, vol. I, pages 43 - 46. Lisbon, Portugal, 2004.
    Abstract: This paper presents SVMTool, a simple, flexible, effective and efficient part–of–speech tagger based on Support Vector Machines. SVMTool offers a fairly good balance among these properties which make it really practical for current NLP applications. SVMTool may be freely downloaded at `http://www.lsi.upc.edu/~nlp/SVMTool`.
  - **[T2] SVMTool: A general POS tagger generator based on Support Vector Machines (Technical Manual)**. Jesús Giménez and Lluís Màrquez. *LSI Departament Research Report (LSI-04-34-R), Technical University of Catalonia*, 2004.
    Abstract: This report is a detailed technical manual for the SVMTool.
  - **[T1] Fast and Accurate Part-of-Speech Tagging: The SVM Approach Revisited**. Jesús Giménez and Lluís Màrquez. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*, pages 158 - 165. Borovets, Bulgary, 2003. Selected as a chapter in volume 260 of CILT series (Current Issues in Linguistic Theory). John Benjamins Publishers, Amsterdam.
    Abstract: In this paper we present a very simple and effective part-of-speech tagger based on Support Vector Machines (SVM). Simplicity and efficiency are achieved by working with linear separators in the primal formulation of SVM, and by using a greedy left-to-right tagging scheme. As a result, we developed the SVMTool.

- **Generation of MT Resources:**

  - **[G8] LC-STAR: XML-coded Phonetic Lexica and Bilingual Corpora for Speech-to-Speech Translation**. Folkert de Vriend, Núria Castell, Jesús Giménez and Giulio Maltese. In *Proceedings of the Papillon Workshop on Multilingual Lexical Databases*. Grenoble, France, 2004.
    Abstract: This paper describes XML encoding of lexica and multilingual corpora and their validation in the framework of the LC-STAR project.

- **[G7] Bilingual Connections for Trilingual Corpora: An XML Approach**. Victoria Arranz, Núria Castell, Josep Maria Crego, Jesús Giménez, Adrià de Gispert and Patrik Lambert. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, vol. IV, pages 1459 - 1462. Lisbon, Portugal, 2004.
  Abstract: An XML representation for a trilingual spontaneous speech corpus for statistical speech-to-speech translation is suggested.

- **[G6] Creació de recursos lingüístics per a la traducció automàtica**. Victoria Arranz, Núria Castell i Jesús Giménez. In *2n Congrés d'Enginyeria en Llengua Catalana*. (CELC'04). Andorra, 2004. (presented also in III Jornadas en Tecnología del Habla. Valencia, Spain. 2004.)
  Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.

- **[G5] Development of Language Resources for Speech-to-Speech Translation**. Victoria Arranz, Núria Castell and Jesús Giménez. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP'03)*, pages 26-30. Borovets, Bulgary, 2003.
  Abstract: This paper describes the design and development of a trilingual spontaneous speech corpus for statistical speech-to-speech translation.

- **[G4] Lexica and Corpora for Speech-to-Speech translation: A Trilingual Approach**. David Conejero, Jesús Giménez, Victoria Arranz, Antonio Bonafonte, Neus Pascual, Núria Castell and Asunción Moreno. In *Proceedings of the 8th European Conference on Speech Communication and Technology (EuroSpeech 2003)*. Geneva, Switzerland, 2003.
  Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.

- **[G3] Description of Language Resources Used for Experiments**. Victoria Arranz, Núria Castell, Jesús Giménez, Hermann Ney and Nicola Ueffing. *Technical Report Deliverable D4.2, LC-STAR project by the European Community (IST project ref. No. 2001-32216)*, 2003.
  Abstract: This documents describes the language resources used in the first experiments as well as the experiments themselves, in the frame of the LC-STAR project. These experiments are described in detail, providing information on both acquisition and expansion of already existing language resources.

- **[G2] Description of Raw Corpora**. Victoria Arranz, Núria Castell, Jesús Giménez and Asunción Moreno. *Technical Report Deliverable 5.3, LC-STAR project by the European Community (IST project ref. No. 2001-32216)*, 2003.
  Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.

- **[G1] Speech Corpora Creation for Tourist Domain**. Victoria Arranz, Núria Castell and Jesús Giménez. *LSI Department Technical Report (LSI-03-2-T), Technical University of Catalonia*, 2003.

Abstract: Creation of lexica and corpora for Catalan, Spanish and US-English is described.

All papers are publicly available at `http://www.lsi.upc.edu/~jgimenez/pubs.html`.

Appendix B

Linguistic Processors and Tag Sets

B.1 Shallow Syntactic Parsing

Shallow parsing is performed using several state-of-the-art performance tools.

B.1.1 Part-of-speech Tagging

PoS and lemma annotation is automatically provided by the SVMTool (Giménez & Màrquez, 2004a; Giménez & Màrquez, 2004b)[1]. We use the Freeling (Carreras et al., 2004)[2] package only for lemmatization.

