Dependency-based translation equivalents for factored machine translation

Irimia Elena, Alexandru Ceauşu

Reasearch Centre for Artificial Intelligence, Bucharest, Romania
{elena, aceausu}@racai.ro
www.racai.ro

Abstract. One of the major concerns of the machine translation practitioners is to create good translation models: correctly extracted translation equivalents and a reduced size of the translation table are the most important evaluation criteria. This paper presents a method for extracting translation examples using the dependency linkage of both the source and target sentence. To decompose the source/target sentence into fragments, we identified two types of dependency link-structures - super-links and chains - and used these structures to set the translation example borders. The option for the dependency-linked n-grams approach is based on the assumption that a decomposition of the sentence in coherent segments, with complete syntactical structure and which accounts for extra-phrasal syntactic dependency would guarantee "better" translation examples and would make a better use of the storage space. The performance of the dependency-based approach is measured with the BLEU-NIST score and in comparison with a baseline system.

Keywords. Lexical attraction model, statistical machine translation, translation model

1. Introduction

Corpus-based paradigm in machine translation has seen various approaches for the task of constructing reliable translation models,

– starting from the naïve "word-to-word" correspondences solution which was studied in the early works ([1], [2])
– continuing with the chunk-bounded n-grams ([3], [4], [5]) which were supposed to account for compounding nouns, collocations or idiomatic expressions,
– passing through the early approach of the bounded-length n-grams IBM statistical translation models and the following phrase-based statistical translation models ([6], [7], etc.),
– exploring the dependency-linked n-grams solutions which can offer the possibility of extracting long and sometimes non-successive examples and are able to catch the structural dependencies in a sentence (e.g., the accord between a verb and a noun phrase in the subject position), see [8],
– and ending with the double-sided option for the sentence granularity level, which can be appealing since the sentence boundaries are easy to identify but brings the

additional problem of fuzzy matching and complicated mechanisms of recombination.

Several studies were dedicated to the impact of using syntactical information in the phrase extraction process over the translation accuracy. Analyzing by comparison the constituency-based model and the dependency based model, [9] concluded that “using dependency annotation yields greater translation quality than constituency annotation for PB-SMT”. But, as previous works ([10] and [11]) have noted, the new phrase models, created by incorporating linguistic knowledge, do not necessarily improve the translation accuracy by themselves, but in combination with the “old–fashioned” bounded-length phrase models.

The process of extracting syntactically motivated translation examples varies according to the different resources and tools available for specific research groups and specific language pairs. In a detailed report over the syntactically-motivated approaches in SMT, focused on the methods that use the dependency formalism, [12] distinguishes the situations when dependency parsers are used for both source and target languages from those in which only a parser for the source side is available. In the latter case, a direct projection technique is usually used to do an annotation transfer from the source to the target translation unit. This approach is motivated by the *direct correspondence assumption* (DCA, [13]), that states that dependency relations are preserved through direct projection. The projection is based on correspondences between the words in the parallel sentences, obtained through the lexical alignment (also called word alignment) process. Obviously, the quality of the projection is dependant of the lexical alignment quality. Furthermore, [13] notes that the target syntax structure obtained through direct projection is isomorphic to the source syntax structure, thus producing isomorphic translation models. This phenomenon is rarely corresponding to a real isomorphism between the two languages involved.

In the experiments we describe in this paper, we had the advantage of a probabilistic non-supervised dependency analyzer which depends on the text’s language only through a small set of rules designed to filter the previously identified links. As both source and target dependency linking analysis is available, there is no need of direct projection in the translation examples extraction and the problem of the “compulsory isomorphism” is avoided.

2. Research Background

In previous experiments with an example-based approach on machine translation for the English-Romanian language pair, we developed a strategy for extracting translation examples using the information provided by a dependency-linker described in [14]. We then justified our opting for the dependency-linked n-grams approach based on the assumption in [15] that the EBMT potential should rely on exploiting text fragments shorter than the sentence and also on the intuition that a decomposition of the source sentence in “coherent segments”, with complete syntactical structure, would be “the best covering” of that sentence.

