Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages

Alexandru Ceauşu[1] and Dan Tufiş[2]

[1] Centre for Next Generation Localisation, Dublin City University
[2] Research Institute for Artificial Intelligence, Romanian Academy
aceausu@computing.dcu.ie; tufis@racai.ro

Abstract. The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. However, they show less improvement on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with limited word order and morphological variations to a highly inflected language. We describe an experiment that uses morpho-syntactic descriptions to translate and generate morphological information in factored machine translation. We show that from English to a morphologically rich language this setting has better performance than the baseline phrase-based system, when only a small parallel corpus is available. Also, we show that it scales well to a large parallel corpus when additional target monolingual corpus is available.

Keywords: statistical machine translation, morphologically-rich languages

1 Introduction

The phrase-based translation approach has overcome several drawbacks of the word-based translation methods and proved to significantly improve the quality of translated output. However, it shows less improvement on translating between languages with very different syntax and morphology, especially when the translation direction is from a language with limited word order and morphological variations to a highly inflected language. Tree-based models were introduced to handle long range reordering – what is believed to be the most difficult part to model in in statistical machine translation (SMT). The rich morphology of a highly inflected language permits a flexible word order, thus shifting the focus from long-range reordering to the selection of a morphological variant. Translating the correct surface form realization of a word is dependent not only on the source word-form, but it also depends on additional morpho-syntactic information.

Morphologically rich languages have a large number of surface forms in the lexicon to compensate for a flexible word order. The large number of word-forms can make very difficult to establish translation equivalents classes between the lexicons.

Both Transfer and Interlingua MT employ a generation step to produce the surface form, from a given context and a dictionary form of the word. In order to allow the same type of flexibility in using the morpho-syntactic information in translation, factored translation models [1] provide the possibility to integrate the linguistic information into the phrase-based translation model.

Most of the SMT approaches that have as target a morphologically rich language employ factored translation models. Our approach is similar to several other factored machine translation experiments such as adding the morphological features as factors [2], adding supertags on source language [3], and mapping syntax to morphology [4]. Our results are comparable with the ones reported in papers describing Arabic-English SMT experiments. For large amounts of training data, applying only a minimal segmentation in the Arabic part of the corpus yields better results than the baseline; however, when only a limited amount of training data is available, better results are achieved with part-of-speech tags and complex morphological analysis [5]. The importance of the generation model is highlighted in [6] through its usage in a hybrid (rule based and SMT) Arabic-English system.

2 Morpho-syntactic Description Codes

In highly inflectional languages, encoding the morpho-syntactic properties of the word-forms requires a large set of description codes. The Multext European project in co-operation with EAGLES Lexical Specification Group developed a set of recommendations [7] for the languages in Western Europe. Starting with these specifications, the Multext-East Copernicus project further developed them so as to account for the specificity of six other languages from Central and Eastern Europe – Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene [8]. The size of the tag-set greatly differs among languages: from a tag-set of around 100 tags in English to more than 2000 tags in Slovene.

Data sparseness in tagging highly inflectional languages with large tag-sets and scarce training resources is a problem that cannot be addressed using only common tagging techniques. Tiered tagging [9] is a two-stage technique addressing the issue of data-sparseness. It uses an intermediary tag-set of a smaller size on the basis of which a common POS tagging technique can be used. In a second phase, it replaces the tags from the small tag-set with tags from the fully-specified morpho-syntactic tag-set (MSD tag-set) also taking into consideration the context. Tiered tagging relies on the assumption that the values of a part of the attributes in a MSD tag (the determinant attributes) and the word-form are sufficient to infer the rest of the attribute values. Based on this assumption, in the second phase of tiered tagging, the original MSD tag-set is recovered using a lexicon and a set of hand-written rules. The original idea of tiered tagging has been extended in [10], so that the second phase is replaced with a maximum entropy-based MSD recovery. In this approach, the rules for POS to MSD conversion are automatically learnt from the corpus. Therefore, even the POS labels assigned to unknown words can be converted into MSD tags. If an MSD-lexicon is available, replacing the POS label for the known words by the appropriate MSD tags is almost 100% accurate.

