Improving Machine Translation Performance Using Comparable Corpora

Andreas Eisele, Jia Xu

{Andreas.Eisele,Jia.Xu}@dfki.de

DFKI GmbH, Language Technology Lab
Stuhlsatzenhausweg 3
D-66123 Saarbrücken Germany

Abstract

The overwhelming majority of the languages in the world are spoken by less than 50 million native speakers, and automatic translation of many of these languages is less investigated due to the lack of linguistic resources such as parallel corpora. In the ACCURAT project we will work on novel methods how comparable corpora can compensate for this shortage and improve machine translation systems of under-resourced languages. Translation systems on eighteen European language pairs will be investigated and methodologies in corpus linguistics will be greatly advanced. We will explore the use of preliminary SMT models to identify the parallel parts within comparable corpora, which will allow us to derive better SMT models via a bootstrapping loop.

1. Introduction

State-of-the-art machine translation based on the statistical approach is a data-driven process. The quality and quantity of the training data is crucial for the performance of a translation system. However, the increasing amount of training corpora can still not meet the demand of automatic translation on different language pairs and in various domains. Rich data are mostly available for few languages and only certain domains. There are still a great number of under-resourced languages. Thousands of languages are spoken by less than 50 million native speakers, with a big group of more than 200 languages that have between 1 and 50 million native speakers. Most of these languages are lacking sufficient linguistic resources. This brings difficulties to improve the translation qualities on these languages.

For instance, the majority of the European languages are under-resourced and lack both parallel corpora and language technologies for MT. The project ACCURAT (Analysis and Evaluation of Comparable Corpora for Under-Resourced Areas of Machine Translation) will focus on developing and evaluating language pairs of English-Latvian, English-Lithuanian, English-Estonian, English-Greek, English-Croatian, Croatian-English, English-Romanian, English-Slovenian, Slovenian-English, English-German, German-English, German-Romanian, Romanian-German, Greek-Romanian, Lithuanian-Romanian, Romanian-Greek, Romanian-English and Latvian-Lithuanian. We also work on the language pair of German and English which is well investigated previously. This can help us find the impact of comparable corpora on translations between language pairs with both rich and poor resources. More details can be found in (Skadina et al., 2010). The participants include organizations of Tilde, USFD, CTS, LISP, FFZG, DFKI, RACAI, Linguatec and Zemanta.

The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality for under-resourced languages and narrow domains. The work will be carried out on the listed European language pairs and adapted to narrow domains, e.g. automotive engineering. We expect an enhancement of language and domain coverage in MT.

The ACCURAT project will provide novel methodologies and models that exploit comparable corpora to enhance the translation quality of current MT systems, which are universal and can be used to new language pairs and domains. We will define criteria to measure the comparability of texts in comparable corpora. Methods for automatic acquisition of a comparable corpus from the Web will be analyzed and evaluated. Advanced techniques of obtaining parallel sentences and phrases from comparable corpora will be applied and extended to provide training and customization data for MT. Domain dependent MT will be exploited by automatic clustering of training data into genres according to their contents. Given limited amounts of available in-domain data, we will also perform the adaptation of domain specific translation systems to enhance the system performance in specific domains. Improvements from applying acquired data will be measured against baseline results from MT systems and validated in practical applications. As a summary, the most important results of ACCURAT will be

- Criteria and metrics of comparability
- Tools for building comparable corpora
- Tools for multi-level alignment and information ex-

Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC 2010, pages 35–41
Malta, 22 May 2010

traction from comparable corpora

- Multilingual comparable corpora for under-resourced languages and narrow domains
- Improved baseline translation systems for under-resourced European language pairs using data extracted from comparable corpora
- Report on requirements, implementation and evaluation of usability in applications for specialists in narrow domain and specific languages

2. State of the Art

Machine translation, in particular the statistical approach to it, has undergone significant improvements in recent years. However SMT research has been mainly focused on widely used languages, such as English, French, Arabic, Chinese, Spanish, and German. Languages with less native speakers such as Romanian are not as well developed due to the lack of linguistic resources. This results in a technical gap between the translation on widely spoken languages and on other languages.