The dependency-linker used is based on Yuret’s Lexical Attraction Model (LAM, [16]), in who’s vision the lexical attraction is a probabilistic measure of the combining affinity between two words in the same sentence. Applied to machine translation, the lexical attraction concept can serve as a mean of guaranteeing the translation examples usefulness. If two words are “lexically attracted” to one another in a sentence, the probability for them to combine in future sentences is significant. Therefore, two or more words from the source sentence that manifest lexical attraction together with their translations in the target language represent a better translation example than a bounded length n-gram.

The choice for the Yuret’s LAM as the base for the dependency analyzer application was motivated by the lack of a dependency grammar for Romanian. The alternative was to perform syntactical analysis based on automatically inducted grammatical models. A basic request for the construction of this type of models is the existence of syntactically annotated corpora from which machine learning techniques could extract statistical information about the ways in which syntactical elements combine. As no syntactically annotated corpus for Romanian was available, the fact that Yuret’s method could use LAM for finding dependency links in a not-annotated corpus made this algorithm a practical choice.

LexPar[14], the dependency links analyzer we used for the experiments described in this paper, is extending Yuret’s algorithm by a set of syntactical rules specific to the processed languages (Romanian and English) that constraints the links’ formation. It also contains a simple generalization mechanism for the link properties, which eliminates the initial algorithm inadaptability to unknown words. However, the LexPar algorithm does not guarantee a complete analysis, because the syntactic filter can contain rules that forbid the linking of two words in a case in which this link should be allowed. The rules were designed by the algorithm’s author based on his observations of the increased ability of a certain rule to reject wrong links, with the risk of rejecting good links in few cases.

In our research group, significant efforts were involved in experimenting with statistical machine translation methodologies, focused on building accurate language resources (the larger the better) and on fine-tuning the statistical parameters. The aim was to demonstrate that, in this way, acceptable MT prototypes can be quickly developed and the claim was supported by the encouraging Bleu scores we obtained for the Romanian<->English translation system. The translation experiments employed the MOSES toolkit, an open source platform for development of statistical machine translation systems (see next section).

One of the goals of this paper was to analyze the impact of incorporating syntactic information in the translation model by means of a probabilistic dependency link analyzer. Although the non-supervised nature of the analyzer is affecting its recall, using this tool brings the advantage of having syntactic information available for translation without the need for training syntactically annotated corpora. We feed the Moses decoder with the new translation model and we compare the translation results with the results of the baseline system. In the remaining sections we will make a short survey of the resources and tools used in the SMT experiments (section 3), we will describe the dependency-motivated translation examples extraction process (section 4) and we will present the experiments and the results with the dependency-based translation model (section 5).

3. Factored Phrase-Based Statistical Machine Translation

The corpus. The Acquis Communautaire is the total body of European Union (EU) law applicable in the EU Member States. This collection of legislative text changes continuously and currently comprises texts written between the 1950s and 2008 in all the languages of EU Member States. A significant part of these parallel texts have been compiled by the Language Technology group of the European Commission's Joint Research Centre at Ispra into an aligned parallel corpus, called JRC-Acquis [17], publicly released in May 2006. Recently, the Romanian side of the JRC-Acquis corpus was extended up to a size comparable with the dimensions of other language-parts (19,211 documents)).

For the experiments described in this paper, we retained only 1-1 alignment pairs and restricted the selected pairs so that none of the sentences contained more than 80 words and that the length ratio between sentence-lengths in an aligned pair was less than 7. Finally, the Romanian-English parallel corpus we used contained about 600,000 translation units.

Romanian and English texts were processed based on the RACAI tools [18] integrated into the linguistic web-service platform available at http://nlp.racai.ro/webservices. After tokenization, tagging and lemmatization, this new information was added to the XML encoding of the parallel corpora. Figure 1 shows the representation of the Romanian segment encoding for the translation unit displayed in Figure 2. The tagsets used were compliant with the MULTEXT-East specifications Version3 [19] (for the details of the morpho-syntactic annotation, see http://nl.ijs.si/ME/V3/msd/).

```
<tu id="3936">
    ...
         <seg lang="ro">
              <s id="31985L0337.n.83.1">
                    <w lemma="informație" ana="Ncfpry">Informațiile</w>
                   <w lemma="culege" ana="Vmp--pf">culese</w>
                    <w lemma="conform" ana="Spsd">conform</w>
                   <w lemma="art." ana="Yn">art.</w>
                   <w lemma="5" ana="Mc">5</w>
                   <c>,</c>
                   <w lemma="6" ana="Mc">6</w>
                   <w lemma="şi" ana="Crssp">şi</w>
                   <w lemma="7" ana="Mc">7</w>
                   <w lemma="trebui" ana="Vmip3s">trebuie</w>
                   <w lemma="să" ana="Qs">să</w>
                   <w lemma="fi" ana="Vasp3">fie</w>
                   <w lemma="lua" ana="Vmp--pf">luate</w>
                   <w lemma="în" ana="Spsa">în</w>
                   <w lemma="considerare" ana="Ncfsrn">considerare</w>
                   <w lemma="în cadrul" ana="Spcg">în cadrul</w>
                   <w lemma="procedură" ana="Ncfsoy">procedurii</w>
                   <w lemma="de" ana="Spsa">de</w>
                   <w lemma="autorizare" ana="Ncfsrn">autorizare</w>
                   <c>.</c>
```

```
                </s>
            </seg>
        ...
</tu>
```