English

The SVMTool for English has been trained on the Wall Street Journal (WSJ) corpus (1,173K words). Sections 0-18 were used for training (912K words), 19-21 for validation (131K words), and 22-24 for test (129K words), respectively. 2.81% of the words in the test set are unknown to the training set. Best other results so far reported on this same test set are (Collins, 2002) (97.11%) and (Toutanova et al., 2003) (97.24%). Table B.1 shows the SVMTool performance as compared to the TnT tagger. 'known' and 'unk.' refer to the subsets of known and unknown words, respectively. 'amb' refers to the set of ambiguous known words and 'all' to the overall accuracy.

	known	amb.	unk.	all.
TnT	96.76%	92.16%	85.86%	96.46%
SVMTool	97.39%	93.91%	89.01%	97.16%

Table B.1: Performance of the SVMTool for English on the WSJ corpus

Table B.2 and Table B.3 show the PoS tag set for English, derived from the Penn Treebank[3] tag set (Marcus et al., 1993). Several coarse classes are included.

[1] http://www.lsi.upc.es/~nlp/SVMTool/
[2] http://www.lsi.upc.es/~nlp/freeling/
[3] http://www.cis.upenn.edu/~treebank/

Type	Description
CC	Coordinating conjunction, e.g., and,but,or...
CD	Cardinal Number
DT	Determiner
EX	Existential there
FW	Foreign Word
IN	Preposition or subordinating conjunction
JJ	Adjective
JJR	Adjective, comparative
JJS	Adjective, superlative
LS	List Item Marker
MD	Modal, e.g., can, could, might, may...
NN	Noun, singular or mass
NNP	Proper Noun, singular
NNPS	Proper Noun, plural
NNS	Noun, plural
PDT	Predeterminer, e.g., all, both ... when they precede an article
POS	Possessive Ending, e.g., Nouns ending in 's
PRP	Personal Pronoun, e.g., I, me, you, he...
PRP$	Possessive Pronoun, e.g., my, your, mine, yours...
RB	Adverb. Most words that end in -ly as well as degree words like quite, too and very.
RBR	Adverb. comparative Adverbs with the comparative ending -er, with a strictly comparative meaning.
RBS	Adverb, superlative
RP	Particle
SYM	Symbol. Should be used for mathematical, scientific or technical symbols
TO	to
UH	Interjection, e.g., uh, well, yes, my...

Table B.2: PoS tag set for English (1/2)

Type	Description
VB	Verb, base form subsumes imperatives, infinitives and subjunctives
VBD	Verb, past tense includes the conditional form of the verb to be
VBG	Verb, gerund or present participle
VBN	Verb, past participle
VBP	Verb, non-3rd person singular present
VBZ	Verb, 3rd person singular present
WDT	Wh-determiner, e.g., which, and that when it is used as a relative pronoun
WP	Wh-pronoun, e.g., what, who, whom...
WP$	Possessive wh-pronoun
WRB	Wh-adverb, e.g., how, where why
#	
$	
”	
(	
)	Punctuation Tags
,	
.	
:	
“	

COARSE TAGS	
N	Nouns
V	Verbs
J	Adjectives
R	Adverbs
P	Pronouns
W	Wh- pronouns
F	Punctuation

Table B.3: PoS tag set for English (2/2)

Spanish

The SVMTool for Spanish has been trained on the 3LB [4] corpus (75K words). It was randomly divided into training set (59K words) and test set (16K words). 13.65% of the words in the test set are unknown to the training set. See results in Table B.4.

	known	amb.	unk.	all.
TnT	97.73%	93.70%	87.66%	96.50%
SVMTool	98.08%	95.04%	88.28%	96.89%

Table B.4: Performance of the SVMTool for Spanish on the 3LB corpus

Tag set for Spanish, derived from the PAROLE tag set, is shown in Table B.5, Table B.6 and Table B.7.

B.1.2 Lemmatization

Word lemmas have been obtained by matching word-PoS pairs against two lemmaries available inside the Freeling package. The English lemmary contains lemmas for 185,201 different word-PoS pairs, whereas the Spanish lemmary contains lemmas for 1,039,365 word-PoS pairs.

B.1.3 Chunking

Partial parsing information (i.e., base phrase chunks) is obtained using the Phreco software based on global on-line learning via the Perceptron algorithm (Carreras et al., 2005).

English

English models have been trained on the Penn Treebank (300K words). We randomly split data into train (211,727 words), development (47,377 words) and test (40,039 words). Best performance (F_1 = 93.72%) was obtained using averaged perceptrons up to epoch 8. Table B.8 shows phrase chunking tag sets for English.

Spanish

Models for Spanish have been trained on the 3LB corpus (95K words), randomly split into training (76,115 words) and test (18,792 words). Best performance (F_1 = 94.55%) was obtained using regular perceptrons after epoch 20. Table B.9 shows phrase chunking tag sets for Spanish.

[4] The 3LB project is funded by the Spanish Ministry of Science and Technology (FIT-15050-2002-244), visit the project website at `http://www.dlsi.ua.es/projectes/3lb/`.