Building statistical machine translation system requires a great amount of parallel corpora for model training. Good results can be easily achieved when the domain of the training corpus is closer to that of the test data. Rule-based machine translation can also profit from the data-driven technique: a MT system can have better translation quality, when bilingual lexical data has been extracted from parallel resources and imported into an RBMT system dictionary (Eisele et al., 2008). Nowadays parallel corpora are still limited in quantity, genre and language coverage.

There have been many investigations to exploit comparable corpora. Whereas early work on alignment such as the sentence aligners described in (Gale and Church, 1993) and (Brown et al., 1991) assumed parallel corpora, models that incorporated lexical information to increase performance on noisy data were investigated early after, e.g. in (Chen, 1993; Fung and McKeown, 1994; Jones and Somers, 1995; Fung, 1995; Rapp, 1995). In (Zhao and Vogel, 2002), sentence length models and lexicon-based models are combined under a maximum likelihood criterion. Specific models are proposed to handle insertions and deletions that are frequent in bilingual data collected from the web. Using the mined data, word-to-word alignment accuracy machine translation modeling is improved as shown in the experiments. In (Utiyama and Isahara, 2003), language information retrieval and dynamic programming methods are applied to align the Japanese and English articles and sentences. In (Munteanu and Marcu, 2005) the parallel sentences are discovered using a maximum entropy classifier, where similar sentence pairs are analyzed using a signal processing-inspired approach. The extracted data have been shown to improve the performance of a state-of-the-art translation system. In (Shi et al., 2006), a new web mining scheme for parallel data acquisition is presented based on the document object model. A comparison of different alignment methods and more approaches considering non-monotone sentence alignments are described in (Khadivi, 2008) and (Xu et al., 2006).

One very promising approach for the iterative bootstrapping of improved translation models from comparable corpora is given in (Rauf and Schwenk, 2009) for the case of English and French. We will apply these methods for all the 18 language pairs investigated in the project and report on the question how well the methods generalize to language pairs from different families.

Also, a number of techniques have been developed for automatically assembling domain specific corpora from the web, e.g. BootCaT in (Baroni and Bernardini, 2004), Corpógrafo in (Maia and Matos, 2008). However, state-of-the-art fully automatic extraction results in noisy output and requires human processing. To select similar documents from comparable corpora, CLIR techniques are applied in selection process for widely used languages, e.g. (Quirk et al., 2007) and (Munteanu and Marcu, 2005).

Furthermore, several phrasal alignment methods have been researched for parallel corpora: IBM Models 1-6 (Brown et al., 1993); applying lexico-syntactic categories for word tagging and the identification of semantically equivalent expressions (Aswani and Gaizauskas, 2005); Phrase-based joint probability model (Marcu and Wong, 2002); factored phrase-based alignments (Koehn and Hoang, 2007).

There are only a few parallel corpora publicly available for the languages we work on. The JRC-Acquis is a huge collection of European Union legislative documents translated into more than twenty official European languages (Steinberger et al., 2006) including under-resourced languages such as Latvian, Lithuanian, Estonian, Greek, Croatian and Romanian. The European Parliament Proceedings Parallel Corpus (Europarl corpus) was extracted from the proceedings of the European Parliament (1996-today) and has included versions in 11 European languages: French, Italian, Spanish, Portuguese, English, Dutch, German, Danish, Swedish, Greek and Finnish (Koehn, 2005). The Europarl corpus was aligned at the sentence level using a tool based on the Church and Gale algorithm (Gale and Church, 1991). Other available multilingual parallel corpus are developed in the framework of projects of Multilingual Corpora for Cooperation (MLCC), the Integrated European language data Repository Area (INTERA2) eContent, SEEERAnet and so on. Very interesting corpora are contained in the OPUS collection described in (Tiedemann, 2009).