Figure 1: Linguistically analysed sentence (Romanian) of a translation unit of the JRC-Acquis parallel corpus

Based on the monolingual data from the JRC-Acquis corpus we built language models for each language. For Romanian we used the TTL [20] and METT [21] tagging modelers. Both systems are able to perform tiered tagging [22], a morpho-syntactic disambiguation method that was specially designed to work with large (lexical) tagsets.

In order to build the translation models from the linguistically analyzed parallel corpora we used GIZA++ [23] and constructed unidirectional translation models (EN-RO, RO-EN) which were subsequently combined. After that step, the final translation tables were computed. The processing unit considered in each language was not the word form but the string formed by its lemma and the first two characters of the associated morpho-syntactic tag (e.g. for the wordform "informațiile" we took the item "informație/Nc"). We used for each language 20 iterations (5 for Model 1, 5 for HMM, 1 for THTo3, 4 for Model3, 1 for T2To4 and 4 for Model4). We included neither Model 5 nor Model 6, as we noticed a degradation of the perplexities of the alignment models on the evaluation data.

The MOSES toolkit [24] is a public domain environment, which was developed in the ongoing European project EUROMATRIX, and allows for rapid prototyping of Statistical Machine Translation systems. It assists the developer in constructing the language and translation models for the languages he/she is concerned with and by its advanced factored decoder and control system ensures the solving of the fundamental equation of the Statistical Machine Translation in a noisy-channel model:

$$\text{Target*} = \text{argmax}_{\text{Target}}\, P(\text{Source}|\text{Target})*P(\text{Target}) \quad (1)$$

The P(Target) is the statistical representation of the (target) language model. In our implementation, a language model is a collection of prior and conditional probabilities for unigrams, bigrams and trigrams seen in the training corpus. The conditional probabilities relate lemmas and morpho-syntactic descriptors (MSD), word-forms and lemmas, sequences of two or three MSDs. The P(Source|Target) is the statistical representation of the translation model and it consists of conditional probabilities for various attributes characterizing equivalences for the considered source and target languages (lemmas, MSDs, word forms, phrases, dependencies, etc). The functional *argmax* is called a decoder and it is a procedure able to find, in the huge search space P(Source|Target)*P(Target) corresponding to possible translations of a given Source text, the Target text that represent the optimal translation, i.e. the one which maximizes the compromise between the *faithfulness* of translation (P(Source|Target)) and the *fluency/grammaticality* of the translation (P(Target)). The standard implementation of a decoder is essentially an A* search algorithm. The current state-of-the-art decoder is the factored decoder implemented in the MOSES toolkit. As the name suggests, this decoder is capable of considering multiple information sources (called factors) in implementing the *argmax* search. What is extremely useful is that the MOSES environment allows a developer to provide the MOSES decoder with language and translation models externally developed, offering means to ensure the conversion of the necessary data structures into the expected format and further improve them. Once the statistical models are in the prescribed format, the MT system developer may define his/her own factoring strategy. If the information is provided, the MOSES decoder can use various factors (attributes) of each of the lexical items (words or phrases): occurrence form, lemmatized form, associated part-of-speech or morpho-syntactic tag. Moreover, the system allows for integration of higher order information (shallow or even deep parsing information) in order to improve the output lexical items reordering. For further details on the MOSES Toolkit for Statistical Machine Translation and its tuning, the reader is directed to the EUROMATRIX project web-page http://www.euromatrix.net/ and to the download web-page http://www.statmt.org/moses/.