Type	Description
NOUN	
NC	Noun, Common
NP	Noun, Proper
VERB	
VAG	Verb, Auxiliary, Gerund
VAI	Verb, Auxiliary, Indicative
VAM	Verb, Auxiliary, Imperative
VAN	Verb, Auxiliary, Infinitive
VAP	Verb, Auxiliary, Participle
VAS	Verb, Auxiliary, Subjunctive
VMG	Verb, Main, Gerund
VMI	Verb, Main, Indicative
VMM	Verb, Main, Imperative
VMN	Verb, Main, Infinitive
VMP	Verb, Main, Participle
VMS	Verb, Main, Subjunctive
VSG	Verb, Semi-Auxiliary, Gerund
VSI	Verb, Semi-Auxiliary, Indicative
VSM	Verb, Semi-Auxiliary, Imperative
VSN	Verb, Semi-Auxiliary, Infinitive
VSP	Verb, Semi-Auxiliary, Participle
VSS	Verb, Semi-Auxiliary, Subjunctive
ADJECTIVE	
AO	Adjective, Ordinal
AQ	Adjective, Qualifier
AQP	Adjective, Qualifier and Past Participle
ADVERB	
RG	Adverb, General
RN	Adverb, Negative
PRONOUN	
P0	Pronoun, Clitic
PD	Pronoun, Demonstrative
PE	Pronoun, Exclamatory
PI	Pronoun, Indefinite
PN	Pronoun, Numeral
PP	Pronoun, Personal
PR	Pronoun, Relative
PT	Pronoun, Interrogative
PX	Pronoun, Possessive

Table B.5: PoS tag set for Spanish and Catalan (1/3)

Type	**Description**
ADPOSITON	
SP	Adposition, Preposition
CONJUNCTION	
CC	Conjunction, Coordinate
CS	Conjunction, Subordinative
DETERMINER	
DA	Determiner, Article
DD	Determiner, Demonstrative
DE	Determiner, Exclamatory
DI	Determiner, Indefinite
DN	Determiner, Numeral
DP	Determiner, Possessive
DT	Determiner, Interrogative
INTERJECTION	
I	Interjection
DATE TIMES	
W	Date Times
UNKNOWN	
X	Unknown
ABBREVIATION	
Y	Abbreviation
NUMBERS	
Z	Figures
Zm	Currency
Zp	Percentage

Table B.6: PoS tag set for Spanish and Catalan (2/3)

Type	Description
PUNCTUATION	
Faa	Fat Punctuation, !
Fc	Punctuation, ,
Fd	Punctuation, :
Fe	Punctuation, “
Fg	Punctuation, -
Fh	Punctuation, /
Fia	Punctuation,
Fit	Punctuation, ?
Fp	Punctuation, .
Fpa	Punctuation, (
Fpt	Punctuation,)
Fs	Punctuation, ...
Fx	Punctuation, ;
Fz	Punctuation, other than those

COARSE TAGS	
A	Adjectives
C	Conjunctions
D	Determiners
F	Punctuation
I	Interjections
N	Nouns
P	Pronouns
S	Adpositions
V	Verbs
VA	Auxiliary Verbs
VS	Semi-Auxiliary Verbs
VM	Main Verbs

Table B.7: PoS tag set for Spanish and Catalan (3/3)

Type	Description
ADJP	Adjective phrase
ADVP	Adverb phrase
CONJP	Conjunction
INTJ	Interjection
LST	List marker
NP	Noun phrase
PP	Preposition
PRT	Particle
SBAR	Subordinated Clause
UCP	Unlike Coordinated phrase
VP	Verb phrase
O	Not-A-Phrase

Table B.8: Base phrase chunking tag set for English

Type	Description
ADJP	Adjective phrase
ADVP	Adverb phrase
CONJP	Conjunction
INTJ	Interjection
NP	Noun phrase
PP	Preposition
SBAR	Subordinated Clause
VP	Verb phrase
AVP	Adjectival verb phrase
NEG	Negation
MORFV	Verbal morpheme
O	Not-A-Phrase

Table B.9: Base phrase chunking tag set for Spanish and Catalan

B.2 Syntactic Parsing

Dependency parsing for English is performed using the MINIPAR[5] parser (Lin, 1998). A brief description of grammatical categories and relations may be found in Table B.10 and Table B.11.

Constituency parsing for English is performed using the Charniak-Johnson's Max-Ent reranking parser (Charniak & Johnson, 2005)[6]. A description of the tag set employed is available in Table B.12.

B.3 Shallow Semantic Parsing

Named entities are automatically annotated using the BIOS Suite of Syntactico-Semantic Analyzers (Surdeanu et al., 2005)[7]. The list of NE types utilized is available in Table B.13.

Semantic role labeling is performed using the SwiRL Semantic Role Labeler (Surdeanu & Turmo, 2005; Màrquez et al., 2005)[8]. A list of SR types is available in Table B.14.