3 Data Preparation

3.1 SEE-ERA.net Corpus

This corpus has been compiled within a SEE-ERA.net project [11] and it is based on the much larger JRC-Acquis multilingual corpus [12]. 1200 documents of high quality sentence alignment were extracted from the JRC-Acquis corpus. The documents have translations in all the languages of the project (Bulgarian, Greek, English, Slovene and Romanian) plus Czech, English, French, and German. The SEE-ERA.net corpus has morpho-syntactic description codes for Bulgarian, Czech, Greek, English, Slovene and Romanian. The aligned documents have 60,389 translation units of approximately 1.4 million tokens per language. For the experiments in 4.1 we used the English-Bulgarian, English-Greek, English-Slovene and English-Romanian parts of the corpus. A sample of an aligned translation unit, with English, Romanian and Slovenian parts, is presented in Example (1).

```
<tu id="60389"><seg lang="en"><s id="32005L0004.n.26.1.en"><w
lemma="do" ana="Vmps">Done</w><w lemma="at" ana="Sp">at</w><w
lemma="Brussels" ana="Np">Brussels</w><c>,</c><w lemma="19"
ana="Mc">19</w><w lemma="January" ana="Ncns">January</w><w
lemma="2005" ana="Mc">2005</w><c>.</c></s></seg></tu>

<tu id="60389"><seg lang="ro"><s id="32005L0004.n.26.1.ro"><w
lemma="adopta" ana="Vmp--sf">Adoptat•</w><w lemma="la"
ana="Spsa">la</w><w lemma="Bruxelles"
ana="Np">Bruxelles</w><c>,</c><w lemma="19" ana="Mc">19</w><w
lemma="ianuarie" ana="Ncms-n">ianuarie</w><w lemma="2005"
ana="Mc">2005</w><c>.</c></s></seg></tu>

<tu id="60389"><seg lang="sl"><s id="32005L0004.n.25.1.sl"><w
lemma="v" ana="Sl">V</w><w lemma="Bruselj"
ana="Npmsl">Bruslju</w><c>,</c><w lemma="19." ana="Mdo">19.</w><w
lemma="januar" ana="Ncmsg">januarja</w><w lemma="2005"
ana="Mdm">2005</w></s></seg></tu>
```

(1)

3.2 STAR Corpus

The STAR bilingual parallel corpus (Romanian-English) was developed during the research project STAR[1]. The parallel part of the corpus mainly contains juridical

[1] http://www.racai.ro/star/

documents, but it also includes journalistic type data. STAR also has a Romanian balanced monolingual corpus containing a large range of documents, from literary texts to news and scientific documents. The content of the STAR corpus is sourced from several other corpora:

- the DGT (Directorate-General for Translation) Translation Memory corpus, a juridical corpus based on the Acquis Communautaire [12];
- EMEA (European Medicines Agency), a corpus with medical content from the Opus Corpus [13];
- SE Times (Southeast European Times corpus), a journalistic corpus from the Opus Corpus [13];
- NAACL news, the English-Romanian journalistic corpus used for NAACL 2005 word alignment shared task [14];
- Romanian balanced monolingual corpus (20 million tokens).

The data cleaning stage for this corpus includes understated processing steps like: deleting the data that is duplicated, removing lines of text that are in other languages, removing lines or tokens of more than a specified character length, etc. Cleaning Romanian data collected from the web (NAACL, SE Times) was a real challenge. Besides spelling errors, there are three specific types of text distortions occurring in Romanian texts: (i) missing diacritical characters, (ii) different encoding codes for the same diacritical characters and (iii) different orthographic systems. When ignored, they have a negative impact on the quality of translation and language models and thus, on the translation results. For details on the process of diacritics restoration see [15].

Table 1. The contribution of each sub-corpus to the STAR parallel corpus

Corpus	Tokens (millions)		Sentence pairs (thousands)
	English	Romanian	
DGT Translation Memory	12.5	12	621
EMEA (Opus Corpus)	10	11	698
SE Times (Opus Corpus)	4.4	4.7	166
NAACL news	0.8	0.7	39
Raw total	27.7	28,4	1,525
Cleaned total	27.3	27,7	1,495

After Romanian text normalization, in order to create the EN-RO bitext, the processing stages of sentence splitting and tokenization had to be adapted to the respective languages. Sentence splitting and tokenization have shared resources like abbreviations, segmentation rules, token merging rules, etc. In the final stage of data preparation – the bitext cleaning – we removed the sentence pairs that are too short or too long and the sentence pairs of a source to target ratio of more than 1/9. Table 1 shows the contribution of each sub-corpus to the STAR parallel corpus and the amount of the remaining data after cleaning.