4. Extracting Translation Examples from Corpora (ExTRact)

In our approach, based on the availability of a dependency-linker for both the source and the target language, the task of extracting translation examples from a corpus contains two sub-problems: *dividing* the source and target sentences *into fragments* (according to the chosen approach) and *setting correspondences* between the fragments in the source sentence and their translations in the target sentence. The last problem is basically *fragment alignment* and we solved it through a heuristic based on lexical alignments produced by GIZA++.

The remaining problem was addressed using the information provided by LexPar, the dependency linker mentioned above. With a recall of 60,70% for English, LexPar was considered an appropriate starting point for the experiments (extending or correcting the set of rules incorporated as a filter in LexPar can improve it's recall).

Using MtKit, a tool specially designed for the visualization and correction of lexical alignments adapted to allow the graphical representation of the dependency links, we could study the dependency structures created by the identified links inside a sentence and we were able to observe some patterns in the links' behavior: they tend to group by nesting and to decompose the sentence by chaining. Of course, these patterns are direct consequences of the syntactical structures and rules involved in the studied languages, but the visual representation offered by MtKit simplified the task of formalization and heuristic modeling (see Fig. 1).

These properties suggest more possible decompositions for the same sentence, and implicitly the extraction of substrings of different length that satisfy the condition of lexical attraction between the component words.

Example 1: in Figure 1, from the word sequence "made in the national currency" can be extracted the subsequences: "national currency", "the national currency", „in the national currency", „made in the national currency". The irrelevant *sequences and those susceptible of generating errors (like "the national", "in the", "made in the national") are ignored.*

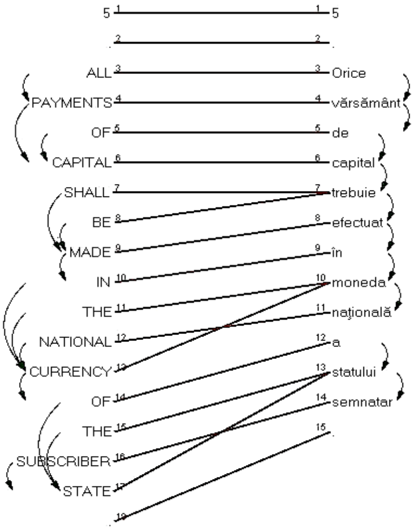

Fig. 2. MtKit visualisation of the alignments and links for an english-romanian translation unit. An arrow marks the existence of a dependence link between the two words it unites. The arrow direction is not relevant for the dependency link orientation.

The patterns observed above were formalized as *superlinks* (link structures composed of at least two simple links which nest, see Figure 3) and as *chains* (link structures composed of at least two simple links or superlinks which form a chain, see Figure 4).

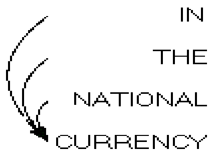

Fig. 3. Superlink structures

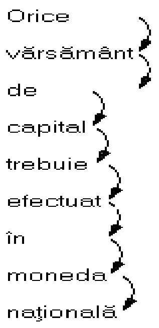

Fig. 4. Chain structures

As input data, *ExTract* (the application that extracts translation examples from corpora) receives the processed corpus and a file containing the lexical alignments produced by GIZA++ [23]. We will describe the extracting procedure for a single translation unit U in the corpus, containing *Ss* (a source sentence) and its trans-lation *Ts* (a target sentence). Starting from the first position in Ss (Ts respectively) we identify and extract every possible chaining of *links* and *superlinks*, with the condition that the number of chain loops is limited to 3. The limitation was introduced to avoid overloading the database. Subsequent experiments showed that increasing the limitation to 4 or 5 chains did not significantly improve the BLEU score of the translation system. Two list of candidate sentence fragment, from Ss and Ts, are extracted.

Every fragment in both sentences is projected through lexical alignment in a word string (note that this is not the direct syntactical structure projection discussed above) in the other language. A projected string of a candidate fragment in Ss is not necessarily part of the list of candidate sentence fragments Ts, and vice versa (LexPar is not able to identify all the dependency links in a sentence, the lexical alignments are also subject to errors). But if a fragment candidate from Ss projects to a fragment candidate from Ts, the pair has a better probability of representing a correct translation example. In this stage, the application extracts all the possible translation examples (<*source fragment candidate, projected word string*>, <*projected word string, target fragment candidate*>*)* but distinguish between them, associating a "trust" flag f="2" to the translation examples of the form <*source fragment candidate, target fragment candidate*>, and a flag f="1" to all the other. Thereby, it is possible to experiment with translation tables of different sizes and different quality levels.

