Hybrid Data-Driven Models of Machine Translation

Declan Groves

A dissertation submitted in fulfilment of the requirements for the award of

Doctor of Philosophy (Ph.D.)

to the

Dublin City University
School of Computing

Supervisor: Prof. Andy Way

January 2007

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ph.D. is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: Declan Lyons (Candidate) ID No.: 99290073

Date: 29/01/07

Contents

Abstract

Corpus-based approaches to Machine Translation (MT) dominate the MT research field today, with Example-Based MT (EBMT) and Statistical MT (SMT) representing two different frameworks within the data-driven paradigm. EBMT has always made use of both phrasal and lexical correspondences to produce high-quality translations. Early SMT models, on the other hand, were based on word-level correpsondences, but with the advent of more sophisticated phrase-based approaches, the line between EBMT and SMT has become increasingly blurred.

In this thesis we carry out a number of translation experiments comparing the performance of the state-of-the-art marker-based EBMT system of Gough and Way (2004a, 2004b), Way and Gough (2005) and Gough (2005) against a phrase-based SMT (PBSMT) system built using the state-of-the-art PHARAOH phrase-based decoder (Koehn, 2004a) and employing standard phrasal extraction heuristics (Koehn et al., 2003). In addition, we describe experiments investigating the possibility of combining elements of EBMT and SMT in order to create a hybrid data-driven model of MT capable of outperforming either approach from which it is derived.

Making use of training and testing data taken from a French–English translation memory of *Sun Microsystems* computer documentation, we find that while better results are seen when the PBSMT system is seeded with GIZA++ word- and phrase-based data compared to EBMT marker-based sub-sentential alignments, in general improvements are obtained when combinations of this 'hybrid' data are used to construct the translation and probability models. While for the most part the baseline marker-based EBMT system outperforms any flavour of the PBSMT systems constructed in these experiments, combining the data sets automatically induced by both GIZA++ and the EBMT system leads to a hybrid system which improves on the EBMT system *per se* for French–English.

On a different data set, taken from the Europarl corpus (Koehn, 2005), we perform a number of experiments making use of incremental training data sizes of 78K, 156K and 322K sentence pairs. On this data set, we show that similar gains are to be had from constructing a hybrid 'statistical EBMT' system capable of outperforming the baseline EBMT system. This time around, although all 'hybrid' variants of the EBMT system fall short of the quality achieved by the baseline PBSMT system, merging elements of the marker-based and SMT data, as in the *Sun Microsystems* experiments, to create a hybrid 'example-based SMT' system, outperforms the baseline SMT and EBMT systems from which it is derived. Furthermore, we provide further evidence in favour of hybrid data-driven approaches by adding an SMT target language model to all EBMT system variants and demonstrate that this too has a positive effect on translation quality.

Following on from these findings we present a new hybrid data-driven MT architecture, together with a novel marker-based decoder which improves upon the performance of the marker-based EBMT system of Gough and Way (2004a, 2004b), Way and Gough (2005) and Gough (2005), and compares favourably with the state-of-the-art PHARAOH SMT decoder (Koehn, 2004a).

Acknowledgements

I would like to start by thanking my supervisor, Andy Way, who has always been approachable and has offered support and encouragement throughout my Ph.D. studies. Thank you to IRCSET who have generously provided me with a Ph.D. Fellowship Award, enabling me to carry out the work contained in this thesis.

Thanks to everyone I've had the pleasure of working with at the National Center for Language Technology in DCU Thanks to Ruth, Mick, Michelle, Ríona, Sara, Karolina, Bart, Stephen and everyone else who has helped make my time in DCU all that more enjoyable. Thanks to Nano for giving me access to the EBMT system used in this work and to Nicolas for answering all my probability questions. I thank Mary who I've really enjoyed working with and has always been more than willing to offer advice whenever I needed it, and for making my time in Baltimore all that more bearable!

I want to thank the great group of people I had the pleasure of working with at Johns Hopkins University in Summer 2005, in particular Anna, for her constant enthusiasm, and Ben for his complete down-to-earth attitude to work, and his endless juggling and comedic talents.

There are numerous people outside of DCU who have been there to remind me there is life beyond this thesis. To all my friends, David, MacDara, John, Eoin, Daire, Conor, Paddy, Deco and the rest of the gang who have been around for longer than I can remember. To Ciarán and Richard for our numerous, and much needed, coffee breaks. Thanks to Vicky, Orla and Louise. Special thanks goes to Dee for being the kind of friend everyone wishes they had, for reminding me to let go every now and then and for always understanding and supporting me.

I also want to thank my family; my Nanny Eileen, my brother Ciarán and my sister Claire, and especially my parents, Martin and Goretti, for giving me the ambition to take my education this far, for their unconditional support, both emotional and financial, and for their constant encouragement.

Finally, to Alan. You have been unbelievably patient and have offered me more support that I was willing (and was too stubborn) to ask for. Thanks for reminding me to relax, for understanding and for making me laugh. I could not have done this without you.

Chapter 1

Introduction

> "...translation is a fine and exacting art, but there is much about it that is mechanical and routine" Kay (1997)

Machine Translation (MT) research has come a long way since the idea to use computers to automate the translation process was first outlined in Warren Weaver's historical memorandum in 1949 (Weaver, 1949). Originally proposed as a fully automated solution to the problem of translating from one natural language into another, it is generally agreed today that, despite their inherent problems caused by the difficulty of the translation task, MT systems can be used to produce output of sufficiently high quality which can greatly reduce the level of human post-editing effort required.

The very first MT systems were based on the direct approach, where the input, or source language sentence is directly processed and converted into the target language output, through the use of a bilingual lexicon. From this early 'brute-force' approach, the paradigm has shifted towards today's dominance of corpus-based approaches. Approaches within the corpus-based paradigm provide an alternative to such early first generation approaches and also to the more advanced second generation MT systems, such as rule-based MT (RBMT) and interlingua approaches. The ever increasing popularity of data-driven MT has been facilitated by the increase in computational power, inexpensive storage and more widely available, machine-

readable bilingual parallel text. The data-driven paradigm can be further broken down into two main approaches: that of Statistical Machine Translation (SMT) and Example-Based Machine Translation (EBMT).

The first word-based models for SMT of Brown et al. (1988) introduced a linguistic-free approach to MT, using only a set of parameters induced from bilingual corpora to perform translation. One huge advantage these SMT systems had over previous MT strategies was their language-independent nature; the only thing a system required to translate a new language pair was, essentially, a bilingual corpus for that particular language pair. This was a great benefit over older transfer-based systems which, although they could produce high-quality translation output, required new transfer modules to be created, not only for each new language pair that was being translated, but also for each new language direction. The early word-based models of SMT (Brown et al., 1990) were quite crude and often produced word salad due to their lack of contextual and syntactic information. However, more recent SMT researchers now make use of phrase-based approaches, resulting in considerably higher translation quality than that of the traditional word-based models (Koehn et al., 2003).

Example-based models of translation, on the other hand, have always made use of both phrasal and lexical correspondences during translation since their very inception (Nagao, 1984), but have been widely ignored by SMT practitioners who focus on the inductive rather than analogical approach. Rather than using models of syntax in a *post hoc* fashion, as is the case with most SMT systems, most EBMT models of translation build in syntax at their core, during both the creation of their knowledge resources and during the translation process itself. However, EBMT systems still tend to suffer from problems of coverage and seem to perform better in controlled language (Way and Gough, 2005b) or sublanguage domains (Sumita et al., 1990).

With the advent of phrase-based SMT systems, the line between EBMT and SMT has become increasingly blurred. It is surprising, therefore, that apart from

the work of Way and Gough (2005a)[1] no substantial research has been performed into comparing these empirical approaches. In this thesis, we wish to compare these approaches both from a theoretical and practical point of view, for the benefit of both SMT and EBMT research communities and to determine whether in fact they constitute two different paradigms, or are, in fact, one and the same. This presents us with our first research question, (RQ1):

(RQ1) *How do state-of-the-art EBMT and SMT methods compare, both theoretically, and also in terms of their translation performance?*

To answer the question in (RQ1), we carry out empirical evaluations on the task of sublanguage translation as well as the translation of text taken from the more commonly used, relatively open domain, Europarl corpus (Koehn, 2005).

One of the main differences between SMT and EBMT lies in the techniques they employ to extract the resources used during the translation process EBMT tends to make use of more linguistically motivated approaches to gain its translation knowledge, such as the marker-based approaches described in this thesis which employ minimal surface syntactic information (Veale and Way (1997); Way and Gough (2003, 2005a, 2005b); Gough and Way (2004a, 2004b); Gough (2005)). SMT, on the other hand, often makes use of only probabilistic information to extract word-level alignments which then feed phrase extraction heuristics which essentially operate on n-grams (Koehn et al., 2003), rather than phrases *per se* (Groves and Way, 2006a). In this work we wish to compare the types of translation resources extracted by both methods.

Additionally, today, more and more MT systems tend to employ hybrid approaches. Hybrid approaches to MT adopt elements from many different MT paradigms, the benefits of which have been well documented; being able to select the best techniques from various paradigms has clear advantages. In previous research example-based methods have been combined with more 'rationalist' approaches to MT, such

[1] And until our work outlined in Groves and Way (2005) and Groves and Way (2006a, 2006b), which form the basis of Chapter 3 and Chapter 4, respectively in this thesis.

as transfer-based, to produce hybrid (e.g. Sumita et al. (1990); Watanabe (1992)) or multi-engine MT systems (e.g Frederking and Nirenburg (1994); Frederking et al. (1994)), as has SMT (e.g. Chen and Chen (1995); Hassan et al (2006), Marcu et al. (2006)). In addition to comparing the types of translation resources extracted by both methods, we also wish to investigate the possible combination of EBMT and SMT data-driven resources, thus leading us to our second research question:

(RQ2) *What contributions do EBMT and SMT translation resources make to the quality of translations produced by an MT system and can EBMT and SMT approaches to translation be combined into a hybrid data-driven model of MT?*

Consequently, we present a number of experiments which investigate whether either approach can benefit from hybrid data-driven techniques

In addition to investigating the use of hybrid data-driven translation resources, we describe further work that has been carried out into the creation of a hybrid data-driven MT architecture, together with the development of a novel hybrid data-driven decoder, which can take full advantage of hybrid EBMT-SMT techniques. The aim of this additional work is to see·

(RQ3) *Can a novel hybrid data-driven decoder take advantage of EBMT approaches, together with SMT search strategies and probabilistic models to improve translation results?*

For the range of experiments reported in this work, it must be noted that all variants of the phrase-based SMT system were left untuned and employed translation models making use of relative frequency information alone. Due to the extensive number of comparative experiments we present in this work, tuning was not feasible and making use of standard model weights allowed us to produce easily replicable results.

1.1 Thesis Outline

The remainder of the thesis is broadly organised as follows. In Chapter 2 we describe the EBMT and SMT models of translation, together with supporting background research and details of the particular approaches used in our work. Chapters 3 and 4 outline a number of experiments comparing the performance of EBMT and SMT, as well as experiments involving various hybrid data-driven models. Finally, in Chapter 5 we describe new and ongoing research into the area of hybrid models of data-driven MT. The following gives a more detailed outline of what we present in this thesis:

Chapter 2 : In this chapter we give a general outline of the two main data-driven approaches to MT: EBMT and SMT. We describe the main processes carried out when performing EBMT and outline the particular approach to EBMT used in our work (Groves and Way, 2005, 2006a, 2006b), based on the Marker Hypothesis (Green, 1979). For SMT, we introduce the translation and language models used within the SMT framework and describe the state-of-the-art phrase-based approach (Koehn et al., 2003) and give details of the phrasal extraction algorithm used in our work (Och and Ney, 2003). In this chapter we also compare EBMT and SMT from a theoretical point of view, attempting to situate the marker-based approach and phrase-based SMT in terms of other related approaches.

Chapter 3: In this chapter we describe a number of experiments comparing the performance of state-of-the-art SMT against state-of-the-art EBMT on a French–English sublanguage corpus. Hybrid approaches to MT are becoming ever more popular, therefore in addition to this comparative work we discuss relevant hybrid approaches to MT and describe experiments making use of a hybrid 'example-based SMT' system, created by combining both EBMT and SMT-induced bilingual alignments and feeding the resulting data set to the SMT system (cf. Groves and Way, 2005).

Chapter 4: In this chapter we compare the performance of EBMT and SMT on a larger, wider domain data set. We investigate the effect of increasing the training set size on translation performance. We perform similar hybrid 'example-based SMT' experiments as carried out in Chapter 3 and perform a number of hybrid 'statistical EBMT' experiments by supplying the EBMT system with combinations of the EBMT and SMT-induced bilingual alignments. In an additional step towards full EBMT-SMT integration, we make use of a statistical language model to investigate whether this can further improve the performance of the EBMT system. In order to fully understand the contributions the EBMT and SMT translation data make to the overall translation performance, we investigate and compare the types of phrasal alignments induced via EBMT and SMT methods (cf. Groves and Way, 2006a, 2006b).

Chapter 5: In this thesis we demonstrate how making use of both EBMT and SMT knowledge sources contribute positively to translation quality. Consequently, in this chapter we describe a new hybrid data-driven architecture which we have developed to enable further research into hybrid data-driven models of MT. We outline a new chunk alignment strategy, and using this new architecture, we evaluate the alignment strategy on a new language pair and larger data set, thus testing the scalability of the system and the adaptability of the marker-based approach to chunking (cf. Armstrong et al., 2006). In addition, we perform further experiments on the difficult task of Basque–English translation, and perform manual evaluation of phrasal alignments to determine the extent to which EBMT techniques aid the translation process (cf. Stroppa et al., 2006). Finally we outline some preliminary experiments which have been carried out into the development of our marker-based decoder.

Chapter 6: In this chapter we present the conclusions of our work and suggest a number of avenues for possible future research.

Chapter 2

State of the Art in Data-Driven MT

Although the memorandum of Weaver (1949) first brought the idea of Machine Translation (MT) to the attention of the general research community, the idea of using mechanical approaches to translate languages was first suggested as early as the 17th century, with the use of mechanical dictionaries.

Today, the field of MT research is largely dominated by corpus-based, or data-driven approaches. The two main data-driven approaches of MT are Example-Based MT (EBMT) and Statistical Machine Translation (SMT). The corpus-based paradigm provides an alternative to earlier first ('direct') and second generation (e.g. rule-based (RBMT)) MT systems, in that approaches within the corpus-based paradigm are largely empirical, making use of bilingual aligned corpora as a basis to the translation process. Ever increasing computational power, inexpensive storage and more widely available, machine-readable bilingual parallel corpora have enabled the processing of the enormous amounts of data required by these data-intensive methods for good quality translation to ensue.

During the 1980s, MT research was predominantly concerned with the use of linguistic rules for analysis, most immediately apparent in transfer-based systems, but also present in interlingua systems. However, the shortcomings of these approaches,

such as the problems of defining a true interlingua and the cost of developing rules for transfer-based systems, led researchers to look at empirical approaches. During this time, the field of MT research experienced a paradigm shift where new corpus-based MT methods emerged. This time saw a revival of statistical methods, with researchers borrowing ideas heavily from the quickly developing Speech Processing community (Brown et al., 1988). At the same time, research involving the use of examples emerged from groups in Japan and these experiments came to be known as EBMT (Nagao, 1984).

The attractiveness of such corpus-based approaches, in particular SMT, was their ability to perform translation without the need of explicit linguistic information. This meant that systems could be developed relatively quickly and inexpensively compared to previous costly transfer-based approaches.

In this chapter we outline the two main data-driven approaches to MT: EBMT and SMT. In Section 2.1 we describe the EBMT approach, including an outline of how knowledge resources are represented and exploited in EBMT systems, while in Section 2.2 we describe the marker-based approach which is the particular instantiation of EBMT used in our work. The SMT framework is discussed in Section 2.3, with particular reference to the recent shift towards phrase-based SMT models such as those employed by the SMT system used in our experiments discussed in later chapters of this thesis (cf. Chapter 3 and Chapter 4). We conclude the chapter by comparing both data-driven approaches in Section 2.4, which, despite becoming increasingly similar, still retain a number of fundamental differences.

2.1 Example-Based Machine Translation

Nagao (1984)[1] was the first to outline the example-based approach to MT, or "machine translation by example-guided inference" stating:

"Man does not translate a simple sentence by doing deep linguistic anal-

[1] Although Nagao first presented his ideas at a conference in 1981, it was not until 1984 that they were published.

> ysis, rather, man does translate, first, by properly decomposing an input sentence into certain fragmental phrases, ... then by translating these phrases into other language phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference" (Nagao, 1984:178f.)

EBMT implements the idea of machine translation by the analogy principle and is based on the intuition that humans construct translations for new unseen input by making use of previously seen translation examples, rather than performing "deep linguistic analysis".

A prerequisite for the induction of sub-sentential fragments in EBMT is a bilingual corpus (or 'bitext') of sententially-aligned examples, such as the Europarl corpus (Koehn, 2005). Assuming such a corpus of aligned source–target sentence pairs, EBMT models of translation perform three distinct processes in order to transform a new input string into a target language translation:

- searching the source side of the bitext for 'close' matches and retrieving their translations;
- determining the sub-sentential translation links in those retrieved examples;
- recombining relevant parts of the target translation links to derive the final translation.

The EBMT model shares similarities in structure with that of the three-stage transfer-based RBMT model. The transfer-based model is made up of three stages: analysis, transfer and generation, as illustrated in Figure 2 1 taken from Somers (2003). In EBMT the search and matching process replaces the analysis stage, transfer is replaced by the extraction and retrieval of examples and recombination takes the place of the generation stage (Somers, *op cit*). However, in Figure 2.1, 'direct translation' does not correspond exactly to 'exact match' in EBMT, as exact

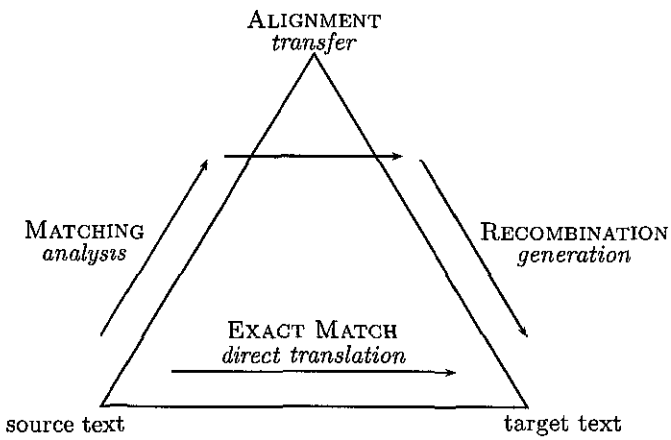


Figure 2.1: The 'Vauquois pyramid' adapted for EBMT (taken from Somers (2003) Figure 1.1). The traditional labels are shown in italics, while the EBMT labels are shown in capitals.

match is a perfect translation and does not require any adaption at all, unlike direct translation.

An illustration of how an EBMT system performs translation is given in Figure 2.2. Here we can see that when translating the source string S into the target string T, the system first searches through the source side of the bitext, selecting the examples $\hat{s}_1$ and $\hat{s}_2$ from the corpus as close matches for S. The equivalent target language fragments $\hat{t}_1$ and $\hat{t}_2$ for the retrieved examples are then extracted and fed to the recombination process to create the final translation T.

To further illustrate the EBMT process, consider that we wish to translate the sentence *John went to the baker's on Monday* and we have the corpus in (1), consisting of just 3 simple sentences:

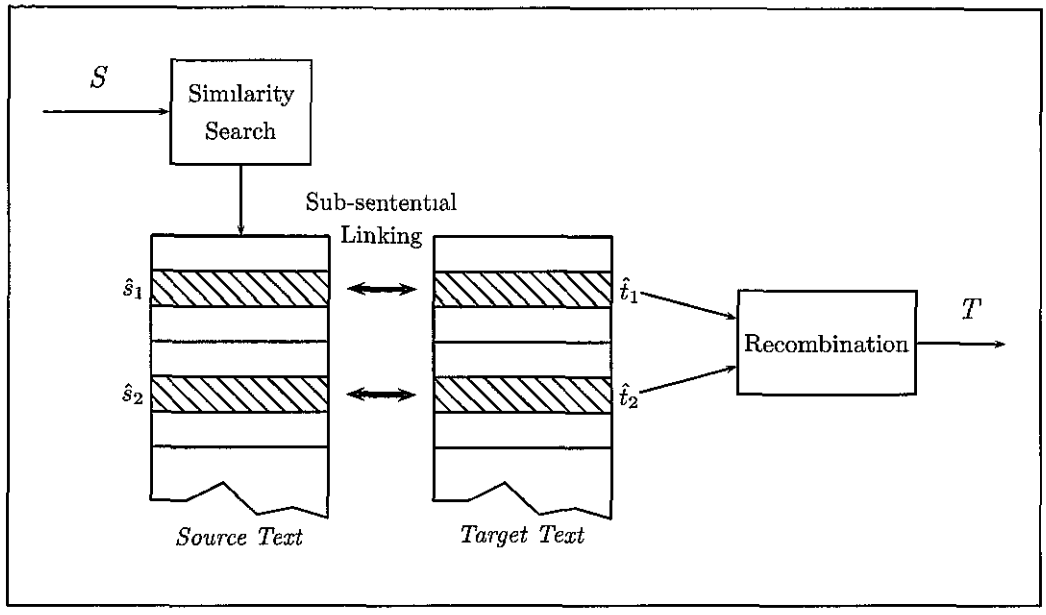


Figure 2.2· The EBMT Translation Process

(1) The butcher's is next to the baker's ⇔La boucherie est à côté de la boulangerie

The shop is open on Monday ⇔Le magasin est ouvert lundi

John went to the swimming pool ⇔Jean est allé à la piscine

Taking the sentences in (1) and applying a bilingual fragment extraction algorithm such as that of Nirenburg et al. (1993) or Somers et al. (1994), we can then identify and extract the useful bilingual fragments given in (2), using a very simple subsequence similarity measure, such as the Levenshtein distance (Levenshtein, 1965), during the matching process.

(2) the baker's ⇔la boulangerie

on Monday ⇔lundi

John went to ⇔Jean est allé à

We can then combine the fragments in (2) to produce a translation for the new input sentence as shown in (3):

(3) John went to the baker's on Monday ⇔Jean est allé à la boulangerie lundı

Note that the sentence pair in (3) did not appear in the original corpus in (1). The sentence pair in (3) can now be added to the example base so that if this same source sentence is encountered subsequently it can simply be retrieved in its entirety via exact sentence matching and the corresponding target language translation output, thus by-passing the recombination step.

The production of the translation ın (3) illustrates how example-based models of translation concentrate their operations at the phrasal level, followıng Becker's theory of language production (Becker, 1975). Instead of producing language by combining words or morphemes, Becker (1975) identified phrases as the building blocks of language which can be combined together to generate new expressions, similar in spirit to how EBMT operates. In EBMT, phrasal units are looked up in a phrasal lexicon (either created at run-time or during a pre-processing stage, as in the marker-based approach used in our work (cf. Section 2.2)) and translated by combining already translated phrases stored in this lexicon, very much along the lines proposed originally by Becker, and applied by Schäler (1996). The use of such sub-sentential phrasal information enables EBMT systems to be particularly useful for capturing complex translation relations, such as idiomatic expressions, and as Cranias et al. (1994) point out, the potential of EBMT relies on this ability to exploit smaller sub-sentential units. Phrases also lend themselves more easily to the matching and subsequent translation process while still minimizing the risk of increasing the level of ambiguity during both stages.

The level of granularıty of examples is ımportant for both the matching and translation stages and selecting the correct level represents a significant problem within itself (Nırenburg et al., 1993). If examples are too long, we decrease the likelihood of finding a complete match within the database of examples. On the other hand, if examples are too short we increase the level of translation ambiguity. For instance, making use of only word-level alignments results in examples that

are too fine-grained; word-based translation is particularly ambiguous due to the many possible translation choices available for individual words and such word-for-word direct translation has been shown to perform very poorly in practice. Using too fine-grained examples, such as those consisting only of words, also increases the risk of encountering boundary friction during the recombination process (Way, 2003) resulting in increased errors in the output. Using shorter examples also introduces problems of coverage as the EBMT system needs to select the best cover of an input text in terms of a set of matched sub-sentential fragments, which is particularly problematic if there are many possible choices available (Maruyama and Watanabe, 1992). In addition, evidence from translation studies suggests that, in practice, human translators work with units shorter than the sentence but longer than individual words (Gerloff, 1987) which also further strengthens the appeal of the example-based approach.

2.1.1 Similarity Search & the Representation of Examples

During the initial search process, the matching algorithm and corresponding similarity metric used by an EBMT engine heavily depend on the form of the examples exploited by the system.

Where examples are stored as simple strings (e.g. Somers et al., 1994), character-based distances may be employed. The problem of determining the distance between one string of characters and another is analogous to the edit-distance problem (Wagner and Fischer, 1974). Determining such string distances can be carried out using well-established dynamic programming techniques, such as that employed by the commonly used Levenshtein distance algorithm (Levenshtein, 1965). String-based metrics, although easily implemented, have their disadvantages, as often candidate strings closer in meaning to the source string that we are attempting to match would be overlooked in favour of less desirable matches, due to a larger edit-distance score.

Taking the example in (4), if we consider (4a) to be the sentence we are attempting to match, using just a character-based distance the system would choose (4b),

rather than the more favourable match in (4c) due to the smaller distance between *agrees* and its antonym *disagrees*, than between *agrees* and its synonym *concurs*

(4) a. The President agrees with the decision.

b. The President disagrees with the decision.

c. The President concurs with the decision.

In order to address such problems, along with string-based distances, many EBMT systems make use of a word-based measure of similarity, employing information from dictionaries and thesauri to determine relative word distances in terms of a semantic hierarchy (e.g. Nagao, 1984; Sumita et al., 1990). Using such matching techniques would correctly select (4c) in the above example as being the better match due to the closer relative semantic distance between *concurs* and *agrees* in (4a). This type of matching is particularly useful where we have competing examples, as in the examples of Nagao (1984). Given the examples in (5), the system correctly produces the Japanese translation of the English verb *eats* as *taberu* (eats food) in the sentence in (6) due to the semantic relationship between *A man* and *He*, and between *vegetables* and *potatoes*.

(5) a. A man eats vegetables ⇔Hito wa yasai o taberu

b. Acid eats metal ⇔San wa kinzoku o okasu

(6) He eats potatoes ⇔Kare wa jagaimo o taberu

In many EBMT systems, similar examples are collected to create variabilised translation templates, where patterns are created and stored by replacing elements of chunks with general substitution variables. Using these generalised template patterns is similar in spirit to rules used in traditional RBMT systems and increases the flexibility of the matching process.

Kaji et al. (1992) generalise by syntactic category, employing source and target language parsers and aligning syntactic units aided by a bilingual dictionary. Taking

the template example in Figure 2.3, the generalised examples in (7a) and (7b) can be created by replacing coupled pairs by variables incorporating information about their syntactic categories

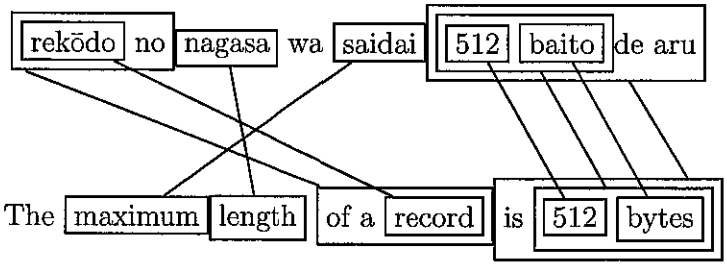


Figure 2 3: Aligned example from Kaji et al. (1992), with coupled Japanese–English word and phrase pairs identified by corresponding links

(7) a. X[NP] no nagasa wa saidai 512 baito de aru ⇔

The maximum length of X[NP] is 512 bytes.

b. X[NP] no nagasa wa saidai Y[N] baito de aru ⇔

The maximum length of X[NP] is Y[N] bytes.

For generalising examples, Brown (1999a) uses placeables, essentially special tokens indicating word class, replacing certain words which are members of a particular equivalence class, such as PERSON, CITY and TIME, by these placeables to create templates. Taking the example sentence in (8a), Brown (1999a) generalises by recursively replacing tokens by their equivalence classes, giving (8b) in the first pass and the final template (8c) in the second pass.

(8) a. John Miller flew to Frankfurt on December 3rd

b. <FIRSTNAME-M> <LASTNAME> flew to <CITY> on <MONTH> <ORDINAL>.

c. <PERSON-M> flew to <CITY> on <DATE>.

The template in (8c) can match any sentence that follows this pattern, such as *Mary Byrne flew to Dublin on January 20th*, by replacing each equivalence class

with an instance of the class. More recent work by Brown (2000) involves making use of clustering techniques to identify equivalence classes.

Approaches based on more traditional EBMT models make use of structured examples consisting of annotated tree structures (e.g. Sata and Nagao, 1990: Watanabe, 1992; Matsumoto et al., 1993) and therefore employ more complex tree-structure matching During the matching process, parts of the parsed input string are matched against the annotated fragments and the corresponding annotated target fragments are extracted and recombined to produce the final translation.

A more recent structure-based approach is that of Data-Oriented Translation (DOT) (Poutsma, 2000, 2003), based on Data-Oriented Parsing (Bod, 1992), which exploits a bilingual treebank consisting of parsed source and language trees with explicit node links indicating translational equivalence (Groves et al., 2004). The DOT methods of Hearne and Way (2003,2006) and Hearne (2005) combine examples, statistics and linguistics to perform simultaneous translation and parsing, and thus can be considered a hybrid approach to MT. An example DOT representation for the English–French sentence pair in (9) is given in Figure 2.4 (Figure 4.2 in Hearne (2005)).

(9) press and release the left button.
⇔exercez une pression brève sur le bouton de gauche.

Way (2003) presents an extension to the DOT framework where examples take a similar form to those in DOT, but also include functional information present in Lexical Functional Grammar (LFG) f-structures (Kaplan and Bresnan, 1982), particularly useful in reducing problems of boundary friction. In LFG-DOT, representations consist of linked c-structures (constituent structures), which are linked phrase-structures (as in Figure 2.4 for DOT) together with f-structures (functional structures) – attribute value matrices with information about grammatical relations contained within the string in question. Representations also contain mappings be-

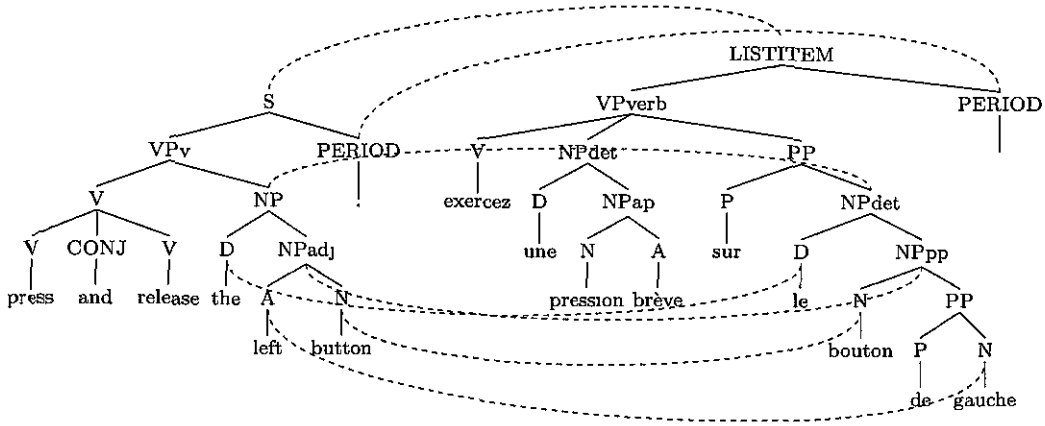


Figure 2.4: Linked source and target language phrase-structure trees in DOT, where the links between source and target nodes indicate that the substrings dominated by these nodes are translationally equivalent.

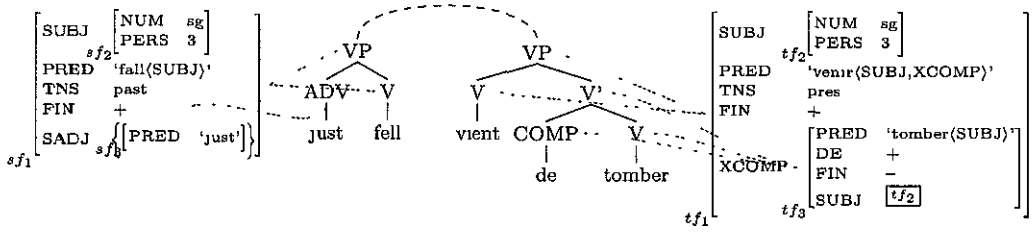


Figure 2.5: An example LFG-DOT representation for an English–French fragment with ϕ links between c-structure and f-structure nodes

tween nodes within the c-structures and f-structures, known as ϕ links. An example of an LFG-DOT representation, taken from Example 8.8 in Hearne (2005), for the English–French verb fragment pair *just fell* ⇔*vient de tomber* is given in Figure 2.5.

The examples used in the work of Planas and Furuse (2003) make use of even deeper linguistic information, employing TELA (*Treillis Etagés et Liés pour le traitement Automatique*, i.e. Floored and Linked Lattices for Automatic Processing) multi-layered structures, which include information such as word, lemma, POS and syntactic structure information together with additional knowledge The methods of Planas and Furuse (2003), although developed for use in translation memories, can also be applied to EBMT. During the matching process each layer of both TELA structures under consideration are compared using an adaptation of the

edit-distance algorithm (Wagner and Fischer, 1974) implemented by a Multi-level Similar Segment Matching (MSSM) algorithm. Firstly each layer in the structure is checked in terms of equality and the resulting score is then added to the overall deletion and insertion scores over all layers in the TELA structure.

Rather than making use of deep syntactic information, alternative approaches to EBMT make use of more superficial syntactic information in their examples, namely part-of-speech (POS) tags, such as the ReVerb system of Collins (1998) which also makes use of explicit links between corresponding chunks and functional syntactic information. One approach which makes use of surface syntactic information during the creation of its examples, is the marker-based approach used in our work and which we describe in more detail in Section 2.2.

2.1.2 Recombination

Once suitable source language subsequences have been identified, the corresponding target language fragments are extracted and are combined, adapted and manipulated where appropriate, in order to generate the output translation. This final recombination process is perhaps the most difficult task for any EBMT system and heavily depends on the nature of the examples used in the first place. We first need to identify the parts of the extracted fragments which are relevant for producing the translation of the input string, the difficulty of which depends on the level of granularity of the examples. In some cases (as with the marker-based approaches described in Section 2.2), more fine-grained examples are created in a preprocessing step and it is this set of examples that are retrieved during the matching process and are applied intact during the recombination stage. This is one of the advantages of using such pre-computed chunk alignments. In other cases, this type of alignment between retrieved source and target examples is performed at run-time.