3. Domain Adaptation

Here we will focus on methods of sentence, paragraph and phrasal alignment and domain adaptation. The discussion on comparability metrics and building comparable corpora is described in (Skadina et al., 2010).

To select similar documents from a comparable or parallel corpus and to find multilingual comparable corpora for certain domains, the cross language information retrieval (CLIR) techniques will be proposed. Bootstrapped bilingual lexical resources will be explored for document selection.

Given a comparable corpus consisting of documents in two languages, L1 and L2, the first step is to find similar documents in L1 and L2. Typical approaches involve treating a document in the L1 collection as a query and then using CLIR techniques to retrieve the top n documents from the L2 collection as described in (Munteanu et al., 2004) and (Quirk et al., 2007). This approach requires some sort of bilingual dictionary in query translation.

After similar documents are selected, similar text fragments need to be identified. These fragments may be sentences or possibly only phrases. Recent research results have shown that in most cases methods designed for parallel texts perform poorly for comparable corpora. For example, most standard sentence aligners exploit the monotonic increase of the sentence positions in a parallel corpus, which is not observed in comparable corpora. ACCURAT will investigate how successful the sentence aligner developed at the Romanian Academy (Tufiş et al., 2006) is in aligning similar sentences in comparable corpora. This sentence aligner, based on SVM technology, builds feature structures characterizing a pair of sentences considered for alignment, including number of translation equivalents, ratio between their lengths, number of non-lexical tokens, such as dates, numbers, abbreviations, etc., and word frequency correlations. These feature structures are afterwards classified to describe how well sentence alignments corresponds to experimentally determined thresholds. This aligner has been evaluated and has an excellent F measure score on parallel corpora, being able to align N-M sentences. It is much better than Vanilla aligner and slightly better than HunAlign. A state-of the-art sentence aligner is described in (Moore, 2002), but this aligner produces only 1-1 alignments loosing N-M alignments. As comparable corpora do not exhibit the monotonic increase of aligned sentence positions, we anticipate that many of the alignments will be of the type 0-M, N-0 and N-M sentences, thus this alignment ability is a must. The SVM approach to sentence alignment has the advantage that it is fully trainable. Another promising method to identify similar sentence pairs within comparable corpora, proposed by (Munteanu et al., 2004), will be also investigated. To select candidate sentences for alignment they propose a word-overlap filter together with a constraint on the ratio of lengths of the two sentences. Given two sentences that meet these criteria, the final determination of whether they are or are not assumed to be parallel sentences is made by a maximum entropy classifier trained over a small parallel corpus, using such features as percentage of words with translations, length of sentences, longest connected and unconnected substrings. We will expand this method to paragraphs/sentences which are only to some extent translations of each other, thus adapting the proposed method to comparable corpora. A challenging research avenue for detecting meaning-equivalent sentence pairs within comparable corpora is using cross-lingual Q&A techniques. The main idea is to exploit dependency linking and the concepts of superlinks and chained links (Irimia, 2009) for determining the most relevant search criteria. The keywords extracted from the dependency linking of a source sentence/paragraph will be translated into a target language and available search engines will look for the most relevant candidate paragraphs/sentences. The possible pairs of translation equivalent textual units will be scored by a reified sentence aligner and will be accepted or rejected based on previously determined thresholds.

4. Sentence, Paragraph and Phrasal Alignment

We will research on multi-level alignment and information extraction methods from comparable corpora, specially building parallel sentence aligned corpora for SMT. We expect to develop pre-processing tools, a search module for detecting similar sentences/paragraphs in given collections of documents, the proper alignment tools for paragraph, sentence and phrase as well as a user-friendly alignment editor allowing the users to view and correct the wrong alignments. By promoting web service architecture, it will integrate the existing tools , especially for the required preprocessing steps such as language identification, tokenization, tagging, lemmatization, chunking etc., and it will allow for easy integrating of new tools and new languages.