B.4 Semantic Parsing

Semantic parsing is performed using the BOXER component (Bos, 2005) available inside the C&C Tools (Clark & Curran, 2004)[9]. BOXER elaborates DRS representations of input sentences parsed on the basis of a Combinatory Categorial Grammar (CCG) parser (Bos et al., 2004).

There are two types of DRS conditions:

basic conditions: one-place properties (predicates), two-place properties (relations), named entities, time expressions, cardinal expressions and equalities.

complex conditions: disjunction, implication, negation, question, and propositional attitude operations.

Tables B.15 to B.19 describe some aspects of the DRS representations utilized. For instance, Tables B.15 and B.16 respectively show basic and complex DRS conditions. Table B.17 shows DRS subtypes. Tables B.18 and B.19 show symbols for one-place and two-place relations.

[5] `http://www.cs.ualberta.ca/~lindek/minipar.htm`
[6] `ftp://ftp.cs.brown.edu/pub/nlparser/`
[7] `http://www.surdeanu.name/mihai/bios/`
[8] `http://www.surdeanu.name/mihai/swirl/`
[9] `http://svn.ask.it.usyd.edu.au/trac/candc`

Type	Description
Det	Determiners
PreDet	Pre-determiners
PostDet	Post-determiners
NUM	numbers
C	Clauses
I	Inflectional Phrases
V	Verb and Verb Phrases
N	Noun and Noun Phrases
NN	noun-noun modifiers
P	Preposition and Preposition Phrases
PpSpec	Specifiers of Preposition Phrases
A	Adjective/Adverbs
Have	verb ‘to have’
Aux	Auxiliary verbs, e.g. should, will, does, ...
Be	Different forms of verb ‘to be’: is, am, were, be, ...
COMP	Complementizer
VBE	‘to be’ used as a linking verb. E.g., I am hungry
V_N	verbs with one argument (the subject), i.e., intransitive verbs
V_N_N	verbs with two arguments, i.e., transitive verbs
V_N_I	verbs taking small clause as complement

Table B.10: Grammatical categories provided by MINIPAR

Type	**Description**
appo	“ACME president, –appo-> P.W. Buckman”
aux	“should <-aux– resign”
be	“is <-be– sleeping”
by-subj	subject with passives
c	clausal complement “that <-c– John loves Mary”
cn	nominalized clause
comp1	first complement
desc	description
det	“the <-det ‘– hat”
gen	“Jane’s <-gen– uncle”
fc	finite complement
have	“have <-have– disappeared”
i	relationship between a C clause and its I clause
inv-aux	inverted auxiliary: “Will <-inv-aux– you stop it?”
inv-be	inverted be: “Is <-inv-be– she sleeping”
inv-have	inverted have: “Have <-inv-have– you slept”
mod	relationship between a word and its adjunct modifier
pnmod	post nominal modifier
p-spec	specifier of prepositional phrases
pcomp-c	clausal complement of prepositions
pcomp-n	nominal complement of prepositions
post	post determiner
pre	pre determiner
pred	predicate of a clause
rel	relative clause
obj	object of verbs
obj2	second object of ditransitive verbs
s	surface subject
sc	sentential complement
subj	subject of verbs
vrel	passive verb modifier of nouns
wha, whn, whp	wh-elements at C-spec positions (a|n|p)

Table B.11: Grammatical relationships provided by MINIPAR

Type	Description
	Clause Level
S	Simple declarative clause
SBAR	Clause introduced by a (possibly empty) subordinating conjunction
SBARQ	Direct question introduced by a wh-word or a wh-phrase
SINV	Inverted declarative sentence, i.e. one in which the subject follows the tensed verb or modal
SQ	Inverted yes/no question, or main clause of a wh-question, following the wh-phrase in SBARQ
	Phrase Level
ADJP	Adjective Phrase
ADVP	Adverb Phrase
CONJP	Conjunction Phrase
FRAG	Fragment
INTJ	Interjection
LST	List marker
NAC	Not a Constituent; used to show the scope of certain prenominal modifiers within a NP
NP	Noun Phrase
NX	Used within certain complex NPs to mark the head of the NP
PP	Prepositional Phrase
PRN	Parenthetical
PRT	Particle. Category for words that should be tagged RP
QP	Quantifier Phrase (i.e. complex measure/amount phrase); used within NP
RRC	Reduced Relative Clause
UCP	Unlike Coordinated Phrase
VP	Verb Phrase
WHADJP	Wh-adjective Phrase
WHAVP	Wh-adverb Phrase
WHNP	Wh-noun Phrase
WHPP	Wh-prepositional Phrase
X	Unknown, uncertain, or unbracketable

Table B.12: Clause/phrase level tag set for English

Type	**Description**
ORG	Organization
PER	Person
LOC	Location
MISC	Miscellaneous
O	Not-A-NE
DATE	Temporal expressions
NUM	Numerical expressions
ANGLE_QUANTITY DISTANCE_QUANTITY SIZE_QUANTITY SPEED_QUANTITY TEMPERATURE_QUANTITY WEIGHT_QUANTITY	Quantities
METHOD MONEY LANGUAGE PERCENT PROJECT SYSTEM	Other