For the majority of EBMT systems, one of the main problems that arises during recombination is that of boundary friction (Nirenburg et al., 1993; Way, 2003). Boundary friction can occur at the meeting of two phrasal translations, where we

observe cases of disfluency. These disfluencies often take the form of agreement and case errors. Revisiting the example corpus in (1), we can extract the fragments given in (10).

(10) the shop ⇔le magasin
on Monday ⇔lundi
John went to ⇔Jean est allé à

Given the fragments in (10), we can attempt to produce a translation for a new input sentence, as given in (11):

(11) John went to the shop on Monday ⇔Jean est allé à le magasin lundi

Unlike the previous example in (3), when attempting to translate the sentence in (11) we encounter the problem of boundary friction, failing to produce the correct French contracted form of the English preposition *to* - producing *à le* instead of *au*

For structure-based EBMT approaches, boundary friction does not occur as often, as the linguistic information contained within the structures themselves helps to ensure that translations produced are grammatical, as shown for LFG-DOT by Way (2003). The recombination stage is less problematic for such EBMT systems, where the problem presents itself as a tree or graph unification problem (Hearne and Way, 2003, 2006; Watanabe et al., 2003; Hearne, 2005). It is in these cases that EBMT seems most closely related to RBMT.

Some approaches, such as that of Somers et al. (1994), take advantage of additional contextual information to help inform the recombination process. As example chunks are generally extracted from larger segments of texts, we can hold on to information about the original context in which the chunks occurred. Somers et al. (1994) make use of "hooks" which indicate words and POS tags that can occur before and after the chunk, according to the chunk's original context. When combining chunks, the most probable hook connections can help indicate the most likely chunk combination.

The work of Brown et al. (2003) and Carbonell et al. (2006) offers an alternative solution to help prevent disfluencies in the target language output, similar in spirit to the idea of using "hooks" In their work, during translation matched fragments retrieved by the EBMT engine are placed into a lattice. Translation proceeds by finding a path through the lattice that combines the retrieved fragments but favours overlapping fragments. Those target fragments that overlap with the previous and following fragments are contextually anchored both left and right, and thus are less likely to cause instances of boundary friction in the final translation To make the best use of the contextual information of the surrounding fragments, during the recombination process the target language fragments that have the maximal overlap are combined to produce the output translation.

2.2 Marker-Based EBMT

As mentioned in Section 2.1, one of the main advantages of EBMT methods is their ability to make use of phrasal information in their matching and recombination processes (Cranias et al., 1994). The extraction of such information can be performed at run-time (such as in the more traditional EBMT systems of Nagao (1984), Sumita et al. (1990) and Sumita (2003)). Alternatively this information can be extracted during a preprocessing step. One approach for extracting such information is to use a set of closed-class words to segment aligned source and target sentences and to derive an additional set of lexical and phrasal resources. In this section we give details of such techniques employed by the EBMT system used in our work.

The work of Veale and Way (1997) and Way and Gough (Way and Gough, 2003, 2005a, 2005b; Gough and Way, 2004a, 2004b; Gough, 2005) is based on the 'Marker Hypothesis' (Green, 1979). The Marker Hypothesis is a universal psycholinguistic constraint which posits that languages are 'marked' for syntactic structure at surface level by a closed set of specific lexemes and morphemes:

> "The Marker Hypothesis states that all natural languages have a closed

set of specific words or morphemes which appear in a limited set of grammatical contexts and which signal that context" (Green, 1979)

As an example, consider the sentence in (12), randomly selected from the *Wall Street Journal* section of the Penn-II Treebank (Marcus et al., 1994):

(12) The Dearborn Mich., energy company stopped paying a dividend in the third quarter of 1984 because of troubles at its Midland nuclear plant.

In this example, out of the underlined noun phrases (NPs), three begin with determiners, and one with a possessive pronoun (the words contained within boxes in (12)). The sets of determiners and possessive pronouns are both very small, and predict that some nominal element will (usually) occur in the right-context of these closed-class items and act as the head of those phrases In addition, there are four prepositional phrases, each headed by a preposition, and which (generally) indicate that some time soon thereafter an NP will be encountered which acts as the object of that preposition. Note also that the set of prepositions is similarly small.

For the experiments reported in this thesis (Chapter 3 and Chapter 4), we use the same 7 sets of closed-class (or 'marker') words for English and French as in this related work.[2] The list of marker word categories and their associated tags are given in Table 2.1.

Determiners	<DET>
Quantifiers	<QUANT>
Prepositions	<PREP>
Conjunctions	<CONJ>
WH-Adverbs	<WH>
Possessive Pronouns	<POSS>
Personal Pronouns	<PRON>

Table 2.1: The set of marker categories and their associated labels

In a preprocessing stage, the source–target aligned sentences in the parallel corpus are segmented at each new occurrence of a marker word, subject to the constraint

[2]In Section 5.1.1 we describe more recent work which additionally makes use of punctuation information as markers

that each 'marker chunk' contains at least one non-marker (or 'content') word (as will be explained in Section 2.2.1). This constraint ensures that the derived chunks are not too fine-grained and consist of more than just function words.

After performing this initial segmentation stage, sets of bilingual chunks are created by aligning based on marker tags and relative source and target chunk positions. Source–target chunks assigned the same marker category and which occur in similar positions within the bilingual sentences are considered for alignment. In addition, cognate and mutual information (MI) scores are employed to help guide the chunk alignment algorithm. The identification of cognates is implemented using the Levenshtein distance algorithm (Levenshtein, 1965), with pairs of <source,target> words with a distance below an empirically set threshold considered to be cognates. MI scores for word pairs are collected over the entire corpus and are also used to help inform the alignment process. A <source,target> word pair with a high MI score indicates that they co-occur frequently within the corpus, whereas a low MI score indicates such co-occurrence is infrequent and that the words in question are in complimentary distribution, with an MI score < 0 indicating that the words do not co-occur together anywhere within the corpus. The formula for calculating the MI score for a particular word pair $< x, y >$ is given in Equation (2.1) (Church and Hanks, 1990):

$$MI(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)} = \log_2 \frac{Nf(x,y)}{f(x)f(y)} \tag{2.1}$$

In Equation (2.1), $P(x)$ and $P(y)$ are calculated by counting the number of times source word x and target word y occur in the corpus and normalising by N, the size of the source language corpus. $P(x,y)$ joint probabilities are estimated by counting over the entire corpus the number of time x and y occur in the same aligned sentence pair ($f(x,y)$) and normalising by N.

When aligning chunks for use in the marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) employed in our work,

the MI scores for <source,target> word pairs within the chunks under consideration are examined. Links are then created between source and target words, with the restriction that the MI score for a particular word pair under consideration is higher than the MI score for all alternative word pairs containing the same target word. During the alignment process if a chunk pair under consideration has the same marker category, but does not share any lexical equivalences according to MI scores and cognate information, they are not linked. Similarly, if a chunk pair does not have the same marker category, but share many lexical equivalences, they can be aligned based on MI scores and cognate information alone.

2.2.1 Creation of the Example Database

In order to describe this resource creation in more detail, consider the English–French example in (13) (from Koehn (2005), Figure 2):

(13) that is almost a personal record for me this autumn!
⟶c' est pratiquement un record personnel pour moi , cet automne!

The first stage involves automatically tagging each closed-class word in (13) with its marker tag, as in (14):

(14) <DET> that is almost <DET> a personal record
<PREP> for <PRON> me <DET> this autumn!

⟶<DET> c' est pratiquement <DET> un record personnel
<PREP> pour <PRON> moi , <DET> cet automne!

Taking into account marker tag information (label, and relative sentence position), and lexical similarity (via mutual information), the marker chunks in (15) are automatically generated from the marker-tagged strings in (14):

(15) a. <DET> that is almost : <DET> c' est pratiquement

b <DET> a personal record : <DET> un record personnel

c. <PREP> for me this autumn : <PREP> pour moi cet automne

The chunk pair in (15c) is an example of how the chunking constraint mentioned previously (that every marker chunk must contain at least one content word) comes into play, as segmentation is not performed at the marker words *me* and *this* in English, and *moi* and *cet* in French.

2.2.2 Generalised Templates

A set of generalised templates which, as Gough (2005) demonstrates, can improve both coverage and translation quality are automatically derived from the marker chunks in (15) by simply replacing the marker word by its relevant tag, in effect deleting the closed-class word. From the examples in (15), the generalised templates in (16) are derived:

(16) a. <DET> is almost : <DET> est pratiquement

b. <DET> personal record : <DET> record personnel

c. <PREP> me this autumn : <PREP> moi cet automne

Generalised templates enable more flexibility in the matching process, as now any marker word can be inserted after the relevant tag if it appears with its translation in the lexicon. For example, assuming it to be absent from the set of marker chunks, the string *this is almost* can now be translated by recourse to the template in (16a) by inserting a (or all) translation(s) for *this* in the system's lexicon.

2.2.3 Word-Level Lexicon

The system lexicon can be constructed in two ways. Firstly, deleted marker words in generalised templates are assumed to be translations of each other. For instance, in deriving the generalised templates in (16) from the marker chunks in (15), the lexical alignments in (17) are created:

(17) a. <DET> that : <DET> c'

b. <DET> a : <DET> un

c. <PREP> for : <PREP> pour

Secondly, in source–target marker chunks or generalised templates, where there is just one content word in both source and target, these are assumed to be translationally equivalent. Taking (15c) as an example, the lexical entry in (18) is automatically created:

(18) <LEX> autumn : <LEX> automne

This marker-derived word-level lexicon can be combined with the lexicon derived from previously calculated MI scores to produce the final word-level lexicon used during the translation process. Additional lexical resources, such as statistically-derived lexicons, can also be added to the EBMT lexicon.

2.2.4 Recombination

When a new sentence is submitted for translation, it is segmented into all possible n-grams that might be retrieved from the system's memories. For each n-gram (with the exception of those ending with a marker word—given our marker-based segmentation method, new chunks are created when marker words are found), these resources are searched in the order below, going from maximal context (specific source–target sentence-pairs) to minimal context (word-for-word translation):

1. The original aligned source–target sentence pairs

2. The marker-aligned chunks

3. The generalised marker chunks

4. The word-level lexicon

Each translation retrieved is assigned a weight using the formula in (19):

(19) $$\text{weight} = \frac{\textit{no occurrences of the proposed translation}}{\textit{total no translations produced for SL phrase}}$$

For example, given the source language phrase *the house*, assuming that *la maison* is found 8 times and *le domicile* is found twice, then weight(*la maison*, *the house*) = 8/10 and weight(*le domicile*, *the house*) = 2/10. In this way, multiple translations can be created by the system and output to the user in a ranked list.

The retrieved target language n-grams are recombined based on the original source language n-gram order, with any reorderings that may be required occurring within the chunks themselves. As there is no restriction on the length of the EBMT chunks (in our experiments the maximum chunk length for chunks within the database was 10 words for English and 12 words for French, with average chunk lengths of 4.5 words and 4.6 words, respectively) this ordering process is sufficient for language pairs where long-distance reordering occurs less frequently (such as French–English). However, for certain linguistic phenomena, such as argument-switching, this linear recombination technique would be insufficient and more sophisticated reordering techniques would be required

2.3 Statistical Machine Translation

Statistical approaches to MT were in fact first proposed by the father of MT (Weaver, 1949). Weaver suggested that statistical methods, such as those that were becoming commonplace in the fields of cryptography and information theory at the time, could be applied to the task of automatically translating text from one language to another. However, due to the limitations of computers at the time, the lack of

machine-readable texts and various theoretical objections, these early ideas were quickly abandoned in favour of approaches founded more in linguistic than probability theory, such as transfer- and interlingua-based systems. Such objections are summed up by the much quoted statement of Chomsky (1969).

> "It must be recognized that the notion of a 'probability of a sentence' is an entirely useless one, under any interpretation of this term" (Chomsky, 1969)

It wasn't until four decades after the original proposition of statistical methods that the statistical approach to MT re-emerged. SMT was first properly outlined by researchers coming from speech recognition (Brown et al., 1988, 1990). Statistical methods developed within the speech community proved to be so successful that they presented themselves as an alternative methodology to solve the MT problem. Researchers at the time were interested in moving away from more traditional, linguistic-based approaches, such as those employed by earlier second generation systems, to translation which could operate with little or no linguistic information, approaching MT as more of a challenge of engineering than of linguistic theory. The idea of doing away with the need for deep linguistic information required by rule-based and interlingua MT systems was an attractive one as it proposed a way to overcome many of the inherent limitations and problems these second generation systems suffered from. In addition the increased availability of machine-readable texts and computational resources increased the feasibility of the purely statistical approach.

2.3.1 The Noisy Channel SMT model

Traditionally, SMT models are based on a Bayesian inference approach using the noisy channel model as commonly used in speech recognition probabilistic models of pronunciation (Jurafsky and Martin, 2000). Working in a somewhat counter-intuitive fashion, in this model the source sentence S is treated as though it has

been passed through a channel of noise which makes it difficult to recognise the equivalent target language translation T from which S originates (Figure 2.6).

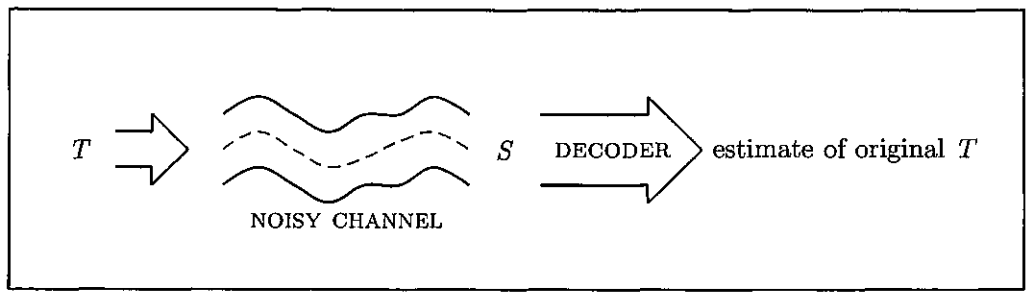


Figure 2.6: Pictorial representation of the Noisy Channel process

In the noisy channel framework, it is the job of the decoder to retrieve the original target language sentence which the given source language sentence was generated from. Applying Bayes' Rule, the noisy channel model in Figure 2.6 is represented by the equation given in (2.2).[3]

$$P(T|S) = \frac{P(S|T)P(T)}{P(S)} \tag{2.2}$$

In Equation (2.2) $P(T|S)$ represents the probability that a translator will produce T in the target language when presented with S in the source language. $P(S)$ is independent of T, remaining constant for each sentence T under consideration, as we are looking for the most likely translation T for the same source sentence S. Therefore, the equation to find the most probable T can be simplified to give us the equation in (2.3).

$$\hat{T} = \underset{T}{\operatorname{argmax}} \left(P\left(S|T\right) P\left(T\right) \right) \tag{2.3}$$

To minimize the chance of error, the system now has to maximize the product of the remaining probabilities – $P(T)$, the probability that the sentence T would be produced in the target language and $P(S|T)$, the probability of a particular

[3]Note that in the literature, the problem is more commonly viewed as searching for the S that is most probable given T. We here reverse the direction, as we are in fact translating from S into T, rather than the other way around.

candidate translation T being translated as S (note how due to Bayes' rule, the translation direction has been reversed from a modeling standpoint). These two models are known as the language model and translation model, respectively. The translation model assigns probabilities to the set of target language words which are most likely to be useful for translating the source language string (attempting to ensure the *faithfulness* of the translation), whereas it is the job of the language model to arrange these corresponding target language words into the best possible order, thus ensuring some sort of translation *fluency* Translation thus realises itself as a search problem, where searching for the T which maximizes the product in (2.3) is known as decoding. Since the number of possible translations is potentially exponential, a beam-search or pruned Viterbi algorithm is usually used to reduce the size of the translation search space.

An SMT system therefore requires three main components: a translation model to calculate $P(S|T)$, a language model to calculate $P(T)$ and a decoder to search for the $\hat{T}$ which estimates T by maximizing the product of the language and translation models, as given by Equation (2.3). The basic architecture for a typical SMT system, as used in the work presented in this thesis, is illustrated in Figure 2.7

More recent research in SMT has begun to move away from the classical source–channel approach, instead applying a log-linear model to compute $P(T|S)$ directly (Och and Ney, 2002; Zens and Ney, 2004; Zens et al., 2005). The formula for the log-linear SMT model is given in Equation (2.4), where we have M feature functions $h_m(T, S), m = 1, ..., M$, and for each feature function a scaling factor, λ_m

$$\hat{T} = \underset{T}{\operatorname{argmax}} \{ \sum_{m=1}^{M} \lambda_m h_m(T, S) \} \tag{2.4}$$

The log-linear model enables the combination of several different models with the additional benefit over the noisy channel approach of being able to easily integrate new additional models into the system. From Equation (2.4) we can see that each feature function $h_m(T, S)$ in the log-linear approach is multiplied by a scaling factor,

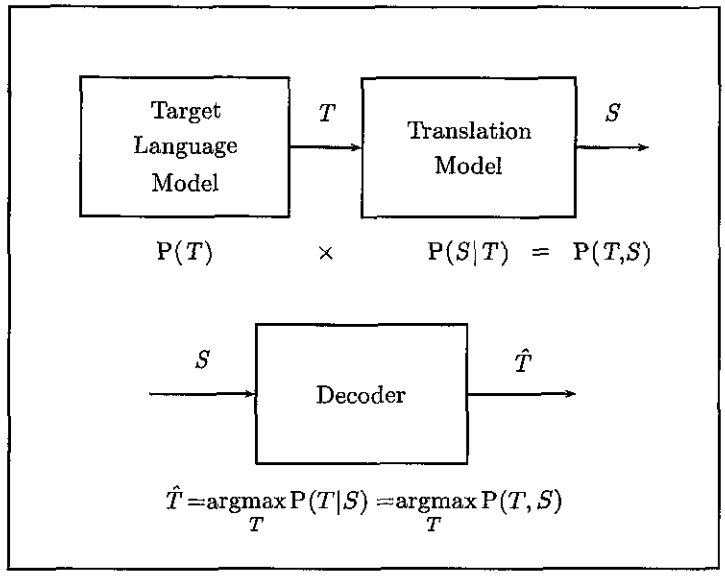


Figure 2.7: A Statistical Machine Translation System (adapted from Figure 1 in Brown et al. (1993))

λ_m. In order to optimize the performance of the system these scaling factors are trained with respect to the final translation quality measured by an error rate (Och, 2003). Feature functions commonly used in such log-linear systems include the logarithms of n-gram language models, reordering models, phrase-translation models (for both source–target and target–source) and word probability models (again, for both source–target and target–source directions). The noisy-channel approach used in the SMT systems employed in our work represents a special case of the log-linear framework, where the two models used in the system are the language model and translation model. The noisy channel SMT approach expressed within the log-linear framework is given in Equation (2.5), where $\lambda_1 = \lambda_2 = 1$.

$$\hat{T} = \underset{T}{\text{argmax}}\{\lambda_1 logP(T) + \lambda_2 logP(S|T)\} \tag{2.5}$$

2.3.2 Language Modeling: *Calculating P(T)*

In order to calculate the value of $P(T)$ within the noisy channel framework as expressed in the Equation (2.3) and the log-linear model in Equation (2.5), an n-gram language model is used. The language model calculates the probability of a particular word sequence occurring in a particular language, with the hope that syntactically correct strings will receive a higher probability than those which are less well formed. In generative language modeling, to calculate the probability of a given string $w_1, w_2, w_3 ... w_n$, we need to calculate the probability of each word in the string occurring in the language, given its history (i.e. all the preceding words in the string), and get the resulting product of these probabilities, as given in Equation (2.6).

$$P(t_1, t_2, t_3 .... t_n) = P(t_n|t_1 t_2 t_3 ... t_{n-1}) ... P(t_2|t_1) P(t_1) \tag{2.6}$$

Considering Equation (2.6), this is not a trivial calculation as for any string of reasonable length there are far too many histories to consider. As a result, instead of computing the probability of a word in a string given all preceding words, this probability is approximated by conditioning the probability of a given word on the preceding n words. For most SMT systems a bigram or trigram (as in Equation (2.7)) language model is used.

$$P(w_1^n) \approx \prod_{k=1}^{n} P(w_k|w_{k-2} w_{k-1}) \tag{2.7}$$

Although the language model is usually trained from a very large (target language) monolingual corpus, we still can encounter sparse data problems, as there is still a very strong chance that we may encounter an n-gram which occurs in our test set that is not present in our training data, especially with higher-order n-grams (in practice it is common to employ a trigram language model, as given in Equation (2.7)). As any given corpus is finite whereas language itself is infinite, we are bound to encounter zero-counts occurring at the unigram or higher n-gram level for

which we wish to assign a non-zero probability. A number of smoothing methods are available to help solve this problem. The simplest of these smoothing methods is add-one smoothing where each n-gram probability is augmented by adding one to the n-gram count and normalizing (Lidstone, 1920). Usually a more sophisticated smoothing method is preferred, such as weighted linear interpolation (Jelinek and Mercer, 1980).

The weighted linear interpolation model is an example of a discounted backoff model, combining trigram, bigram and unigram probabilities in the calculation of the language model score for a particular string. Essentially, in backoff, if we cannot find a trigram, we search for a bigram and if no bigrams exist we estimate the probability using the unigram score. In weighted linear interpolation, in contrast to backoff where make use of only the highest order model that occurs, we make use of all n-gram probabilities available to us and multiply each n-gram probability by a discounting factor λ_n, where all λ values sum to 1 and λ_n is greater than λ_{n+1}. The formula for the weighted linear interpolation model is given in Equation (2.8).

$$\begin{aligned}\hat{P}(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) \\ +\lambda_2 P(w_n|w_{n-1}) \\ +\lambda_3 P(w_n); \\ \sum_i \lambda_i = 1\end{aligned} \tag{2.8}$$

The SRILM toolkit (Stolcke, 2002)[4] enables the efficient computation of a language model over a monolingual corpus, and is used in the experiments we describe in later chapters of this thesis. For our experiments we used a discounted interpolated language model (such as that given in Equation (2.8)) employing Kneser-Ney discounting (Kneser and Ney, 1995). Kneser-Ney discounting is an extension of absolute discounting. Absolute discounting is similar in spirit to the weighted linear interpolation model, but instead of multiplying the higher-order n-gram probabilities

[4] `http://www.speech.sri.com/projects/srilm/`

by a λ, we subtract a fixed discount, δ, chosen during held-out estimation. Kneser-Ney smoothing provides an elegant way of constructing the lower-order backoff models by considering the lower-order model to be significant only when the count is small (or zero) for the higher n-gram model. For example, if *New York* frequently occurs in the corpus, but *York* only ever occurs after *New*, then we give *York* a low unigram probability, as the bigram model fits well.

2.3.3 From Word- to Phrase-Based Translation Models

The translation model is used to calculate $P(S|T)$ and is trained over a large bilingual, sentence-aligned corpus. It measures word co-occurrence frequencies as well as relative source and target word positions, along with sentence lengths, to determine word translation probabilities.

In order to estimate word translation probabilities the problem of word alignment first needs to be addressed. The alignment problem is central to SMT as in order to estimate word translation probabilities, we first need to define correspondences between words in the source language sentence with words in the target language sentence. An example of such an alignment for an English–French sentence pair is given in Figure 2.8, where the connecting lines represent an alignment, indicating translational equivalence, between words in the English sentence with words in the French sentence. As can be seen from the example, often words in one language can align with one or more words in the other (e.g. *soutiendra* in French being aligned with *will*, *be* and *supporting* in English), and sometimes may even align to nothing (in these cases the word in question is said to align to the special token 'NULL').

Assuming a sententially aligned corpus, a number of methods are available for establishing such word-level correspondences between source and target. One of the more commonly used methods for word alignment is based on the expectation-maximization (EM) algorithm (Dempster et al., 1977), efficiently implemented by the GIZA++ statistical word alignment tool (Och and Ney, 2003)[5] which is com-

[5] `http://www.fjoch.com/Giza++.html`

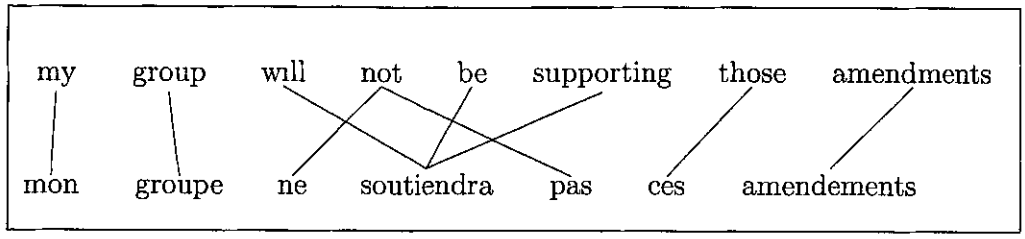


Figure 2.8: An example of English–French word alignment taken from the Europarl corpus (Koehn, 2005)

monly used to extract word alignment sets from bilingual sententially-aligned text. A number of word alignment models can be calculated with this tool These models are collectively known as the IBM models (Brown et al., 1990, 1993).

The most primitive word alignment model, that of IBM model 1, is a simple model where all connections for each target position are assumed to be equally likely, so the order of the words in the source and target sentences does not affect the translation model probability. EM for model 1 results in a global maximum, so the resulting initial probabilities can be used to seed the EM process for subsequent models. IBM model 2 takes into account word order and the length of both source and target strings Models 3 through 5 consider the possibility of a word being aligned to zero, one or more words in the other language and also the position of the translated word. These models also take into account positioning probabilities. In our work we make use of the standard IBM Model 4.

As well as word translation probabilities, the translation model calculates fertility (the number of target words generated by a source word) (e.g. in the alignment example in Figure 2 8, the English word *not* has a fertility of 2 as it is aligned with the two French words *ne* and *pas*) and distortion (changes in word position) probabilities.

The very first SMT models of Brown et al. (1990) and Brown et al. (1993) made use of word-based translation models, capturing relationships between individual source–target word pairs. These systems did not make use of any syntactic information. In fact, only the relatively weak distortion probabilities influence the ordering of the target words in the final translation and most of the work to en-

sure grammaticality is left up to the language model. Their fertility models are asymmetric meaning that they are insufficient to accurately model many linguistic phenomena, such as multi-word units, as they can only model one-to-one and a limited number of one-to-many mappings; the benefits of making phrasal information available during translation are obvious, as alignments such as *will be supporting* $\Longrightarrow$*soutiendra*, from the example in Figure 2.8, are extremely difficult to capture by such approaches. Consequently, these earlier word-based SMT systems often produced word salad, with translation quality not that far removed from the first 'direct' MT systems. Without any notion of syntax, such as phrases, non-local dependencies are extremely difficult to capture and without any morphological analysis, related words are treated as completely separate types. Huge improvements were seen even when simple morphological analysis was added to the translation models (Brown et al., 1991).

The relative shortcomings of the earlier word-based SMT systems can be mainly attributed to the lack of linguistic knowledge present in the models. In order to overcome these problems, more recently SMT researchers make use of phrase-based translation models. In phrase-based SMT systems, the translation model contains not only relationships between individual words, but also between sequences of words and phrases. With this added contextual information, unsurprisingly such SMT systems can produce much higher quality translations than the earlier word-based systems.

2.3.4 Phrasal Extraction Heuristics

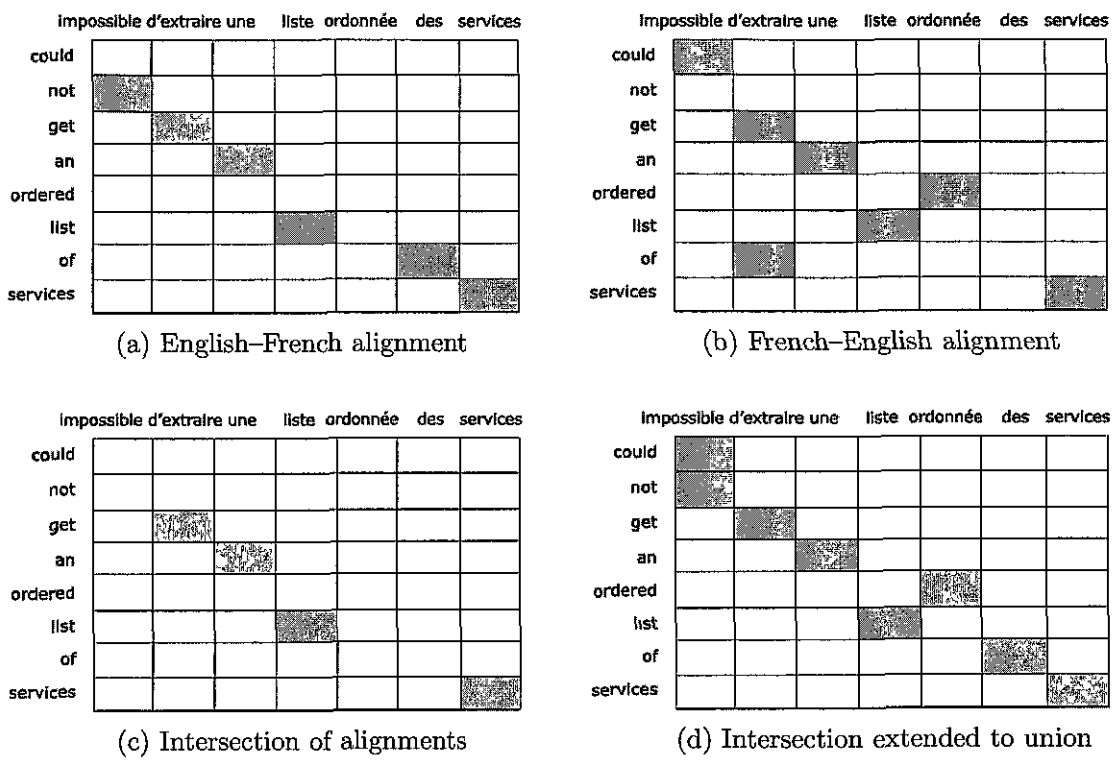
In order to model phrase translation information, we first need to define a method for identifying phrasal correspondences in bilingual text. Koehn et al. (2003) compare a number of different methods for phrasal extraction, evaluating their performance in terms of translation quality. From their results, phrasal extraction heuristics based on initial word alignments, such as the method outlined by Och and Ney (2003), proved to be the most successful for translation. Koehn et al. (2003) found

that more sophisticated syntax-based extraction heuristics, where only syntactically valid constituents were deemed to be valid phrases, do not actually lead to higher quality phrases, but instead only overconstrain the extraction process, resulting in fewer phrases and ultimately lower quality translations and reduced coverage

Following the methods outlined by Och and Ney (2003), in order to extract phrasal alignments for use in the phrase-based SMT system employed in our work, we first perform word alignment in both directions, from source–target and target–source, as the IBM word alignment models (Brown et al., 1993) only allow one-to-one and one-to-many mappings. An example of such alignments for English–French and French–English mapped onto a two dimensional bitext grid are given in Figure 2.9(a) and Figure 2.9(b). Performing word alignment in both directions gives us a large set of alignments containing one-to-one, one-to-many and many-to-many alignments (due to the one-to-many mappings from both language directions). Taking this large set of alignments we can produce a more highly confident set by first taking the intersection of these two unidirectional alignment sets, as shown in Figure 2.9(c) As these alignments are common to both directions, they are likely to be of high precision. We proceed by following suggestions put forward by Koehn et al. (2003), who extend this intersection set into the union by iteratively adding adjacent[6] alignments present within the alignment matrix, where either the source or target word under consideration are unaligned. Adding only adjacent alignments allows us to gradually build up contiguous phrase alignments. In a final step to improve recall, any remaining alignments present within the union alignment set, where both source and target words remain unaligned, are added to the alignment set, as shown in Figure 2.9(d).

Such 'refined' alignments provide the best balance between coverage and recall and so are ideally suited to the MT task (Tiedemann, 2004). From the resulting alignment set we can extract all possible n-gram pairs which correspond to these

[6]By adjacent, we mean horizontally or vertically adjacent word alignments when mapped onto a two-dimensional bitext grid space



(a) English–French alignment

(b) French–English alignment

(c) Intersection of alignments

(d) Intersection extended to union

Figure 2.9: Extracting Phrase Alignments from Word Alignments

alignments, giving us our set of phrase alignments. A subset of the possible phrase alignments that could be extracted from the example in Figure 2.9 are given in (20).

(20) could not ⇔impossible

could not get ⇔impossible d'extraire

get an ⇔d'extraire une

ordered list ⇔liste ordonnées

get an ordered list ⇔d'extraire une liste ordonnées

could not get an ordered list ⇔impossible d'extraire une liste ordonnées

of ⇔des

of services ⇔des services

ordered list of services ⇔liste ordonnées des services

an ordered list of services ⇔une liste ordonnées des services

could not get an ordered list of services ⇔impossible d'extraire une liste ordonnées des services

...

Applying this n-gram collection method on a sentence-by-sentence basis, phrase pairs are collected over the entire corpus. The resulting source and target phrase pairs are counted and probabilities are then estimated from relative frequencies (Equation (2.9)). Note that given Bayes' rule and the formula in Equation (2.3) the modeling direction is reversed and so our translation model needs to provide us with $P(S|T)$. For a given source–target phrase pair p_s, p_t, we need to calculate $P(p_s|p_t)$.

$$P(p_s|p_t) = \frac{C(p_s, p_t)}{C(p_t)} \tag{2.9}$$

This type of phrasal extraction can result in phrases of arbitrary length, even incorporating entire sentences. However, to increase the efficiency of translation, the length of the possible phrases extracted is commonly limited to 5, 6 or 7 words. It has been shown in previous research by Koehn et al. (2003) that using phrases of 3 words is actually sufficient to achieve almost optimal performance, as increasing

phrase length only results in slight improvements.

In contrast to the phrase extraction heuristics of Koehn et al. (2003) and Och and Ney (2003), Marcu and Wong (2002) propose a joint probability model which directly calculates phrase translation probabilities, rather than relying on a pre-existing set of word-level alignments. Their work, like the original IBM models of Brown et al. (1990, 1993), uses EM to align and estimate the probabilities of bilingual n-gram sequences in a parallel corpus. This task is computationally expensive as when considering all possible segmentations of phrases, the number of possible phrase alignments between two sentences is exponential with relation to the length of the shorter sentence (as shown in Birch et al. (2006)). Consequently the model is not scalable to large data sets.