Language independent methods in the spirit of those proposed in (Munteanu and Marcu, 2005) will be further investigated and elaborated for English-Latvian, English-Lithuanian, English-Estonian, English-Greek, English-Croatian, English-Romanian, English-Slovenian, German-Romanian, Lithuanian-Romanian, Romanian-Greek and Latvian-Lithuanian, allowing sentence/paragraph alignment of comparable corpora. Such methods are knowledge-poor but there is no reason for not using current language technology to embed easy to access knowledge sources. Since all partners have tools for basic preprocessing of their languages, such as tokenizers, POS-taggers, lemmatizers, the linguistic information revealed by these tools will be relied on heavily in order to decrease the danger of data sparseness and to increase the reliability of the statistical judgments.

When sentence/paragraph level alignment is established, the next step is to compute phrasal alignment, which is a central issue to exploit comparable corpora in MT applications. ACCURAT will start with the evaluation of existing methods for phrasal alignment, such as IBM Models1-6 as described in (Brown et al., 1993) and (Och and Ney, 2003),

lexico-syntactic categories for word tagging and the identification of semantically equivalent expressions (Aswani and Gaizauskas, 2005) and reified word alignment in (Tufiş et al., 2006) and (Tufiş et al., 2008) as well as their combinations. Since in many cases under-resourced languages lack linguistic resources, we will research on possibilities to extract phrasal alignments directly from similar document pairs in comparable corpora, without the use of dictionaries or pre-processing of the training data. Phrase-based joint probability model (Marcu and Wong, 2002) will be extended with the aim to overcome the sparseness of linguistic resources for under-resourced languages. We will use log-likelihood ratio statistics to assess the reliability of alignment (Kumano et al., 2007) which allows phrasal alignments to be produced just for parallel parts of the comparable corpora. To prevent alignments being produced between unrelated phrases while searching for optimal alignments, log-likelihood ratio (LLR) statistics will be applied.

Another novel way information extraction techniques can assist in aligning comparable corpora is through the identification of cross-language mappings between relation-expressing contexts. (Hasegawa et al., 2004) propose a technique for unsupervised relation discovery in texts, whereby contexts surrounding pairs of NEs of given types are extracted and then clustered, the clusters correspond to particular relations. This technique achieves impressive results and could be used to align relation expressing contexts as follows: First, relation clusters could be established monolingually given NERC tools in each language; These clusters could then be aligned cross-lingually using aligned sentence pairs containing NE pairs present found in the clusters, the aligned sentences coming either from the small amount of parallel data or from high confidence alignments in the comparable corpus; Once relation clusters where aligned cross-lingually, then presence of a pair of NEs from an aligned relation cluster in an L1 and L2 sentence pair would constitute evidence that the sentences should be aligned. ACCURAT will also investigate potential of unsupervised discovery of relations in text using NERC tools for monolingual clustering and perform cross-lingual alignment to improve fragment alignment in comparable corpora. Orthographic and phonetic-based approaches will be explored to develop adaptive HMM and/or CRF-based techniques e.g. (Zhou et al., 2008) trained on name pairs gathered initially from parallel training data and then bootstrapped using lexicons derived in the project. New advances in adaptive, semi-supervised NE recognition e.g. (Nadeau, 2007) will be explored and applied for languages other than English. Existing named entity recognition and classification systems for Croatian, English, German, Greek and Romanian will be deployed. First NERC systems for the Baltic languages will be developed, too.

Q&A techniques will be further researched and elaborated to find most relevant candidate paragraphs/sentences in comparable corpora. Cross-lingual Q&A techniques are highly relevant for this task. Queries formulated in one language and translated in another language may be used for searching the comparable corpora to find the paragraphs or sentences which are most likely to contain similar information.

5. Comparable Corpora for Machine Translation

The impact of comparable corpora on MT quality will be measured for seventeen language pairs, and detailed studies involving human evaluation will be carried out for six language pairs. Existing baseline SMT systems based on the Moses decoder will be coupled with data extracted from comparable corpora. Comparative evaluation will be performed to measure improvements by applying data extracted from comparable corpora. Comparable corpora will be used to update the linguistic knowledge of RBMT systems by applying terminology and named entity extraction technology.