Table B.13: Named Entity types

Type	Description
A0 A1 A2 A3 A4 A5	arguments associated with a verb predicate, defined in the PropBank Frames scheme.
AA	Causative agent
AM-ADV	Adverbial (general-purpose) adjunct
AM-CAU	Causal adjunct
AM-DIR	Directional adjunct
AM-DIS	Discourse marker
AM-EXT	Extent adjunct
AM-LOC	Locative adjunct
AM-MNR	Manner adjunct
AM-MOD	Modal adjunct
AM-NEG	Negation marker
AM-PNC	Purpose and reason adjunct
AM-PRD	Predication adjunct
AM-REC	Reciprocal adjunct
AM-TMP	Temporal adjunct

Table B.14: Semantic Roles

Type	Description
pred	one-place properties (predicates)
rel	two-place properties (relations)
named	named entities
timex	time expressions
card	cardinal expressions
eq	equalities

Table B.15: Discourse Representation Structures. Basic DRS-conditions

Type	Description
or	disjunction
imp	implication
not	negation
whq	question
prop	propositional attitude

Table B.16: Discourse Representation Structures. Complex DRS-conditions

Type	**Description**
Types of anaphoric information	
pro	anaphoric pronoun
def	definite description
nam	proper name
ref	reflexive pronoun
dei	deictic pronoun
Part-of-speech type	
n	noun
v	verb
a	adjective/adverb
Named Entity types	
org	organization
per	person
ttl	title
quo	quoted
loc	location
fst	first name
sur	surname
url	URL
ema	email
nam	name (when type is unknown)
Cardinality type	
eq	equal
le	less or equal
ge	greater or equal

Table B.17: Discourse Representation Structures. Subtypes

Type	Description
topic,a,n	elliptical noun phrases
thing,n,12	used in NP quantifiers: 'something', etc.)
person,n,1	used in first-person pronouns, 'who'-questions)
event,n,1	introduced by main verbs)
group,n,1	used for plural descriptions)
reason,n,2	used in 'why'-questions)
manner,n,2	used in 'how'-questions)
proposition,n,1	arguments of propositional complement verbs)
unit_of_time,n,1	used in 'when'-questions)
location,n,1	used in 'there' insertion, 'where'-questions)
quantity,n,1	used in 'how many')
amount,n,3	used in 'how much')
degree,n,1	
age,n,1	
neuter,a,0	used in third-person pronouns: it, its)
male,a,0	used in third-person pronouns: he, his, him)
female,a,0	used in third-person pronouns: she, her)
base,v,2	
bear,v,2	

Table B.18: Discourse Representation. Symbols for one-place predicates used in basic DRS conditions

Type	Description
rel,0	general, underspecified type of relation
loc_rel,0	locative relation
role,0	underspecified role: agent,patient,theme
member,0	used for plural descriptions
agent,0	subject
theme,0	indirect object
patient,0	semantic object, subject of passive verbs

Table B.19: Discourse Representation. Symbols for two-place relations used in basic DRS conditions

Appendix C

Metric Sets

1-WER	= {	1-WER }
1-PER	= {	1-PER }
1-TER	= {	1-TER }
BLEU	= {	BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4 }
GTM	= {	GTM-1, GTM-2, GTM-3 }
METEOR	= {	METEOR$_{exact}$, METEOR$_{stem}$, METEOR$_{wnstm}$, METEOR$_{wnsyn}$ }
NIST	= {	NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5 }
ROUGE	= {	ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE$_L$, ROUGE$_{S\star}$, ROUGE$_{SU\star}$, ROUGE$_W$ }
LEX	= {	1-PER, 1-WER, 1-TER, BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, GTM-1, GTM-2, GTM-3, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE$_L$, ROUGE$_{S\star}$, ROUGE$_{SU\star}$, ROUGE$_W$, METEOR$_{exact}$, METEOR$_{stem}$, METEOR$_{wnstm}$, METEOR$_{wnsyn}$ }