The corpus was tokenised, POS-tagged and received morpho-syntactic annotation using our publicly available web services [16]. Each token is composed of four factors: (i) the word-form, (ii) lemma disambiguated with its lexical category, (iii) the POS-tag from the reduced tag-set, and (iv) the MSD. Table 2 provides an example for the annotations available in the STAR corpus.

Table 2. Example of annotated sentence pair

English	Romanian
Grounds / ground^Nc / NNS / Ncnp	Motive / motiv^Nc / NPN / Ncfp-n
of / of^Sp / PREP / Sp non-recognition / recognition^Nc / NN / Ncns	de / de^Sp / S / Spsa refuz / refuz^Nc / NSN / Ncms-n al / al^Ts / TS / Tsms recunoaşterii / recunoaştere^Nc / NSOY / Ncfsoy
for / for^Sp / PREP / Sp judgments / judgment^Nc / NNS / Ncnp	hotărârilor_judecătoreşti / hotărâre_judecătorească^Nc / NSRN / Ncfsrn
relating / relate^Vm / PPRE / Vmpp to / to^Sp / PREP/ Sp	în / în^Sp / S / Spsa materia / materie^Nc / NSRY / Ncfsry
parental_responsibility / parental_responsibility^Nc / NN / Ncns	răspunderii_părinteşti / răspundere_părintească^Nc / NSOY / Ncfsoy

4 Factored Translation with Morpho-Syntactic Description Codes

Factored translation models extend the phrase-based translation by taking into account, not only the surface form of the phrase, but also, additional information like the dictionary form (lemma), the part-of-speech tag or the morphological specification. It also provides, on the target side, the possibility to add a generation step. All these new features accommodates well in the log-linear model employed by the decoder:

$$P(e|f) = exp \sum_{i=1}^{n} \lambda_i h_i(e,f) \tag{1}$$

where $h_i(e,f)$ is a function associated with the pair e, f and λ_i is the weight of the function.

Factored translation offers great possibilities on modeling translation: (i) there can be several translation steps; (ii) the fluency of the output can be checked on different levels with several language models; (iii) long-range word reordering can be achieved with more than one reordering model; and (iv) on the target side, there can be different generation steps.

To improve the translation into morphologically-rich languages, the multitude of options provided by the factored translation can help validate the following assumptions:

- Aligning and translating *lemma* could significantly reduce the number of translation equivalency classes, especially for languages with rich morphology;
- *Part of speech affinities*. In general, the translated words tend to keep their part of speech and when this is not the case, the part-of-speech chosen is not random;
- The *re-ordering* of the target sentence words can be improved if language models over POS or MSD tags are used.

4.1 Multilingual Setting

Based on the SEE-ERA.net corpus, we tested, using the MOSES factored framework [17], several configurations of translation, generation and reordering steps. The language pairs tested were English-Greek, English-Bulgarian, English-Slovene and English-Romanian. After cleaning, we split the corpus into training, development and test sets resulting in almost 57,000 sentence pairs for training, 500 for the development test and 1000 for testing. The 4-gram word-form language models and the 5-gram POS or MSD language models were built only using the training data sets. Considering current practices for training SMT systems, our training corpus is very small, but, as we will show, the additional linguistic information, made available in the pre-processing phase, compensates for the scarcity in raw data.

In order to test the improvement of the factored model over the phrase-based approach, we built strong baseline systems for each language pair. The baseline systems were trained using word alignment on lemmas and they employ an additional lexicalised reordering model. The default distance reordering model operates in a window of tokens and provides a reordering cost given the difference between source and target positions. We choose to use a better reordering model for the baseline system. The lexicalised reordering model has a probability assigned for the position change (monotone, swap or discontinuous) of a target phrase given the source phrase.

The baseline phrase-based translation systems and the different factored configurations had their parameters tuned on the development set using MERT [18].

We found that translating lemmas and morpho-syntactic descriptors and then generating, accordingly, the word-forms achieved better results than the baseline phrase-based translation model. Table 3 presents BLEU scores [19] for some of the factored configurations tested for the English-Romanian part of the SEE-ERA.net corpus. The BLEU scores for the English-Romanian translation direction are consistent with the scores for the other language pairs in the SEE-ERA.net corpus, although some of the configurations could not be tested because the intermediary tag-set (POS tag-set) was not available for the Bulgarian, Greek and Slovene parts.