However, more recently, Birch et al. (2006) present a number of improvements to the original joint probability model of Marcu and Wong (2002). They constrain the model search by making use of the intersected set of source–target and target–source word alignments. Higher probability is assigned to those phrase alignments which are consistent with the intersected word alignment set during the initialization stage of the EM algorithm, rather than starting off with a uniform probability distribution over all phrase alignments as in Marcu and Wong (2002). Any phrase pairs that are not consistent with the word alignments are given a small non-zero probability, thus allowing the phrase alignments that do not contradict these high probability word alignments to be considered first. Birch et al. (2006) also investigate the use of linguistic information, identifying words that are identical orthographically in both source and target, and use bilingual dictionary entries to further constrain the search. They find that when training on reasonably small data sets (approx. 10,000 German–English sentence pairs from the Europarl corpus (Koehn, 2005)), their joint model can outperform the more standard phrase-extraction method of Koehn et al. (2003), in terms of translation performance (21.69 BLEU score[7] for the method of Koehn et al. (2003) vs. 22.79 BLEU for Birch et al. (2006)'s approach,

[7]cf. Section 3 2.2 for an explanation of the BLEU metric

with the additional linguistic constraints providing further improvements (BLEU score increases to 23.30). However, when scaling to a larger training set of 730,740 Spanish–English sentence pairs, although their joint probability model scales up to the larger training set, the phrase extraction model of Koehn et al. (2003) actually outperforms their model in terms of BLEU score (26.17 BLEU for Birch et al. (2006) vs. 28.35 for Koehn et al. (2003)). Birch et al. (2006) attribute this poor performance to data sparseness and over-fitting of their model due to the fact that only a small proportion of the overall alignment space is searched.

2.3.5 Decoding

Given the language model probability $P(T)$ and translation model probability $P(S|T)$, we now need to search for the $\hat{T}$ which maximizes the product of these probabilities. In other words, we need to search through all possible Ts which are likely to have produced S, and select the T which is most likely.

Searching through all possible Ts is not feasible in practice, especially due to the number of possible target language translations for a given input sequence, and, in fact, the problem has been shown by Knight (1999) to be NP-complete for word-based models In order to make the decoding process efficient and implementable a beam-search strategy is usually employed (Germann et al., 2001).

The PHARAOH decoder[8] of Koehn (2004a) is a widely-used phrase-based decoder which employs a beam-search strategy, similar to that used by Jelinek (1998) for speech recognition. During the search process PHARAOH uses a priority queue, organising partial hypotheses under investigation into stacks based on the number of input words that the hypotheses cover, as illustrated in Figure 2.10.

These hypothesis stacks are pruned during the search process based on histogram pruning, where only the top n-scoring partial hypotheses are kept. The phrase translation table used during translation is also pruned using threshold pruning where the probability of the highest-scoring candidate phrase is multiplied by an

[8] http://www.isi.edu/publications/licensed-sw/pharaoh/

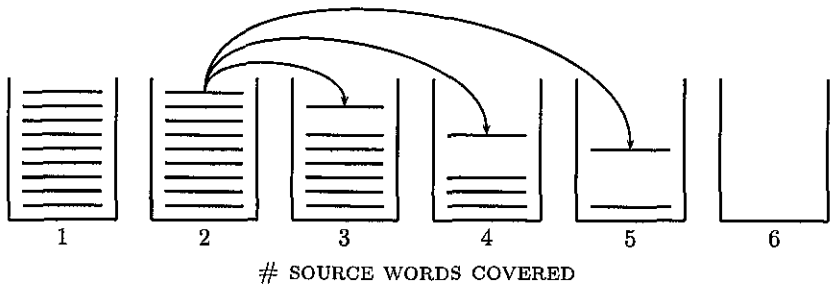


Figure 2.10: Organisation of hypotheses into stacks. If a particular hypothesis is expanded into new hypotheses, these are placed into stacks (indicated by the arrows) according to the number of source words covered so far.

empirically set factor and only those phrase translation candidates whose probability lies above this threshold are considered in the search.

During translation PHARAOH proceeds as follows:

1. A sequence of n words is chosen in the input string (all segmentations of the input are considered equally probable, therefore a uniform probability distribution is assumed).

2. All possible target translations for this input sequence are attached to the end of the current hypotheses, thus generating a set of new hypotheses (the stack is initialised with an empty hypothesis with a score of 1).

3. The probabilities of the new hypotheses are updated.

4. The weaker hypotheses are pruned based on their score so far multiplied by an estimated future cost. The future cost estimates are calculated from the product of the language model and translation model probabilities for the remaining sequence of untranslated words. For the remaining sequence of untranslated words, many possible overlapping translation options exists, therefore the future cost estimator selects the cheapest path i.e. the path with the lowest

probability according to the translation and language models. These costs are calculated before decoding using dynamic programming and stored in a table which can then be accessed during the decoding process. As there are only $n(n+1)/2$ possible sequences for any segment of length n, this calculation is trivial.

5. The cheapest hypothesis covering the entire input is chosen as the final translation. If the entire input has not yet been covered, we return to step 1.

During decoding, the output target language phrases can be reordered. Reordering is based on a relative distortion probability distribution, relative to the source position of the current translated phrase and the equivalent source position of the previous target language phrase. The distortion probability can be estimated using the joint probability model described by Marcu and Wong (2002) (as mentioned previously in Section 2 3 4), which along with direct phrase translation probabilities, also yields distortion probabilities for a phrase in source position i being translated as a phrase in target position j. Similarly, the more recent work of Birch et al (2006) could also be applied to directly estimate phrase distortion probabilities.

Generally, the type of reordering models used during phrase-based decoding are weak and so the quality of translations produced by the system relies heavily on the quality of the phrases extracted during the translation stage. However, recently much research has been carried out on the development of more sophisticated phrase-based reordering models which are applied only at phrase boundaries. The reordering models of Zens et al. (2004,0) make use of position classes (e.g. 1 position to the left, 1 position to the right etc.), word classes and local contextual information in the form of POS tags to calculate the target language phrase order in the final translation

2.4 Comparing EBMT and SMT

EBMT and SMT approaches to translation represent two frameworks within the one data-driven paradigm. EBMT research was first presented in the work of Nagao on translation by analogy (Nagao, 1984), whereas SMT approaches were first introduced by Brown et al. (1988). As described in Section 2.3.3, earlier SMT systems used only word-level information whereas EBMT has made use of phrasal information since its very inception. Relatively recent research in SMT however, has led to the development of SMT models which now make use of both phrasal and word-level information when carrying out translation (Koehn et al , 2003). With the advent of such phrase-based approaches, the line between statistical methods and EBMT approaches has become ever more blurred. In order to compare both methods in terms of their similarities and differences, we first need to be able to define what classifies a particular data-driven approach as being example-based or statistical and to determine whether or not these approaches are simply variations of each other.

2.4.1 Defining EBMT

The definition of what is classified as SMT is somewhat clearer than that of EBMT. SMT methods are easily identifiable by their use of distinct probabilistic models and their relative lack of linguistic knowledge. EBMT, on the other hand, is less clearly defined.

A number of researchers have attempted to identify what exactly defines an approach as being example-based and what characteristics separate it from other MT paradigms such as RBMT and SMT. Somers (1999, 2003) points out that the problem of defining EBMT is exacerbated by the fact that today there are many different flavours of EBMT, most incorporating techniques from other paradigms, in contrast to the earlier 'pure' analogical systems of Nagao (1984).

The use of a bilingual corpus is considered part of the definition, but not sufficient to qualify an approach as being example-based, as this does not separate it from

SMT, and almost all modern-day approaches to MT make use of information taken from bilingual corpora in one way or another. Refining the criterion, Somers (1999, 2003) considers any system that uses a set of examples as its primary knowledge base as EBMT. Consequently, one could argue that on the one hand, phrase-based (and word-based) SMT methods are clearly example-based as they essentially make use of bilingual examples in their translation models, but the matching and recombination approaches of EBMT are implemented in quite a different way Somers (1999, 2003) goes on to say that an additional defining characteristic of EBMT systems is that they make use of their example databases implicitly at run-time (such as the approaches of Sumita et al. (1990), Brown (1999a), Planas and Furuse (2003) and Sumita (2003)), but this excludes many approaches that are claimed to be example-based, such as the approaches of Watanabe et al. (2003), Matsumoto et al. (1993) and the marker-based approach described in Section 2.2.

In their definition of EBMT, Turcato and Popowich (2003) concentrate on the knowledge content of a system instead of the way that the knowledge is expressed or acquired. Contrary to Somers (1999, 2003), they state that the existence of a database of examples is not justification in itself for labeling a system as EBMT, as the way in which system knowledge is acquired is irrelevant; rather, what matters is how knowledge is actually used in practice. Turcato and Popowich (2003) agree somewhat with Somers (1999, 2003), in that they believe the only true EBMT systems are those where the information is not preprocessed and is available intact and unanalysed throughout the matching and extraction process.

However, as with the definition of Somers (1999, 2003), Hutchins (2005) points out that this type of restriction to the use of 'implicit knowledge' only is too narrow as it excludes much research described as EBMT. Hutchins (2005) follows the lead of Turcato and Popowich (2003) and considers that the way systems make use of knowledge during the core translation process is what separates them from each other. Hutchins (2005) states that the core process of any given MT system is the conversion of elements of the input text into equivalent elements of the output text.

In EBMT, this core process, as described in Section 2.1, is the selection and extraction of target language fragments which correspond to input source language fragments. The analysis of the input text may be as simple as in SMT, consisting of dividing sentences into phrases or word strings based on predetermined closed-class words, as in the marker-based approach. However, the majority of EBMT systems also perform further analysis to create templates or generalised tree structures before proceeding to the matching process. In SMT, the core process centres on the translation model which produces target language word sequences from input word sequences from predetermined source–target sub-sentential alignments (similar in spirit to methods in EBMT). However, unlike EBMT methods, the SMT translation model selects translation candidates for input word sequences based solely on corresponding statistical frequency data

Following this, and according to Hutchins (2005), phrase-based SMT and EBMT can be considered as variants of a single framework, but still with significant differences to set them apart. The main theoretical differences are that while SMT systems base the translation process heavily (and almost exclusively) on statistical methods, EBMT systems work on the basis of linguistic fragments and text examples. With both approaches, information about the well-formedness of translations is implicitly contained within their bilingual databases. However, EBMT relies much more heavily on the information contained within retrieved aligned examples to ensure well-formedness during the recombination approach. This in turn motivates the use of linguistic information, in particular syntactic information, during the identification and selection of its examples. In SMT, in addition to information contained within the pre-computed phrasal alignments, information is implicity utilised in the decoding stages by referring to a monolingual statistical language model

2.4.2 Marker-Based EBMT within the MT Model Space

Wu (2006) makes perhaps the most successful attempt at differentiating between SMT and EBMT methods by defining a three dimensional space in which all possible

approaches to MT sit. The axes of this model space represent the degree of example-based (the x-axis), compositional (the y-axis) and statistical (the z-axis) techniques employed (see Figure 2.11) The x-axis ranges from schema-based MT to example-based MT, the y-axis from purely lexical MT, through collocational MT and onto fully compositional MT, and the z-axis ranges from purely logical techniques to purely statistical techniques.

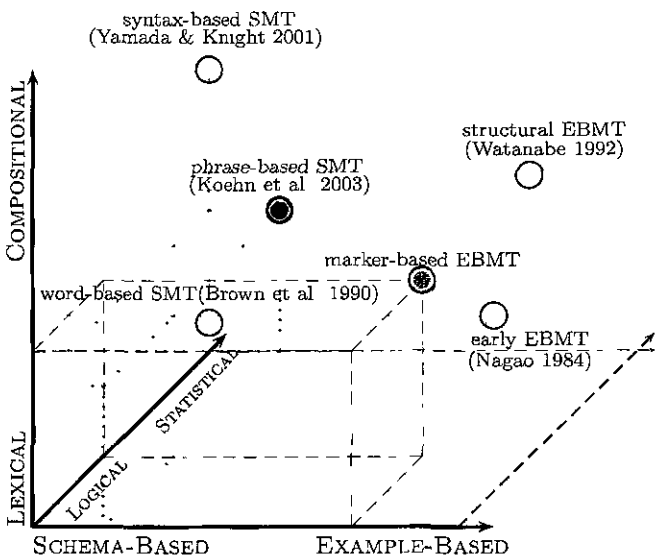


Figure 2.11: The MT model space of Wu (2006) Phrase-based SMT lies along the statistical end of the z-axis and also makes use of some example-based and collocational (lying midway between fully lexical and fully compositional on the y-axis) techniques. Marker-based EBMT is placed closer to the schema-based techniques than its earlier 'pure' counterparts and is also collocational (with some more compositional techniques than phrase-based SMT).

Wu (2006) states that classical EBMT models make nontrivial use of a large library of examples at runtime rather than during training, essentially memorizing data rather than abstracting away from it, thus agreeing with the definitions of Somers (1999, 2003) and Turcato and Popowich (2003). SMT models, defined as

making nontrivial use of mathematical statistics and probability, on the other hand, attempt to do just that, being more schema-based.

As EBMT systems begin to make use of example databases created during pre-processing stages and move away from memorization, they become more schema-based and thus more SMT-like although still remaining closer to the EBMT family of models. The marker-based approach to EBMT extracts sets of examples, generalised templates and lexical alignments during a preprocessing stage, which has much in common with phrase-based SMT approaches (indicated by its position with regards to the x-axis in Figure 2.11). In addition, the use of weights, despite not qualifying as true probabilities, during its recombination stage also represents an SMT element, according to Wu's definition (Wu, 2006) (indicated by its position on the z-axis in Figure 2.11). However, the presence of generalised templates, together with a distinctive recombination stage and the lack of true probability models, make it more akin to EBMT that to SMT. At the same time, as SMT models begin to make use of phrase alignments, which can also be considered as types of lexical collocation translation rules, the more EBMT-like they become (Wu, 2006) (cf. Figure 2.11). In the SMT alignment template approach of Och et al. (1999), bilingual source-target word classes are trained using the method of Och (1999), and templates representing an alignment between a source class sequence and a target class sequence are automatically created. The use of classes allows for better generalisation. Making use of such alignment templates is similar in fashion to the use of generalised templates in EBMT, and further increases the proximity of such SMT approaches to the EBMT family of models.

2.4.3 Remaining Differences

With regards to EBMT and SMT models of translation, despite their similarities some differences still remain. As Way and Gough (2005a) state, one advantage EBMT systems have over their SMT counterparts is their re-use of examples. When translating an input sentence, if an EBMT system has previously encountered the

sentence it can simply perform an exact sentence match, retrieving the source language sentence and outputting the corresponding target language sentence. Additionally, after successfully translating a new unseen input sentence, this new sentence along with its translation can then be added to the example database, thus allowing it to be re-used during translation. On the other hand, when an SMT system encounters a sentence it has seen previously, it proceeds by repeating the complicated search required to find the translation which maximises the probabilities of the translation and language models. Exact sentence matching, despite being a relatively simple technique, is still not generally employed in SMT systems. If such techniques are in fact used in some SMT approaches, something which is extremely feasible, there is still no evidence within the relevant published literature.

Generally speaking, EBMT systems are founded on linguistic principles, the level of which varies from system to system, with the majority of approaches making use of explicit syntactic information during the creation of their example bases (e.g. Sato and Nagao (1990), Planas and Furuse (2003), Watanabe et al. (2003), Gough (2005), Hearne (2005)). This in turn means that EBMT systems, in general, by their very nature, incorporate some syntax at their core, in contrast to SMT systems which generally only make use of syntax in a *post hoc* fashion The *post hoc* reranking experiments of Koehn et al. (2003) actually demonstrated that adding syntactic information was detrimental to the translation quality. In general, syntax-based SMT approaches, such as the approaches of Yamada & Knight (2001, 2002), Charniak et al. (2003) and Burbank et al. (2005), have yet to result in significant improvements in translation quality. However, more recent experiments by Chiang (2005), who makes use of hierarchical phrase probabilities, are promising. Chiang (2005) demonstrates that making use of such hierarchical information for dealing with higher-level dependencies, in particular for the ordering of NP-modifying relative clauses, results in improved translation results; a 7.5% relative increase in BLEU score was observed for Mandarin–English translation when compared to a baseline phrase-based SMT system built using the PHARAOH decoder (Koehn et al., 2003).

However, as we described in Mellebeek et al. (2006), although the method of Chiang (2005) is language independent and can deal with certain problematic tasks, it does not rely on any linguisitc annotations or assumptions during the induction of its grammar, and is therefore not completely justified from a linguistic perspective.

In addition, most SMT phrasal extraction heuristics, and those shown to be most effective (Koehn et al., 2003), are only based on word alignments and do not make use of any syntactic constraints on the phrasal extraction process. Consequently, SMT systems essentially learn n-grams, rather than phrases *per se*. As a result, they are likely to learn word sequences that EBMT systems do not (for further discussion and examples, see Section 4 5).

2.5 Summary

Data-driven, empirical approaches to MT now dominate the MT research field, presenting an alternative to earlier rule-based and direct approaches. Corpus-based approaches provide a way of performing translation, without a need for the enormous amounts of linguistic expertise that was required in earlier transfer-based approaches, resulting in the ability to develop systems quickly and inexpensively. In this chapter we introduced the two main data-driven approaches to MT: EBMT and SMT.

From Section 2.1 we can see how EBMT models of translation, based on the principle of analogy, have always made use of both phrasal and lexical information, in addition to information taken from more generalised template structures. One state-of-the-art approach to EBMT as used in our work is an approach based on the marker hypothesis (Green, 1979). The marker-based approach is a *linguistic lite* approach (Veale and Way, 1997), making use of surface syntactic information in the form of predefined sets of closed-class words to segment source and target language sentences in order to create its databases of examples, generalised templates and word-level lexicon.

Early SMT methods borrowed heavily from techniques coming from the Speech Processing community at the time, with their translation models based on relationships between individual words (Brown et al., 1998, 1990). However, recent SMT systems now make use of both phrasal and lexical information within their translation models and, unsurprisingly, perform much better than their word-based counterparts (Och and Ney, 2002; Koehn et al., 2003).

As both SMT and EBMT methods of translation make use of phrasal information in addition to other translation knowledge sources, the line between them has become ever more blurred. However, despite this seeming convergence, some differences still remain allowing us to classify a given data-driven approach as belonging to the SMT or EBMT model family. In addition to discussing the theoretical differences between EBMT and SMT, it is also worthwhile to investigate how these two different approaches compare in terms of translation performance and in the following chapter we describe a number of experiments to this end.

Following the definitions outlined by Wu (2006), all MT systems seem to fall somewhere within the three dimensional MT model space, employing a mixture of hybrid techniques. In the following chapter, in addition to comparing EBMT and SMT in terms of their translation performance, we outline a number of investigations into the combination of elements of EBMT and SMT techniques in a move towards creating a hybrid system which can capitalise on the advantages of either approach.

Chapter 3

Hybrid 'Example-Based SMT': Experiments on the *Sun Microsystems* TM

It is clear that data-driven approaches to Machine Translation have now become the norm for producing state-of-the-art translation results. However, how do the state-of-the-art approaches to corpus-based MT compare in terms of translation performance? In this Chapter we address this question. We continue the work described by Way and Gough (2005a) by comparing the performance of the two main corpus-based approaches of example-based MT (EBMT) and statistical phrase-based MT (PBSMT) when trained and tested on reasonably large amounts of data taken from the *Sun Microsystems* TM (described in Section 3.1). After describing a range of automatic metrics used in the evaluation of MT in Section 3.2, these comparative experiments, along with their results are presented in Section 3 3

All approaches to MT have their advantages and disadvantages,therefore it is unsurprising that many practitioners have investigated combining various MT approaches into hybrid approaches. An MT system is considered to be hybrid if it integrates relatively autonomous subsystems (which often implement various different computational techniques) to achieve different tasks in the the overall MT

process.

SMT and EBMT make use of very different techniques in order to extract the resources they use during the translation process. EBMT tends to make use of more linguistically motivated approaches to gain its translation knowledge, such as the marker-based approaches described in Section 2.2 which employ minimal surface syntactic information to extract chunks, generalised templates and word-level correspondences. On closer inspection, it can be said that there are in fact very few 'pure' MT systems, which make use of techniques coming from only one paradigm. For instance, from Chapter 2 we can see that most methods of EBMT consist of hybrid approaches, exploiting not only example-based techniques, but also syntactic and generalised rules along with statistical and direct dictionary-based translation techniques

SMT, on the other hand, often makes use of only probabilistic information to extract word-level alignments which then feed phrase extraction heuristics which essentially operate on n-grams, rather than phrases *per se*. As SMT methods move more towards the use of phrasal information, they too have become more hybrid-based most obviously in the application of syntactic rules to aid the reordering process.

In our work we wish to investigate whether EBMT and SMT techniques can be combined to produce a hybrid data-driven system capable of outperforming both approaches In addition to our initial experiments, we decided to examine the contributions these different translation resources make to the quality of the translations produced by an MT system. In Section 3.4 we describe a number of empirical investigations into the possibility of combining example-based and statistical machine translation resources in a move towards improving the translation performance of the under-performing baseline phrase-based SMT (PBSMT) system on the *Sun Microsystems* data set, as part of our investigations into the creation of a hybrid 'example-based SMT' system.

3.1 Experimental Setup

3.1.1 Data Resources

For our initial experiments described in this Chapter we made use of a large translation memory, obtained from *Sun Microsystems*. This data set was used in the previous experiments of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) and consists of 207,468 English–French sentence pairs. The *Sun Microsystems* data is made up of English computer documentation along with their French translations. A sample of English-French sentence pairs taken from the corpus is given in Figure 3.1.

support services and warranty upgrades	⇔	services de support et de mise à jour de la garantie
these procedures describe how to perform a clean install of the software	⇔	les méthodes présentées ci-après indiquent comment effectuer l'installation initiale du logiciel
to build the java hotspot vms perform the following steps	⇔	pour construire les machines virtuelles java hotspot suivez la procédure ci-dessous
the network cable is disconnected	⇔	le câble réseau est déconnecté
make sure the location exists before you download Netscape	⇔	assurez-vous de l'existence de l'emplacement avant de télécharger Netscape
Sun storedge n8600 filer release notes	⇔	notes de version de Sun storedge filer n8600

Figure 3.1: Sample English–French sentence pairs from *Sun Microsystems* data set

As the corpus comes from a particular sublanguage domain, is it particularly suited to MT. In MT, a sublanguage usually refers to a particular type of text written with a particular communicative purpose, written using particular language constructions and specialised vocabulary (e.g. *Netscape*, *java*, *download* and *install* in Figure 3.1). Within such a domain, there is likely to be much more repetition in terms of words and phrases, with a lower occurrence of *hapax legomena*, making it easier for any corpus-based system trained on the data to learn from regularities, both in terms of lexical content and syntactic structure, that occur within it.

From the full corpus of 207,468 sentence pairs, 3,939 sentence pairs were randomly extracted as a test set, with the remaining 203,529 sentences used as training data. The average sentence length for the English test set was 13.1 words and 15.2 words for the corresponding French test set. Futher details of the training and test

corpus sizes are given in Table 3.1.

		#Sentences	#Tokens	Avg	Max	Min
Training Corpus	EN	203,318	2,206,566	10.85	112	1
	FR	203,318	2,450,260	12.05	134	1
Test Corpus	EN	3,939	51,608	13.10	87	1
	FR	3,939	59,723	15.16	91	1

Table 3.1: Details of *Sun Microsystems* training and test corpora. Avg, Max and Min refer to the average, maximum and minimum sentence lengths, respectively, in terms of words

At each stage of our experiments we performed translation from French into English and from English into French. In order to determine the performance of each system at each stage, we made use of a number of automatic evaluation metrics We describe these metrics in the Section 3.2.

3.1.2 MT System Details

For our experiments we used the marker-based system used in the work of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a). From the *Sun Microsystems* training set we used the marker-based techniques, as described in Section 2.2 to extract databases of marker chunks, generalised templates and a word-level lexicon.

To build our baseline PBSMT system, we made use of the following freely available tools:

- GIZA++ EM word alignment toolkit (Och and Ney, 2003)[1], to extract the word-level correspondences;
- The SRI language modelling toolkit (Stolcke, 2002);[2]
- The PHARAOH phrase-based decoder.(Koehn, 2004a)[3]

[1]http://www.fjoch com/Giza++.html
[2]http.//www.speech.sri com/projects/srilm/
[3]http·//www.isi.edu/publications/licensed-sw/pharaoh/

A refined set of word alignments were collected from the combined bidirectional GIZA++ word alignment results. This refined word alignment set was then used to extract a set of phrasal alignments (Och and Ney, 2003; Koehn et al., 2003), using the phrasal extraction method described in Section 2.3.4. The aligned phrases along with their translation probabilities estimated from relative frequencies made up the SMT system's phrase translation table. We made use of an interploated trigram language model, employing Kneser-Ney discounting (Goodman, 2001), an extension of absolute discounting with a more sophisticated method to calculate backoff distributions The language model made no use of additional monolingual data and was trained solely on the relevant monolingual portion of the *Sun Microsystems* training data. As stated previously (cf. Chapter 1), for the experiments described in this chapter, together with those described in later chapters, the SMT system was left untuned, employing default tuning parameters.

3.2 Evaluation Metrics

Nowadays, evaluation of MT output is deemed essential in the development of any MT system. In terms of evaluation, translated texts need to be judged on their clarity, their style (the extent to which the translation uses the language appropriate to its content), and their accuracy (the extent to which the translated text contains the same information) (Hutchins and Somers, 1992). Carrying out this type of judgement is a difficult task due to the degree of ambiguity present in languages, making an objective human manual evaluation difficult to carry out. In addition, manual evaluation is a costly and extremely time-consuming process prone to inconsistencies. This is particularly true when evaluating the quality of a large number of translations produced by various systems.

Automatic metrics, on the other hand, provide a fast, efficient, inexpensive, consistent and objective way of assessing the quality of translations. In our experiments we make use of a number of freely available MT automatic metrics: Sentence Error

Rate (SER), Word Error Rate (WER), BLEU (Papineni et al., 2001, 2002), Precision and Recall (Turian et al., 2003). These automatic evaluation metrics attempt to judge the quality of MT output by comparing each candidate translation against one or more manually produced human reference, or gold standard, translations

3.2.1 Sentence and Word Error Rate

Sentence Error Rate (SER) is a measure of the number translations produced which exactly match the reference translation. To calculate SER for any given test set we simply count the number of output translations which match their corresponding reference translations exactly and express this count as a percentage of the total number of sentences in the original test set. As SER is an error rate, we subtract this percentage from 100 in order to give us our final figure For example, if we have a test set of 10 sentences and our MT system translates 8 of these sentences to reproduce exactly their corresponding gold standard translations, the SER for this particular test set will be 20%. The lower the SER, the better the performance of the system. SER is a crude metric, but allows us to see the ability of a system to correctly produce the reference translation in its entirety.

Word Error Rate (WER) is a slightly more sophisticated metric, also commonly used in the field of speech recognition. WER is based on the Levenshtein distance (Levenshtein, 1965), an instance of the minimum edit distance algorithm (Wagner and Fischer, 1974). The Levenshtein distance between two individual strings is a measure of the least amount of insertions, substitutions and deletions that need to be made to transform one string into the other. The standard Levenshtein distance gives a penalty of 1 for each insertion, substitution and deletion of a single character that is required for this type of transformation. WER is implemented in a similar fashion, but is word-, rather than character-based, and is calculated from the Equation in (3.1)

$$WordErrorRate = 100\frac{Insertions + Substitutions + Deletions}{Total\ \#\ Words\ in\ Reference\ Translation} \tag{3.1}$$

An illustrative example of the calculation of the WER between a candidate translation and its reference is given in Figure 3.2.

REF:	check	system	console	logs	for	detailed	error	messages	
CAND:	check	the		logs	to	see	the	messages	error
EVAL:		*Subs.*	*Del.*		*Subs.*	*Subs.*	*Subs.*		*Ins.*

Figure 3.2: Insertions, Substitutions and Deletions between a candidate translation and its reference.

In order to transform the candidate string in Figure 3.2 into its reference a minimum of 1 insertion, 4 substitutions and 1 deletion are needed. Therefore the WER can be calculated from the formula in Equation (3.1):

$$WordErrorRate = 100\frac{1+4+1}{8} = 75\% \tag{3.2}$$

As with SER, WER is expressed as an overall percentage, and as it is an error rate, the lower the WER the better the quality of the translation produced. WER penalises not only translations which contain mistranslated words, but also translations that may contain the correct words, but in the wrong order, as can be seen with the example in Figure 3.2, where *the messages error* in the candidate translation is penalised twice; once for containing the incorrect word (with respect to the reference translation) *the*, being counted as a substitution, and again for having *messages* and *error* occurring the wrong order by considering *error* as an inserted word If, instead, *the error messages* occurred in the candidate translation, the candidate would only be penalised once, for the insertion of the word *the*, resulting in a lower WER.

3.2.2 BLEU

The BLEU metric (Papineni et al., 2001, 2002) is a widely used MT evaluation metric which compares the output of an MT system against one, or possibly more, reference translations. It is an n-gram co-occurrence metric and calculates the number of n-grams in the output translation which also occur in one or more of the reference translations. BLEU rewards translations which contain longer contiguous subsequences of matching words.

BLEU takes inspiration from the highly successful WER metric, as described in Section 3.2.1, but allows for differences in word choice and word order by calculating a weighted average of variable length n-gram matches for a particular candidate translation against multiple reference translations. It employs a **modified n-gram precision** score which avoids problems of double counting which would otherwise give inflated precision for candidate translations which overgenerate or repeat words.

The modified n-gram precision score counts the maximum number of times a given candidate n-gram occurs in any reference translation, then clips the total count for each candidate n-gram by its maximum reference count In other words, if an n-gram w_i^j occurs x times in a candidate translation and it is seen y times in the reference (i.e its maximum reference count is y) and $x \geq y$, then w_i^j is given a count of y. The modified n-gram precision for any given candidate translation is calculated using the formula in 3 3.

$$p_n = \frac{Count_{clip}(c_n \cap r_n)}{Count(c_n)} \tag{3.3}$$

where

c_n is the multiset of n-grams occurring in the candidate translation

r_n is the multiset of n-grams occurring in the reference translation.

$Count(c_n)$ is the number of n-grams occurring in the candidate translation

$Count_{clip}(c_n \cap r_n)$ is the number of n-grams occurring in c_n that also occur in r_n, such that elements occurring x times in c_n and y times in r_n occur maximally y times

From the example in Figure 3.3 taken from Papineni et al. (2002) we can see how overgeneration results in the candidate translation receiving a unigram precision of 7/7. Whereas applying the modified n-gram precision formula from Equation (3.3), resulting in only the underlined unigrams in Figure 3.3 being considered, giving a much more reasonable unigram precision of 2/7 to this improbable translation.

Cand: the the the the the the the.
Ref1: The cat is on the mat.
Ref2: There is a cat on the mat

Figure 3.3. English candidate translation displaying overgeneration and its two corresponding references.

As with human judgements, scores for individual sentences can vary from judge to judge, so evaluation is normally performed on a reasonably large test set. The BLEU score for a multi-sentence test set can be calculated by adapting the formula in Equation (3.3) by simply summing all the sentence-based clipped n-gram matches and dividing by the total number of candidate n-grams in the entire translated test set.

To calculate p_n any individual value of n can be used. However, Papineni et al. (2002) state that it is more robust to calculate p_n for a range of n values and combine these modified precision scores into a single metric. As larger values of n are considered, however, the number of possible n-gram matches for any candidate translation decreases. This in turn results in p_n decaying almost exponentially with n. BLEU takes this decay into account by calculating a weighted average of the

logarithm of modified precisions for a range of values of n[4], using a uniform weight $\frac{1}{N}$, as given by the equation in (3.4):

$$p_N = \exp(\sum_{n=1}^{N} \frac{1}{N} \log(p_n)) \tag{3.4}$$

In addition, BLEU also introduces a brevity penalty BP. Whereas p_n penalises a candidate translation for not having the correct word choice or word order, this brevity penalty penalises candidate translations that do not match its references in terms of length. The BP is implemented as a decaying exponential in the length of the reference translation over the length of the candidate translation. The use of a decaying exponential means effectively that for a candidate translation that is the same length as any reference translation, the BP is 1.0 and a BP of greater than 1.0 is given to any candidate translation that has a length less than any of its corresponding references. To avoid punishing shorter sentences more harshly, the brevity penalty is computed over the entire corpus, rather than on a sentence-per-sentence basis. Therefore, to calculate the BLEU score for any test set, p_N, the geometric mean of the test corpus' modified precision score, is multiplied by an exponential brevity penalty factor, as in Equation (3.5).

$$\text{BLEU} = BP.p_N \tag{3.5}$$

By incorporating the brevity penalty BP, BLEU now rewards candidate translations for having the same length as their corresponding references as well as rewarding candidates for containing the correct words in the correct order Candidate translations can now be given a score ranging from 0 to 1, with higher scores indicating higher quality translations in terms of adequacy and fluency.

[4]In their experiments, Papineni et al. (2001, 2002) found that a maximum value of $n = 4$ to be sufficient for adequate correlation with human evaluation.

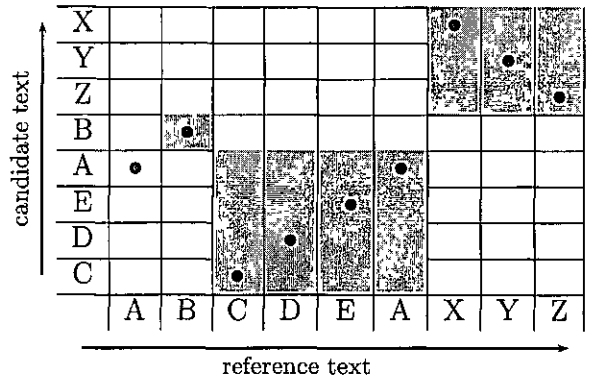


Figure 3.4: Bitext grid illustrating the relationship between an example candidate translation and its corresponding reference translation. Each bullet, called a *hit*, indicates a word contained in both the candidate and reference texts.

3.2.3 Precision and Recall

Turian et al. (2003) adapt the more traditional metrics of Precision and Recall for the evaluation of MT output. For a given set of candidate items C and a set of reference items R, precision and recall can be calculated by the formulae in 3.6 and 3.7 respectively.

$$precision(C|R) = \frac{|C \cap R|}{|C|} \tag{3.6}$$

$$recall(C|R) = \frac{|C \cap R|}{|R|} \tag{3.7}$$

Turian et al. (2003) define a method for calculating the intersection $|C \cap R|$ of a candidate text C and a reference text R. The two texts can be viewed as a grid, with filled entries in the grid representing word matches (or **hits**) between the candidate translation and its reference. An example of such a grid, similar to the example given in Turian et al. (2003), is given in Figure 3.4

If we naively count $|C \cap R|$ for a particular bitext, we run the risk of unfairly double-counting items which occur more than once in the reference text, e.g. the item A in Figure 3.4 receives an inflated count by getting two hits in the reference.