Table C.1: Metrics at the Lexical Level

SP = { SP-NIST$_c$-1, SP-NIST$_c$-2, SP-NIST$_c$-3, SP-NIST$_c$-4, SP-NIST$_c$-5,
SP-NISTi$_c$-2, SP-NISTi$_c$-3, SP-NISTi$_c$-4, SP-NISTi$_c$-5, SP-NIST$_{iob}$-1,
SP-NIST$_{iob}$-2, SP-NIST$_{iob}$-3, SP-NIST$_{iob}$-4, SP-NIST$_{iob}$-5, SP-NISTi$_{iob}$-2,
SP-NISTi$_{iob}$-3, SP-NISTi$_{iob}$-4, SP-NISTi$_{iob}$-5, SP-NIST$_l$-1, SP-NIST$_l$-2,
SP-NIST$_l$-3, SP-NIST$_l$-4, SP-NIST$_l$-5, SP-NISTi$_l$-2, SP-NISTi$_l$-3,
SP-NISTi$_l$-4, SP-NISTi$_l$-5, SP-O_c-$\star$, SP-O_c-ADJP, SP-O_c-ADVP,
SP-O_c-CONJP, SP-O_c-INTJ, SP-O_c-LST,
SP-O_c-NP, SP-O_c-O, SP-O_c-PP, SP-O_c-PRT,
SP-O_c-SBAR, SP-O_c-UCP, SP-O_c-VP, SP-O_p-#,
SP-O_p-$, SP-$O_p$-”, SP-$O_p$-(, SP-$O_p$-), SP-$O_p$-$\star$,
SP-O_p-, , SP-O_p-., SP-O_p-:, SP-O_p-CC, SP-O_p-CD,
SP-O_p-DT, SP-O_p-EX, SP-O_p-F, SP-O_p-FW, SP-O_p-IN,
SP-O_p-J, SP-O_p-JJ, SP-O_p-JJR, SP-O_p-JJS, SP-O_p-LS,
SP-O_p-MD, SP-O_p-N, SP-O_p-NN, SP-O_p-NNP,
SP-O_p-NNPS, SP-O_p-NNS, SP-O_p-P, SP-O_p-PDT,
SP-O_p-POS, SP-O_p-PRP, SP-O_p-PRP$, SP-$O_p$-R,
SP-O_p-RB, SP-O_p-RBR, SP-O_p-RBS, SP-O_p-RP,
SP-O_p-SYM, SP-O_p-TO, SP-O_p-UH, SP-O_p-V,
SP-O_p-VB, SP-O_p-VBD, SP-O_p-VBG, SP-O_p-VBN,
SP-O_p-VBP, SP-O_p-VBZ, SP-O_p-W, SP-O_p-WDT,
SP-O_p-WP, SP-O_p-WP$, SP-$O_p$-WRB, SP-$O_p$-“,
SP-NIST$_p$-1, SP-NIST$_p$-2, SP-NIST$_p$-3, SP-NIST$_p$-4, SP-NIST$_p$-5,
SP-NISTi$_p$-2, SP-NISTi$_p$-3, SP-NISTi$_p$-4, SP-NISTi$_p$-5 }

Table C.2: Metrics based on Shallow Parsing

DP = { DP-O_c-⋆, DP-O_c_a, DP-O_c_as, DP-O_c_aux, DP-O_c_be, DP-O_c_c,
DP-O_c_comp, DP-O_c_det, DP-O_c_have, DP-O_c_n, DP-O_c_postdet,
DP-O_c_ppspec DP-O_c_predet, DP-O_c_saidx, DP-O_c_sentadjunct, DP-O_c_subj,
DP-O_c_that, DP-O_c_prep, DP-O_c_u, DP-O_c_v, DP-O_c_vbe, DP-O_c_xsaid,
DP-HWC$_c$-1, DP-HWC$_c$-2, DP-HWC$_c$-3, DP-HWC$_c$-4, DP-HWC$_r$-1,
DP-HWC$_r$-2, DP-HWC$_r$-3, DP-HWC$_r$-4, DP-HWC$_w$-1, DP-HWC$_w$-2,
DP-HWC$_w$-3, DP-HWC$_w$-4, DP-HWCi$_c$-2, DP-HWCi$_c$-3, DP-HWCi$_c$-4,
DP-HWCi$_r$-2, DP-HWCi$_r$-3, DP-HWCi$_r$-4, DP-HWCi$_w$-2, DP-HWCi$_w$-3,
DP-HWCi$_w$-4, DP-O_l-⋆, DP-O_l_1, DP-O_l_2, DP-O_l_3, DP-O_l_4, DP-O_l_5,
DP-O_l_6, DP-O_l_7, DP-O_l_8, DP-O_l_9, DP-O_r-⋆, DP-O_r_amod,
DP-O_r_amount-value, DP-O_r_appo, DP-O_r_appo-mod, DP-O_r_as-arg,
DP-O_r_as1, DP-O_r_as2, DP-O_r_aux, DP-O_r_be, DP-O_r_being,
DP-O_r_by-subj, DP-O_r_c, DP-O_r_cn, DP-O_r_comp1, DP-O_r_conj, DP-O_r_desc,
DP-O_r_dest, DP-O_r_det, DP-O_r_else, DP-O_r_fc, DP-O_r_gen, DP-O_r_guest,
DP-O_r_have, DP-O_r_head, DP-O_r_i, DP-O_r_inv-aux, DP-O_r_inv-have,
DP-O_r_lex-dep, DP-O_r_lex-mod, DP-O_r_mod, DP-O_r_mod-before, DP-O_r_neg,
DP-O_r_nn, DP-O_r_num, DP-O_r_num-mod, DP-O_r_obj, DP-O_r_obj1, DP-O_r_obj2,
DP-O_r_p, DP-O_r_p-spec, DP-O_r_pcomp-c, DP-O_r_pcomp-n, DP-O_r_person,
DP-O_r_pnmod, DP-O_r_poss, DP-O_r_post, DP-O_r_pre, DP-O_r_pred, DP-O_r_punc,
DP-O_r_rel, DP-O_r_s, DP-O_r_sc, DP-O_r_subcat, DP-O_r_subclass,
DP-O_r_subj, DP-O_r_title, DP-O_r_vrel, DP-O_r_wha, DP-O_r_whn, DP-O_r_whp }