The first row in Table 3 is the baseline system. It has a translation model trained on word-forms (column 2), no generation model (column 3), a word-form language model (column 4) and a lexicalised reordering model trained on word-forms (column 5).

While the use of linguistically informed language models (wordform +MSD/POS) and translation and generations models (lemma+MSD/POS) ensured improvements over the baseline, we noticed a significant drop in performance when the system used the word-form or MSD reordering models instead of the distance model. One possible explanation for the drop in performance for configurations 5 and 6 is that the lexicalized reordering model is made redundant when using a language model over MSD (or POS tags).

Table 3. Different factored configurations and their BLEU scores for the English-Romanian part of the SEE-ERA.net corpus

Config	Translation model	Generation model	Language model	Reordering model	BLEU score
1	word-form	-	word-form	word-form	51.76
2	lemma	lemma -> word-form	word-form	distance	51.79
3	lemma POS	lemma -> POS lemma,POS -> word-form	POS word-form	distance	52.31
4	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	distance	52.76
5	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	word-form	46.39
6	lemma MSD	lemma -> MSD lemma,MSD -> word-form	MSD word-form	MSD	45.77

As a side-note, on this particular corpus, the high BLEU scores might be explained as a consequence of the nature of the corpus – juridical texts have limited vocabulary, with long sequences of words repeated across the entire corpus. Although the scores are higher than the ones reported on news test corpora, the differences in absolute BLEU points can still offer good indices on the performance of the different factored configurations.

One particular configuration of factored translation (configuration 4 in Table 3) has provided better results than others. The proposed configuration (see Fig. 1) can be summarized as: (i) translate lemmas, (ii) generate all possible word forms and associated morpho-syntactic descriptions corresponding to a given lemma, (iii) translate the associated morpho-syntactic descriptions and (iv) generate the target surface forms given the lemma and the morpho-syntactic description. In this configuration, the decoder uses two language models: one for the word-forms and another one for the morpho-syntactic descriptions.

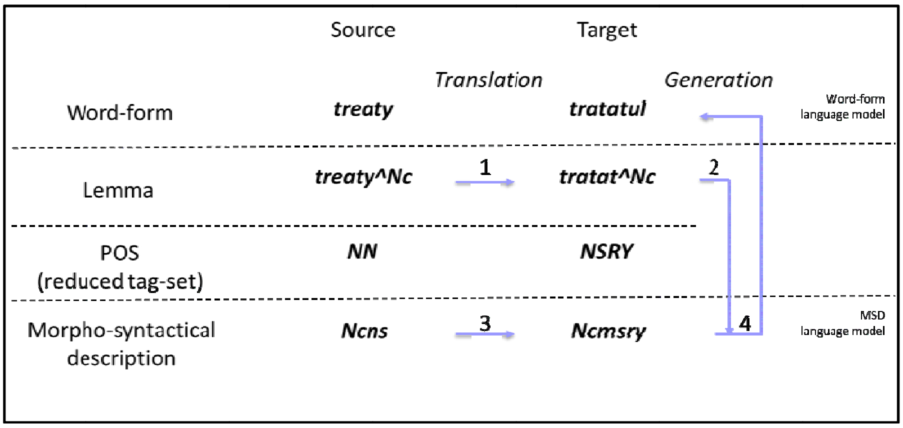

Fig. 1. Factored translation configuration with generation steps using lemma and MSD translation steps

We tested the systems using the BLEU score and we observed (see Table 4) improvements in accuracy ranging between 1% for Romanian and 2% for Slovene. Better handling of long-distance dependencies based on the MSD language model, a robust lemma translation equivalents table and a more precise selection of morphological variants are all possible explanations for the improvement in translation accuracy.

Table 4. BLEU scores for English-Bulgarian, English-Greek, English-Romanian and English-Slovene parts of the SEE-ERA.net corpus

Direction	Baseline	Factored
English-Bulgarian	38.94	39.60
English-Greek	42.22	43.07
English-Romanian	51.76	52.76
English-Slovene	40.73	42.68

The difference in BLEU scores between English-Romanian systems and the other systems are inherent to better lexical resources used for the tokenization and tagging of the English and Romanian texts. The idiomatic expressions and terminology tokenization were correlated between the English and the Romanian parts of the corpus.