5. Experiments and results

Taking into account results from previous works ([12],[13]) that proved that dependency-based translation models give improved performance in combination with a phrase-based translation model, we decided to conduct our experiments in a mixed frame: we extracted from the dependency-based translation model only the translation examples longer than *2 source words <-> 2 target words,* creating a reduced dependency-based translation model and we combined it with the phrase-based translation model generated with the Moses toolkit.

Starting from the reduced D-based translation model, we can develop two different translation tables, based on the "trust" flags we introduced before:

- a *trustful D-based translation table* (if we keep only the examples with the flag f="2")
- a *relaxed D-based translation table* (if we accept all the examples, irrespective of the flags).

As we previously mentioned, the initial working corpus contained around 600,000 translation units. From this number, 600 were extracted for tuning and testing. The tuning of the factored translation decoder (the weights on the various factors) was based on the 200 development sentence pairs using MERT [25] method. The testing set contains 400 translation units.

The evaluation tool was the last version of the NIST official mteval script[1] which produces BLEU and NIST scores [26]. For the evaluation, we lowered the case in both reference and automatic translations. The results are synthesized in the following table, where you can notice that our assumption that the *trustful* table would produce better results than the *relaxed* one was contradicted by evidence. We thus learned that a wider range of multi-word examples is preferable to a restricted one, even if their correctness was not guaranteed by the syntactical analysis.

Table 1. Evaluation of the dependency translation table compared with the translation table generated with Moses (on unseen data)

Language pair	**Moses translation table**		**Dependency translation table**			
			Trustful table		**Relaxed table**	
	NIST score	**BLEU score**	**NIST. score**	**BLEU score**	**NIST. score**	**BLEU score**
English to Romanian	8.6671	*0.5300*	8.4998	0.5006	8.6900	*0.5334*
Romanian to English	10.7655	*0.6102*	10.3122	0.5812	10.3235	*0.6191*

As can be seen in the table, the translation accuracy obtained with the dependency-based translation table is very close to the one manifested by Moses, but still lesser. Therefore, we took a closer look at the translations and we noticed an important number of cases in which the dependency-based translation was more accurate in terms of human evaluation. Because of the space restriction, we will present here only a few of these cases and only for one direction of translation (English to Romanian). It can be noticed that the exact n-gram matching between the dependency-based translation and the reference is not as successful as the one between the Moses translation and the reference. But a flexible word matching, allowing for morphological variants and synonyms to be taken into account as legitimate correspondences, shows that the dependency-based translation is also very legitimate in terms of human translation evaluation.

English original:

the insurance is connected to a contract to provide assistance in the event of accident or breakdown involving a road vehicle;

whereas, in the light of experience gained, it is necessary to reconsider the consequences of the disposal of products from intervention on the markets of third countries other than those intended at the time of exportation;

the competent authorities of the member states shall afford each other administrative assistance in all supervisory procedures in connection with legal provisions and quality standards applicable to foodstuffs and in all proceedings for infringements of the law applicable to foodstuffs.

any administrative measure taken against an individual, leaving aside any consideration of general interest referred to above, on one of the grounds mentioned in article 1a, which is sufficiently severe in the light of the criteria referred to in section 4 of this joint position, may be regarded as persecution, in particular where it is intentional, systematic and lasting.

Romanian original:

[1] ftp://jaguar.ncsl.nist.gov/mt/resources/mteval-v12.pl

asigurarea priveşte un contract de acordare de asistenţă în caz de accident sau defecţiune a unui vehicul rutier;

întrucât, luând în considerare experienţa dobândită, este necesar să se reconsidere consecinţele desfacerii produselor de intervenţie asupra pieţelor din ţări terţe altele decât cele prevăzute în cazul exportului;

autorităţile competente din statele membre trebuie să îşi acorde reciproc asistenţă administrativă în toate procedurile de supraveghere prevăzute în dispoziţiile legale şi în normele de calitate aplicabile alimentelor, precum şi în toate procedurile privind încălcarea legislaţiei în domeniul produselor alimentare.

orice măsură administrativă luată împotriva unui individ, în afara considerentelor de interes general evocate mai sus, datorită unuia dintre motivele menţionate în art. 1a, care este suficient de severă potrivit criteriilor enunţate în secţiunea 4 din prezenta poziţie comună, poate fi considerată ca persecuţie, în special când aceasta prezintă un caracter intenţional, sistematic şi durabil.