To avoid such problems of double-counting, Turian et al. (2003) instead identify **matchings** in the grid. A matching is defined as a subset of the hits in the grid, such that no two hits are in the same row or column. The matching for a particular bitext which is of maximum possible size can be selected, as indicated by the shaded regions in Figure 3.4. In order to calculate the precision for a particular bitext, we can take the size of this maximum matching and divide it by the length of the candidate translation ($|C|$); for recall we divide the maximum matching size (MMS) by the length of the reference text ($|R|$), as given by Equations (3.8) and (3.9), respectively.

$$precision(C|R) = \frac{MMS(C,R)}{|C|} \tag{3.8}$$

$$recall(C|R) = \frac{MMS(C,R)}{|R|} \tag{3.9}$$

As an example, the MMS for the grid in Figure 3.4 is 8 (calculated by summing the sizes for the individual smaller matchings of 1, 4 and 3, as indicated by the shaded areas in the grid), the length of the candidate text is 8 and the length of the reference text is 9, so Precision in this case is $8/8 = 1.0$, whereas Recall is $8/9 = 0.89$. However, the precision and recall methods of Equations (3.8) and (3 9) do not reward correct word order in a candidate translation. Figure 3.5 shows grids for two different candidate texts against the same reference text. Applying the precision and recall metrics of Equations (3.8) and (3.9) results in both candidates being given the same precision and recall scores, despite the fact that the candidate in Figure 3.5(b) has a better word order, successfully realising the suffix XYZ in the correct order.

In order to take word order into account, Turian et al. (2003) generalise their definition of match size to reward the existence of **runs** – contiguous sequences of matching words within a bitext grid (appearing as diagonally adjacent hits in the grid running parallel to the main diagonal). The minimum enclosing square of a run is considered to be an aligned block. We can calculate precision and recall figures by

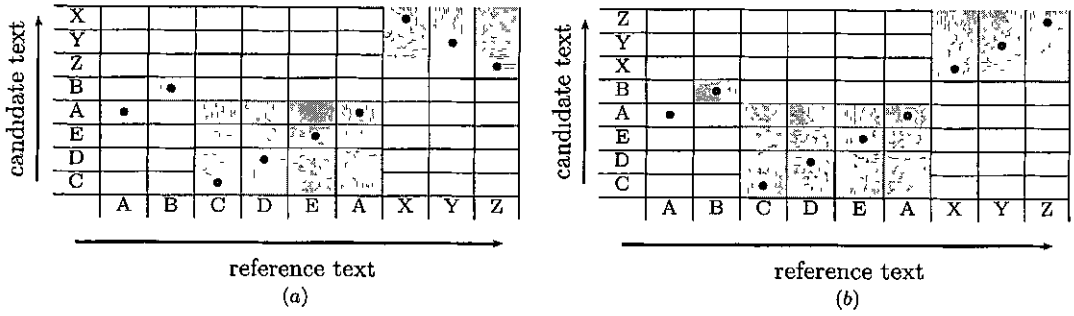


Figure 3.5: Bitext representing two different candidate texts for the same reference text. Using the original MMS definition, the two candidates score the same in terms of precision and recall, despite the candidate on the right grid containing more accurate word order

calculating the intersection of the candidate translation and its reference in terms of the area of the aligned blocks, with the weight of a run defined as the area of its minimum enclosing square. Thus, the generalised definition of match size is given by Equation (3.10)

$$MMS(M) = \sqrt{\sum_{r \in M} length(r)^2} \tag{3.10}$$

Identifying the runs (hits occuring diagonally adjacent in the grid running parallel to the main diagonal) and corresponding aligned blocks of the two candidate texts, as indicated by the shaded areas in Figures 3.6(a) and 3.6(b), we can use the formula in Equation (3.10) to calculate the MMS for each candidate text and their corresponding precision and recall scores. Looking at Figure3.6, the MMS for the candidate in Figure 3.6(a) is $\sqrt{1^2 + 4^2 + 1^2 + 1^2 + 1^2} \approx 4.5$ and $\sqrt{1^2 + 4^2 + 3^2} \approx 4.9$ for the candidate in Figure 3.6(b), giving Figure 3.6(a) precision of $4.5/8 = 0.5625$ and recall of $4.5/9 = 0\ 5$, whereas Figure 3.6(b) scores a higher precision of $4.9/8 = 0.6125$ and higher recall of $4.9/9 = 0.5445$, reflecting the higher quality of this particular candidate text.

The MMS for a particular candidate text can also be calculated when we have multiple reference texts. By concatenating the various reference texts into a single

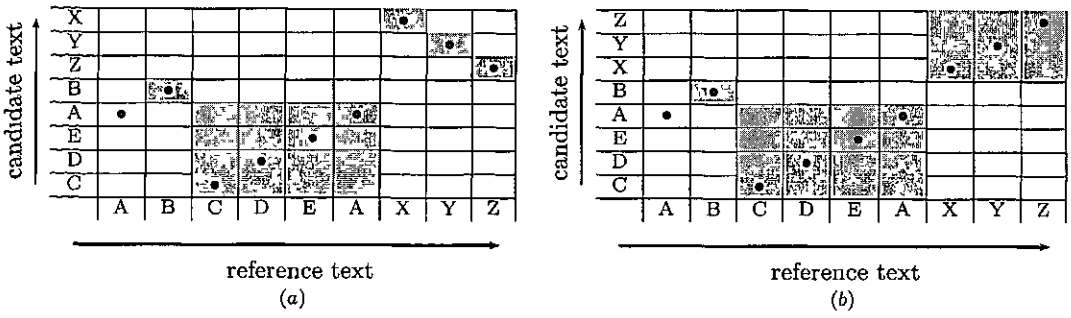


Figure 3.6: Bitext representing two different candidate texts for the same reference text, with the shaded boxes indicating aligned blocks containing the runs used in the calculation of the generalised MMS, as calculated from Equation (3.10).

grid, the MMS can be easily computed as before, employing the restriction that runs may not span more than one reference text and capping the MMS with respect to the mean reference length. During the capping step, excess hits are removed from the matching until the number of hits is equal to the mean reference length, removing hits in the order that maximizes the size of the remaining matching (Turian et al., 2003).

3.3 EBMT vs SMT

Way and Gough (2005a) provide what are to our knowledge the first published results comparing Example-Based and Statistical models of MT. Given that most MT research carried out today is corpus-based, it is somewhat surprising that until quite recently no qualitative research existed on the relative performance of the two approaches. This may be due to a number of factors: the relative unavailability of EBMT systems, the lack of participation of EBMT researchers in competitive evaluations or the dominance in the MT research community of the SMT approach—whenever one paradigm finds favour with the clear majority of MT practitioners, the assumption made by most of the community is that this way of doing things is clearly better than the alternatives.

Like Way and Gough (2005a), we find this regrettable: the only basis on which such views should be allowed to permeate our field is following extensive testing and evaluation. Nonetheless, given that no EBMT systems are freely available, very few research groups are in the position of being able to carry out such work.

Following on from the work of Way and Gough (2005a) we decided to test EBMT against state-of-the-art phrase-based models of SMT, rather than the poorer word-based models used in this previous work. In so doing, it provides a more complete evaluation of whether an SMT system is capable of outperforming an EBMT system on reasonably large training and test sets (Groves and Way, 2005).

In our experiments we compared the performance of the marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) against two variants of the PBSMT system: the first SMT system's translation table was made up of phrases extracted using the standard statistical GIZA++ methods; the second employed a translation table made up of the same marker chunks and word alignments as used in the marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a). This enabled us to directly compare the performance of both systems when using the same translation resources [5] The results of these experiments are given in Sections 3.3.1 and 3.3.2

3.3.1 English–French Translation

Using the full 203K training set of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a), and testing on their near 4K test set, the results for English–French are given in Table 3 2. In Table 3.2, we also provide the results for the word-based SMT system (WBSMT) as given in Way and Gough (2005a) and reproduced in Groves and Way (2005) This WBSMT system was created

[5]Although it must be noted that the EBMT system has the ability to use generalised templates that are not used in any of the SMT system configurations. Ideally, the set of templates should be excluded from the EBMT system, or alignment templates, such as those used by Och et al (1999) could be employed by the SMT system, which although not identical to marker-based templates, would give both systems the capability of using generalised examples and render any comparison more accurate.

		BLEU	Prec.	Recall	WER	SER
WBSMT		.3223	.6513	.5704	53.50	89.10
PBSMT	EBMT-DATA	3643	.6661	5759	61.33	87.99
	GIZA-DATA	.3753	.6598	.5879	58.50	86.82
EBMT		.4409	.6727	.6877	52.40	65.60

Table 3.2: Seeding PHARAOH with GIZA++ and EBMT sub-sentential alignments for English–French.

using only the word-level alignments extracted using GIZA++ together with the ISI word-based ReWrite decoder [6] and CMU Statistical Modeling toolkit.[7]

From Table 3.2 we can see that for the same training and test data, the PBSMT system outperforms the WBSMT system on most metrics, considerably so with respect to BLEU score (0.3753 vs. 0.3223), as is to be expected. WER, however, is somewhat worse (58.50% vs. 53.50%), and SER, although lower (89.10% for WBSMT vs. 86.62% for the PBSMT system), still remains disappointingly high. The WER for PBSMT is particularly surprising as one would expect with the additional contextual information the phrase-based system would have greater success than the WBSMT system in terms of lexical choice as well as in terms of word order. This result seems to indicate that the PBSMT system may have to resort to word-to-word translation quite frequently on this particular data set, and thus does not benefit from any of its additional phrasal contextual information.

With regards to the performance of SMT, it is clear to see that the SMT system making use of the GIZA++ alignments (GIZA-DATA) obtains better scores than the SMT system using the EBMT sub-sentential data (EBMT-DATA). The increase in performance of the GIZA-DATA SMT system over the same system seeded with the EBMT-DATA is statistically significant, with $p < 0.05$ (Koehn, 2004b). This would seemingly indicate the higher quality of the SMT data induced via GIZA++, however, before one considers the full impact of these results, one should take into account that the size of the EBMT data set (word- and phrase-alignments) is 403,317,

[6] `http://www.isi.edu/natural-language/software/decoder/`
[7] `http://www.mi.eng.cam.ac.uk/ prc14/toolkit.html`

while there are over four times as many SMT sub-sentential alignments (1,732,715). With such a significant difference in the size of these data sets, one would expect the use of the EBMT-data to result in a much greater deterioration in translation performance than is actually observed in Table 3.2 However, the marker-based EBMT system still beats both phrase-based system configurations with respect to BLEU score (0.4409 for EBMT vs. 0.3573 for SMT) and notably for SER (0.656 EBMT, 0.868 SMT).

3.3.2 French–English Translation

		BLEU	Prec.	Recall	WER	SER
WBSMT		.4462	.7035	.7240	46.80	80.8
PBSMT	EBMT-DATA	3952	6151	.6643	74.77	86.21
	GIZA-DATA	.4198	.6527	.7100	62.93	82.84
EBMT		.4611	.6782	.7441	50.80	51.20

Table 3.3. Seeding PHARAOH with GIZA++ and EBMT sub-sentential alignments for French–English

Again, as for English–French, for the French–English experiments the PBSMT system was seeded with the GIZA++ (GIZA-DATA) and EBMT (EBMT-DATA) word and sub-sentential alignments, trained on the full 203K-sentence training set, and tested on the 4K test set. The results are given in Table 3.3.

As for English–French, the PBSMT system employing GIZA++ alignments obtain better scores than when the EBMT sub-sentential data is used, with the difference in system performance judged to be statistically significant (Koehn, 2004b). However, we see that both flavours of the PBSMT system actually perform worse than WBSMT, which is an unexpected, and even contradictory result.[8] Accordingly, the results for PBSMT here are worse still compared to the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a), for both versions of

[8]The PHARAOH system is untuned, so as to provide an easily replicable baseline for other similar research It is quite possible that with minimum-error rate training (Och, 2003) the PBSMT system will outperform the word-based system.

the SMT system.

We should note here that the majority of the results for French–English are somewhat higher than the equivalent results for English–French (apart from the precision and WER results for both PBSMT systems which are slightly worse for French–English than English–French, more so for the EBMT-DATA SMT system which achieves a 7.7% relative decrease in precision and a 13.44% increase in WER). This behaviour is to be expected as our previous work (Groves and Way, 2005) and the work of Way and Gough (2005a) suggests that translating from French–English is inherently 'easier' than for English–French as far fewer agreement errors (and therefore fewer problems in making lexical choices) and cases of boundary friction are likely. For instance, translating *le* as a determiner into English can only realise the word *the*, but in the reverse direction *the* has the possible translations *le, la, l'* and *les*, only one of which will usually be correct in a particular context depending on number and gender.

It is also interesting to note that the EBMT system's performance appears to be much more consistent for both translation directions. In fact, even the PBSMT system seeded with the EBMT data appears to perform much more consistently than the PBSMT seeded with SMT data. Looking at the results in Table 3.2 and Table 3.3, in terms of relative BLEU score, the WBSMT system sees an increase of 38.44% when translating into English than when translating into French, with the PBSMT seeded with the GIZA++ alignments we observe a relative increase in BLEU score of 11.86% whereas the same system seeded with EBMT alignments achieves a lower increase of 8.48%. The EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) appears to be the most consistent, experiencing only a 4.58% relative increase in BLEU score when translating from French–English than when translating from English–French.

3.3.3 Discussion

From the results from Sections 3.3.1 and 3 3.2 it is quite apparent that the marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) outperforms the untuned baseline PBSMT system when seeded with alignments induced via GIZA++, and when fed the chunks and word-alignments from the EBMT system. The increase in performance of the SMT system seeded with GIZA++ data over the same system seeded with EBMT data would seem to indicate that the quality of the SMT data is greater than that of the EBMT data However, it may be attributed more to the fact that the SMT data contains many more alignments than that of the EBMT data. This is turn may mean that most of the gains achieved by the SMT system seeded with GIZA++ alignments are in fact made in recall, as the coverage of this SMT system configuration is higher than its EBMT-seeded counterpart. This is corroborated by the recall results in Tables 3.2 and 3.3, which are significantly higher for the GIZA-DATA SMT system than the EBMT-DATA SMT system. However, the EBMT-DATA system seems to achieve higher precision results for English–French, the more difficult of the two language directions, indicating that when the EBMT-DATA system has sufficient coverage it can produce more precise translations due to the higher quality of its translation resources. In order to make use of this marker-based data without suffering from problems of recall, we decided to investigate how elements of the EBMT data could be incorporated into the PBSMT system in order to contribute to its translation performance. We describe these experiments in the following section (Section 3.4).

3.4 Hybrid 'Example-Based SMT': Integrating EBMT and SMT

The results in Section 3.3 show how the performance of the marker-based EBMT system (Gough and Way, 2004b; Gough, 2005, Way and Gough, 2005a) is significantly

greater than that of the PBSMT system, both when seeded with the alignments induced via GIZA++ and when seeded with the EBMT chunk and word alignments.

In order to attempt to boost the performance of the SMT system we decided to investigate whether elements of the EBMT data and SMT data could be combined and used by the original baseline SMT system (in Chapter 4 we describe experiments investigating the use of combined EBMT and SMT data in the baseline marker-based EBMT system).

3.4.1 Related Hybrid MT Approaches

Although hybrid EBMT-SMT research remains relatively unexplored, there exists a body of work which described alternative hybrid data-driven approaches. One approach is to merge translation memory (TM) resources with SMT A TM is a type of translation database that is used by software programs to aid human translators. Similar to EBMT, it makes use of sets of examples to suggest possible partial translations for an input text, but does not perform fully automatic translation, in contrast to EBMT. Vogel and Ney (2000) automatically derive a hierarchical TM from a parallel corpus, comprising a set of transducers encoding a simple grammar. In a similar manner, Marcu (2001) uses an SMT model (Brown et al., 1993) to automatically derive a statistical TM. In addition, he adapts the SMT decoder of Germann et al. (2001) to avail of both the statistical TM resources and the translation model itself. Marcu points out the benefit of using such hybrid corpus-based techniques, stating that:

> "if a sentence to be translated or a very similar one can be found in the TM, an EBMT system has a good chance of producing a good translation. However, if the sentence to be translated has no close matches in the TM, then an EBMT system is less likely to succeed. In contrast, an SMT system may be able to produce perfect translations even when the sentence given as input does not resemble any sentence from the training

corpus". (Marcu, 2001), p.379.

Unlike the system of Vogel and Ney (2000), for which no evaluation is provided, Marcu demonstrates that his hybrid system outperforms two (unnamed) commercial systems. the hybrid French–English system translated 58% of a 505-sentence test set perfectly, while the commercial systems did so for only 40–42% of the sentences.

In similar work, Langlais and Simard (2002) also attempt to merge TM and SMT resources. Somewhat disappointingly, WER actually increases when the SMT system is augmented with TM data. Nonetheless, the authors observe "many cases where the translation obtained by merging the extracted examples with the decoder clearly improved the results obtained by the engine alone".

The work of Paul et al. (2005a, 2005b) presents a multi-engine hybrid approach to MT, making use of statistical models to select the best possible output from various MT systems. When using an SMT model to select the best output from multiple initial hypothesis produced by a number of SMT and EBMT systems Paul et al. (2005a) found that a Hidden Markov Model phrase-based SMT system provided the best translation results (0.327 WER vs 0.346 for the best-performing EBMT-based system, for Chinese-English translation). Paul et al. (2005b) found comparable results for Japanese–English, when using a decision-tree technique to pass the best initial hypothesis from multiple RBMT and EBMT systems to an SMT decoder (0.458 WER vs. 0.496 for the best-performing EBMT system). As with Marcu (2001), they too note the benefits of hybrid MT approaches:

> "Combining multiple MT systems has the advantage of exploiting the strengths of each MT engine. Quite different initial translation hypotheses are produced due to particular output characteristics of each MT engine." (Paul et al., 2005b:117)

They show that their multi-engine selection algorithm is capable of outperforming all in-house MT engines, reducing WER by 4-5% for the *C-STAR* track submission at IWSLT 2005 for Japanese-English and Chinese-English.

There also exist previous attempts to link TMs with EBMT. Carl and Hansen (1999) show that when the fuzzy match score of a TM falls below 80%, translation quality is likely to be higher using EBMT than with TM. Planas and Furuse (2003) extend TMs in the direction of EBMT by allowing sub-sentential matches, and providing a multi-level structuring of TMs.

3.4.2 Experimental Setup

In order to attempt to improve the performance of the untuned SMT system, we combined elements of the EBMT data with elements of the SMT data. We then seeded the PBSMT system with these hybrid data sets. For the experiments in this section, we seeded the PBSMT system with two versions of the resulting merged data sets, namely:

- the EBMT phrase-alignments combined with the GIZA++ word-alignments (referred to as SEMI-HYBRID in what follows) in order to implicitly compare the quality of the EBMT phrases against the GIZA++ phrases in terms of their contribution to overall translation quality;
- all the EBMT and GIZA++ sub-sentential alignments (both word and phrase alignments — referred to as HYBRID in what follows) to allow the SMT system to make full use of all of the translation resources that are available.[9]

During the merging process, the original alignments involved in the combination (word and/or phrase alignments) were collected together along with their counts, including repetitions. The counts for the data sets were recomputed and probabilities were recalculated from relative frequencies. These new hybrid data sets, along with their probabilities were then fed to the PHARAOH decoder and trigram language model, as used in Section 3.3, to perform French–English and English–French translation. The results for these 'hybrid' system configurations are given in Sections 3.4.3 and 3.4.4

[9]Note, however, that the EBMT system also makes use of a set of generalised templates that were not used in any of the SMT system configurations

3.4.3 GIZA++ Words and EBMT Phrases

The experiments investigating the performance of the PBSMT system when seeded with the GIZA++ and EBMT translation resources as described in Section 3.3 seem to indicate that the EBMT data, although not providing as much coverage, as reflected in the recall results in Tables 3.2 and 3.3, may be of somewhat higher quality than that of the SMT data.

For these experiments we wish to investigate whether the quality of the EBMT chunks is in fact better than that of the GIZA++ phrases. In order to do so, we seeded the PHARAOH decoder with the word-alignments induced by GIZA++ and the EBMT phrasal chunks only (i.e. we did not include any of the GIZA++ phrases or any of the EBMT lexical alignments). Essentially replacing the GIZA++ phrase alignments by the EBMT phrasal chunks allows us to see whether the EBMT phrasal chunks are in fact of greater quality than the GIZA++ phrases and, as a result, can improve the quality of the translations produced by the SMT system.

English–French Results

		BLEU	Prec.	Recall	WER	SER
PBSMT	EBMT-DATA	.3643	.6661	5759	61.33	87.99
	GIZA-DATA	.3753	.6598	.5879	58.50	86 82
	SEMI-HYBRID	**.3962**	**.6773**	**.5913**	**59.32**	**85.43**
EBMT		.4409	.6727	.6877	52.40	65.60

Table 3.4: Seeding PHARAOH with GIZA++ word and EBMT phrasal alignments for English–French.

Using the full 203K-sentence training set of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a), and testing on their near 4K-sentence test set, the results for English–French translation when using the EBMT chunks and GIZA++ word alignments are included in Table 3 4 (labeled in the table as SEMI-HYBRID) From the results in this table we can see that all automatic evaluation metrics improve with this particular hybrid system configuration. The improvement

of the SEMI-HYBRID system over the GIZA-DATA system is statistically significant, with $p < 0.05$ (Koehn, 2004b). Note that the data set size (the size of the resulting phrase-translation table) remains relatively small at 430,336 entries, compared to 1.73M for the PBSMT system seeded solely with GIZA++ alignments. However, the total number of entries contained in the translation table is not completely relevant, as many of the alignments contained within the table may not be applicable to the current test set. Instead, we need to consider the set of alignments which may actually be used during translation, i.e. those alignments which contain sub-sequences which are also present in the test set. Filtering the phrase-translation tables based on the test set leaves us with 78,654 entries in the SEMI-HYBRID table and 164,726 for the GIZA-DATA translation table, indicating that the quality of the EBMT chunks is in fact superior to that of the SMT phrases, as even with fewer translation examples, the SEMI-HYBRID system is capable of producing higher quality translations.

With respect to scores for the original marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a), these results remain slightly below those figures, except for precision, where the SEMI-HYBRID PBSMT system configuration slightly outperforms the EBMT system (0.6773 for the SEMI-HYBRID system vs. 0.6727 for the EBMT system)

French–English Results

		BLEU	Prec	Recall	WER	SER
PBSMT	EBMT-DATA	.3952	.6151	.6643	74.77	86.21
	GIZA-DATA	.4198	.6527	.7100	62.93	82.84
	SEMI-HYBRID	**.4265**	**.6424**	**.6918**	**68.05**	**83.40**
EBMT		.4611	.6782	.7441	50.80	51.20

Table 3.5: Seeding PHARAOH with GIZA++ word and EBMT phrasal alignments for French–English.

Running the same experimental set up for the reverse language direction gives the results in Table 3.5. While recall drops slightly from 0.7100 to 0.6918, all the other

metrics show a slight increase for the SEMI-HYBRID SMT system compared to the performance obtained when PHARAOH is seeded with GIZA++ word- and phrase-alignments. These results again show that the EBMT phrases contribute more to translation quality than the SMT phrases, resulting in statistically significant improvements in system performance (Koehn, 2004b).

As for English–French, the marker-based EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) outperforms the SEMI-HYBRID SMT system configuration, most significantly so for WER, where we see a 17.25% absolute gain (almost 33% relative) for the EBMT system, and SER, where the EBMT system manages to correctly reproduce 32.20% (absolute) (60% relative) more of the reference translations.

3.4.4 Merging All Data

After observing improvements in the performance of the PBSMT system when the EBMT chunks were used in place of the GIZA++ phrases, we decided to make full use of the translation resources at our disposal. The following two experiments were carried out by seeding PHARAOH with *all* the EBMT chunk and word alignments and *all* of the GIZA++ sub-sentential alignments, i.e. both words and phrases.

English–French Results

		BLEU	Prec.	Recall	WER	SER
PBSMT	EBMT-DATA	.3643	.6661	.5759	61.33	87.99
	GIZA-DATA	.3753	.6598	.5879	58.50	86.82
	SEMI-HYBRID	.3962	.6773	.5913	59.32	85.43
	HYBRID	**.4259**	**.7026**	**.6099**	**54.26**	**83.63**
EBMT		.4409	.6727	.6877	52.40	65.60

Table 3.6: Seeding PHARAOH with all GIZA++ and EBMT sub-sentential alignments for English–French.

Inserting all GIZA++ and EBMT data into PHARAOH's knowledge sources gives the results in Table 3.6. The scores for this hybrid system configuration (referred

to in Table 3.6 as HYBRID) give further statistically significant increases over the the 'semi-hybrid' system configuration, as described in Section 3.4.3, most significantly for BLEU where we see a 7.5% relative increase and WER decreases by 5.06% absolute, and thus are considerably better than those for the baseline SMT system seeded with GIZA++ data. This indicates that a PBSMT system is likely to perform better when EBMT word- and phrase-alignments are used in the calculation of the translation and target language probability models. Note, however, that the size of the translation table for the HYBRID system increases to over 2 million items (with 206,772 entries in the filtered phrase-translation table). Despite this, compared to the results for the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a), these results for the 'fully hybrid' SMT system still fall somewhat short (except for Precision: 0.6727 vs. 0.7026).

French–English Results

		BLEU	Prec.	Recall	WER	SER
PBSMT	GIZA-DATA	4198	.6527	.7100	62.93	82.84
	EBMT-DATA	.3952	6151	.6643	74.77	86.21
	SEMI-HYBRID	.4265	.6424	.6918	68.05	83.40
	HYBRID	**.4888**	**.6927**	**.7173**	**56.37**	**78.42**
EBMT		.4611	.6782	.7441	50.80	51.20

Table 3.7: Seeding PHARAOH with all GIZA++ and EBMT sub-sentential alignments for French–English.

Running the same experimental set up for French–English gives the results in Table 3.5. These results show an even greater improvement than was seen for English–French for the 'hybrid' SMT system over the 'semi-hybrid' system for all evaluation metrics. This improvement is most apparent for BLEU score and WER, observing a relative increase in BLEU score of 14.61% and a decrease of 11.68% absolute in WER over the SMT system incorporating the EBMT marker-based chunks and SMT word alignments. This indicates that both the EBMT chunks and EBMT word alignments contribute positively to the quality of translations produced by the PBSMT system.

In contrast to the English–French results presented in Table 3.6, for this translation direction the 'fully-hybrid' system does outperform the EBMT system (Gough and Way, 2004b; Gough, 2005; Way and Gough, 2005a) with respect to BLEU score (0.4888 vs. 0.4611) and Precision (0.6927 vs. 0.6782). Even with this increase in BLEU score and precision, the EBMT system still wins out on Recall, WER and SER. Regarding the latter, it seems that the correlation between low SER and high BLEU score is not as important as is claimed in the work of Way and Gough (2005a).

3.4.5 Discussion

From the results in Section 3.4.3 and Section 3.4.4 we can see that seeding the PB-SMT system with a 'hybrid' data set comprising of phrases and word alignments induced using GIZA++ along with marker-based EBMT chunks and lexical correspondences improves over the highest performing baseline SMT system primed solely with GIZA++ data.

This improvement in translation results for both French–English and English–French indicates that the marker-based EBMT data can contribute positively to the performance of PBSMT. For French–English using the fully hybrid set actually produces a system which outperforms the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) in terms of BLEU score and Precision. This research demonstrates how merging SMT and EBMT data can improve overall translation quality, as phrases extracted by both methods that are more likely to function as syntactic units and are of better quality are given a higher statistical significance. Conversely, the probabilities of those SMT n-grams that are not also generated by the (more syntactically-motivated) EBMT system are reduced. Essentially, the EBMT data helps the SMT system to make better use of phrase alignments during translation.

3.5 Summary

Way and Gough (2005a) carried out a number of experiments designed to test their large-scale marker-based EBMT system described in Gough and Way (2004b) against a WBSMT system constructed from publicly available tools. While the results were a little mixed, the EBMT system won out overall.

Nonetheless, as mentioned in Chapter 2, WBSMT has long been abandoned in favour of more sophisticated phrase-based models. In this Chapter we presented research which extended the work of Way and Gough (2005a) by performing a range of experiments using the PHARAOH phrase-based decoder trained on an French–English parallel corpus extracted from the *Sun Microsystems* TM.

From our first set of experiments we observed that seeding PHARAOH with word- and phrase-alignments induced via GIZA++ generates better results than if EBMT sub-sentential data alone is used. This initial result would seemingly indicate that the original GIZA++ SMT word and phrase alignments are a richer source of translation information than the EBMT chunk and alignments alone extracted from the same training data, and thus produce higher quality translations.

In order to investigate this hypothesis more fully, we performed a number of investigations into combining elements of the EBMT sub-sentential alignments with elements of the data induced using GIZA++ with a view towards creating a hybrid 'example-based SMT' system which could improve upon the performance of the baseline GIZA++-seeded system. From our experiments we observed that seeding PHARAOH with a 'hybrid' dataset of GIZA++ word alignments and EBMT phrases improves over the baseline PBSMT system primed solely with GIZA++ data. This would appear to indicate that the quality of the EBMT phrases, contrary to our initial assumption, is in fact better than that of the SMT phrases, and that SMT practitioners should use EBMT phrasal data in the calculating of their language and translation models, if available.

In addition to this discovery, we found that seeding PHARAOH with *all* data in-

duced by GIZA++ and the EBMT system leads to the best-performing hybrid SMT system. Using this system configuration for English–French, as well as the EBMT phrasal data, the EBMT word alignments also contribute positively However, for this language direction the original marker-based EBMT system still manages to outperform the best-performing hybrid SMT system (except for Precision). For French–English, however, our hybrid 'example-based SMT' system actually outperforms the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) in terms of BLEU score and precision.

Following on from the extremely promising results outlined in this Chapter, we decided to evaluate this hybrid methodology on a more complex data set, that of the Europarl corpus, to see whether we could observe the same trend on these more widely used corpora as was observed with the *Sun Microsystems* training and test data. In addition to investigating the improvements a hybrid 'example-based SMT' system could yield over a baseline PBSMT system, we wished to see if similar improvements could be made to the performance of an EBMT system by supplying it with similar hybrid translation resources, thus creating a hybrid 'statistical EBMT' system. In the next Chapter we describe a number of experiments which we carried out to this effect making use of training and test data extracted from the Europarl corpus.

Chapter 4

Hybrid 'Statistical EBMT': Experiments on the Europarl Corpus

In Chapter 3 we demonstrated how the performance of a PBSMT system could be improved through the use of hybrid translation data sets. Merging phrase and word alignments induced via SMT methods with those extracted using marker-based EBMT approaches resulted in the creation of a hybrid 'example-based SMT' system, capable of outperforming a baseline PBSMT system seeded with only alignments induced via GIZA++. In addition, this hybrid SMT system was capable of matching the performance of the marker-based EBMT system of Gough and Way (2004b), Gough (2005) andWay and Gough (2005a) for English–French.

In these previous experiments we made use of a large data set consisting of sentence pairs extracted from the sublanguage corpus of *Sun Microsystems* computer documentation. However, an important question is how do both methods of MT compare on more general and widely used corpora. For the experiments described in this chapter we decided to make use of data taken from the Europarl corpus (Koehn, 2005) which is quickly becoming one of the standard data sources for data-driven approaches to MT. We describe this corpus along with some of its characteristics in

Section 4.1.

For this particular data set we wish to see whether, as on the *Sun Microsystems* training and test sets, the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) can still outperform state-of-the-art PBSMT. In addition we investigate the effect increasing the amount of training data has on the performance of both approaches to see whether there is in fact a correlation between training set size and translation performance for EBMT, as has been widely observed for SMT. We describe these baseline comparative experiments in Section 4.2.

In addition to these initial experiments, we perform a number of experiments making use of similar hybrid system configurations as were used in the experiments described in Chapter 3. This time around, however, not only do we feed the merged data sets to the baseline PBSMT system, creating a hybrid 'example-based SMT' system as before (cf. Section 3.4), but we also feed the resulting merged translation data sets to the baseline marker-based EBMT system. The purpose of this is to see whether we can boost the performance of the baseline system and consequently achieve similar improvements for this new hybrid 'statistical EBMT' system as were seen for the hybrid PBSMT system in previous experiments. We describe these hybrid experiments in Section 4.3.

In a step towards increased hybridity, in Section 4 4 we outline experiments where we make use of a statistical language model to rerank the output of the 'fully hybrid' EBMT system in order to see whether this results in even further improvements to the quality of the translations produced by the system.

Following on from the results of these experiments and in an effort to understand the benefits of the example-based and phrase-based approaches to translation, in Section 4.5 we perform some detailed manual analysis of the types of phrase alignments that are extracted via marker-based EBMT methods and those extracted using standard statistical methods In this section we also discuss how the type of data used for this phrasal extraction affects the quality of the chunks and word

alignments produced, and consequently the quality of the final output translations. Elements of the work presented in this chapter is also presented in Groves and Way (2006a) and Groves and Way (2006b).

4.1 Data Resources: the Europarl Corpus

For the experiments described in this chapter we made use of data taken from a larger and more commonly used multilingual corpus than the *Sun Microsystems* corpus that was used in the previous experiments described in Chapter 3. The Europarl corpus (Koehn, 2005) consists of European parliamentary proceedings taken from the European Parliament website and contains data in 11 of the official languages of the European Union, with on average, 645,000 sentences approximately, or 20M words approximately for each of the languages contained in the collection. The last quarter of the corpus spanning November–December 2000 is generally reserved for testing, while the remainder is used for training systems. A sample of English–French sentence pairs taken from the Europarl corpus are given in Figure 4.1.

Resumption of the session	⇔	Reprise de la session
I believe that both parties will have to make great efforts to achieve this	⇔	Je pense que les deux parties doivent produire à cette fin des efforts considérables
Meetings took place in Stockholm on 14 March and in Brussels on 26 April	⇔	des rencontres ont eu lieu à Stockholm le 14 mars et Bruxelles le 26 avril
It is a human rights issue that we are discussing here	⇔	Il s'agit pourtant bien d'un débat sur les droits de l'homme
The traditional family is no longer the norm	⇔	la famille traditionelle n'est plus la norme
Mr President, this time we are discussing Russia's financial crisis	⇔	Monsieur le Président, nous discutons aujourd'hui de la crise financière en Russie
Thirdly, the role of a self-assured dynamic European Union in the world	⇔	Troisièmement le rôle d'une europe dynamique et forte sur l'échiquier international
If you think the legislation is no good let us revise it	⇔	Si vous êtes d'avis que la législation ne convient pas présentez-la-nous pour révision

Figure 4.1: Sample English–French sentence pairs taken from the Europarl corpus

As the data is taken from parliamentary speeches, the Europarl corpus contains quite complex language on a very wide range of topics. Along with quite a specialised vocabulary containing parliamentary and political terms (such as the terms *session*, *parties* and *legislation* from the sample sentences in Figure 4 1), the subject matter of

the corpus can cover a broad range of topics such as economics, human rights, trade, transport, employment and health. Accordingly, the corpus consists of relatively open domain, heterogenous data.