4.2 English-Romanian Factored Translation

Using the STAR corpus (1.5 million sentence pairs) we tested if the factored configuration maintains its improvement over the baseline when a larger amount of training data is available. Similar to the experiments with the SEE-ERA.net corpus, we built a strong baseline system (lemma word alignment and lexicalised reordering model) that scored 53.82 BLEU points on 1000 sentences test-set (see Table 5).

For the bigger corpus, the English to Romanian factorized system achieves a BLEU score of 53.41, showing no improvement over the baseline system, on par with the results reported on other experiments with factored translation models [2], [20].

Table 5. English-Romanian factored translation on the STAR corpus

MT System	*BLEU score*
Phrase-based	53.82
Factored only parallel data	53.41
Factored plus monolingual data	54.52

Leveraging the fact that the generation step only deals with the target language, we used the STAR monolingual corpus in addition to the Romanian part of the parallel data to build the generation models and to train the MSD language model. The size of the statistical generation table trained on the 1.5 million sentences of the parallel data has almost 270,000 entries of the format shown in Example 2 (the format of an entry in the generation table is: <a b p(a|b) p(b|a)>):

```
complot^Nc/Ncms-n    complot   1.0000000   1.0000000
prăjit^Af/Afpfsrn    prăjită   1.0000000   0.9583333
absurd^Af/Afpms-n    absurd    1.0000000   0.3076923       (2)
încărca^Vm/Vmip1s    încarc    1.0000000   1.0000000
punctare^Nc/Ncfsoy   punctării 1.0000000   1.0000000
```

The size of the generation table that was trained on the additional Romanian monolingual data has almost 620,000 entries. Not all the additional entries are valid Romanian words. The monolingual corpus was built from texts of different domains (newspaper articles, old Romanian literature, contemporary literature, scientific articles, etc.) with different types of diacritical representations. We estimate that cleaning up the generation table would further improve the BLEU score. In the new configuration, the larger training data for the generation model and the MSD language model produced an increase of 0.7 absolute BLEU points over the baseline system (see Table 5).

4.3 Analysis of the Results

Although the factorized model (with additional monolingual data) has a marginal increase in BLEU score (at the cost of lower translation speed), we estimate that the actual improvements are higher from a human evaluation point of view. We observed that the factorized model frequently produces translations of better word order and more accurate morphological variant selection over the baseline model.

In order to assess in how many cases the translation system chooses the correct morphological variant, we investigated a difficult case of morphological attributes translation: the agreement of the words in a noun-phrases that include a conjunction. The baseline system, in this particular case of the test set, has a correct agreement in 61 of the 81 (75%) noun phrases that include a conjunction. The factored system with a generation model trained on more monolingual data has the correct agreement in 75 of the cases (92%).

In Table 6 we present a case of noun phrase agreement in which the baseline system misses the correct morphological variant. The noun *prelucrarea* (processing) is (wrongly) a definite form while the noun *export* (export) is (correctly) an indefinite form.

Table 6. Example of noun phrase agreement in the English-Romanian phrase-based and factored translation systems

Token	*English*	*Romanian phrase-based*	*Romanian factored*
1	representative ↘	piețe	piețe
2	markets ↗	reprezentative	reprezentative
3	for →	pentru	pentru
4	processing →	prelucrarea (def)	prelucrare (indef)
5	and →	şi	şi
6	export →	export (indef)	export (indef)

5 Conclusion and Further Research

The paper presented two scenarios in which factored machine translation for morphologically-rich languages can show improvements in performance over the baseline phrase-based translation: (i) when there is very little amount of parallel data available and (ii) for a larger parallel corpus, when an additional, target-side, monolingual corpus with automatic annotations is available. The experiments described in this paper showed that an additional generation step on the target-side can prove as useful for statistical machine translation as it is for rule-based approaches to MT.

One major research priority in SMT is to overcome the scarcity of parallel data for less-resource language pairs. As future research, we are considering extending the factored experiment with comparable parallel data. The comparable data is available through the ACCURAT project. The aim of the ACCURAT project is to research methods and techniques to overcome one of the central problems of machine translation (MT) – the lack of linguistic resources for under-resourced areas of machine translation. The main goal is to find, analyze and evaluate novel methods that exploit comparable corpora in order to compensate for the shortage of linguistic resources, and ultimately to significantly improve MT quality for less-resourced languages and narrow domains.