Moses translation :

asigurarea este conectat la un contract să furnizeze asistenţă în caz de accident sau defecţiune a unui vehicul rutier ;

întrucât, ţinând seama de experienţa dobândită, este necesar să se reconsidere consecinţele comercializării produselor de intervenţie pe pieţele din ţările terţe, altele decât cele prevăzute în momentul exportului;

autorităţile competente ale statelor membre îşi acordă reciproc asistenţă administrativă în toate procedurile de supraveghere legate de dispoziţiile legale şi standardele de calitate se aplică produselor alimentare şi în toate procedurile pentru încălcarea legii aplicabile produselor alimentare

orice măsură administrativă luată împotriva unui individ, lăsând din circuitul agricol orice cauză de interes general menţionat anterior, pe unul din motivele menţionate în art. 1a, care este suficient de grave în lumina criteriilor menţionate la punctul 4 din prezenta poziţie comună, pot fi considerate ca persecuţie, în special atunci când s-a intenţionat, sistematic şi de durată.

Dependency-based translation :

asigurarea priveşte un contract de asistenţă în caz de accident sau defecţiune a unui vehicul rutier;

întrucât, în lumina experienţei acumulate, este necesar să se reconsidere consecinţele comercializării produselor de intervenţie pe pieţele ţărilor terţe altele decât cele prevăzute în cazul exportului;

autorităţile naţionale competente din statele membre acorde reciproc asistenţă administrativă în toate procedurile prevăzute în dispoziţiile financiare şi ale standardelor de calitate aplicabile produselor alimentare şi în toate procedurile privind încălcarea legii aplicabile produselor alimentare.

orice măsură administrativă luată împotriva unui individ, exclusiv, în afara considerentelor de interes general menţionat anterior, pentru unul din motivele menţionate la articolul 1a, care este suficient de grave ţinând seama de criteriile enunţate în secţiunea 4 din prezenta poziţie comună, poate fi considerată ca persecuţie, în cazul în care este intenţionat, sistematic şi durabile.

5. Conclusions

We described in this paper our method of extracting translation examples from corpora based on the links identified with a statistical non-supervised dependency-linker. Although the evaluation results did not overcome the performance of the

Moses translation model, the scores are promising and they can be improved by increasing LexPar's recall. We also intend to evaluate the results using metrics more sensitive to morphology variations and synonymy (e.g. METEOR, [27]).

7. Acknowledgements

The work reported here is funded by the STAR project, financed by the Ministry of Education, Research and Innovation under the grant no 742.

6. References

1. Gale, W. and K. Church, 1991. Identifying Word Correspondences in Parallel Texts. In Proceedings of the 4th DARPA Speech and Natural Language Workshop, Pacific Grove, CA., pp. 152-157.
2. Melamed, I.D. 1995. Automatic Evaluation and Uniform Filter Cascades for Inducing N-best translation lexicons. In proceedings of the Third Annual Workshop on Very Large Corpora, Cambridge, England, pp. 184-198.
3. Kupiec, J. 1993. *An Algorithm for Finding Noun Phrases Correspondences in Bilingual Corpora.* In 31st Annual Meeting of the Association for Computational Linguistics, Columbus, OH., pages 23-30.
4. Kumano, A. and H. Hirakawa. 1994. *Building an MT dictionary from parallel texts based on linguistic and statistical information.* In COLING-94: Proceedings of the 15th International Conference on Computational Linguistics, Kyoto, Japan, pages 76-81.
5. Smadja, F., K.R. McKeown and V. Hatzivassiloglou. 1996. *Translating Collocations for Bilingual Lexicons: A Statistical Approach.* Computational Linguistics 22(1):1-38.
6. Och, F.-J., Ch. Tillmann and H. Ney. 1999. *Improved Alignment Models for Statistical Machine Translation.* In Proc. of the Joint Conf. on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC 99), pages 20–28, College Park, MD, June.
7. Marcu D. and W. Wong. 2002. *A Phrased-Based, Joint Probability Model for Statistical Machine Translation.* In Proceedings Of the Conference on Empirical Methods in Natural Language Processing (EMNLP 02); pages 133-139, Philadelphia, PA, July.
8. Yamamoto, K. and Y. Matsumoto. 2003. *Extracting translation knowledge from parallel corpora.* In: Michael Carl & Andy Way (eds.) Recent advances in example-based machine translation (Dordrecht: Kluwer Academic Publishers, 2003); pages 365-395.
9. Hearne, M., S. Ozdowska, and J. Tinsley. 2008. *Comparing Constituency and Dependency Representations for SMT Phrase-Extraction.* In Proceedings of TALN '08, Avignon, France.
10. Groves D. & Way A. 2005. *Hybrid Example-Based SMT: the Best of Both Worlds?* In Proceedings of ACL 2005 Workshop on Building and Using Parallel Texts: Data-Driven Machine Translation and Beyond, p. 183–190, Ann Arbor, MI.
11. Tinsley J., Hearne M. & Way A. 2007. *Exploiting Parallel Treebanks to Improve Phrase-Based Statistical Machine Translation.* In Proceedings of The Sixth