From the original designated French–English training section of the corpus, one of the larger sections of the Europarl corpus comprising over 960K sentence pairs (Koehn (2005) reports slightly higher figures for the corpus size, but this may be attributed to the inclusion of sentences in one language aligned with an empty line in the other, which were excluded in our calculations), we randomly extracted training sets consisting of approximately 78K, 156K and 322K sentence pairs made up of an average of 1.49M, 2.98M and 6.12M words, respectively. The use of the various training sets of incremental sizes enables us to investigate the effect of increasing the amount of training data on the quality of translations produced by the MT systems. As the sentences contained in the various training sets were extracted randomly from the original training section of the French–English Europarl corpus, they are not necessarily supersets of each other.

In general, corpora that are obtained largely through automatic methods are likely to contain elements of noisy data Many sentences within the corpus may not be very useful for training and may in fact harm the training process. In order to help reduce the amount of noise in the data, and thus improve the quality of the training data available, filtering techniques are usually applied to the original data Length-based filtering techniques, together with the exclusion of sentences in one language aligned with nothing in the second language, are commonly used to reduce the level of noise contained in training data (Nie and Cai, 2001), often along with more sophisticated methods such as translation likelihood-based filtering which makes use of word-level translation probabilities to estimate how likely it is that an aligned sentence pair are in fact translations of each other based on an empirically-set threshold (Khadivi and Ney, 2005).

During the training set extraction process, we excluded sentence pairs that were over 40 words in length for either French or English. We also excluded sentence pairs

with a relative sentence length ratio greater than 1.5. This empirically-set threshold allowed us to create high-quality training sets by filtering out those aligned sentence pairs which were unlikely to be good translations of each other, e.g. a 20-word English sentence is extremely unlikely to be translated into a French sentence that is longer than 30 words, or vice versa. This threshold was relaxed for sentences less than or equal to 5 words in length, as using the original ratio resulted in filtering out many of the shorter sentences within the corpus.[1]

For testing, we created a test set by randomly selecting 5,000 sentences from the Europarl common test set, again limiting sentence length to 40 words for French and English. For this test set, the average sentence length was 20.50 words for French and 18.99 words for English. Details of the various training sets and the test set used in our experiments are given in Table 4.1.

			#Sentences	#Tokens	Avg	Min	Max
Training Corpus	78K	EN	78,454	1,452,190	18.46	1	40
		FR	78,454	1,538,395	19.55	1	40
	156K	EN	156,440	2,896,122	18.49	1	40
		FR	156,440	3,066,712	19.58	1	40
	322K	EN	321,965	5,945,163	18 47	1	40
		FR	321,965	6,293,923	19.55	1	40
Test Corpus		EN	5,000	94,952	18.99	1	40
		FR	5,000	102,508	20.50	1	40

Table 4.1: Details of the Europarl training sets and test corpora. Avg, Max and Min refer to the average, maximum and minimum sentence lengths, respectively, in terms of words

4.2 Performance Comparison of EBMT and PB-SMT

In order to evaluate the performance of PBSMT against the baseline marker-based EBMT system (Gough and Way, 2004b; Gough, 2005; Way and Gough, 2005a), we

[1]For sentences of length less than 5 words, a threshold of 2.5 was set. The majority of English sentences of length 5 or less consisted largely of noun compounds which when translated into French resulted in much longer sentences, due mainly to the insertion of prepositions (e.g. *de*) for which there was no equivalent in the English.

built a baseline PBSMT system using the PHARAOH phrase-based decoder (Koehn, 2004a) along with the SRI language modeling toolkit (Stolcke, 2002) in a similar fashion as described in Section 3.1.2.[2] The translation model of the system was created using the phrasal extraction technique as described in Section 2.3.4 and we made use of a trigram language modeling with Kneser-Ney smoothing (Goodman, 2001, cf. Section 2.3.2) .

In the wider SMT community it is the general consensus that the translation performance of any SMT system can be improved by simply making use of more training data. In order to investigate this claim, and to see whether the correlation between training size and translation performance also holds for EBMT, we trained both the PBSMT and the EBMT system on the incremental training set sizes of 78K, 156K and 322K sentence pairs and tested at each stage on our 5,000 test set. We performed translation for French–English and English–French, evaluating translations in terms of BLEU score (Papineni et al., 2001, 2002) , Precision and Recall (Turian et al., 2003), WER and SER (see Section 3.2 for more details on these metrics).

4.2.1 French–English Translation

The results for French–English translation on the various training sets are given in Table 4.2. Note that doubling the amount of training data improves the performance of both systems for all metrics, demonstrating clearly that in these experiments the quality of translations produced using EBMT methods, as with the SMT approach, increases with the amount of training data used.

Increasing the amount of training data results in a 3% to 5% increase in relative BLEU score for the PBSMT system, whereas we see a higher increase for EBMT, with a 6.2% to 10.3% relative BLEU score improvement. The SER remains consistently high for both systems indicating that we rarely manage to produce an exact trans-

[2]Again, note that all variants of the PBSMT system were left untuned, with default parameter weighting.

lation in terms of the reference However, in accordance with the increase in BLEU score, we observe a drop in WER as the amount of training data increases (dropping from 85 63% to 82.43% for EBMT and from 70.74% to 68.55% for PBSMT, when going from the the 78K to the 322K training data set).

However, it is clear to see from Table 4.2 that the PBSMT system considerably outperforms the EBMT system across all metrics and for all training data sizes, with the difference between the performance of both systems deemed to be statistically significant (Koehn, 2004b). The PBSMT system, on average, achieves 0.07 BLEU score higher than the EBMT system and achieves a significantly lower WER (70.74 vs. 85.63 for the 78K data set, 69.41 vs. 83.55 for the 156K data set and 68.55 vs. 82.43 for the 322K data set).

		BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	.1217	.4556	.5315	85.63	98.94
	PBSMT	.1943	.5289	.5477	70.74	98.42
156K	*EBMT*	.1343	.4645	.5368	83.55	99.02
	PBSMT	.2040	.5369	.5526	69.41	98.30
322K	*EBMT*	.1427	.4734	.5419	82.43	99.06
	PBSMT	2102	.5409	.5539	68.55	98.72

Table 4.2: Comparing the EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a) with a PBSMT system for French–English.

4.2.2 English–French Translation

Results for the same experiments for the reverse language direction are given in Table 4.3. For this language direction the PBSMT continues to outperform the EBMT system of Gough and Way (2004b), Gough (2005) andWay and Gough (2005a) by some distance across all metrics, with the difference again being statistical significant (Koehn, 2004b). Again, as for French–English, WER is lower for the PBSMT system than the EBMT system (e.g. 68.30 vs. 77.73 on the 322K data set), but the difference between the two systems is somewhat less than for French–English. On average the BLEU score for PBSMT is 0.05 absolute above that of the EBMT system

for English–French across data sets, whereas PBSMT achieves an average increase of 0.07 absolute BLEU score above the EBMT system for the reverse language direction.

Doubling the amount of training data improves BLEU score by about 0.8 absolute (that is, between 4% and 4.7% relative improvement) for the PBSMT system Consequently and as expected, precision and recall rise and WER and SER fall linearly as the amount of training data increases. For EBMT, as with French–English, we see a greater increase in BLEU score as we increase the amount of training data, with relative BLEU score improving by 10.8% when going from the 78K to the 156K training set, and by 8.3% when increasing the training data size from 156K to 322K sentence pairs.

		BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	.1240	4422	4365	79.09	99.10
	PBSMT	.1771	5046	4696	70.44	98 54
156K	*EBMT*	.1374	.4548	.4476	77.66	98.96
	PBSMT	1855	.5120	.4724	69 37	98.20
322K	*EBMT*	.1488	.4587	4530	77.73	99.22
	PBSMT	.1933	5180	4751	68.30	98.12

Table 4.3: Comparing the EBMT system of Gough and Way (2004b), Gough (2005) andWay and Gough (2005a) with a PBSMT system for English–French.

4.2.3 Translation Direction & the BLEU Metric

It should be noted that, as with our experiments on the *Sun Microsystems* data described in Chapter 3, the performance of the EBMT system remains much more consistent for both language directions than the baseline PBSMT system. Looking at the results in Table 4.2 and Table 4.3 we can see that the PBSMT system performs on average 1.75% BLEU score worse (6.28% relative) for English–French than for French–English translation across training sets. The EBMT system, by contrast, performs better, on average, for English–French than for French–English in terms of BLEU score, but remains more consistent, only performing 0.38% BLEU score better (2.82% relative) for English–French translation.

Koehn (2005) creates 110 SMT systems for each possible language pair and direction for the eleven different languages in the Europarl corpus. He too observes that "some languages are more difficult to translate *into* than *from*" (p. 83, emphasis original), and provides some observations as to why this might be the case for German and Finnish, with particular reference to the morphological richness of such languages which can increase the level of ambiguity during translation. Our results, as given in Table 4.2 and Table 4.3, contrast with those reported by Koehn (2005). His results (albeit on larger training sets taken from the full Europarl corpus) indicate that translation into French is easier, observing a 1.1% higher BLEU score compared to translation into English. In contrast, the work of Way and Gough (2005a) together with our results on the *Sun Microsystems* data set presented in Chapter 3 suggest that translating from French to English is actually easier than English–French (in fact, Koehn (2005) also says that, in general, English is one of the easiest languages to translate into, a statement which appears to conflict somewhat with the results he presents in the same paper). In Tables 4.2 and 4.3 we can see that the results for the PBSMT system supports this hypothesis, performing better for French–English than for English–French. However, in accordance with our previous results from Chapter 3, the EBMT system performs better for French–English than for English–French, in terms of precision (achieving an average increase of 3.14%) and recall (an average increase of 20.44%).

However, we get a conflicting picture when we consider BLEU scores for the EBMT system, which are 2.82% higher, on average, for English–French than for French–English. We conjecture that this apparent contradiction may be more an indication that BLEU (or even any alternative n-gram-based evaluation metric, as a similar discordance is reflected in WER results) may be unsuitable as an evaluation metric when comparing systems which use different translation strategies. Callison-Burch et al. (2006) consider this exact problem by questioning the validity of BLEU when scoring a PBSMT system and a comparative rule-based system, both trained and tested on the same data. Following on from the results of the 2006 NIST MT

evaluation exercise, where for the Arabic–English task the translation system that was ranked 1st by human evaluators was ranked 6th according to the BLEU metric, Callison-Burch et al. (2006) compared the output of two PBSMT systems against that of Systran[3], the leading rule-based MT system Making use of French–English Europarl data to train their PBSMT systems, 300 resulting English translations were evaluated by three human judges in terms of adequacy and fluency. Comparing the human scores for the systems against their automatically generated BLEU scores, Callison-Burch et al. (2006) found that the BLEU score for (the non n-gram-based) Systran underestimated the actual quality of the translations it produced, throwing into doubt the validity of BLEU for scoring such systems. As we see in further experiments in Section 4.3 and Section 4.4, the more SMT-like the EBMT system becomes, the more the BLEU scores seem to correlate with other automatic evaluation metrics, indicating that BLEU is in fact more suited for evaluating MT systems which base their translation strategies heavily on the use of n-grams.[4]

4.3 Hybrid Experiments

From the results presented in Section 4.2 we can see that the performance of the EBMT system falls considerably short of that of the PBSMT system when trained and tested on data taken from the Europarl corpus. Previous experiments on the *Sun Microsystems* dataset (cf Chapter 3) showed that the performance of the baseline PBSMT system could be improved through the use of hybrid translation data Following on from these findings we decided to merge the EBMT marker-based alignments with the PBSMT phrases and words induced from the GIZA++ word alignments for each of the training sets. This time around however, not only did we feed the merged alignment sets to the baseline PBSMT system, thus creat-

[3]http //www.systransoft.com

[4]Of course a manual evaluation of a sample of translations produced by the various systems would provide us with more concrete evidence to support the claim that BLEU underestimates the performance of non n-gram-based systems, but, given the range of experiments carried out here, such an evaluation was beyond the scope of this thesis.

ing a hybrid 'example-based SMT' system as before (Chapter 3), but also to the marker-based EBMT system in order to see whether this new hybrid 'statistical example-based' MT system could also produce translations of higher quality than those produced by either the baseline EBMT or baseline PBSMT systems.

As with the *Sun Microsystems* hybrid translation experiments, we combined the EBMT and PBSMT translation resources in a number of ways in order to improve the performance of our baseline systems:

SEMI-HYBRID EBMT vs. SEMI-HYBRID PBSMT - *Making use of the PBSMT lexicon*: For these experiments we replaced the lexicon of the baseline EBMT system with the higher-quality PBSMT word alignments, in an attempt to improve coverage and lower WER. Both SEMI-HYBRID systems, therefore, make use of the same translation data consisting of EBMT marker chunks and SMT word alignments (with the exception that the SEMI-HYBRID EBMT system, due to the EBMT system design, has the additional ability to make use of generalised templates during translation).

HYBRID EBMT vs. HYBRID PBSMT - *Hybrid 'Statistical EBMT' System vs. Hybrid 'Example-Based SMT' System*: For these experiments we merged all of the PBSMT data (words and phrases) induced from the GIZA++ alignments with the EBMT data (chunks and word alignments) extracted via the marker hypothesis, in order to see if these 'fully hybrid' systems could outperform their baseline equivalents. Again note that the HYBRID EBMT system, as with all EBMT system variants, has the additional ability to use generalised templates which are not used by any variant of the PBSMT system.[5]

As before, we trained on the incremental Europarl training sets of 78K, 156K and 322K sentence pairs and using our 5,000-sentence test set performed translation for French–English and English–French. We describe these experiments in the following

[5]As mentioned previously in Section 3.3, rather than removing the generalised templates from the EBMT system variants, one alternative option would be to include alignment templates (Och et al., 1999) in the PBSMT systems.

Sections 4.3.1 and 4.3.2.

4.3.1 Improving the Lexicon: *SEMI-HYBRID EBMT vs. SEMI-HYBRID PBSMT*

In contrast to the domain-specific *Sun Microsystems* data which consists of rather homogenous sublanguage data, the Europarl corpus consists of much more diverse and complex language. The increased diversity of the language reduces the coverage of the EBMT chunk database, as there is less repetition and the chance of finding a chunk match for an input segment is reduced. When translating the Europarl data, the baseline EBMT system had to backoff to its word-level translation database much more often than when translating the *Sun Microsystems* data. On the Europarl data, the EBMT system had to backoff to its word-level lexicon more often with the result that, on average, 13 words per sentence were considered for direct translation by the word-level lexicon compared to only 7 words per sentence on average which were candidates for direct word-for-word translation on the *Sun Microsystems* data set.

As was expected, the EBMT system seems to perform most poorly when it needs to resort to its lexicon to perform word-for-word translations as the system cannot benefit from the extra contextual information contained within the chunk alignments, reflected in the poor WER scores in the tables, in particular for French–English (cf. Table 4.2), especially when compared to the PBSMT system WER scores (e.g. 83.55% for the baseline EBMT system vs. 69.41% for the baseline phrase-based system, trained on the 156K set).

In order to help address this problem we decided to use the SMT word alignments in place of the EBMT lexicon and repeated the previous experiments for French–English and English–French. From initial empirical investigations into the quality of the lexical alignments, the SMT word alignments appear to be of higher quality than those created via EBMT methods and as there are many more lexical alignments

created by SMT methods, we expect that these alignments will also provide us with broader coverage. In addition, we fed the resulting hybrid data sets, comprising of *EBMT chunks and SMT word alignments*, to the baseline PBSMT system in the same vein as the semi-hybrid 'example-based SMT' system experiments carried out on the *Sun Microsystems* data, to see if we could similarly improve the performance of the system.

French–English Translation

The results for French–English, along with the baseline system results, are given in Table 4.4. From this table we can see that the use of the improved SMT lexicon in the EBMT system leads to only small improvements in the translation quality of the equivalent baseline system. We see a slight improvement in BLEU score for the semi-hybrid EBMT system above the baseline, with an average increase of 2.9% relative BLEU score across all training sets. Disappointingly, we also observe only a slight decrease in WER with an average reduction of just 0.5% absolute, indicating that only improving the lexicon is not enough to improve translation quality sufficiently. Precision and recall are actually worse on the 78K training set, but indicate small improvements when we increase the training set size to 156K and 322K.

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1217	.4556	.5315	85.63	98.94
		SEMI-HYBRID	**.1256**	**.4537**	**.5280**	**85.19**	**98.92**
	PBSMT	BASELINE	.1943	.5289	.5477	70 74	98.42
		SEMI-HYBRID	**.1861**	**.5217**	**.5464**	**73.77**	**98.36**
156K	*EBMT*	BASELINE	.1343	.4645	.5368	83.55	99.02
		SEMI-HYBRID	**.1387**	**.4670**	**.5394**	**82.95**	**99.10**
	PBSMT	BASELINE	.2040	.5369	.5526	69.41	98.30
		SEMI-HYBRID	**.1970**	**.5318**	**.5518**	**72.37**	**98.38**
322K	*EBMT*	BASELINE	.1427	.4734	.5419	82.43	99.06
		SEMI-HYBRID	**.1459**	**.4750**	**.5434**	**81.92**	**99.04**
	PBSMT	BASELINE	.2102	.5409	.5539	68.55	98.72
		SEMI-HYBRID	**.2089**	**.5406**	**.5542**	**70.55**	**98.38**

Table 4.4: Comparing the 'semi-hybrid' EBMT system and 'semi-hybrid' PBSMT system for French–English.

Whereas we see slight improvements for the semi-hybrid EBMT system in gen-

eral, results for the PBSMT system seeded with the EBMT chunks and GIZA++ word alignments are slightly below those for the baseline PBSMT system seeded with SMT phrasal alignments and GIZA++ word alignments, reflected in the increase in WER scores. This indicates that, somewhat in contrast to the results for the *Sun Microsystems* experiments, that for the PBSMT system, the PBSMT phrases are more beneficial to translation performance than the EBMT chunk alignments. BLEU score decreases on average by 2.7% relative across the three training sets. Recall and precision also experience a very slight drop, with recall results leveling off when we reach the 322K sentence training set.

English–French Translation

From Table 4.5 we can see again that BLEU scores increase slightly for the semi-hybrid EBMT system seeded with marker-based chunks and SMT word alignments over the baseline system seeded with EBMT chunk and word alignments, with relative BLEU score increasing by 3.3% on average. Again WER only drops slightly, with an average decrese of just 0.48%.

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1240	.4422	.4365	79.09	99.10
		SEMI-HYBRID	**.1304**	**.4515**	**.4445**	**78.20**	**99.10**
	PBSMT	BASELINE	.1771	.5045	.4696	70.44	98.54
		SEMI-HYBRID	**.1670**	**.4968**	**.4617**	**73.32**	**98.46**
156K	*EBMT*	BASELINE	.1374	.4548	.4476	77.66	98.96
		SEMI-HYBRID	**.1404**	**.4592**	**.4517**	**77.34**	**98.96**
	PBSMT	BASELINE	.1855	.5120	.4724	69.39	98.20
		SEMI-HYBRID	**.1773**	**.5048**	**.4681**	**72.27**	**98.26**
322K	*EBMT*	BASELINE	.1448	.4587	.4530	77.73	99.22
		SEMI-HYBRID	**.1486**	**.4632**	**.4575**	**77.51**	**99.56**
	PBSMT	BASELINE	.1933	.5180	.4751	68.30	98.12
		SEMI-HYBRID	**.1850**	**.5107**	**.4708**	**71.22**	**98.22**

Table 4.5: Comparing the 'semi-hybrid' EBMT system and 'semi-hybrid' PBSMT system for English–French.

As for French–English, for English–French the semi-hybrid PBSMT system actually performs slightly worse than the equivalent PBSMT baseline system. On average BLEU score falls by 4.87% relative, with precision and recall scores also de-

creasing, and WER increases by 2.89% on average. These results again show that although the SMT lexicon can help improve the performance of the EBMT system, it is not enough to increase translation quality of the EBMT system sufficiently.

4.3.2 Merging All Data: *HYBRID EBMT vs. HYBRID PBSMT*

From the results in Section 4.3.1 it is clear to see that using the higher quality SMT lexicon in place of the EBMT word-level database within the EBMT system results in improvements over the baseline system. However, for the PBSMT system, making use of EBMT chunk alignments in place of SMT phrasal alignments in the semi-hybrid system configuration actually decreases translation quality

Experiments carried out on the *Sun Microsystems* data, as described in Chapter 3, show that making use of both EBMT sub-sentential alignments (words and phrases) along with SMT word and phrase alignments can improve the performance of a PBSMT system. Following on from these results and the somewhat disappointing results of the experiments described in Section 4.3.1, we merged all of the EBMT marker-based alignments and the PBSMT phrases (and words) induced from the GIZA++ word alignments. We fed the resulting merged translation resources to both the baseline EBMT and PBSMT systems to create our hybrid 'statistical EBMT' system and our 'example-based SMT' system, referred to as HYBRID-EBMT and HYBRID-PBSMT, respectively, in what follows. These experiments allow us to see whether we can achieve similar improvements when using hybrid data sets to those seen on the *Sun Microsystems* corpus. They also allow us to implicitly investigate the possible differences in quality of these data resources and their effect on translation (cf. Section 4.5 for further discussion).

French–English Translation

The results for French–English are given in Table 4.6. From these results it is clear that adding the hybrid data improves over all baseline results. Most importantly, as we showed for the *Sun* data in Chapter 3, incorporating the EBMT marker chunks and the PBSMT sub-sentential alignments in a hybrid 'example-based SMT' system improves on the baseline PBSMT performance. Increases in BLEU score are consistent across training sets, rising from 0.194 to 0.207 (a 6.7% relative improvement) for the HYBRID-PBSMT system on the 78K-sentence training set, from 0.204 to 0.218 on the 156K set (6.9% improvement) and from 0.210 to 0.224 on the 322K set (a 6.7% relative improvement). These improvements were deemed to be statistically significant (with $p < 0.05$) Koehn (2004b). Precision and Recall rise, and WER falls; somewhat surprisingly, SER rises, but for all bar this metric, the hybrid system outperforms both system variants on which it is based.

One very interesting result is that the HYBRID-PBSMT system achieves a higher BLEU score when trained on the 78K data set compared with the baseline system trained on twice as much data (0.207 vs. 0.204). We obtain a similar result for the 156K set (0.218 for the HYBRID-PBSMT system vs. 0.210 for the baseline system trained on the 322K data set). Using the hybrid data sets, it is possible to achieve much higher translation performance with much less data, which may be particularly useful for language pairs where smaller amounts of training data are available.

For EBMT, we see a greater increase over the baseline system; on the 78K training set, BLEU scores rise from 0.122 to 0.152 (24% relative improvement), from 0.134 to 0.162 on the 156K set (20.6% relative increase) and from 0.143 to 0.170 on the 322K training set (a 19.6% relative improvement). We also observe a drop in WER (e.g. from 82.43% to 74.99% for the 322K set). The improvements in the performance of the HYBRID-EBMT system is also reflected in the increase in chunk coverage. The SMT phrases add robustness to the EBMT system by improving this coverage. A further 6% of test sentences are successfully translated by the HYBRID-

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1217	.4556	.5315	85.63	98.94
		SEMI-HYBRID	.1256	.4537	.5280	85.19	98.92
		HYBRID	**.1519**	**.4997**	**.5349**	**76.12**	**98.98**
	PBSMT	BASELINE	.1943	.5289	.5477	70 74	98.42
		SEMI-HYBRID	.1861	.5217	.5464	73.77	98.36
		HYBRID	**.2070**	**.5368**	**.5551**	**69.56**	**98.20**
156K	*EBMT*	BASELINE	.1343	.4645	.5368	83.55	99.02
		SEMI-HYBRID	.1387	.4670	.5394	82.95	99.10
		HYBRID	**.1620**	**.5081**	**.5457**	**75.14**	**99.16**
	PBSMT	BASELINE	.2040	.5369	.5526	69.41	98.30
		SEMI-HYBRID	.1970	.5318	.5518	72.37	98.38
		HYBRID	**.2176**	**.5437**	**.5579**	**68.17**	**98.10**
322K	*EBMT*	BASELINE	.1427	.4734	.5419	82.43	99.06
		SEMI-HYBRID	.1459	.4750	.5434	81 92	99.04
		HYBRID	**.1699**	**.5145**	**.5558**	**74.99**	**99.20**
	PBSMT	BASELINE	.2102	.5409	.5539	68.55	98.72
		SEMI-HYBRID	2089	.5406	.5542	70 55	98.38
		HYBRID	**.2236**	**.5483**	**.5592**	**67.40**	**98.58**

Table 4.6: Comparing the 'hybrid' EBMT system and 'hybrid' PB-SMT system for French–English.

EBMT system using chunks alone (i.e. not having to resort to the word-level lexicon during translation) and we see an average relative increase of 76% for the number of possible chunk translation candidates contained in the HYBRID-EBMT database for a particular input sentence.

English–French Translation

The results for the reverse language direction are given in Table 4.7. Here again we can see that adding the hybrid data improves over all baseline results The addition of the hybrid data into the baseline EBMT system results in BLEU scores rising from 0.124 to 0 146 for the 78K data set (17.7% relative improvement), from 0.137 to 0.157 for the 156K data set (14.6% relative increase) and from 0.149 to 0.167 for the 322K data set (12.1% relative improvement). Again, these improvements are statistically significant (Koehn, 2004b). It is interesting to note that for this language direction the HYBRID-EBMT system trained on only 78K sentence pairs performs almost as well as the baseline system trained on over four times as much data.

As with the HYBRID-EBMT system, the HYBRID-PBSMT system improves over its baseline equivalent, on average achieving a relative increase in BLEU score of 6.2%. As with French–English, the SER rate remains high over all training sets, but we see a decrease in WER for the hybrid systems compared to their baseline equivalents (an average decrease of 4.05% for the HYBRID-EBMT system and 1.09% for the HYBRID-PBSMT system).

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1240	.4422	.4365	79 09	99.10
		SEMI-HYBRID	.1304	.4515	.4445	78 20	99.10
		HYBRID	**.1462**	**.4783**	**.4580**	**74.54**	**99.20**
	PBSMT	BASELINE	.1771	.5045	.4696	70.44	98.54
		SEMI-HYBRID	.1670	.4968	.4617	73.32	98.46
		HYBRID	**.1898**	**.5152**	**.4787**	**69.20**	**98.50**
156K	*EBMT*	BASELINE	.1374	.4548	.4476	77.66	98.96
		SEMI-HYBRID	.1404	.4592	.4517	77.34	98.96
		HYBRID	**.1573**	**.4870**	**.4682**	**73.86**	**99.16**
	PBSMT	BASELINE	1855	.5120	.4724	69.39	98.20
		SEMI-HYBRID	1773	.5048	.4681	72 27	98.26
		HYBRID	**.1965**	**.5208**	**.4810**	**68.36**	**98.24**
322K	*EBMT*	BASELINE	.1448	.4587	.4530	77.73	99 22
		SEMI-HYBRID	.1486	.4632	.4575	77.51	99.56
		HYBRID	**.1668**	**.4912**	**.4798**	**73.94**	**99.38**
	PBSMT	BASELINE	.1933	.5180	.4751	68.30	98.12
		SEMI-HYBRID	.1850	.5107	.4708	71.22	98.22
		HYBRID	**.2040**	**.5284**	**.4851**	**67.28**	**98.12**

Table 4.7: Comparing the 'hybrid' EBMT system and 'hybrid' PB-SMT system for English–French.

Here we can see that with the inclusion of the SMT data, the BLEU scores now fall more in line with the remaining evaluation metrics (as compared to the results in Table 4.2 and Table 4.3). It appears that the more SMT-like the EBMT system becomes, the more the BLEU scores tend to reflect the same trend as displayed by the remaining automatic evaluation metrics, with BLEU scores now higher for French–English (Table 4.6) than English–French (Table 4.7). As BLEU is an n-gram statistic (cf. Section 3.2.2 for more details on BLEU), the more n-gram-based the system becomes, the more reliable BLEU becomes, further strengthening the claim that BLEU is not a suitable evaluation metric for reflecting translation quality when

we are dealing with non-n-gram-based systems (Callison-Burch et al., 2006).[6]

4.4 Language Model Reranking: Further Hybridity

The results in Section 4.3.2 show that although making use of both EBMT and SMT translation resources in a hybrid 'statistical EBMT' system result in statistically significant improvements over the baseline EBMT system, translation quality for the hybrid system configuration is still somewhat short of that for the baseline PBSMT system for both French–English and English–French.

To understand more fully why this is the case, we performed an analysis of how the EBMT system actually generates a translation. As mentioned previously in Section 4.3.1, unlike the *Sun Microsystems* corpus which consists of rather homogeneous data, the Europarl corpus consists of very diverse and much more complex language. This is reflected in the differences in chunk coverage of the EBMT system on the Europarl and *Sun* test sets. Due to the more repetitive nature of the *Sun* data, many more sentences were translated fully by EBMT chunks alone (approx 6% of translations) than on the Europarl data (approx 1% of test sentences). Note also that the possibility of producing a perfect translation (or an exact sentence match) was also much higher for the *Sun* experiments. This is reflected in the SER rates achieved (65.6% for English–French and 51.2% for French–English, when trained on 203K *Sun* sentence pairs (cf. Tables 3.2 and 3.3, respectively) vs. 98-99% SER for both language directions when training on the Europarl data sets (cf. Table 4.2 for French–English and Table 4.3 for English–French).

As we pointed out in Section 4.3.1, the EBMT system performs most poorly when it cannot make use of its chunk or template databases and needs to resort to using its lexical database, performing word-for-word translation. However, even improving

[6]Ideally we need to carry out manual evaluation to further support this claim, something which is beyond the scope of this thesis.

the quality of this lexicon by replacing it with the higher quality set of SMT word alignments was not adequate to sufficiently increase translation performance. The reasons for this lie with how the EBMT system's recombination stage operates.

During recombination, the EBMT system of Gough and Way (2004a, 2004b), Way and Gough (2005) and Gough (2005) takes a completely naive approach, simply positioning translated input segments (chunks and words from the input sentence) according to their original source language order. For languages which share similar syntactic structure, where translationally equivalent constituents are generally realised in the same or similar positions in both source and target, such as for French–English, this technique, in most instances, is completely adequate as any reorderings that occur when translating between these two languages generally occur at a local level and can be encapsulated within the aligned chunks themselves. Without any guide as to the correct target language word order, whenever the EBMT system cannot avail of its chunk alignments and has to resort to word-for-word translation it simply follows the order of the words in the original input sentence and thus often fails to produce a syntactically well-formed output translation.

In addition, the EBMT system lacks any sophisticated target language probabilistic models to aid the lexical selection process, and often when a number of equally weighted lexical choices are available for a particular input segment, the system outputs all of the possible alternate final translations as part of an n-best list, with many translations assigned the same score.

Therefore, although the improved word alignments may provide more suitable lexical translation candidates, the EBMT system requires more information to improve both lexical selection and final word order.

In order to see if we could further improve the performance of the hybrid EBMT system we decided to integrate a statistical language model, following the work of Bangalore et al. (2002), who demonstrate that using a trigram language model to select the final translation output from multiple candidates improves system performance. For these experiments we ranked the output of the hybrid EBMT system

using the PBSMT system's equivalent language model (e.g. the 78K language model to rerank the output of the EBMT system training on the 78K sentence test set) in a *post hoc* reranking stage. Making use of the SRILM toolkit's n-best rescoring tool[7] (Stolcke, 2002), we rescored the n-best list produced by the hybrid EBMT system for each input sentence, multiplying the resulting statistical language model score by the original score assigned to the translation candidate by the EBMT system. We took the resulting top-scoring translation for each input sentence to be the final translation which was subsequently passed to the automatic evaluation metrics. The statistical language model should help the EBMT system in particular where it has to resort to performing direct word-for-word translation where it cannot benefit from the contextual information contained within its chunk database, thus improving the translation fluency.

Apart from Bangalore et al. (2002), other hybrid systems have previously integrated language models, but in different ways to our approach. For instance, Aue et al. (2004) add a dependency (or 'logical form') tree-based statistical language model into their EBMT system, but with little improvement over the baseline system. The hybrid system of Quirk and Menezes (2006) also contains a target language model, but in a much more straightforward SMT-like experiment. With a similar aim to our work here, Imamura et al. (2004) employ syntactic transfer strategies, using a language model (together with a lexicon model) as part of their statistical generation module for final lexical selection.

4.4.1 French–English Translation

Looking at the results in Table 4.8, we can see that using the language model does improve the performance of the 'fully-hybrid' 'statistical EBMT' system. These results illustrate how the language model guides the reordering and lexical selection of these word-to-word translations to improve overall translation quality.

For the hybrid 'statistical EBMT' system, the BLEU score rises by 6–7% relative

[7] http://www.speech.sri.com/projects/srilm/

across training sets. Precision rises, recall stays about the same, but WER improves by about 2% absolute on average across training sets. SER falls only slightly, by less than 1% absolute on average across training sets. These improvements were deemed to be statistically significant (Koehn, 2004b) ($p < 0.01$).

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1217	.4556	.5315	85.63	98.94
		SEMI-HYBRID	.1256	.4537	.5280	85 19	98.92
		HYBRID	.1519	.4997	.5349	76.12	98.98
		HYBRID-LM	**.1624**	**.5091**	**.5341**	**74.15**	**98.80**
	PBSMT	BASELINE	.1943	.5289	.5477	70.74	98 42
		SEMI-HYBRID	.1861	.5217	.5464	73.77	98.36
		HYBRID	.2070	.5368	.5551	69.56	98 20
156K	*EBMT*	BASELINE	.1343	.4645	.5368	83.55	99 02
		SEMI-HYBRID	.1387	.4670	5394	82.95	99.10
		HYBRID	.1620	.5081	5457	75.14	99.16
		HYBRID-LM	**.1722**	**.5177**	**.5463**	**73.19**	**98.68**
	PBSMT	BASELINE	.2040	.5369	.5526	69.41	98.30
		SEMI-HYBRID	.1970	.5318	.5518	72.37	98.38
		HYBRID	.2176	.5437	.5579	68.17	98.10
322K	*EBMT*	BASELINE	.1427	4734	.5419	82.43	99.06
		SEMI-HYBRID	.1459	.4750	.5434	81.92	99.04
		HYBRID	.1699	5145	.5558	74.99	99 20
		HYBRID-LM	**.1773**	**.5224**	**.5530**	**72.85**	**98.80**
	PBSMT	BASELINE	.2102	.5409	.5539	68 55	98.72
		SEMI-HYBRID	.2089	.5406	.5542	70.55	98.38
		HYBRID	.2236	.5483	.5592	67 40	98.58

Table 4.8: Re-ranking the output of the 'fully-hybrid' EBMT system for French–English.