Acknowledgments. This work has been supported by the ACCURAT project (http://www.accurat-project.eu/) funded by the European Community's Seventh Framework Program (FP7/2007-2013) under the Grant Agreement n° 248347. It has also been partially supported by the Romanian Ministry of Education and Research through the STAR project (no. 742/19.01.2009).

References

1. Koehn, P., Hoang, H.: Factored Translation Models. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 868–876, Prague, (2007)
2. Avramidis E., Koehn, P.: Enriching morphologically poor languages for statistical machine translation. In: Proceedings of ACL-08/HLT, pp. 763–770, Columbus, Ohio (2008)
3. Haque, R., Naskar, S.K., Ma Y., Way, A.: Using Supertags as Source Language Context in SMT. In: Proceedings of the 13th Annual Meeting of the European Association for Machine Translation, pp. 234-241, Barcelona (2009)
4. Yeniterzi, R., Oflazer, K.: Syntax-to-Morphology Mapping in Factored Phrase-Based Statistical Machine Translation from English to Turkish. In: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pp. 454–464, Uppsala (2010)
5. Habash, N., Sadat, F.: Arabic Preprocessing Schemes for Statistical Machine Translation. In *Proc. of NAACL*, New York. (2006)
6. Habash, N., Dorr, B., Monz, C.: Challenges in Building an Arabic-English GHMT System with SMT Components. In *Proceedings of AMTA'06*, Cambridge, MA, USA (2006)
7. Monachini, M., Calzolari, N. (Eds.): EAGLES Synopsis and Comparison of Morphosyntactic Phenomena Encoded in Lexicons and Corpora A Common Proposal and Applications to European Languages http://www.ilc.pi.cnr.it/ EAGLES96/morphsyn/ morphsyn.html (1996)
8. Erjavec, T., Monachini, M. (Eds.): Specifications and Notation for Lexicon Encoding. Deliverable D1.1 F. Multext-East Project COP-106. http://nl.ijs.si/ME/CD/docs/ mte-d11f/ (1997)
9. Tufiş, D.: Tiered Tagging and Combined Classifiers. In: F. Jelinek, E. Nth (eds) *Text, Speech and Dialogue* LNCS vol. 1692, pp. 28-33 Springer-Verlag Berlin Heidelberg, (1999)
10. Alexandru Ceauşu: Maximum Entropy Tiered Tagging, Janneke Huitink & Sophia Katrenko (eds), *Proceedings of the Eleventh ESSLLI Student Session*, ESSLLI 2006, pp. 173-179 (2006)
11. Dan Tufiş, Svetla Koeva, Tomaž Erjavec, Maria Gavrilidou, and Cvetana Krstev: Building Language Resources and Translation Models for Machine Translation focused on South Slavic and Balkan Languages. In Marko Tadić, Mila Dimitrova-Vulchanova and Svetla Koeva (eds.) *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages* (FASSBL 2008), pp. 145-152, Dubrovnik, Croatia, September 25-28 (2008)
12. Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D.: The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: Proceedings of the 5th LREC Conference, Genoa, pp.2142-2147 (2006)
13. Tiedemann, J.: News from OPUS - A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In: N. Nicolov and K. Bontcheva and G. Angelova and R. Mitkov (eds.) Recent Advances in Natural Language Processing (vol V), pp. 237-248, John Benjamins, Amsterdam/Philadelphia (2009)
14. Martin, J., Mihalcea, R., Pedersen T. (eds.): Proceedings of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond", Ann Arbor, Michigan*,* Association for Computational Linguistics (2005)

15. Tufiş, D., Ceauşu, A.: DIAC+: A Professional Diacritics Recovering System, in *Proceedings of LREC 2008* (Language Resources and Evaluation Conference), May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association (2008)
16. Tufiş, D., Ion, R., Ceauşu, A., Ştefãnescu, D.: RACAI's Linguistic Web Services, in *Proceedings of LREC 2008* (Language Resources and Evaluation Conference), May 26 - June 1, Marakkech, Morocco. ELRA - European Language Resources Association (2008)
17. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, demonstration session, Prague, (2007)
18. Och, F. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of ACL*, pp 160-16, Association for Computational Linguistics (2003).
19. Papineni, K., Roukos, S., Ward, T., Zhu W.J.: BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 20th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 311-318 (2002)
20. Birch, A., Osborne, M., Koehn, P.: CCG Supertags in Factored Statistical Machine Translation. In *Proceedings of the Second Workshop on Statistical Machine Translation, Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Prague (2007)