Table C.3: Metrics based on Dependency Parsing

CP = { CP-O_c-⋆, CP-O_c-ADJP, CP-O_c-ADVP, CP-O_c-CONJP, CP-O_c-FRAG, CP-O_c-INTJ, CP-O_c-LST, CP-O_c-NAC, CP-O_c-NP, CP-O_c-NX, CP-O_c-O, CP-O_c-PP, CP-O_c-PRN, CP-O_c-PRT, CP-O_c-QP, CP-O_c-RRC, CP-O_c-S, CP-O_c-SBAR, CP-O_c-SINV, CP-O_c-SQ, CP-O_c-UCP, CP-O_c-VP, CP-O_c-WHADJP, CP-O_c-WHADVP, CP-O_c-WHNP, CP-O_c-WHPP, CP-O_c-X, CP-O_p-#, CP-O_p-$, CP-$O_p$-”, CP-$O_p$-(, CP-$O_p$-), CP-$O_p$-⋆, CP-$O_p$-,, CP-$O_p$-., CP-$O_p$-:, CP-$O_p$-CC, CP-$O_p$-CD, CP-$O_p$-DT, CP-$O_p$-EX, CP-$O_p$-F, CP-$O_p$-FW, CP-$O_p$-IN, CP-$O_p$-J, CP-$O_p$-JJ, CP-$O_p$-JJR, CP-$O_p$-JJS, CP-$O_p$-LS, CP-$O_p$-MD, CP-$O_p$-N, CP-$O_p$-NN, CP-$O_p$-NNP, CP-$O_p$-NNPS, CP-$O_p$-NNS, CP-$O_p$-P, CP-$O_p$-PDT, CP-$O_p$-POS, CP-$O_p$-PRP, CP-$O_p$-PRP$, CP-O_p-R, CP-O_p-RB, CP-O_p-RBR, CP-O_p-RBS, CP-O_p-RP, CP-O_p-SYM, CP-O_p-TO, CP-O_p-UH, CP-O_p-V, CP-O_p-VB, CP-O_p-VBD, CP-O_p-VBG, CP-O_p-VBN, CP-O_p-VBP, CP-O_p-VBZ, CP-O_p-W, CP-O_p-WDT, CP-O_p-WP, CP-O_p-WP$, CP-$O_p$-WRB, CP-$O_p$-“, CP-STM-1, CP-STM-2, CP-STM-3, CP-STM-4, CP-STM-5, CP-STM-6, CP-STM-7, CP-STM-8, CP-STM-9, CP-STMi-2, CP-STMi-3, CP-STMi-4, CP-STMi-5, CP-STMi-6, CP-STMi-7, CP-STMi-8, CP-STMi-9 }

Table C.4: Metrics based on Constituency Parsing

NE = { NE-M_e-⋆, NE-M_e-ANGLE_QUANTITY, NE-M_e-DATE, NE-M_e-DISTANCE_QUANTITY, NE-M_e-LANGUAGE, NE-M_e-LOC, NE-M_e-METHOD, NE-M_e-MISC, NE-M_e-MONEY, NE-M_e-NUM, NE-M_e-ORG, NE-M_e-PER, NE-M_e-PERCENT, NE-M_e-PROJECT, NE-M_e-SIZE_QUANTITY, NE-M_e-SPEED_QUANTITY, NE-M_e-SYSTEM, NE-M_e-TEMPERATURE_QUANTITY, NE-M_e-WEIGHT_QUANTITY, NE-O_e-⋆, NE-O_e-⋆⋆, NE-O_e-ANGLE_QUANTITY, NE-O_e-DATE, NE-O_e-DISTANCE_QUANTITY, NE-O_e-LANGUAGE, NE-O_e-LOC, NE-O_e-METHOD, NE-O_e-MISC, NE-O_e-MONEY, NE-O_e-NUM, NE-O_e-O, NE-O_e-ORG, NE-O_e-PER, NE-O_e-PERCENT, NE-O_e-PROJECT, NE-O_e-SIZE_QUANTITY, NE-O_e-SPEED_QUANTITY, NE-O_e-SYSTEM, NE-O_e-TEMPERATURE_QUANTITY, NE-O_e-WEIGHT_QUANTITY }