International Workshop on Treebanks and Linguistic Theories (TLT-07), Bergen, Norway.

12. Ambati, V. 2008. *Dependency Structure Trees in Syntax Based Machine Translation*, 11-734 Spring 2008, Survey Report, http://www.cs.cmu.edu/~vamshi/publications/DependencyMT_report.pdf
13. Hwa, R., Ph. Resnik, A. Weinberg, C. Cabezas and O. Kolak. 2005. *Bootstrapping parsers via syntactic projection across parallel texts*. Nat. Lang. Eng., 11(3):311–325, September.
14. Ion, R. 2007. *Metode de dezambiguizare automată. Aplicații pentru limbile engleză şi română.* Teză de doctorat. Academia Română. Bucureşti.
15. Cranias, L., H. Papageorgiou and S. Piperidis. 1994. *A Matching Technique in Example-Based Machine Translation.* In Proceedings of the 15th conference on Computational linguistics - Volume 1, Kyoto, Japan 100–104.
16. Yuret, D. 1998. *Discovery of linguistic relations using lexical atrraction.* PhD thesis, Department of Computer Science and Electrical Engineering, MIT
17. Steinberger R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D. 2006. *The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages.* In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, pp. 2142-2147
18. Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu, D. (2008). *RACAI's Linguistic Web Services*. In Proceedings of the 6th Language Resources and Evaluation Conference - LREC 2008, Marrakech, Morocco. ELRA - European Language Ressources Association. ISBN 2-9517408-4-0.
19. Erjavec, T. 2004. *MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora*. In: Proc. of the Fourth Intl. Conf. on Language Resources and Evaluation, LREC'04, ELRA, Paris, pp. 1535 – 1538
20. Ion, R. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis (in Romanian), Romanian Academy, Bucharest, 138 p.
21. Ceausu Al. 2006. *Maximum Entropy Tiered Tagging*. In Janneke Huitink & Sophia Katrenko (editors), Proceedings of the Eleventh ESSLLI Student Session, pp. 173-179
22. Tufiş, D. 1999. *Tiered Tagging and Combined Language Models Classifiers*. In Václav Matousek, Pavel Mautner, Jana Ocelíková, and Petr Sojka, editors, *Text, Speech and Dialogue (TSD 1999)*, Lecture Notes in Artificial Intelligence 1692, Springer Berlin / Heidelberg,. ISBN 978-3-540-66494-9, pp. 28-33.
23. Och, F. J., Ney H. 2000. *Improved Statistical Alignment Models*. In Proceedings of the 38th Conference of ACL, Hong Kong, pp. 440-447
24. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Wade, S., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. 2007. *MOSES: Open Source Toolkit for Statistical Machine Translation.* Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic.
25. Och, F. J. 2003. *Minimal Error Rate Training in Statistical Machine Translation*. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167.
26. Banerjee S., Lavie A., *An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments*, Proceedings Of The ACL Workshop On Intrinsic And Extrinsic Evaluation Measures For Machine Translation And/Or Summarization, Pages 65-72, Ann Arbor, June 2005.