4.4.2 English–French Translation

Similar statistically significant improvements can also be seen for English–French (Table 4.9). From these results we can see that we get an average relative increase of 4.8% BLEU score and again WER scores improve, falling from an average of 74.11% for the hybrid EBMT system to 73.55% with the addition of a language model. As for French–English, the improvements in results for English–French are statistically significant (with $p < 0.05$) Koehn (2004b).

A summary of the BLEU score results for the various EBMT and PBSMT systems,

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1240	.4422	.4365	79 09	99.10
		SEMI-HYBRID	.1304	.4515	.4445	78.20	99 10
		HYBRID	.1462	.4783	.4580	74.54	99.20
		HYBRID-LM	**.1527**	**.4871**	**.4611**	**73.23**	**99.08**
	PBSMT	BASELINE	.1771	.5045	.4696	70.44	98.54
		SEMI-HYBRID	.1670	.4968	.4617	73.32	98.46
		HYBRID	.1898	.5152	.4787	69.20	98.50
156K	*EBMT*	BASELINE	.1374	.4548	.4476	77.66	98.96
		SEMI-HYBRID	.1404	.4592	.4517	77 34	98.96
		HYBRID	.1573	.4870	.4682	73.86	99.16
		HYBRID-LM	**.1635**	**.4955**	**.4709**	**72.50**	**98.88**
	PBSMT	BASELINE	.1855	.5120	.4724	69 39	98.20
		SEMI-HYBRID	.1773	.5048	.4681	72.27	98.26
		HYBRID	.1965	.5208	.4810	68 36	98.24
322K	*EBMT*	BASELINE	.1448	.4587	.4530	77.73	99.22
		SEMI-HYBRID	.1486	.4632	4575	77.51	99.56
		HYBRID	.1668	.4912	4798	73.94	99.38
		HYBRID-LM	**.1744**	**.5014**	**.4818**	**71.96**	**98.80**
	PBSMT	BASELINE	.1933	.5180	.4751	68.30	98.12
		SEMI-HYBRID	.1850	.5107	.4708	71.22	98.22
		HYBRID	.2040	5284	.4851	67.28	98.12

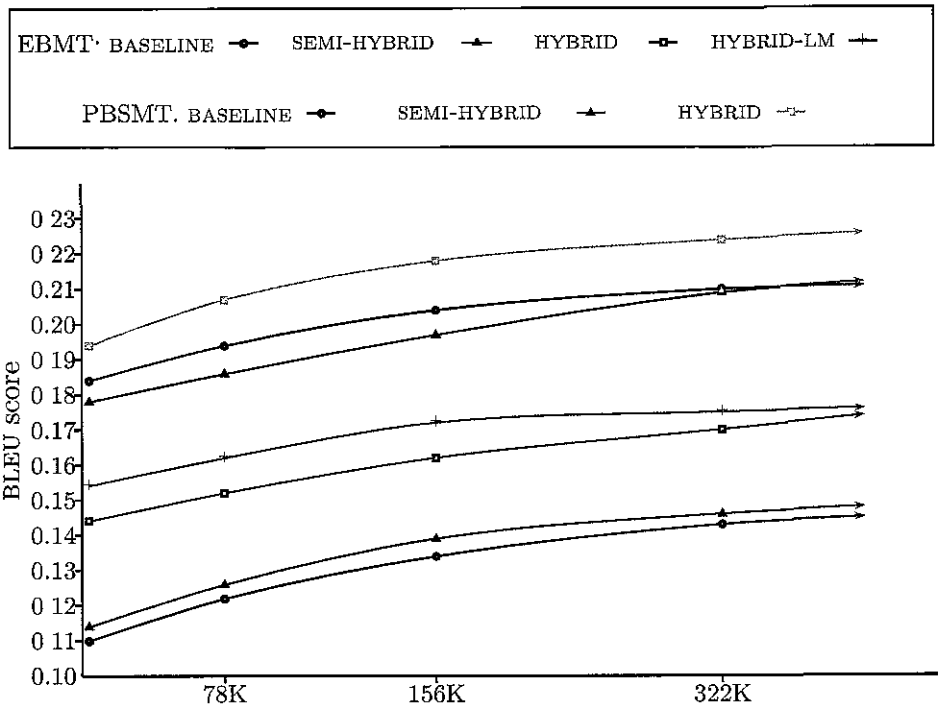
Table 4.9· Re-ranking the output of the baseline EBMT system and 'hybrid' EBMT system for English–French.

for French–English and for English–French translation, are given in the graphs in Figure 4.2.

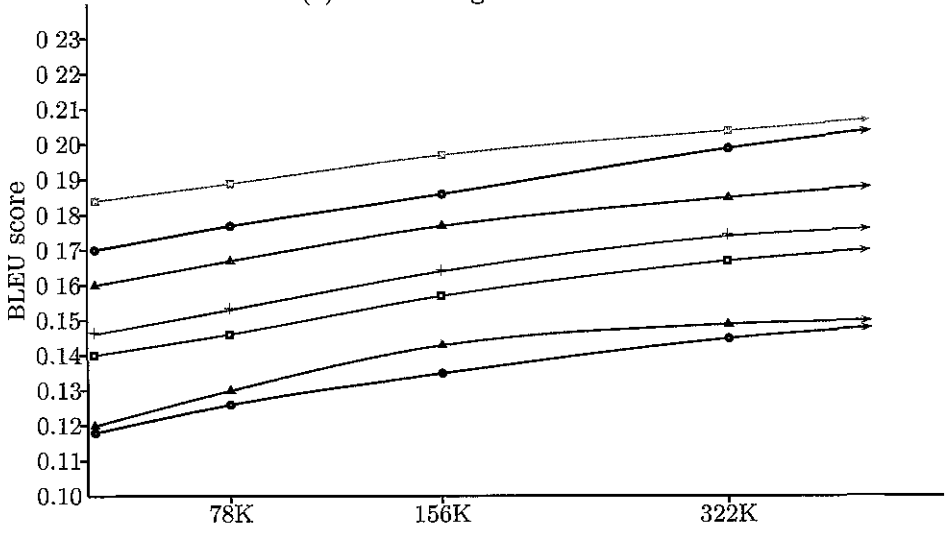
4.5 Comparing EBMT Chunks and SMT Phrases

The experiments in Section 4.3 and Section 4.4 show that adding various PBSMT resources to an EBMT system results in a novel 'statistical EBMT' model of translation that is capable of outperforming the baseline EBMT system of Gough and Way (2004b), Gough (2005) and Way and Gough (2005a). However, in contrast to the experiments on the *Sun Microsystems* corpus, even the highest performing hybrid EBMT system was unable to match the performance of the baseline PBSMT system.

Despite this somewhat surprising result, the results in Table 4.6 and Table 4.7



(a) French–English translation



(b) English–French translation

Figure 4.2: BLEU scores for the various configurations of the EBMT and PBSMT systems when trained on incremental training data sets taken from the Europarl corpus (Koehn, 2005)

demonstrate that incorporating the marker chunks with the PBSMT sub-sentential alignments in an 'example-based SMT' system can outperform both baseline systems from which it is created.

From these results it is clear that both EBMT and PBSMT translation resources can contribute positively to translation quality In this section, we examine these translation resources further by providing a detailed description of the types of phrasal alignments that each system generates, in a bid to increase our understanding of the contributions both EBMT and SMT methods can make towards translation performance and the apparent contradiction the results in this chapter have with those present in Chapter 3.

For the various training sets, the numbers of chunks derived by both systems is shown in Table 4.10.

	SMT	EBMT	BOTH	SMT ONLY	EBMT ONLY
78K	1.17M	242,907	47,311	1.12M	195,596
156K	2 45M	470,588	92,662	2.36M	378,026
322K	5.15M	928,717	181,669	4.97M	747,048

Table 4.10. The number of chunks derived by the PBSMT and EBMT baseline systems, over the different training sets

From the information in Table 4.10, one can see that there are many more SMT chunks than EBMT fragments: over five times as many on average for all training set sizes. As SMT methods extract all possible n-grams that correspond to initial word alignments (which may even include complete sentences) and place no further restrictions on the content or context of these n-grams, this result is to be expected; EBMT marker-based methods, by their very definition, are much more restrictive in their identification of a valid chunk, as they base their decisions heavily on specific contextual factors, namely the presence of marker words.

It is interesting to note, however, that despite the huge differences in the amount of SMT chunks vs. EBMT chunks, this is not reflected in the results in the previous sections. It would be expected that the EBMT system would fare much worse due to its seeming lack of extracted resources, but it performs reasonably well compared

with the equivalent PBSMT system with over five times as many alignments (cf. Tables 4.2 and 4.3). This result can be seen as an indication of the higher quality of the information encapsulated within the EBMT chunks, and even the possible superfluous nature of many of the SMT n-grams.

Note that doubling the amount of training data leads to about twice as many sub-sentential alignments for each system type. This is a further indication of the heterogeneous nature of the Europarl corpus, which, despite becoming the standard in MT training and testing, is still perhaps quite a challenging domain for MT. The lack of repetition within the corpus makes it difficult for any type of MT system to automatically learn from it.[8] What is interesting, of course, given the number of chunks created, as given in Table 4.10, is to see what the overlap is between the chunks that the two systems create, as well as which chunks are derived by just one of the two systems. In this way we can see how different the two types of approaches are in terms of the translation knowledge they extract.

For the 322K-sentence training set, over 5.71 million chunks are created in total. About 182K chunks, or just over 3% of the total number, are found by both systems. Of the remainder, about 87% are SMT-only chunks, and 13% are derived solely by the EBMT system. These figures alone indicate the huge differences between the phrasal alignments extracted via SMT and EBMT methods, as we observe an extremely small amount of alignments common to both systems. Of those alignments which are found exclusively by either system, interestingly, 93% of the SMT-only chunks are found just once, and 99.4% occur less than 10 times; for the EBMT chunks, 96.63% are found once, while 99.8% are seen less than 10 times.

Taking the 322K-sentence training set, the ten most frequent chunks found by both systems (i.e. the most frequent chunks occurring within the intersection set)are those in (21):

[8]Incidentally, for the *Sun Microsystems* 203K-sentence training set, approx 1.99M phrase alignments were extracted in total between SMT and EBMT methods. Almost 1M more alignments were extracted on the 156K Europarl training set (cf. Table 4.10, even though it contains 47K fewer sentence pairs, further indication of the lack of repetition in the Europarl data compared with the *Sun* TM

(21) a. monsieur le président ⟹mr president

b au nom ⟹on behalf

c. je pense ⟹i think

d. madame le président ⟹madam president

e. je crois ⟹i believe

f. je pense ⟹i believe

g. et messieurs ⟹and gentlemen

h. par exemple ⟹for example

i. la commission ⟹the commission

j. madame la présidente ⟹madam president

It can be seen from the chunks in (21) that they are generally of good quality and apart from (21d) which incorrectly contains the masculine form *le président* in French of the equivalent English noun *president*, rather than the correct feminine form *la présidente* (cf (21j), for example), consist of well-formed syntactic units. These are exactly the type of alignments we would expect to be most useful during translation.

The ten most frequent chunks found by just the EBMT system are those in (22):

(22) a. de l union européenne ⟹of the european union

b. le vote aura lieu demain ⟹the vote will take place tomorrow

c. à la commission ⟹the commission

d. du jour appelle ⟹the next item is

e. d accord ⟹i agree

f. dans l union européen ⟹in the european union

g. de l union européen ⟹in the european union

h. tout d abord ⟹first

i. la séance est levée ⟹the sitting was closed

j à l avenir ⟹in future

Again, as with the chunks in (21), the chunks in (22) are generally of good quality (apart from 22(f–g), which contain the French masculine form *européen* of the

English adjective *european*, rather than the appropriate feminine form *européenne* required by the French feminine noun *union*, although these errors would be simple to post-edit) and contain important contextual information which is useful for translation, such as determiner–noun agreement (e g. 22(i), between the determiner *la* and the feminine noun *séance*) and noun–adjective agreement (e.g. 22(a), between *union* and *européenne*)

The ten most frequent chunks found by just the SMT system are those in (23):

(23) a. du ⟹of the

b. de la ⟹of the

c. union européenne ⟹union

d. états membres ⟹member states

e. de l ⟹of the

f. dans le ⟹in the

g. n est ⟹is

h parlement européen ⟹parliament

i. que nous ⟹that we

j. que la ⟹that the

From the chunks in (23) we can see that the chunks extracted by just the SMT system are also largely correct, but contain different types or errors than those made by the EBMT system, e.g. (23c) and (23g) contain actual semantic errors. The remaining chunks, although they can be considered well-formed, are not particularly informative and contain very little contextual information which would greatly affect the resulting translation. Most consist of groups of marker or function words, without the accompanying content words which modify them. These are, of course, notoriously difficult to translate, and constitute one of the main reasons why marker-based translation is worthwhile in the first place.

Leading on from this observation, one possibility that we will consider for future work would be to use elements of the statistical method of chunk alignment to automatically identify and extract marker-word translation candidates, which could

then be fed to the marker-based chunker to fully automate the chunk alignment and extraction process.

Given that both system types derive some chunks that the other does not, it is unsurprising that our claim that better systems can be built when combining both sets of sub-sentential alignments is borne out to be correct. Given the results from Chapter 3, and those presented in this Chapter, the implications for designers of data-driven MT systems are obvious: if available, incorporating sub-sentential resources from both SMT and EBMT into novel hybrid 'example-based SMT' and 'statistical EBMT' systems guarantee that systems will be derived that are capable of higher translation quality than either standalone translation variant. That is, while there is an obvious convergence between both paradigmatic variants (Way and Gough, 2005a; Wu, 2006), more gains are to be had from combining their relative strengths in novel hybrid systems.

4.6 Summary

In Chapter 3 we demonstrated that the marker-based EBMT system of Gough and Way (2004b), Gough (2005) andWay and Gough (2005a) is capable of outperforming a PBSMT system constructed from freely available resources for the task of sublanguage translation. However, perhaps more importantly, we showed that a hybrid 'example-based SMT' system incorporating marker chunks together with SMT sub-sentential alignments (words and phrases) is capable of outperforming both baseline example-based and statistical phrase-based translation models on which it is based.

On a different data set – the Europarl corpus – in this chapter we demonstrated that the baseline PBSMT system achieves higher translation quality than the EBMT system of Gough and Way (2004b), Gough (2005) andWay and Gough (2005a). For the most part, we feel that this is due to the heterogeneous nature of the training data compared to the *Sun Microsystems* TM that was used in previous experiments; due to its domain and wide range of subject matter, the Europarl contains much more

complex language and varied vocabulary than the *Sun Microsystems* sublanguage data.

We demonstrated that in a number of novel improvements, the baseline EBMT system can be improved by adding in resources derived from a PBSMT system. We observed that making use of the higher-quality SMT word alignments improves translation quality slightly for the EBMT system; although making use of the same data set of EBMT chunk alignments and SMT word alignments has an adverse negative effect on the performance of the PBSMT system. However, we observe statistically significant improvements over the baseline EBMT system, when we create a novel hybrid 'statistical EBMT' system by combining all of the EBMT translation data (chunk and word alignments) with the data (phrases and words) induced via GIZA++. Incorporating a statistical target language model in a *post hoc* reranking stage improves translation quality still further. However, the novel hybrid 'statistical EBMT' systems continue to fall short of the translation quality achieved by the baseline PBSMT system.

Nevertheless, as with the experiments in Chapter 3 and the work presented in Groves and Way (2005), we confirmed in a further experiment that adding the EBMT marker chunks and word-level lexicon to the baseline SMT system derived an 'example-based SMT' system that was capable of improving translation quality compared to the baseline PBSMT system, for a range of automatic evaluation metrics. In fact, both this hybrid 'example-based PBSMT' system and the hybrid 'statistical EBMT' system configuration perform as well as their baseline equivalents trained on twice, and even in certain cases, four times as much data (e.g. Table 4.7 HYBRID-EBMT vs. EBMT baseline). This particular outcome has significant implications for languages with scarce resources, where large amounts of bilingual corpora are not readily available.

Given these results, we explored the nature of the chunks produced automatically by both underlying system types. Perhaps unsurprisingly, a number of chunks remain derivable by just one of the system types: either EBMT, or SMT. Looking

at the intersection and mutually exclusive alignment sets, it seems that for the hybrid system configurations, during translation, phrases extracted by both methods that are more likely to function as syntactic units (and therefore more beneficial during the translation process) are given a higher statistical significance, whereas, conversely, those superfluous (possibly even 'less useful') SMT n-grams that are not also generated by the EBMT system are reduced. Essentially, the EBMT data helps the SMT system to make the best use of phrase alignments during translation, whereas the SMT data, containing many more alignments, helps to improve the coverage of the system.

Given that these two sets of alignments can contribute positively to the overall translation quality, the consequences for the field of data-driven MT are clear; by incorporating sub-sentential resources from both SMT and EBMT into novel hybrid systems, translation quality will improve compared to the baseline variants. That is, while there is an obvious convergence between both paradigmatic variants, more gains are to be had from combining their relative strengths in novel hybrid systems.

Following on from the success of the hybrid approaches presented in this chapter and also in Chapter 3, in Chapter 5 we outline a number of further developments which have been made into the creation of a hybrid data-driven architecture which can take advantage of the benefits of both SMT and EBMT methods. We also describe a number of experiments which have been made involving different language pairs and data sets, including investigations into the development of a data-driven marker-based EBMT decoder.

Chapter 5

Further Applications and Development

The results presented in Chapter 3 and Chapter 4 demonstrated clearly how making use of both SMT and EBMT lexical, and in particular, phrasal information can significantly improve translation results. In addition, in Chapter 4 we demonstrated how making use of a statistical language model to rerank the n-best lists output by the marker-based EBMT system improved upon all system variants. These experiments are further indications that both EBMT and SMT methods can be employed to produce hybrid system configurations capable of outperforming their baseline equivalents.

Following on from these results, a number of other research avenues have presented themselves. In this Chapter we describe a number of different hybrid data-driven experiments which have been carried out as a result of our significant findings. In Section 5.1, we describe how we have begun to develop a hybrid data-driven system architecture, MaTrEx, which can take advantage of the scientific findings presented in this work and which can also facilitate further research into the area of data-driven MT.

In order to help determine the best configuration for the new marker-based alignment strategy used in the MaTrEx system, we carried out some developmental

experiments as part of the Spanish–English OpenLab 2006 shared task on MT. We present these experiments in Section 5.2. This translation task not only presents us with a new language pair but also provides us with a very large data set to test the scalability of our new MT architecture.

In Section 5.3 we describe a set of experiments performed on Basque–English translation. As Basque is an agglutinative and highly inflected language, these experiments involve carrying out novel morpheme-to-word alignment using the marker hypothesis in addition to chunk alignment. This set of experiments also demonstrate the use of existing tools for Basque chunking, which given the MaTrEx system's modular design, are easily utilised. In these experiments we investigate the benefits of incorporating EBMT-style chunk alignments within the system's translation model in a bid to further strengthen our claim that SMT and EBMT systems can benefit within a hybrid framework.

In addition to describing the development of our new hybrid data-driven architecture and experiments on new language pairs, we discuss some preliminary work that has carried out on the use of additional hybrid data-driven techniques. In Section 5.4, we propose a new "example-based" decoder for use within this new data-driven architecture which employs marker-based segmentation strategies together with SMT decoding techniques.

5.1 The MaTrEx System

In order to make use of the findings from Chapter 3 and Chapter 4, we decided to develop a new software framework for data-driven MT capable of taking advantage of hybrid knowledge resources. The MaTrEx (**Ma**chine **Tr**anslation using **Ex**amples) system used in our work and described in Armstrong et al. (2006), Stroppa et al. (2006) and Stroppa and Way (2006) is a modular data-driven MT engine, built following established Design Patterns (Gamma et al., 1995).

An overview of the system architecture is given in Figure 5.1.

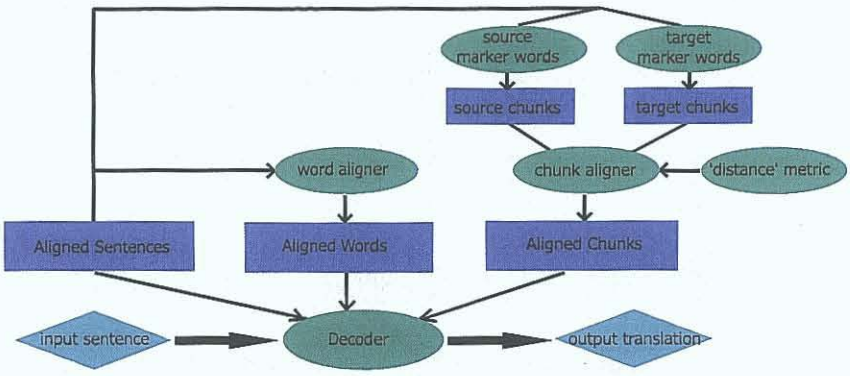


Figure 5.1: The MaTrEx system architecture

From Figure 5.1, we can see that the system comprises a number of easily extendible and re-implementable modules, the most important of which are:

- *Word Alignment Module*: takes as its input a bilingual sententially-aligned corpus and outputs a set of word alignments with their associated probabilities
- *Chunking Module*: takes in a bilingual sententially-aligned corpus and produces source and target chunks.
- *Chunk Alignment Module*: takes the source and target chunks produced by the chunking module and aligns them on a sentence-by-sentence basis producing a final set of source–target chunk alignments together with their translation probabilities (based on relative frequencies).
- *Decoder*: searches for the translation of the input text, making use of the original sententially-aligned corpus together with the derived chunks and word alignments.

The system's modular design means that all of these above modules, indicated by the highlighted components in Figure 5.2, can be completely reimplemented, extended or even adapted to allow the integration of existing software (i.e. the use of wrapper technologies).

In its current configuration, the system employs marker-based chunking techniques to derive source and target chunks. For new language pairs, we need only

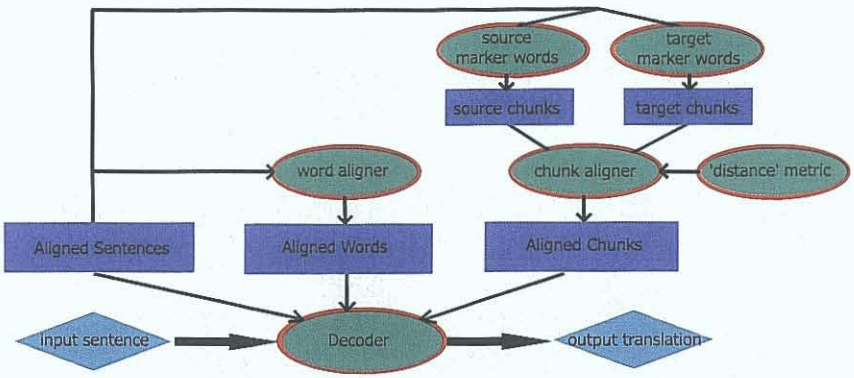


Figure 5.2: Reimplementable modules within the MATREX system.

provide the system with the appropriate list of source and target marker words. However, other chunking strategies are also implementable, such as the use of existing chunking tools as described in Section 5.3.2. During training, the aligned source–target sentences are passed in turn to the word alignment, chunking and chunk alignment modules to create our chunk and lexical databases. In addition, SMT phrasal alignments extracted based on the set of derived word alignments can be easily incorporated into the system's architecture as an additional module (cf. Section 2.3.4).

For the experiments described in this chapter the word alignment and decoder modules are simply wrappers around existing tools. To extract word-level alignments from the original bilingual corpus, the GIZA++ statistical word alignment tool is used (Och and Ney, 2003).[1] From the resulting word alignment sets produced for source–target and target–source language directions, the refined word alignment set is computed following the methods of Koehn et al. (2003) and Och and Ney (2003) as described in Section 2.3.3. This type of word alignment technique has been shown to produce high-quality alignments with equally high recall making it particularly suited to the MT task (Tiedemann, 2004). For decoding we made use of the PHARAOH phrase-based SMT decoder (Koehn, 2004a), as described in Section 2.3.5.

In its current configuration, the MATREX system is, therefore, essentially a hy-

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

brid 'example-based SMT' system, as it makes use of a traditional phrase-based SMT decoder, together with both example-based and SMT-derived translation knowledge, which when combined, make up the system's translation table.

5.1.1 Adding Punctuation to the Marker Set

The MaTrEx system's chunking module takes as its input a sententially-aligned corpus, and chunks each sentence pair on a sentence-by-sentence basis. The current instantiation of the chunking module implements marker-based chunking techniques, similar to those of Way and Gough (2003, 2005a, 2005b), Gough and Way (2004a, 2004b) and Gough (2005) The marker-based approach to chunking makes use of a set of closed-class words to segment source and target sentences, as described in Section 2.2.

In previous experiments on the *Sun Microsystems* and Europarl data sets, as described in Chapter 3 and Chapter 4 respectively, we used 7 marker word categories during chunking The list of marker word categories and their associated tags are given in Table 2.1.

For these earlier experiments, we did not make use of any punctuation during chunking or subsequent chunk alignment. However, subsequent developmental experiments demonstrated that punctuation could in fact be useful when performing chunking, thus aiding the overall translation. Punctuation can in fact be considered as markers, following the original definition of Green (1979); they represent a closed-class set of items which are easily identifiable and can be used to indicate context, such as the end of a sentence, clause or phrase. Therefore, in addition to the marker categories in Table 2.1, we added punctuation to the list of possible markers, assigning instances of punctuation the tag <PUNC>. The revised set of marker categories used in the experiments described in this Chapter are given in Table 5.1.

The marker tags contained in Table 2.1 are used to indicate the possible start of a new chunk. Rather than being used in chunk-initial position, punctuation marks

Determiners	<DET>
Quantifiers	<QUANT>
Prepositions	<PREP>
Conjunctions	<CONJ>
WH-Adverbs	<WH>
Possessive Pronouns	<POSS>
Personal Pronouns	<PRON>
Punctuation Marks	<PUNC>

Table 5.1: The revised set of marker categories and their associated labels

occur in chunk-final position, indicating the end of a chunk. Punctuation can be particularly useful if we encounter long sequences of words in the source or target training sentences that do not contain any instances of marker words (e.g lists, addresses etc.). In these cases, the use of punctuation allows us to identify more fine-grained examples for use during translation, an important task for EBMT and corpus-based MT in general (Nirenburg et al., 1993) (cf. Section 2 1). To give a simple illustration of the effectiveness of using punctuation during the segmentation process, consider the English–French sentence pair in (24), taken from the Europarl corpus (Koehn, 2005).

(24) Mr. President, Ladies and Gentlemen, read the documents
⇔Monsieur le Président, Mesdames et Messieurs, lisez les documents

If we consider that we only have the set of punctuation marks as our possible markers, traversing and tagging the sentence pair in (24) produces the tagged sentence pair in (25).

(25) Mr. President ,<PUNC> Ladies and Gentlemen ,<PUNC> read the documents
⇔Monsieur le Président ,<PUNC> Mesdames et Messieurs ,<PUNC> lisez les documents

Note in (25) that, unlike the original set of marker tags, the <PUNC> tag indicates the end of a chunk and therefore appears after the relevant punctuation mark.

From the tagged sentence pair in (25), we can see how it is possible to successfully segment the sentence pair in (24) based solely on punctuation information Assuming, a very naive alignment algorithm making use of only marker tag and chunk position information, we can produce the set of aligned chunks in (26).

(26) Mr. President ⇔Monsieur le Président
Ladies and Gentlemen ⇔Mesdames et Messieurs
read the documents ⇔lisez les documents

5.1.2 Sub-sentential Alignment

After the initial chunking stage is perfomed by the MaTrEx system's chunking module, the resulting source–target chunks are passed to the chunk alignment module. Working on a sentence-by-sentence basis, this alignment module takes the set of source chunks and the set of target chunks and outputs a set of corresponding bilingual chunks, similar to those given in (26) above, together with chunk translation probabilities.

Currently, we use a dynamic programming "edit-distance style" alignment algorithm, based on the classical edit-distance algorithm of Wagner and Fischer (1974). The alignment algorithm aligns those chunks present within the source–target sentence pair which are closest to each other, or more strictly, 'less distant', in terms of a particular distance metric.

Following the notation of Stroppa et al. (2006), in the following, a denotes an alignment between a target sequence e and a source sequence f, with $I = |e|$ and $J = |f|$. Given two sequences of chunks, we are looking for the most likely alignment $\hat{a}$:

$$\hat{a} = \underset{a}{\operatorname{argmax}}\, P(a|e, f) = \underset{a}{\operatorname{argmax}}\, P(a, e|f).$$

We first consider alignments such as those obtained by an edit-distance algo-

rithm. The alignment consists of a set of tuples:

$$a = (t_1, s_1)(t_2, s_2) \dots (t_n, s_n),$$

with $\forall k \in [1, n]$, $t_k \in [0, I]$ and $s_k \in [0, J]$, and $\forall k < k'$:

$$t_k \leq t_{k'} \text{ or } t_{k'} = 0,$$

$$s_k \leq s_{k'} \text{ or } s_{k'} = 0,$$

$$I \subseteq \cup_{k=1}^{n} \{t_k\}, J \subseteq \cup_{k=1}^{n} \{s_k\},$$

where $t_k = 0$ (resp. $s_k = 0$) denotes a non-aligned target (resp. source) chunk.

We then assume the following model to calculate $P(a, e|f)$, the probability of a particular alignment:

$$P(a, e|f) = \Pi_k P(t_k, s_k, e|f) = \Pi_k P(e_{t_k}|f_{s_k}), \tag{5.1}$$

where $P(e_0|f_j)$ (resp. $P(e_i|f_0)$) denotes an "insertion" (resp. "deletion") probability.

Assuming that the parameters $P(e_{t_k}|f_{s_k})$ (the probability of the source chunk in position t_k being aligned with the target chunk in position s_k) in Equation (5.1) are known, the most likely alignment is computed by a simple dynamic-programming algorithm.[2] Moreover, this algorithm can be simply adapted to allow for block movements or "jumps", following ideas introduced by Leusch et al. (2006) in the context of MT evaluation. Leusch et al. (2006) extend the Levenshtein distance (Levenshtein, 1965) by combining an additional parameter to account for block movements with the traditional insertion, deletion and substitution operators.

This type of adaptation is potentially very useful for taking into account differences between the order of constituents between certain languages. An example of such a difference in constituent order between Basque and English is given in Figure

[2]This algorithm is actually a classical edit-distance algorithm in which distances are replaced by inverse-log-conditional probabilities

5.3. Here we can see how translationally equivalent constituents between Basque and English often do not occur in a similar linear order. The use of "jumps" allows us to consider the possibility of aligning such crossing corresponding constituents, as indicated by the arrows in Figure 5.3.

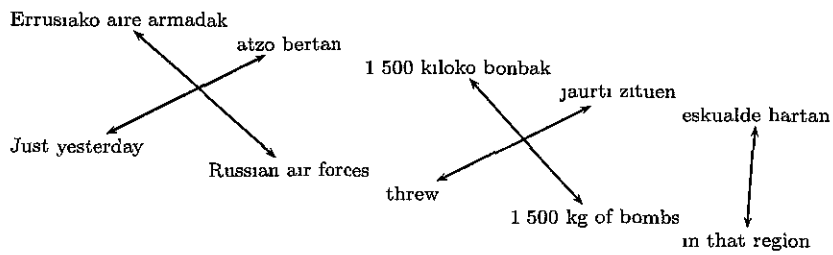


Figure 5.3: Equivalence between components in Basque and English

Instead of using an Expectation-Maximization algorithm to estimate the parameters $P(e_{t_k}|f_{s_k})$[3] in Equation (5.1), as commonly done when performing word alignment (Brown et al., 1993; Och and Ney, 2003), we directly compute these parameters by relying on the information contained within the chunks. For our experiments, when aligning chunks we make use of three sources of knowledge: word-to-word translation probabilities, source–target cognate information and marker tags.

As stated above, word-to-word translation probabilities are extracted based on a refined set of word alignments induced from the GIZA++ statistical word alignment toolkit. These pre-calculated word translation probabilities can then be used to help determine relationships between source–target chunks based on the model given in Equation (5.2).

$$P(e_i|f_j) = \sum_{a_c} P(a_c, e_i|f_j) \simeq \max_{a_c} P(a_c, e_i|f_j) \tag{5.2}$$

$$= \Pi_k \max_l P(e_{i_l}|f_{j_k}). \tag{5.3}$$

[3]Note in our experiments we make use of *features* combined into a log-linear model, rather than true probabilities *per se*, as described below, to give us the "distances" used in the edit-distance style algorithm.

For each word, e_{i_l} in the source chunk under consideration, we select the single target word (f_{j_k}) with the highest translation probability given the source word, according to the word-to-word translation probabilities. We then calculate the product over the resulting selected set of word translation probabilities to give us our final (word translation probability-based) chunk alignment feature.

In order to extract the set of source–target cognate pairs for the chunk pairs under consideration, we first calculate standard string edit-distance scores for all possible pairings of source–target words. The string edit-distance is used as an early pass filter to identify those word pairs which may be possible cognates. Those word pairs with an edit-distance score < 10 are then allowed through to the second pass of the cognate extraction algorithm [4] During the second pass we make use of Dice coefficient scores and the Lowest Common Subsequence Ratio (LCSR) For a given word pair, Dice's coefficient can be defined as the ratio of the number of shared character bigrams to the total number of bigrams in both words, multiplied by 2. For example, the English word *colour* and equivalent French translation *couleur* have three bigrams in common – *co*, *ou*[5] and *ur* – and altogether contain 11 bigrams. This produces a Dice coefficient of $2(3/11) = 0.55$.

A subsequence of a particular sequence (in our case, a sequence of characters) is a new sequence which is formed from the original by deleting some of the elements without disturbing the relative positions of the remaining elements For example, the string *cognate* has many possible subsequences which include *cogne* (*COGNatE*), *oae* (*cOgnAtE*) and *gnt* (*coGNaTe*). The LCSR of two words is computed by dividing the length of their longest common subsequence by the length of the longer word. For example, LCSR(colour,couleur) $= 5/7 = 0.71$, as their longest common subsequence is made up of *co*, *l* and *ur*.

[4]For our experiments, this threshold was empirically established during the initial development stage.