Table C.5: Metrics based on Named Entities

SR = { SR-O_r, SR-O_{rv}, SR-N-$_v$, SR-O_v, SR-M_r-$\star$,
SR-M_r-A0, SR-M_r-A1, SR-M_r-A2, SR-M_r-A3,
SR-M_r-A4, SR-M_r-A5, SR-M_r-AA, SR-M_r-AM-ADV,
SR-M_r-AM-CAU, SR-M_r-AM-DIR, SR-M_r-AM-DIS,
SR-M_r-AM-EXT, SR-M_r-AM-LOC, SR-M_r-AM-MNR,
SR-M_r-AM-MOD, SR-M_r-AM-NEG, SR-M_r-AM-PNC,
SR-M_r-AM-PRD, SR-M_r-AM-REC, SR-M_r-AM-TMP,
SR-M_{rv}-$\star$, SR-M_{rv}-A0, SR-M_{rv}-A1,
SR-M_{rv}-A2, SR-M_{rv}-A3, SR-M_{rv}-A4,
SR-M_{rv}-A5, SR-M_{rv}-AA, SR-M_{rv}-AM-ADV,
SR-M_{rv}-AM-CAU, SR-M_{rv}-AM-DIR, SR-M_{rv}-AM-DIS,
SR-M_{rv}-AM-EXT, SR-M_{rv}-AM-LOC, SR-M_{rv}-AM-MNR,
SR-M_{rv}-AM-MOD, SR-M_{rv}-AM-NEG, SR-M_{rv}-AM-PNC,
SR-M_{rv}-AM-PRD, SR-M_{rv}-AM-REC, SR-M_{rv}-AM-TMP,
SR-O_r-$\star$, SR-O_r-A0, SR-O_r-A1,
SR-O_r-A2, SR-O_r-A3, SR-O_r-A4,
SR-O_r-A5, SR-O_r-AA, SR-O_r-AM-ADV,
SR-O_r-AM-CAU, SR-O_r-AM-DIR, SR-O_r-AM-DIS,
SR-O_r-AM-EXT, SR-O_r-AM-LOC, SR-O_r-AM-MNR,
SR-O_r-AM-MOD, SR-O_r-AM-NEG, SR-O_r-AM-PNC,
SR-O_r-AM-PRD, SR-O_r-AM-REC, SR-O_r-AM-TMP,
SR-O_{rv}-$\star$, SR-O_{rv}-A0, SR-O_{rv}-A1,
SR-O_{rv}-A2, SR-O_{rv}-A3, SR-O_{rv}-A4,
SR-O_{rv}-A5, SR-O_{rv}-AA, SR-O_{rv}-AM-ADV,
SR-O_{rv}-AM-CAU, SR-O_{rv}-AM-DIR, SR-O_{rv}-AM-DIS,
SR-O_{rv}-AM-EXT, SR-O_{rv}-AM-LOC, SR-O_{rv}-AM-MNR,
SR-O_{rv}-AM-MOD, SR-O_{rv}-AM-NEG, SR-O_{rv}-AM-PNC,
SR-O_{rv}-AM-PRD, SR-O_{rv}-AM-REC, SR-O_{rv}-AM-TMP,
SR-M_r-$\star$-b, SR-M_r-$\star$-i, SR-M_{rv}-$\star$-b, SR-M_{rv}-$\star$-i,
SR-O_r-$\star$-b, SR-O_r-$\star$-i, SR-O_{rv}-$\star$-b, SR-O_{rv}-$\star$-i }

Table C.6: Metrics based on Semantic Roles

DR = {	DR-O_r-$\star$, DR-O_r-alfa, DR-O_r-card, DR-O_r-dr, DR-O_r-drs, DR-O_r-eq, DR-O_r-imp, DR-O_r-merge, DR-O_r-named, DR-O_r-not, DR-O_r-or, DR-O_r-pred, DR-O_r-prop, DR-O_r-rel, DR-O_r-smerge, DR-O_r-timex, DR-O_r-whq, DR-O_{rp}-$\star$, DR-O_{rp}-alfa, DR-O_{rp}-card, DR-O_{rp}-dr, DR-O_{rp}-drs, DR-O_{rp}-eq, DR-O_{rp}-imp, DR-O_{rp}-merge, DR-O_{rp}-named, DR-O_{rp}-not, DR-O_{rp}-or, DR-O_{rp}-pred, DR-O_{rp}-prop, DR-O_{rp}-rel, DR-O_{rp}-smerge, DR-O_{rp}-timex, DR-O_{rp}-whq, DR-STM-1, DR-STM-2, DR-STM-3, DR-STM-4, DR-STM-5, DR-STM-6, DR-STM-7, DR-STM-8, DR-STM-9, DR-STMi-2, DR-STMi-3, DR-STMi-4, DR-STMi-5, DR-STMi-6, DR-STMi-7, DR-STMi-8, DR-STMi-9, DR-O_r-$\star$-b, DR-O_r-$\star$-i, DR-O_{rp}-$\star$-b, DR-O_{rp}-$\star$-i, DR-STM-4-b, DR-STM-4-i }

Table C.7: Metrics based on Discourse Representations

Index