[5]This example illustrates the potentially haphazard nature of the Dice coefficient as we are considering the bigram *ou* to be common between the French and English word, even though *ou* in the French word *couleur* does not actually correspond to *ou* in *colour*, as they are taken from different parts of their respective words.

To identify and count the number of cognates present, after selecting initial word pairs using the string edit-distance score, we take the mean of the LCSR and Dice coefficient for the reduced set of word pairs to give us our cognate score. Word pairs with a cognate score above an empirically determined threshold are considered cognates. In our case, a cognate threshold of > 0.3 was used. This figure was determined by manually inspecting the output of a series of experiments using a range of threshold values when collecting cognates over the entire corpus. It was determined that below this threshold, the accuracy of the cognate identification algorithm deteriorated rapidly. The cognate feature used in the alignment algorithm is computed by dividing the number of cognates that occur between the source and target chunk, by the number of words contained in the source chunk (cf. Equation (5.4))

$$CogDistance = \frac{\#cognates(e_i, f_i)}{|e_i|} \tag{5.4}$$

For alignment based on marker tags, a simple matching algorithm is used, giving a particular chunk pair a marker tag-based alignment feature of either 1 (for matching tags) or 0.

Within the alignment algorithm it is possible to combine knowledge from marker tag matching, word probabilities and cognate information in a log-linear framework (similar in spirit to the log-linear framework of Och and Ney (2002) used in MT, as described in Section 2.3.1) according to the Equation in (5.5).

$$\log P(e_i|f_j) = \sum \lambda_k \log h_k(e_i|f_j) - \log Z, \tag{5.5}$$

In Equation (5.5), $h_k(.)$ represents a given source of knowledge, λ_k the associated weight parameter and Z a normalization parameter. The log-linear framework also facilitates the potential use of additional knowledge sources, as these knowledge sources can be integrated very easily into the model as additional features. The "distance" between a particular source–target chunk pair, f_j, e_i, is estimated from

the result of Equation (5.5). It is these "distances" that are used in calculating the best chunk alignment according to the "edit-distance-style" algorithm

5.2 OpenLab 2006 Shared MT Task

From Section 5.1.2, we can see a number of parameters are available to us for use during the chunk alignment process. In order to determine the influence of these different distance metrics on translation performance, we carried out some initial preliminary experiments as part of our participation in the TC-STAR OpenLab 2006 shared task on MT (Armstrong et al., 2006). OpenLab 2006 focused on Spanish–English translation and provided a large corpus consisting of 1.28M aligned Spanish–English sentence pairs taken from speeches from the European Parliament Plenary Sessions (EPPS) for training. This data not only provided us with a test-bed for our new alignment algorithm, but also presented us with a new language pair for testing the marker-based chunking approach. In addition, the training data consists of the largest amount of data that the marker-based approach has currently been tested on and allowed us to gauge the scalability of the MATREX system.

5.2.1 Data Resources and Experimental Setup

From the original Spanish–English corpus, we extracted 954,050 sentence pairs for training, filtering out sentence pairs greater than 40 words in length for either Spanish or English and those sentence pairs whose relative sentence length ratio was greater than 1.5. For testing, we translated the official Spanish test set consisting of 840 sentences. The statistics for the training set and test set are given in Table 5 2.

Making use of the filtered corpus, we performed chunking using the marker-based approach described in Section 2.2, including the use of the additional <PUNC> marker category, as described in Section 5.1.1. The resulting chunks were then aligned using the edit-distance-style alignment algorithm described in Section 5 1 2,

		#Sentences	#Tokens	Avg	Max	Min
Training Corpus	EN	954,050	18,319,374	19.20	40	1
	ES	954,050	18,952,444	19.87	40	1
Test Corpus	ES	840	21,770	25.95	40	1

Table 5.2: Details of OpenLab 2006 training and test corpora Avg, Max and Min refer to the average, maximum and minimum sentence lengths, respectively, in terms of words

employing tag-based, cognate-based and word probability-based features in determining the distance between chunks.

Following the findings presented in Chapter 3 and Chapter 4, the marker-based aligned chunks were then combined with SMT phrases, extracted using the method of Koehn et al. (2003) and Och and Ney (2003), as discussed in Section 2.3.4, to give us our set of hybrid EBMT-SMT phrase alignments This set of hybrid phrasal alignments made up the phrase translation table which was passed to the PHARAOH decoder wrapper along with a trigram Kneser-Ney (Kneser and Ney, 1995) smoothed language model estimated from the English section of the training corpus, essentially producing a hybrid 'example-based PBSMT' system.

Translating the 840-sentence testset we evaluated the resulting translations against two provided English reference translation sets in terms of a number of automatic evaluation metrics:

BLEU : An n-gram-based co-occurrence statistic, as described in Section 3.2.2 (Papineni et al., 2001,2002) .

NIST : An n-gram-based metric similar to BLEU, but with some modifications (Doddington, 2002). NIST assigns more weight to co-occurring n-grams which are less frequent in the reference text as they are considered more informative. NIST additionally reformulates the BLEU brevity penalty in order to minimise the impact of scores given small variations in translation length. NIST has been shown to correlate better with human evaluations in terms of *accuracy* than of *fluency*, with the majority of the score for a typical MT system coming

from unigram matches (Zhang and Vogel, 2004).

WER : The traditional word error rate (cf. Section 3.2.1). The higher the number, the worse the quality of the translations.

PER : Position-independent WER (Tillmann et al., 1997). Essentially, this a "bag-of-words" evaluation metric, calculating the percentage of words correctly produced in the output according to the reference, independent of position.

In order to determine the optimal configuration for our new alignment algorithm, we compared three different configurations:

- **Cog + Tag:** Aligning based on cognate information together with marker tags.
- **WordP + Tag:** Aligning based on word probabilities together with marker tags.
- **WordP + Cog + Tag:** Aligning making use of all three distance metrics based on word probabilities, cognate information and marker tags.

During these translation experiments we investigated the use of varying weights for the different knowledge sources, implementing values for λ_k in Equation (5.5) in the range $[0.1 - 1]$. It was observed that for any of the tested combinations, a uniform weighting for all knowledge sources resulted in optimum performance, so for all values of k in Equation (5.5), we set $\lambda_k = 1$.

5.2.2 Translation Results

The results for Spanish–English translation for the various configurations are presented in Table 5.3.

From the results in Table 5.3, we can see that using word translation probability information and marker tag information does not result in much improvement over

	BLEU	NIST	PER	WER
Cog + Tag	0.4039	8.7712	33.37	53.23
WordP + Tag	0.4077	8.8294	33.14	53.34
Cog + WordP + Tag	0.4092	8.8498	33.05	53.12

Table 5.3: Results for Spanish–English translation for the OpenLab 2006 shared MT task

using cognate information together with marker tags. This has interesting implications as calculating word translation probabilities, although it only needs to be performed once, is a difficult, time-consuming and complex task that needs to be carried out in advance as a preprocessing step, whereas the identification of cognates is extremely simple and fast, and can be determined at run-time. The combination of cognate information, word translation probabilities and tag information results in the highest overall performance. However, the use of all three features does not result in statistically significant improvements over either of the remaining feature combinations (in fact we only observe a 1.31% relative increase in BLEU score and a 0.11% relative decrease in WER compared to the system when making use of only cognate and tag information during the marker-based chunk alignment). For this shared MT task at OpenLab2006, the best performing system was produced by Shen et al. (2006) who made use of additional 4-gram and 5-gram language models, minimum error-rate training (Och, 2003) and extensive tuning of their model parameters to achieve a BLEU score of 0.5445 for Spanish–English. The MaTrEx system performed reasonably well by comparison, especially when taking into account that the system was not tuned and did not make use of minimum-error rate training.

5.3 Basque–English Translation

Basque is both a minority and a highly inflected language with free order of sentence constituents. In order to investigate whether EBMT methods, in particular the marker-based approach, can be useful for the translation of Basque we carried out a number of investigative experiments (Stroppa et al., 2006). In this sec-

tion we describe the various chunking resources used to segment and align sets of Basque–English chunks and subsequent translation experiments. In addition to EBMT knowledge resources, following on from results presented in this thesis, SMT knowledge resources are also acquired and implemented within the MATREX system. Following the improvements observed in translation performance of the hybrid configuration of the MATREX system over the baseline, in Section 5.3.5 we perform some additional manual evaluation on the quality of Basque–English chunks extracted via SMT and via EBMT methods, in line with the comparison outlined earlier in this thesis, in Section 4.5, in order to more fully understand the benefits of using both EBMT and SMT resources during translation.

5.3.1 Difficulties with Basque Translation

Particular characteristics of the Basque language increase the complexity of the task of aligning words and phrases with linguistic units in other languages. As mentioned previously, from a morphological point of view, Basque is a highly inflected agglutinative language Individual tokens in Basque often contain complex information and are commonly made up of much more than just single words. In addition, when examining Basque at a surface sentential level, it is a relatively free constituent order language which increases the complexity of alignment, and ultimately, translation.

Since Basque is an agglutinative language, there is no definitive division between morphology and syntax. As a consequence, morphemes are generally used as the basic units of analysis instead of words (Abaitua, 1998). A "word" in Basque can thus correspond to several words in English (cf. Figure 5.4).

pantailan	→	within the screen
atxikitze-puntuaren	→	of the snap point
duen	→	that has

Figure 5.4: Basque morphemes and English words

This separates Basque from most European languages, although the use of such

morphological information for translation is common when translating other highly inflected languages, such as Arabic. Lee (2004) makes use of presegmentation techniques for Arabic when performing Arabic–English phrase-based translation. The presegmentation technique produces stems and affixes from the original Arabic sentences. Lee (2004) concludes that using such morphological analysis helps only for training corpora consisting of up to 350,000 sentence pairs, and does not significantly improve results when their system is trained on a large corpus consisting of 3.3 million sentences. In our work, we make use of a reasonably small corpus consisting of less than 350,000 sentence pairs, so following the findings of Lee (2004) working at the morpheme level for Basque should prove beneficial.

As mentioned previously, in Basque, the order of the main constituents of a sentence is relatively free. For example, the 24 possible permutations obtained by changing the order of the *subject*, *object* and *PP* in the sentence displayed in Figure 5 5 are all well-formed Basque sentences. Consequently, unlike English or French and most other European languages, Basque sentences do not usually start with a subject followed by a verb; the subject and verb do not necessarily appear in the same order and may not even appear consecutively in Basque sentences. It is also important to mention that this general flexibility at the sentence level is much more restricted within other syntactic units (for example, inside NPs or subordinated sentences). Moreover, there is agreement in number and person between verb and subject, and object and indirect object (corresponding roughly to the ergative, absolutive and dative cases) in Basque.

5.3.2 Basque Chunking

The linguistic complexity of Basque, as discussed in Section 5.3.1, makes it much more difficult to apply the relatively simple marker-based chunking techniques, largely due to the difficulty in identifying and defining the possible set of Basque marker words. To overcome this problem, in our experiments we use existing tools to perform Basque chunking, namely the EUSMG toolkit developed by the Ixa group

"The dog brought the newspaper in his mouth" $\Longrightarrow$ *"Txakurrak egunkaria ahoan zekarren"*			
Txakur-rak	egunkari-a	aho-an	zekarren
The-dog	the-newspaper	in-his-mouth	brought
ergative-3-s	absolutive-3-s	inessive-3-s	
Subject	Object	Modifier	Verb
23 other possible orders:			
Txakur-rak	aho-an	egunkari-a	zekarren
Txakur-rak	aho-an	zekarren	egunkari-a
Egunkari-a	txakur-rak	zekarren	aho-an
...			

Figure 5.5: Free constituent order between sentence units in Basque

at the University of the Basque Country. The EUSMG toolkit can be used to perform POS tagging, lemmatisation and subsequent chunking (Aduriz et al., 1997). It recognizes syntactic structures by means of features assigned to word units, following the constraint grammar formalism (Karlsson et al., 1995).

After making use of the EUSMG toolkit to perform chunking, a sentence is treated as a sequence of morphemes, in which chunk boundaries are clearly visible. As a result of the chunking process, morphemes denoting morphosyntactic features are replaced by conventional symbolic strings (cf. the example in (27), in which *++abs++ms* denotes the morphosyntactic features absolute, definite and singular).

(27) Fitxategi zaharra ezin izan da irakurri

$\Downarrow$

[fitxategi zahar ++abs++ms] [ezin izan da irakurri]

([The old file] [could not be read])

5.3.3 Data Resources & Chunk Alignment

In order to test the performance of the MATREX system when preforming Basque–English translation, we made use of a bilingual corpus constructed from a translation memory of software manuals, generously supplied by Elhuyar Fundazioa. In total the corpus contains 320,000 translation units consisting of 3 2 million English words and 2.9 million Basque "words". The average number of words per unit is thus approximately 10 for English and 9 for Basque. Note that the difference in average

sentence length between English and Basque is somewhat less than expected, given the agglutinative nature of the Basque language. However, in Basque, verbs are commonly composed of more than one word due to inflection and the frequent use of periphrastic verbs. In addition, lexical compound terms occur frequently in Basque, resulting in Basque sentences being longer on average than expected.

This full Basque–English corpus was filtered based on absolute sentence length (sentences > 40 words were removed, to improve processing speed) and relative sentence length (sentence pairs with a relative sentence length > 1.5 were removed to help reduce the level of noise in the training data) resulting in 276,000 entries. From our filtered corpus, we randomly extracted 3,000 sentences for testing, using the remainder for training

Making use of the 273,000-sentence training set and testing on the 3000-sentence Basque test set, we performed translation from Basque to English. For English we made use of the marker-based chunker and used the EUSMG chunker for Basque. Following the results from the Spanish–English OpenLab 2006 task (cf. Section 5.2), our alignment strategy made use of a uniformly weighted combination of cognate, word-to-word (or more accurately, morpheme-to-word, as alignment was performed using the morphologically processed Basque text produced by the EUSMG toolkit) translation probability and chunk label information. We fed the resulting Basque–English chunk alignments, together with the SMT word alignments, calculated via GIZA++, and the corresponding SMT bilingual phrases to the PHARAOH phrase-based decoder.[6] The interpolated language model used in our experiments was calculated from the English section of the training corpus and employed Kneser-Ney smoothing (Kneser and Ney, 1995).

[6]The word alignment task was performed after the Basque morphological analysis was carried out, so in reality these word alignments consisted of morpheme-to-word alignments.

5.3.4 Translation Results

In addition to seeding the MaTrEx system with the extracted translation resources, we built a baseline phrase-based SMT system against which we could compare the performance of the MaTrEx system. The phrase-based SMT system was built using the statistical word alignments and extracted corresponding SMT phrasal alignments (following the phrasal extraction methods of Koehn et al. (2003) and Och and Ney (2003) described in Section 2.3.4) which were passed along with a trigram English language model to the Pharaoh phrase-based decoder. We evaluated translation performance in terms of Bleu score, Precision and Recall (Turian et al., 2003), PER and WER (cf. Section 3.2).

	BLEU	Prec.	Rec	PER	WER
phrase-based SMT	17 31	49.63	52 37	68.86	89.91
MaTrEx	22 23	59 99	55 67	49 36	68.25

Table 5 4: Basque–English translation results

The results for Basque-English MT are displayed in Table 5.4. Here we can observe that the MaTrEx system achieves a BLEU score of 22.23%, a 28.42% relative increase over the BLEU score for the baseline phrase-based SMT system. We also see a significant drop in PER and WER (19.5% and 21.66% absolute, respectively, compared with the phrase-based SMT system) and an increase in Precision and Recall. As with the experiments on the *Sun Microsystems* data and Europarl data (cf. Chapter 3 and Chapter 4, respectively), this indicates that the EBMT chunks contribute positively to overall translation quality, enabling the MaTrEx system to outperform the phrase-based baseline SMT system. It also indicates that the EBMT chunks allow the system to correctly translate many more of the input words correctly, reflected particularly in the drops in PER and WER

From these results we can see that, as demonstrated on the *Sun Microsystems* and Europarl experiments, making use of hybrid EBMT-SMT translation data contributes positively to translation performance.

5.3.5 Quality of the Hybrid Chunk Set

In addition to our translation experiments, from the set of Basque–English SMT chunks with frequency greater than 10, we randomly selected a subset consisting of 100 bilingual examples and manually evaluated them in terms of quality. The alignments were classified by hand as either being correct (semantically correct), quasi-correct or incorrect (semantically incorrect). The quasi-correct chunks are those which can be considered correct in restricted situations; i e. they are not *a priori* exact equivalents, but in many cases are possible and can even be optimal translations. These quasi-correct alignments cannot be applied everywhere, but in many cases they reflect interesting phenomena, which otherwise, for example when using rule-based approaches, would not be described accurately. For example, if looking at the translation pair *'your computer'* $\Longrightarrow$ *'zure ordenadore ++gel++ms'* out of context it would appear to be incorrect as a genitive case exists in Basque (equivalent to *'of the computer'*)), that does not occur in the English chunk. However, when translating the chunk *'your computer screen'* into Basque, the genitive case information encapsulated within the chunk is necessary for accurate translation.

The results of the manual evaluation of the chunks extracted via SMT methods are given in Table 5.5. Here we can see that over 94% of the randomly selected subset (consisting of 100 chunk alignments selected from the set of most frequently occurring chunk alignments) are classified as either correct or quasi-correct, indicating that precision is very high, at least with regards to frequently occurring chunks.

Correct	Quasi-Correct	Incorrect
63.45%	31.10%	5.45%

Table 5.5: Evaluating the quality of the SMT phrasal alignments

In order to more fully understand the contribution the EBMT chunks make to the overall increase in performance of the MaTrEx system, we performed a similar evaluation setup, this time selecting 100 examples randomly from the set of chunks which are extracted by both the EBMT chunking and alignment methods and the

SMT phrasal extraction methods (i.e. the intersection of the two sets of phrasal alignments). The results for the evaluation of this set of 100 chunks are given in Table 5.6.

Correct	Quasi-Correct	Incorrect
84 27%	12.36%	3.37%

Table 5.6. Alignment Evaluation results for the Intersection of the SMT phrase and EBMT chunk sets

From Table 5.6, we can see that, as with the phrasal alignments extracted by both marker-based methods and SMT methods from the Europarl data (cf. Section 4.5), those chunks that are found by both methods are of actually higher quality than those found by SMT methods alone. Out of 100 of the most frequent chunks occurring in the intersection of the two sets of chunks, 84.27% can be considered as correct translations of each other, 12.36% as quasi-correct and only 3.37% as completely erroneous. This indicates that these higher quality chunks extracted by SMT methods are given a boost in probability when merged with the EBMT data, resulting in the significant improvements in translation quality observed in Table 5.4.

5.4 Marker-Based Decoding

For the experiments described in Sections 5.2 and 5.3 of this Chapter we made use of the state-of-the-art PHARAOH phrase-based decoder. In previous experiments we found that the marker-based EBMT recombination process was unable to match the performance of the PHARAOH decoder (Koehn, 2004a) when fed the same translation resources extracted from the Europarl corpus (cf. Chapter 4). Even the use of hybrid data sets and the addition of a statistical language model, although both improving results, did not result in translations of as high quality as those produced by the phrase-based decoder.

Following on from this and in an attempt to take further advantages of combining

EBMT and SMT approaches to translation within the MaTrEx system, we have carried out some early preliminary work into the development of an SMT-style decoder employing marker-based segmentation. In this section we present the ideas behind the decoder as well as indicate further possibilities for future development.

The new MaTrEx decoder makes use of techniques used in both the EBMT recombination process and traditional SMT decoding practices. Following the marker-based EBMT recombination approach, the decoder makes use of three data resources, moving from maximal to minimal context: the originally sententially aligned corpus, the phrasal database and the word-level alignments.

The three data resources constitute three individual translation tables. Rather than employing weights, as in the previous recombination algorithm, the decoder makes use of statistical information when choosing translations for input segments, with translation probabilities for entries within the translation tables estimated from relative frequencies.

When translating an input sentence, if an exact sentence match cannot be found, the sentence is segmented into chunks according to the marker-based approach. These segments are then translated based on a combination of language model and translation model probabilities. If a chunk translation cannot be found, word-level translation is performed similarly. Selecting the best possible translation for each input segment given the current partial hypothesis score results in an extremely large search space. In order to reduce the size of this search space, a number of pruning strategies are implemented.

5.4.1 Pruning the Search & Scoring Hypotheses

During the search process, the decoder employs a number of pruning strategies as commonly used in SMT word graph generation and search (Ueffing et al., 2002). As hypotheses are formulated, they are placed in an stack ordered in terms of the hypotheses' current translation probabilities which are stored together with the incomplete hypotheses. Histogram pruning is applied to the stack where after new

hypotheses have been added to the ordered stack, where only the top M best partial hypotheses are retained.

When retrieving translations for an input segment, only those candidate segment translations whose probability is above the probability of the most likely candidate target segment is multiplied by a threshold $\alpha < 1$ (e.g. 0.0001) are considered. In addition, only the top n candidate translations are initially selected and subsequently pruned by the threshold pruning strategy.

The probability of a given hypothesis is made up of the translation model probability (as calculated from the corresponding translation table) and language model probability. As the decoder works with log probabilities, when a given hypothesis is extended, its score is incremented by adding the translation model score and language model score for the extension sequence.

In order to efficiently calculate the language model score for a new hypothesis according to a trigram language model, the last two words (t_{n-1}, t_n) of the current partial hypothesis are stored (Tillmann and Ney, 2000). The score for the new extended hypothesis is therefore calculated from the Equation in (5.6)

$$\begin{aligned} P(\hat{T}) = P({t_1}^n) \quad & +P({t'_1}^i) \\ & = P({s'_1}^J|{t'_1}^i) + P(t'_i|t'_{i-2}t'_{i-1}) + ... + P(t'_1|t_{n-1}t_n) \end{aligned} \tag{5.6}$$

where $\hat{T}$ represents the extended hypothesis, ${t_1}^n$ is the current hypothesis, ${t'_1}^i$ is the proposed extension, with $P({s'_1}^J|{t'_1}^i)$ the translation model score for the proposed extension and $P(t'_i|t'_{i-2}t'_{i-1}) + ... + P(t'_1|t_{n-1}t_n)$ the trigram language model score.

5.4.2 Decoding Algorithm

To summarise the decoding process, taking an input sentence, during translation the marker-based decoder proceeds as follows:

1. Search the sentence-level database for exact matches of the input sentence. If a match is found, output the corresponding target language translation, otherwise proceed to step 2.

2. Segment the input sentence according to the marker-based approach, producing a set of marker chunks. For each input chunk·

   (a) Search the chunk database for translations of the input chunk Return the top n candidate translations of the chunk and apply threshold pruning. If no chunk translations are found, proceed to Step 2(c).

   (b) Append the corresponding target language chunks to the end of the current hypotheses, thus producing a set of new hypotheses. Proceed to Step 3.

   (c) If no chunk translations are found, segment the source chunk into individual words For each source word:

      - Search the word-level database for translations of the source word. Return the list of the top N scoring words. Apply threshold pruning. If no translations are found for the source word, extend the current hypotheses by appending the original source language word.

3. Rescore the new hypotheses based on the new combined language model and translation model scores. Prune by holding onto only the top M scoring hypotheses.

4. If no source chunks are awaiting processing, output the top-ranked hypothesis. Otherwise, return to Step 2(a).

5.4.3 Experimental Setup

To evaluated the performance of the MaTrEx decoder, we replicated the experiments described in Chapter 4 which made use of training and test sets taken from

the Europarl corpus (Koehn, 2005). As the MATREX system makes use of statistical word alignments, in contrast to the experiments in Chapter 4 we performed translation using only two different system configurations:

SEMI-HYBRID . As the MATREX system makes use of statistical word alignments, as described in Section 5.1, making use of marker-based EBMT chunk alignments essentially results in what we labeled previously as the 'semi-hybrid' system configuration.

HYBRID : Making use of all the information resources available to us, namely marker-based EBMT chunks and word alignments, and SMT phrasal and word alignments, results in the 'hybrid' system configuration.

In these experiments, in order to make the results directly comparable to those from Chapter 4, rather than employing the new marker-based 'edit-distance-style' aligner as described above, we reused the exact same EBMT and SMT alignments (words and phrases) as were used in these previous Europarl experiments

Taking the semi-hybrid and hybrid data sets, we performed translation for French–English and for English–French using the new marker-based decoder.

5.4.4 French–English Translation

In Table 5.7 we present the results for French–English for the variants of the original EBMT and PBSMT systems, including the EBMT-LM language model reranking experiments, together with the results for the MATREX decoder.

From these results we can see that the marker-based decoder seeded with the word alignments induced via GIZA++ and the marker-based chunks, outperforms the original EBMT recombination algorithm, both when using the baseline data set and the semi-hybrid data set, across all training data sets.

The semi-hybrid MATREX decoder receives 14.53% relative increase in BLEU score, on average, across the three training sets. It also outperforms the semi-hybrid

			BLEU	Prec.	Recall	WER	SER
78K	*EBMT*	BASELINE	.1217	.4556	.5315	85.63	98.94
		SEMI-HYBRID	.1256	.4537	.5280	85.19	98.92
		HYBRID	.1519	.4997	.5349	76.12	98.98
		HYBRID-LM	1624	.5091	.5341	74.15	98.80
	PBSMT	BASELINE	1943	.5289	.5477	70.74	98.42
		SEMI-HYBRID	.1861	.5217	.5464	73.77	98.36
		HYBRID	.2070	.5368	.5551	69.56	98.20
	MaTrEx	**SEMI-HYBRID**	**.1493**	**.5167**	**.5650**	**76.62**	**98.56**
		HYBRID	**.1589**	**.5294**	**.5675**	**74.21**	**98.48**
156K	*EBMT*	BASELINE	.1343	.4645	.5368	83.55	99.02
		SEMI-HYBRID	.1387	.4670	.5394	82.95	99 10
		HYBRID	.1620	.5081	.5457	75.14	99 16
		HYBRID-LM	.1722	.5177	.5463	73.19	98.68
	PBSMT	BASELINE	.2040	.5369	.5526	69.41	98.30
		SEMI-HYBRID	.1970	.5318	.5518	72 37	98.38
		HYBRID	.2176	.5437	.5579	68.17	98.10
	MaTrEx	**SEMI-HYBRID**	**.1486**	**.4985**	**.5592**	**79.27**	**98.34**
		HYBRID	**.1638**	**.5312**	**.5685**	**73.74**	**98.46**
322K	*EBMT*	BASELINE	.1427	.4734	.5419	82.43	99.06
		SEMI-HYBRID	.1459	.4750	.5434	81.92	99.04
		HYBRID	.1699	.5145	.5558	74.99	99.20
		HYBRID-LM	.1773	5224	.5530	72.85	98.80
	PBSMT	BASELINE	.2102	.5409	.5539	68.55	98.72
		SEMI-HYBRID	.2089	.5406	.5542	70.55	98.38
		HYBRID	.2236	.5483	.5592	67.40	98.58
	MaTrEx	**SEMI-HYBRID**	**.1573**	**.5212**	**.5696**	**75.89**	**98.46**
		HYBRID	**.1703**	**.5373**	**.5738**	**73.02**	**98.46**

Table 5.7: The performance of the marker-based MaTrEx decoder, seeded with the 'semi-hybrid' (EBMT chunks & Giza++ word alignments) and 'hybrid' (EBMT chunks, SMT phrases and Giza++ word alignments) data sets for French–English.

EBMT system by 11 27% average relative BLEU score. Despite this improvement the MaTrEx decoder is still beaten by all variants of the PBSMT system for all metrics, bar recall, which it improves upon by 2.39% relative.

The hybrid MaTrEx decoder configuration, improves upon the hybrid EBMT BLEU score by 1.93% relative, on average across training sets, and also wins out on Precision, Recall and WER. However, it is beaten by the reranked hybrid EBMT system, making use of the statistical language model, in terms of BLEU, but still manages to outperform it in terms of precision and recall. In general, it appears that the marker-based decoder outperforms all other systems in terms of recall but when compared with all variants of the PBSMT system, does not perform as well in terms of BLEU. This seeming contradiction of BLEU score results with the remaining metrics again calls into question the validity of BLEU when evaluating systems that are not heavily n-gram based (Callison-Burch et al., 2006) (cf. Section 4.2 3). A manual evaluation of the translation output would help to determine whether the BLEU metric is in fact inadequate for accurately measuring the performance of such systems. Such an evaluation, however, is beyond the scope of this thesis

5.4.5 English–French Translation

The results for the reverse language direction are given in Table 5.8. Here we can see that, as for French–English, the MaTrEx decoder improves upon the performance of the marker-based EBMT recombination process When seeded with Giza++ word alignments and marker-based chunks, the new decoder outperforms the baseline EBMT system making used of marker-based data alone and also the semi-hybrid EBMT system making use of the same marker-based chunk and Giza++ word alignments. This semi-hybrid MaTrEx configuration achieves a 12.46% relative increase in BLEU score over the baseline EBMT system and a 8.84% relative increase over the semi-hybrid EBMT system on average across training sets, displaying similar improvements as were observed for French–English

As for French–English, for English–French, in general, the semi-hybrid MaTrEx

			BLEU	Prec	Recall	WER	SER
78K	*EBMT*	BASELINE	.1240	.4422	.4365	79.09	99.10
		SEMI-HYBRID	.1304	.4515	.4445	78.20	99.10
		HYBRID	.1462	.4783	.4580	74.54	99.20
		HYBRID-LM	.1527	.4871	4611	73.23	99 08
	PBSMT	BASELINE	.1771	.5045	.4696	70.44	98.54
		SEMI-HYBRID	.1670	.4968	.4617	73.32	98.46
		HYBRID	.1898	.5152	.4787	69.20	98.50
	MaTrEx	**SEMI-HYBRID**	**.1423**	**.4900**	**.4675**	**74.03**	**98.48**
		HYBRID	**.1518**	**.5012**	**.4709**	**72.27**	**98.42**
156K	*EBMT*	BASELINE	.1374	.4548	4476	77.66	98.96
		SEMI-HYBRID	.1404	4592	.4517	77.34	98.96
		HYBRID	.1573	4870	.4682	73 86	99 16
		HYBRID-LM	.1635	.4955	.4709	72 50	98 88
	PBSMT	BASELINE	.1855	.5120	.4724	69.39	98.20
		SEMI-HYBRID	.1773	.5048	.4681	72.27	98.26
		HYBRID	.1965	.5208	.4810	68.36	98 24
	MaTrEx	**SEMI-HYBRID**	**.1531**	**.4963**	**.4751**	**73.37**	**98.18**
		HYBRID	**.1632**	**.5105**	**.4802**	**71.14**	**98.18**
322K	*EBMT*	BASELINE	.1448	.4587	.4530	77 73	99.22
		SEMI-HYBRID	.1486	.4632	.4575	77 51	99.56
		HYBRID	.1668	.4912	.4798	73.94	99.38
		HYBRID-LM	.1744	.5014	.4818	71.96	98.80
	PBSMT	BASELINE	.1933	.5180	.4751	68.30	98.12
		SEMI-HYBRID	1850	.5107	.4708	71 22	98.22
		HYBRID	.2040	.5284	.4851	67.28	98.12
	MaTrEx	**SEMI-HYBRID**	**.1610**	**.5017**	**.4801**	**72.72**	**98.12**
		HYBRID	**.1730**	**.5182**	**.4878**	**70.37**	**98.12**

Table 5.8: The performance of the marker-based MaTrEx decoder, seeded with the 'semi-hybrid' (EBMT chunks & Giza++ word alignments) and 'hybrid' (EBMT chunks, SMT phrases and Giza++ word alignments) data sets for English–French.

decoder outperforms the baseline and semi-hybrid configurations of the PBSMT system in terms of recall, but fails to match the performance of either PBSMT configuration according to the remaining evaluation metrics.

The MaTrEx decoder, making use of all of the marker-based alignments and the alignments extracted via SMT methods, improves further upon the baseline and semi-hybrid EBMT systems, as is to be expected. In general, the hybrid MaTrEx decoder beats the hybrid EBMT system across all metrics, and the reranked hybrid EBMT system ('HYBRID-LM') across all metrics, bar BLEU score and WER, with the most significant improvements visible in recall results.

We see an increase of 3.77% relative BLEU score over the EBMT system seeded with the same data (all of the EBMT and SMT alignments). The hybrid EBMT system making use of the re-ranking statistical language model, outperforms the hybrid MaTrEx system slightly in terms of BLEU score, but performs slightly worse on average across training sets for precision (an average relative decrease of 3.07%) and recall (an average decrease of 1.78%).

With regards to the PBSMT system, in general across the three training set sizes, the hybrid MaTrEx system is outperformed slightly by the hybrid PBSMT system in terms of precision and WER. However the hybrid MaTrEx system manages to *match the performance of the* hybrid PBSMT system in terms of recall. Considering the general trend, the hybrid MaTrEx system falls considerably short in terms of translation performance according to BLEU score when compared to all configuations of the PBSMT system, which *as for French–English, calls into question the use of* BLEU for evaluating those MT systems which are not purely n-gram-based.

5.5 Summary

In this chapter, we presented MaTrEx, a large-scale modular data-driven MT system based on chunking and chunk alignment. In this system, chunk alignment is performed thanks to a simple dynamic programming algorithm which exploits

relationships between chunks. In this context, different kinds of relationships can be considered and even combined. Moreover, this system can be considered a hybrid MT system since it also makes use of aligned phrases extracted using classical SMT techniques.

We performed some experimental evaluation on the various relationships that can be used to determine chunk alignment. Making use of Spanish–English data from the OpenLab 2006 shared task on MT we discovered that making use of cognate information, word-to-word translation probabilities and marker tag information results in the best system performance.

In addition to evaluating alignment strategies, we outlined a number of experiments on Basque to English translation, making use of a reasonably small corpus consisting of software manuals data, similar to that of the *Sun Microsystems* corpus The results we obtained showed significant improvements on state-of-the-art phrase-based SMT (28% relative increase for BLEU and 21.66% absolute drop in WER). Additionally, following the manual evaluation of the quality of the chunks aligned by our method, we discovered that making use of marker-based alignment methods, together with SMT phrasal extraction heuristics, resulting in producing chunks of generally higher quality than those produced using SMT techniques alone. We find this further evidence to strengthen our claim that making use of both EBMT and SMT techniques results in increased translation quality.

Finally, we have outlined some initial experiments into the development of a marker-based SMT-style decoder. Making use of data sets extracted from the Europarl corpus used in previous experiments, we observed that the new statistically-driven decoder was capable of outperforming the marker-based EBMT recombination algorithm of Way and Gough (2003, 2005a, 2005b), Gough and Way (2004a, 2004b) and Gough (2005). In addition, we demonstrated that the new hybrid decoder was capable of matching the performance of the state-of-the art PHARAOH phrase-based SMT decoder in terms of recall with precision scores falling slightly short. With these experiments, as with those on the Europarl corpus, we noted

that BLEU scores for the MaTrEx decoder did not seem to follow the general trend set by the remaining evaluation metrics We feel that this again demonstrates that the BLEU metric may not be the most suitable for evaluating systems that are not purely n-gram-based.

As the decoder is under continuing development, these results are extremely promising and offer further indications of the benefits of using both EBMT and SMT techniques within a hybrid data-driven system to improve translation performance.

Chapter 6

Conclusions

Empirical corpus-based approaches to MT now dominate the MT research field. They present an alternative to earlier first generation, and more expensive second generation approaches, and are based on the use of translation knowledge extracted from existing bilingual corpora. SMT and EBMT models of translation represent two frameworks within the data-driven paradigm The fact that they both now make use of phrasal and word-level information led us to our first research question:

(RQ1) *How do state-of-the-art EBMT and SMT methods compare, both theoretically, and also in terms of their translation performance?*

In this thesis, we have outlined how, although EBMT and SMT methods now have much more in common, they still retain a number of theoretical differences that set them apart. In response to the question presented in (RQ1), we concluded, following the definitions presented by Wu (2006), that both modern EBMT and modern SMT methods constitute hybrid models of MT, as they borrow techniques from various disciplines. As SMT methods now make use of phrasal information (Koehn et al., 2003), they have become more EBMT-like, although they are still easily identifiable by their use of statistical modeling techniques in their distinct translation and language models. Many EBMT methods, such as the marker-based methods (Veale and Way, 1997; Way and Gough, 2003, 2005a, 2005b; Gough and Way, 2004a, 2004b; Gough, 2005) used in our work, extract examples in a pre-processing stage,

and thus are moving further away from memorisation techniques, which are typically example-based, and closer towards schema-based abstraction techniques, such as those employed by SMT methods. In addition, the use of stronger probability models is clearly a statistical technique, yet the use of generalised templates and linguistic motivations make them distinctly EBMT (e.g. Brown (1999a)).

In order to answer the second part of this initial research question (RQ1), in this work we have presented a number of comparative experiments which have shown that EBMT is particularly suited to translation within a sublanguage domain, such as the *Sun Microsystems* data, comprising of computer documentation. We also performed a number of additional experiments, using data taken from the larger Europarl French–English corpus (Koehn, 2005), to see whether our results held for data taken from a different, more open-domain. For these second set of experiments we made use of incremental amounts of training data, consisting of 78K, 156K and 322K sentence pairs. From the experiments on these two data sets we found that:

- For sublanguage translation using the *Sun Microsystems* data, marker-based EBMT is capable of outperforming phrase-based SMT[1] for French–English and English–French translation according to a range of automatic evaluation metrics.

- On the Europarl data sets, the phrase-based SMT system outperformed the marker-based EBMT system for French–English and for English–French for all training set sizes. As expected, both the EBMT and SMT systems benefit equally from increased amounts of training data.

Following on from these initial comparative experiments, we performed a further set of experiments in order to address our second research question.

(RQ2) *What contributions do EBMT and SMT translation resources make to the quality of translations produced by an MT system and can*

[1]Note that, as mentioned previously, all variants of the phrase-based SMT system were left untuned, with default weighting and employing only relative frequency scores within their translation tables.

EBMT and SMT approaches to translation be combined into a hybrid data-driven model of MT?

Making use of the *Sun Microsystems* data, we presented a number of experiments on the use of hybrid data-driven translation resources by feeding the baseline SMT system from our earlier comparative experiments with various combinations of EBMT and SMT-induced data. For translation on this particular data set we found that:

- The phrase-based SMT system seeded with EBMT phrasal alignments and EBMT lexical alignments does not perform as well as the baseline phrase-based SMT system (seeded with SMT phrasal and SMT lexical alignments).
- Within the phrase-based SMT system, replacing the SMT phrasal alignments with the EBMT chunk alignments results in improvements over the baseline phrase-based SMT system, thus indicating the higher quality of the EBMT chunk alignments.
- Combining EBMT and SMT phrasal and lexical data within the phrase-based SMT system, results in significant improvements over the baseline. This hybrid 'example-based SMT' system is capable of outperforming the marker-based EBMT system for French–English translation in terms of BLEU score and precision.

Given these findings for the *Sun Microsystems* hybrid data-driven experiments, we performed additional hybrid experiments on the larger Europarl data set (Koehn, 2005). Previously for this data set, in contrast to the results for the *Sun Microsystems* experiments, the marker-based EBMT system of Gough and Way (2004a, 2004b), Way and Gough (2005a) and Gough(2005) was unable to match the performance of the baseline phrase-based SMT system. On this particular data set we found that:

- Making use of SMT word alignments in the EBMT system results in translation improvements. Further improvements are seen when the EBMT system is seeded with all of the SMT phrasal and word alignments together with the EBMT chunk and lexical information, thus creating a 'statistical EBMT' system.

- Reranking the output of the 'statistical EBMT' system with a statistical language model results in additional improvements in translation performance. However, all variants of the EBMT system still fall short of the performance of the baseline phrase-based SMT system.

- Incorporating EBMT chunks and EBMT word alignments into the phrase-based SMT system improves translation performance over the baseline phrase-based SMT system

These experiments again strengthened our hypothesis, that making use of hybrid system configurations results in better translation performance, thus successfully answering the research question in (RQ2). Examining the SMT phrasal alignments and the EBMT chunk alignments induced from the Europarl corpus, we found that those alignments extracted via EBMT methods are of higher quality than those extracted via SMT methods. Consequently, when combining the two sets of alignments, those phrase pairs which are found by both methods are given a boost in probability, thus explaining the improvements seen for the hybrid system configurations.

Following these findings, we introduced a new hybrid data-driven MT system employing a modular design, thus facilitating the use of hybrid techniques. We described a number of additional experiments carried out making use of this new architecture. Performing Spanish–English translation on a significantly large data set consisting of over 950,000 sentence pairs demonstrated the scalability of the system and showed that making use of cognate information, word-to-word probability information and marker tag information results in extracting the best quality set of chunk alignments for use in translation.

In a set of further experiments, we demonstrated the adaptability of this new architecture to new language pairs, namely Basque–English, and new technologies, namely the EUSMG Basque chunker. For these experiments, as with the experiments on the *Sun Microsystems* and Europarl data sets, we demonstrated that the use of both EBMT and SMT phrasal alignments results in improvements in translation quality.

Finally, in order to address our third and final research question, which asked:

(RQ3) *Can a novel hybrid data-driven decoder take advantage of EBMT approaches, together with SMT search strategies and probabilistic models to improve translation results?*

We proposed a new hybrid marker-based SMT-style decoder, which takes advantage of marker-based segmentation and recombination techniques, together with probabilistic search and statistical modeling strategies to perform translation.

We outlined a number of initial evaluation experiments which showed that the new decoder was capable of outperforming the original marker-based recombination technique for French–English and English–French. We also showed that the hybrid decoder was capable of matching the performance of state-of-the-art PHARAOH (Koehn, 2004a) in terms of recall with precision scores falling only slightly short, despite not employing any reordering models.

In this thesis we have demonstrated, through a series of theoretical and practical explorations, that while there is an obvious convergence between EBMT and SMT approaches to translation, the crucial differences between them contribute positively to the overall translation quality, thus supporting the use of hybrid data-driven models of MT.

6.1 Future Work

In the experiments presented in this work, all variants of the PBSMT system were left untuned due to the large number of experiments carried out, research time con-

straints and to provide results that were easily replicable by other researchers. The PBSMT systems also only made use of phrase translation probabilities calculated from relative frequencies in their translation models. For future work we would like to reproduce the experiments carried out in this work, particularly those experiments on the *Sun Microsystems* data in Chapter 3 and the Europarl data in Chapter 4, making use of minimum error-rate training to tune model parameters and employing additional scores, such as inverse translation model scores, to improve overall PBSMT system results further.

The initial experiments using the marker-based decoder are extremely promising. As the decoder is currently in a very early stage of development a number of possible improvements and extensions present themselves. Currently, the decoder requires a significant amount of memory to store the translation table, language model and search space explored during translation. Being able to access the translation and language models at run-time, rather than storing them temporarily in memory, may speed up the process by freeing some system memory. Also, the use of word-graphs, as implemented by Ueffing et al (2002), could help reduce the size of the search space, as rather than storing translations as strings, they can be stored within a connected word-graph, with nodes covering segments possibly longer than individual words.

As it currently stands, the decoder does not make use of any reordering models. However there has been some research carried out into the reordering of word-graphs which could be applied to the decoder (Ueffing et al., 2002). Alternatively, Zens et al. (2004) describe a discriminative reordering model which operates using a number of features within a log-linear framework. The decoder also does not make use of generalised templates, which in the past have been shown to improve both coverage and quality (Gough, 2005). Generalising on content words using clustering techniques, such as those proposed by Brown (2000), would also provide the system with additional translation knowledge.

Currently, the list of marker words is created manually for each language pair.

One interesting avenue for further research would be the automatic extraction and identification of marker words. This could be performed using POS tag information, which would allow the automatic collection and identification of words corresponding to a particular marker category. A more simple approach would be to make use of frequency data, as the marker words in any language tend to be those words which occur most often. Additionally, word-to-word probability information together with phrasal extraction heuristics could be used to determine pairs of corresponding source-target marker words. From Section 4.5 we can see that the majority of the most frequent phrase pairs extracted via SMT methods are composed almost entirely of marker words.

We would also like to extend our hybrid approaches to other language directions and language pairs. Currently we are investigating the possibility of performing English–Basque translation within our MaTrEx system, as well as Basque–Spanish translation which would allow us to directly compare the performance of the MaTrEx system against the OpenTrad[2] transfer-based MT system (Algeria et al., 2005). In addition, we would like to test our hybrid methods on different and larger data sets. Jaime Carbonell (personal communication at AMTA 2006 Question and Answers session) has recently stated that the research group at CMU have found exactly the same benefits of using hybrid data sets created with their EBMT (Brown, 1996, 1999a, 1999b, 2000; Brown et al., 2003) and SMT systems (Vogel et al., 2003), on the NIST-2006 data set, thus indicating the potential benefits of using hybrid data-driven techniques on additional corpora and language pairs.

[2] http://www.opentrad.org

References

Abaitua, J. (1998). *Complex Predicates in Basque: from Lexical Forms to Functional Structures.* Ph.d. thesis, University of Manchester, UK.

Aduriz, I., Arriola, J. M., Artola, X., de Ilarraza, A. D., Gojenola, K., and Maritxalar, M. (1997) Morphosyntactic Disambiguation for Basque Based on the Constraint Grammar Formalism. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*, pages 282–288, Tzigov Chark, Bulgaria.

Algeria, I., de Ilarraza, A. D., Labaka, G., Lersundi, M., Mayor, A., Sarasola, K., Forcada, M , Oritz, S., and Padró, L. (2005). An Open Architecture for Transfer-Based Machine Translation. In *Machine Translation Summit X: Workshop on Open-Source Machine Translation*, pages 7–14, Phuket, Thailand.

Armstrong, S., Flanagan, M., Graham, Y., Groves, D., Mellebeek, B., Morrissey, S., Stroppa, N., and Way, A. (2006). MaTrEx: Machine Translation Using Examples. In *TC-Star OpenLab on Speech Translation*, Trento, Italy. Slides at `http://tc-star.org/openlab2006/day1/Groves_openlab.pdf`.

Aue, A , Menezes, A., Moore, R., Quirk, C., and Ringger, E. (2004) Statistical Machine Translation Using Labeled Semantic Dependency Graphs. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 125–134, Baltimore, MD.

Bangalore, S., Murdock, V., and Riccardi, G. (2002). Boostrapping Bilingual Data

using Consensus Translation for a Multilingual Instant Messaging System. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002)*, pages 1–7, Taipei, Taiwan

Becker, J. D. (1975). The Phrasal Lexicon. In *Proceedings of Theoretical Issues in Natural Language Processing*, pages 70–73, Cambridge, MA.

Birch, A., Callison-Burch, C., and Osborne, M. (2006). Constraining the Phrase-Based, Joint Probability Statitistical Translation Model. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, pages 10–18, Boston, MA.

Bod, R. (1992). A Computational Model of Language Performance: Data Oriented Parsing. In *Proceedings of the 15th[sic] International Conference on Computational Linguistics (COLING'92)*, pages 855–859, Nantes, France.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., and Roosin, P. (1988). A Statistical Approach to Language Translation. In *Proceedings of the 12th International Conference on Computational Linguistics (COLING-88)*, pages 71–76, Budapest, Hungary.

Brown, P., Cocke, J., Pietra, S. D., Pietra, V. D., Jelinek, F., Mercer, R., and Roossin, P. (1990). A Statistical Approach to Machine Translation. *Computational Linguistics*, 16:79–85.

Brown, P , Pietra, S. D., Pietra, V. D., and Mercer, R. (1991). Word-Sense Disambiguation Using Statistical Methods. In *Proceedings of the 29th Annual Meeting of the Assosiation for Compuatational Lingusitics (ACL-91)*, pages 264–270, Berkeley, CA.

Brown, P , Pietra, S. D., Pietra, V. D., and Mercer, R. (1993). The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2):263–311.

Brown, R. (1996). Example-Based Machine Translation in the Pangloss System In *Proceedings of the 16th International Conference on Computational Linguistics (COLING-96)*, pages 169–174, Copenhagen, Denmark.

Brown, R. (1999a). Adding Linguistic Knowledge to a Lexical Example-based Translation System. In *Proceedings of the Eighth Conference on Theoretical and Methodological Issues in Machine Translation TMI-99*, pages 22–32, Chester, England.

Brown, R. (1999b). Adding Linguistic Knowledge to a Lexical Example-Based Translation System. In *Proceedings of the Eigghth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99)*, pages 22–32, Chester, England

Brown, R. (2000). Automated Generalization of Translation Examples. In *Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000 in Europe)*, pages 125–131, Saarbrücken, Germany.

Brown, R., Hutchinson, R., Bennett, P., Carbonell, J., and Janson, P. (2003). Reducing Boundary Friction Using Translation-Fragment Overlap In *Machine Translation Summit IX*, pages 24–31, New Orleans, LA.

Burbank, A., Carpuat, M., Clark, S., Dreyer, M., Fox, P., Groves, D., Hall, K., Hearne, M., Melamed, D., Shen, Y., Way, A., Wellington, B., and Wu, D. (2005). *Statistical Machine Translation by Parsing: Final Report.* Johns Hopkins Univeristy CLSP Summer Workshop 2005. `http://www.clsp.jhu.edu/ws2005/groups/statistical/documents/finalreport.pdf`.

Callison-Burch, C., Osborne, M., and Koehn, P. (2006). Re-evaluating the Role of BLEU in Machine Translation Research. In *Proceedings of the 11th European Chapter of the Association for Computational Linguistics (EACL-06)*, pages 249–256, Trento, Italy.

Carbonell, J., Klein, S., Miller, D., Steinbaum, M., Grassiany, T., and Frey, J.

(2006). Context-Based Machine Translation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas*, pages 19–28, Cambridge, MA.

Carl, M. and Hansen, S. (1999). Linking Translation Memories with Example-Based Machine Translation. In *Machine Translation Summit VII*, pages 617–624, Singapore.

Charniak, E., Knight, K., and Yamada, K. (2003). Syntax-Based Language Models for Statistical Machine Translation. In *Proceedings of the Ninth Machine Translation Summit*, pages 40–46, New Orleans, LO.

Chen, K. and Chen, H.-H. (1995). Machine Translation: An Integrated Approach In *Proceedings of the Sixth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-95)*, pages 287–294, Leuven, Belgium.

Chiang, D. (2005). A Hierarchical Phrase-Based Model for Statistical Machine Translation. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 263–279, Ann Arbor, MI.

Chomsky, N. (1969). Quine's Empirical Assumptions. In Davidson, D. and Hintikka, J., editors, *Words and Objections. Essays on the Work of W. V. Quine.* Reidel, Dordrecht, The Netherlands.

Church, K. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Leixcography. *Computational Linguistics*, 16(1):22–29.

Collins, B. (1998). *Example-Based Machine Translation: An Adaptation-Guided Retrieval Approach.* Ph.D. Thesis, Trinity College, Dublin, Ireland.

Cranias, L., Papageorgiou, H., and Piperidis, S. (1994). A Matching Technique in Example-Based Machine Translation. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING 1994)*, pages 100–104, Kyoto, Japan.

Dempster, A., Laird, N., and 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. In *Human Language Technology. Notebook Proceedings*, pages 128–132, San Diego, CA.

Frederking, R. and Nirenburg, S. (1994). Three Heads are Better than One. In *Proceedings of the 4th Conference on Applied Natural Language Processing*, pages 95–100, Stuttgart, Germany.

Frederking, R., Nirenburg, S , Farwell, D., Helmreich, S , Hovy, E., Knight, K , Beale, S., Domashnev, C., Attardo, D., Grannes, D., and Brown, R. (1994). Integrating Translations from Multiple Sources within the Pangloss Mark III Machine Translation System. In *Technology Partnerships for Crossing the Language Barrier: Proceedings of the First Conference of the Association for Machine Translation in the Americas*, pages 73–80, Columbia, Maryland.

Gamma, E., Helm, R., Johnson, R., and Vlissides, J. (1995). *Design Patters. Elements of Reusable Objected-Oriented Software.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA.

Gerloff, P (1987). Identifying the Unit of Analysis in Translation. In Faerch, C and Kasper, G., editors, *Introspection in Second Language Research,* pages 135–158 Calvedon; Philadelphia: Multilingual Matters.

Germann, U., Jahr, M , Knight, K., Marcu, D., and Yamada, K. (2001). Fast Decoding and Optimal Decoding for Machine Translation In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter*, pages 228–235, Toulouse, France.

Goodman, J. (2001). A Bit of Progress in Language Modeling. *Computer Speech and Language*, 15(1)·403–434.

Gough, N. (2005). *Example-Based Machine Translation Using the Marker Hypothesis.* PhD Thesis, Dublin City University, Dublin, Ireland.

Gough, N. and Way, A. (2004a). Example-Based Controlled Translation. In *Proceedings of the Ninth Conference of the European Association for Machine Translation Workshop*, pages 73–81, Valetta, Malta.

Gough, N. and Way, A. (2004b). Robust Large-Scale EBMT with Marker-Based Segmentation. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 95–104, Baltimore, MD.

Green, T (1979). The Necessity of Syntax Markers. Two experiments with artificial languages. *Journal of Verbal Learning and Behavior*, 18.481–496.

Groves, D., Hearne, M., and Way, A. (2004). Robust Sub-Sentential Alignment of Phrase-Structure Trees. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 1072–1078, Geneva, Switzerland.

Groves, D. and Way, A. (2005). Hybrid Example-Based SMT: the Best of Both Worlds? In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond*, pages 183–190, Ann Arbor, MI.

Groves, D. and Way, A. (2006a) Hybrid Data-Driven Models of MT. In *Machine Translation, Special Issue on EBMT.* (in press).

Groves, D. and Way, A. (2006b). Hybridity in MT: Experiments on the Europarl Corpus. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, pages 115–124, Oslo, Norway.

Hassan, H., Hearne, M., Sima'an, K., and Way, A. (2006). Syntactic Phrase-Based Statistical Machine Translation. In *Proceedings of the IEEE 2006 Workshop on Spoken Language Translation*, Palm Beach, Aruba (to appear).

Hearne, M. (2005). *Data-Oriented Models of Parsing and Translation.* PhD Thesis, Dublin City University, Dublin, Ireland.

Hearne, M. and Way, A. (2003). Seeing the Wood for the Trees: Data-Oriented Translation. In *Machine Translation Summit IX*, pages 165–172, New Orleans, LA.

Hearne, M. and Way, A. (2006). Disambiguation Strategies for Data-Oriented Translation. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, pages 59–68, Oslo, Norway.

Hutchins, J. (2005). Towards a Definition of Example-Based Machine Translation. In *Machine Translation Summit X: Second Workshop on Example-Based Machine Translation*, pages 63–70, Phuket, Thailand.

Hutchins, J. and Somers, H. (1992). *An Introduction to Machine Translation*, pages 161 – 174. Academic Press, London.

Imamura, K., Okuma, H., Watanabe, T., and Sumita, E. (2004). Example-based Machine Translation Based on Syntactic Transfer with Statistical Models. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING 2004)*, pages 99–105, Geneva, Switzerland.

Jelinek, F. (1998) *Statistical Methods for Speech Recognition*, chapter 5. The MIT Press, Cambridge, MA.

Jelinek, F. and Mercer, R. L. (1980). Interpolated Estimation of Markov Source Parameters from Sparse Data. In *Proceedings of the Workshop on Pattern Recognition in Practice*, pages 381–402, Amsterdam, the Netherlands: North-Holland.

Jurafsky, D. and Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* Prentice Hall, Upper Saddle River, NJ.

Kaji, H., Kida, Y , and Morimoto, Y. (1992). Learning Translation Templates from Bilingual Text. In *Proceedings of the 15th [sic] Internation Conference on Computational Linguistics (COLING-92)*, pages 672–678, Nantes, France.

Kaplan, R. and Bresnan, J. (1982). Lexical Functional Grammar, a Formal System for Grammatical Representation. In Bresnan, J., editor, *The Mental Representation of Grammatical Relations*, pages 173–281. MIT Press, Cambridge, MA.

Karlsson, F., Voutilainen, A., Heikkila, J., and Anttila, A. (1995) *Constraint Grammar: A Language-Independent System for Parsing Unrestricted Text.* Mouton de Gruyter, Berlin, New York.

Kay, M. (1997). The Proper Place of Men and Machines in Language Translation. *Machine Translation*, 12(1):3–23.

Khadivi, S. and Ney, H. (2005). Automatic Filtering of Bilingual Corpora for Statistical Machine Translation In *In Proceedings 10th International Conference on Application of Natural Language to Information Systems, NLDB 2005*, pages 263–274, Alicante, Spain. Springer Verlag

Kneser, R. and Ney, H. (1995). Improved Backing-off for M-gram Language Modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 181–184. Detroit, MI.

Knight, K. (1999) Decoding Complexity in Word-Replacement Translation Models. *Computational Linguistics*, 25(4):607–615.

Koehn, P. (2004a). Pharaoh: A Beam Search Decoder for Phrase-Based Statistical Machine Translation Models. In Frederking, R. and Taylor, K., editors, *Machine Translation: From Real Users to Research; AMTA 2004, LNAI 3265*, pages 115–124. Springer Verlag, Berlin/Heidelberg, Germany.

Koehn, P. (2004b). Statistical Significance Tests for Machine Translation Evaluation.

In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP-2004)*, pages 388–395, Barcelona, Spain.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., Och, F., and Marcu, D. (2003) Statistical Phrase-Based Translation In *Human Language Technology Conference (HLT-NAACL)*, pages 48–54, Edmonton, Canada

Langlais, P. and Simard, M. (2002). Merging Example-Based and Statistical Machine Translation. *Machine Translation: From Research to Real Users, AMTA-2002, LNAI 2499*, pages 104–113.

Lee, Y.-S (2004). Morphological Analysis for Statistical Machine Translation. In *Proceedings of the Joint Meeting of the Human Language Technology Conference and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL-04)*, pages 57–60, Lisbon, Portugal.

Leusch, G., Ueffing, N., and Ney, H. (2006). CDER: Efficient MT Evaluation Using Block Movements. In *Proceedings of 11th Conference of the European Chapter of the Association for Compuatational Linguistics (EACL-06)*, pages 241–248, Trento, Italy.

Levenshtein, V. I. (1965). Binary Codes Capable of Correcting Spurious Insertions and Deletions of Ones. *Problems of Information Transition*, 1:8–17.

Lidstone, G J. (1920). Note on the General Case of the Bayes-Laplace forumla for Inductive or *a posteriori* Probabilities. *Transactions of the Faculty of Actuaries*, 8.192–192.

Marcu, D. (2001). Towards a Unified Approach to Memory- and Statistical-Based Machine Translation. In *Proceedings of the 39th Annual Meeting of the Association*

for Compuatational Linguistics and 10th Conference of the European Chapter, pages 378–385, Toulouse, France.

Marcu, D., Wang, W., Echihabi, A., and Knight, K. (2006). SPMT: Statistical Machine Translation with Syntactified Target Language Phrases. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 44–52, Sydney, Australia.

Marcu, D. and Wong, W. (2002). A Phrase-Based Joint Probability Model for Statistical Machine Translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 133–139, Philadelphia, PA.

Marcus, M , Kim, G., Marcinkiewicz, M. A., MacIntyre, R., Bies, A., Ferguson, M., Katz, K., and Schasberger, B. (1994). The Penn Treebank: Annotating Predicate Argument Structure. In *Proceedings of the ARPA Human Language Technology Workshop*, pages 114–119, Princeton, New Jersey.

Maruyama, H. and Watanabe, H. (1992). Tree Cover Search Algorithm for Example-Based Translation. In *Proceedings of the Fourth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-92)*, pages 173–184, Montréal, Canada.

Matsumoto, Y., Ishimoto, H., and Utsuro, T. (1993). Structural Matching of Parallel Texts. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 23–30, Columbus, OH.

Mellebeek, B., Owczarzak, K., Groves, D , van Genabith, J., and Way, A. (2006). A Syntactic Skeleton for Statistical Machine Translation. In *Proceedings of the 11th Conference of the European Association for Machine Translation*, pages 195–202, Oslo, Norway.

Nagao, M. (1984). A Framework of a Mechanical Translation between Japanese and English by Analogy Principle. In Elithorn, A. and Banerji, R., editors, *Ar-*

tificial and Human Intelligence, pages 173–180. North-Holland, Amsterdam, The Netherlands.

Nie, J.-Y. and Cai, J. (2001). Filtering Noisy Parallel Corpora of Web Pages. In *IEEE Symposium on NLP and Knowledge Engineering*, pages 453–458, Tuscon, AZ.

Nirenburg, S., Domashnev, C., and Grannes, D. J. (1993). Two Approaches to Matching in Example-Based Machine Translation. In *Proceedings of the Fifth International Conference on Theoretical and Methodological Issues in Machine Translation TMI '93: MT in the Next Generation*, pages 47–57, Kyoto, Japan.

Och, F. J. (1999). An Efficient Method to Determine Bilingual Word Classes. In *EACL '99. Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 71–76, Bergen, Norway.

Och, F. J. (2003). Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 160–167, Sapporo, Japan.

Och, F. J and Ney, H. (2002). Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 295–302, Philadelphia, PA.

Och, F. J. and Ney, H. (2003). A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics*, 29.19–51.

Och, F. J., Tillmann, C., and Ney, H. (1999). Improved Alignment Models for Statistical Machine Translation. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 20–28, University of Maryland, College Park, MD.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2001). BLEU: a Method for Automatic Evaluation of Machine Translation. Technical report, IBM T J. Watson Research Center, Yorktown Heights, NY.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02)*, pages 311–318, Philadelphia, PA

Paul, M., Doi, T., Hwang, Y., Imamura, K., and Sumita, E. (2005a). Nobody is Perfect: ATR's Hybrid Approach to Spoken Language Translation. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, pages 55–62, Pittsburgh, PA.

Paul, M., Sumita, E., and Yamamoto, S. (2005b). A Machine Learning Approach to Hypothesis Selection of Greedy Decoding for SMT. In *Machine Translation Summit X: Second Workshop on Example-Based Machine Translation*, pages 117–124, Phuket, Thailand.

Planas, E. and Furuse, O. (2003). Formalizing Translation Memory. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 157–188. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Poutsma, A. (2000). Data-Oriented Translation: Using the Data-Oriented Parsing framework for Machine Translation. Master's thesis, University of Amsterdam, The Netherlands.

Poutsma, A. (2003). Machine Translation with Tree-DOP. In Bod, R., Scha, R., and Sima'an, K., editors, *Data-Oriented Parsing*, pages 339–357. Stanford CA: CSLI Publications.

Quirk, C and Menezes, A (2006). Dependency Treelet Translation: the convergence of statistical and example-based machine translation? In *Machine Translation, Special Issue on EBMT.* (in press).

Sato, S. and Nagao, M. (1990). Toward Memory-Based Translation. In *Proceedings of the 13th International Conference on Computational Linguistics (COLING-90)*, volume 3, pages 247–252, Helsinki, Finland.

Schäler, R. (1996). Machine Translation, Translation Memories and the Phrasal Lexicon: the Localisation Perspective. In *Proceedings of the First Workshop of the European Association for Machine Translation: Language Resources, Terminology, Economics and User Needs*, pages 21–33, Vienna, Austria.

Shen, W., Delaney, B., and Anderson, T. (2006). The MITLL/AFRL TC-Star System: Experiments with Large Vocabulary Speech Translation. In *TC-Star OpenLab on Speech Translation*, Trento, Italy. Slides at `http://tc-star.org/openlab2006/day3/Wade_TC-Star-Openlab-presentation-v1a.pdf`.

Somers, H (1999). Review Article: Example-based Machine Translation. *Machine Translation*, 14(2):113–157.

Somers, H. (2003). An Overview of EBMT. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 3–57. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Somers, H., McLean, I., and Jones, D. (1994). Experiments in Multilingual Example-Based Generation. In *Proceedings of the 3rd Conference on the Cognitive Science of Natural Language Processing (CSNLP-94)*, [pages not numbered], Dublin, Ireland.

Stolcke, A. (2002). SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing*, volume 2, pages 901–904, Denver, CO.

Stroppa, N., Groves, D., Way, A., and Sarasola, K. (2006). Example-Based Machine Translation of the Basque Language. In *Proceedings of the 7th Biennial Conference of the Association for Machine Translation in the Americas*, pages 232–241, Boston, MA.

Stroppa, N. and Way, A. (2006). MaTrEx: DCU Machine Translation System for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 31–36, Kyoto, Japan.

Sumita, E. (2003). EBMT Using DP-Matching Between Word Sequences. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 189–210. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Sumita, E., Iida, H., and Kohyama, H. (1990). Translating With Examples: A New Approach to Machine Translation. In *Third International Conference on Theoretical and Methodological Issues in Machine Translation*, pages 203–212, Austin, TX.

Tiedemann, J. (2004). Word to Word Alignment Strategies. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-2004)*, pages 212–218, Geneva, Switzerland

Tillmann, C. and Ney, H. (2000). Word Re-Ordering and DP-Based Search in Statistical Machine Translation. In *The 18th International Conference on Computational Linguistics (COLING-00)*, pages 850–856, Saarbrücken, Germany.

Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., and Sawaf, H. (1997). Accelerated DP Based Search for Statistical Translation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2667–2670, Rhodes, Greece

Turcato, D. and Popowich, F. (2003). What is Example-Based Machine Translation? In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 59–79. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Turian, J., Shen, L., and Melamed, I. D. (2003). Evaluation of Machine Translation and its Evaluation In *Machine Translation Summit IX*, pages 386–393, New Orleans, LA.

Ueffing, N , Och, F , and Ney, H. (2002). Generation of Word Graphs in Statistical Machine Translation. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pages 156–163, Philadelphia, PA.

Veale, T. and Way, A. (1997). Gaijin. A Bootstrapping Approach to Example-Based Machine Translation. In *Proceedings of the 2nd International Conference on Recent Advances in Natural Language Processing (RANLP-97)*, pages 239–244, Tzigov Chark, Bulgaria.

Vogel, S. and Ney, H. (2000). Construction of a Hierarchical Translation Memory. In *Proceedings of the 18th International Conference on Computational Linguistics: COLING 2000 in Europe*, pages 1131–1135, Saarbrücken, Germany.

Vogel, S., Zhang, Y., Huang, F., Tribble, A , Venugopal, A., Zhao, B., and Waibel, A (2003). The CMU Statistical Machine Translation System In *Proceedings of MT Summit IX*, pages 110–117, New Orleans, LA.

Wagner, R. A. and Fischer, M. J. (1974). The String-to-String Correction Problem. *Journal of the Association of Computing Machinery*, 21(1):168 -173.

Watanabe, H. (1992). A Similarity-Driven Transfer System. In *Proceedings of the 15th [sic] International Conference on Computational Linguistics (COLING-1992)*, pages 770- 776, Nantes, France.

Watanabe, H., Kurohashi, S., and Aramaki, E. (2003). Finding translation patterns from paired source and target dependency structures. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 397–420. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Way, A. (2003). Translating With Examples: the LFG-DOT Models of Translation. In Carl, M. and Way, A., editors, *Recent Advances in Example-Based Machine Translation*, pages 443–472. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Way, A. and Gough, N. (2003). *wEBMT*: Developing and Validating an Example-Based Machine Translation System using the World Wide Web. *Computational Linguistics. Special Issue on the Web as Corpus*, 29(3):421–457.

Way, A. and Gough, N. (2005a). Comparing Example-Based and Statistical Machine Translation. *Natural Language Engineering*, 11(3):295–309.

Way, A. and Gough, N (2005b). Controlled Translation in an Example-Based Environment: what do Automatic Evaluation Metrics tell us? *Machine Translation*, 19(2):1–36.

Weaver, W. (1949). Recent Contributions to the Mathematical Theory of Communication. In Shannon, C. E. and Weaver, W., editors, *The Mathematical Theory of Communication*, pages 94–117. The University of Illinois Press, Urbana, IL.

Wu, D. (2006). MT Model Space: Statistical vs. Compositional vs. Example-Based Machine Translation. In *Machine Translation, Special Issue on EBMT*. (in press).

Yamada, K. and Knight, K. (2001). A Syntax-Based Statistical Translation Model In *Proceedings of the 39th Annual Conference for the Association for Computational Linguistics and 15th Conference of the European Chapter*, pages 523–530, Toulouse, France.

Yamada, K. and Knight, K. (2002). A Decoder for Syntax-Based Statistical SMT. In *Proceedings of the 40th Annual Conference for the Association for Computational Linguistics*, pages 303–310, Philadelphia, PA.

Zens, R., Bender, O., Hasan, S., Khadivi, S., Matsuov, E., Xu, J., Zhang, Y., and Ney, H. (2005). The RWTH Phrase-Based Statistical Machine Translation System. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT-05)*, pages 155–162, Pittsburgh, PA.

Zens, R. and Ney, H. (2004). Improvements in Phrase-Based Statistical Machine Translation. In *Proceedings of the Human Language Technology Conference /*

North American Chapter of the Association of Computational Linguistics Annual Meeting (HLT-NAACL-04), pages 257–264, Boston, MA

Zens, R., Ney, H., Watanabe, T., and Sumita, E. (2004). Reordering Constraints for Phrase-based Statistical Machine Translation. In *In Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, pages 205–211, Geneva, Switzerland.

Zhang, Y. and Vogel, S. (2004). Measuring Confidence Intervals for Machine Translation Evaluation Metrics. In *Proceedings of the Tenth Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pages 4–6, Baltimore, MD.