Comparable corpora in machine translation systems will be created with the goal to evaluate results of data extracted from the comparable corpora. MT systems will be created using existing SMT techniques (Moses decoder) and existing RBMT techniques (Linguatec RBMT engine). Innovation in MT techniques will be in (1) enabling the use of additional data extracted from comparable corpora and (2) adjusting MT systems to under-resourced languages or narrow domains. To evaluate the efficiency and usability of the approach proposed in ACCURAT for under-resourced areas of MT, we will integrate research results into SMT using existing SMT techniques. In Task 4.1 baseline SMT systems will be built using traditional SMT techniques. Translation models will be trained on parallel corpora e.g. Europarl Parallel Corpus and JRC-ACQUIS multilingual Parallel Corpus. Performance of baseline SMT systems will be evaluated using automatic metrics such as BLEU and NIST as well as human metrics including fluency and adequacy. After the baseline SMT systems are built they will be improved by the integration of additional data from the comparable corpora. Data from comparable corpora will be integrated into both the translation model and the language model. Finally, SMT systems will be adjusted for a narrow domain using factored and reified models and will include domain specific knowledge such as terminology, named entities, domain specific language models, etc. Several approaches for the integration of additional data from comparable corpora into SMT will be investigated and evaluated. One option for the integration is to add extracted phrases to the training data and to retrain SMT. Another option is to use factored translation models (Koehn and Hoang, 2007) and to add data from comparable corpora as an additional phrase table.

In the ACCURAT project comparable corpora will be used instead of parallel corpora to extract bilingual lexi-

cal data for feeding rule-based machine translation systems. Comparable corpora will be used to update the linguistic knowledge of RBMT systems by applying terminology and named entity extraction technology. This is a step towards automating the current work flow in MT lexicon for RBMT production. Once these data are imported into a RBMT system, the next problem to solve is when to activate this acquired information in a given text. Automatic topic extraction would help in determining the narrow domain to which a given text belongs (Thurmair, 2006). However, many terms stay ambiguous in the selected domain, as they often have a general meaning which is also used in this narrow domain, and additional data-driven criteria will be used to further select the right translations in the narrow domain. ACCURAT will make use of techniques developed for the enrichment of a RBMT system with new lexical entries acquired automatically from parallel corpora in a specific domain in the framework of an ongoing collaboration with the European Patent Office on hybrid MT. The solution in this case was to construct a hierarchy of lexicons of increasing specificity and to traverse these lexicons from specific to more general for each ambiguous term that arises. These techniques will be generalized in case we do not have a fine-grained mark-up of the document topics but need to infer the topic via automatic classification, and in cases where the alignments are less clean because they are built from comparable instead of parallel data.

6. Conclusions

Lack of sufficient linguistic resources for many languages and domains is one of the major obstacle in further advancement of automated translation currently. The main goal of the ACCURAT research is to find, analyze and evaluate novel methods how comparable corpora can compensate for this shortage of linguistic resources to improve MT quality significantly for under-resourced languages and narrow domains.

The ACCURAT project will provide researchers and developers with reimplemented baseline methods such as that in (Munteanu and Marcu, 2005) along with novel methodologies to exploit comparable corpora for machine translation. We will determine criteria to measure the comparability of texts in comparable corpora. Methods for automatic acquisition of a comparable corpus from the Web will be analyzed and evaluated. Advanced techniques will be elaborated to extract lexical, terminological and other linguistic data from comparable corpora to provide training and customization data for MT. Improvements from applying acquired data will be measured against baseline results from MT systems and validated in practical applications. ACCURAT will provide novel approaches to achieve high quality MT translation for a number of under-resourced EU languages and to adapt existing MT technologies to narrow domains, significantly increasing the language and domain coverage of MT.

7. Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no 248347. We thank our colleagues from the ACCURAT consortium for the inspiration for many of the proposed methods and for the permission to re-use parts of the project's work plan. We apologize for some overlap with the material presented in (Skadina et al., 2010).

8. References

N. Aswani and R. Gaizauskas. 2005. Aligning words in english-hindi parallel corpora. In *Proceedings of the ACL 2005 Workshop on Building and Using Parallel Texts: Data-driven Machine Translation and Beyond*, pages 115–118.

M. Baroni and S. Bernardini. 2004. BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of Language Resources and Evaluation Conference LREC*.

P. F. Brown, J. C. Lai, and R. L. Mercer. 1991. Aligning sentences in parallel corpora. In *Proc. of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 177–184, Berkeley, California, June.

P. F. Brown, S. A. Della Pietra, V. J. Della Pietr, and R. L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311, June.

S. F. Chen. 1993. Aligning sentences in bilingual corpora using lexical information. In *Proc. of the 31th Annual Meeting of the Association for Computational Linguistics*, pages 9–16, Columbus, Ohio, June.

A. Eisele, C. Federmann, H. Uszkoreit, H. Saint-Amand, M. Kay, M. Jellinghaus, S. Hunsicker, T. Herrmann, and Y. Chen. 2008. Hybrid machine translation architectures within and beyond the euromatrix project. In *Proceedings of EAMT*.

P. Fung and K. McKeown. 1994. Aligning noisy parallel corpora across language groups: Word pair feature matching by dynamic warping. In *First Conf. of the Association for Machine Translation in the Americas (AMTA 94)*, pages 81–88, Columbia, MD, October.

P. Fung. 1995. A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. pages 236–243.

W. Gale and K. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*.

W. A. Gale and K. W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–90.

T. Hasegawa, S. Sekine, and R. Grishman. 2004. Discovering relations among named entities from large corpora. In *Proceedings of the 42nd Meeting of the Association*

for Computational Linguistics (ACL'04), Main Volume, pages 415–422, Barcelona, Spain, July.

E. Irimia. 2009. *Methods for Analogy-based Machine Translation. Applications for Romanian and English*. Ph.D. thesis, March.

D. B. Jones and H. L. Somers. 1995. Automatically determining bilingual vocabulary from noisy bilingual corpora using variable bag estimation. In *Recent Advances in Natural Language Processing*, pages 81–86, September.

S. Khadivi. 2008. *Statistical Computer-Assisted Translation*. Ph.D. thesis, RWTH-Aachen University, Aachen, Germany, July.

P. Koehn and H. Hoang. 2007. Factored translation models. In *Proceedings of EMNLP*.

P. Koehn. 2005. Europarl: a parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*.

T. Kumano, H. Tanaka, and T. Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In *Proceedings of TMI*.

B. Maia and S. Matos. 2008. Corpógrafo v.4 – tools for researchers and teachers using comparable corpora. In *Proceedings of the Workshop on Comparable Corpora, LREC*.

D. Marcu and W. Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *Proc. Conf. on Empirical Methods for Natural Language Processing*, pages 133–139, Philadelphia, PA, July.

R. C. Moore. 2002. Fast and accurate sentence alignment of bilingual corpora. In *Proc. of the 5th Conf. of the Association for Machine Translation in the Americas*, pages 135–244, Tiburon, California, October.

D. S. Munteanu and D. Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504.

D. Munteanu, A. Fraser, and D. Marcu. 2004. Improved machine translation performance via parallel sentence extraction from comparable corpora. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, HLT / NAACL*.

D. Nadeau. 2007. *Semi-Supervised Named Entity Recognition: Learning to Recognize 100 Entity Types with Little Supervision*. Ph.D. thesis, DUniversity of Ottawa, Ottawa.

F. J. Och and H. Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, March.

C. Quirk, R. Udupa, and A. Menezes. 2007. Generative models of noisy translations with applications to parallel fragment extraction. In *Proceedings of MT Summit XI, European Association for Machine Translation*.

R. Rapp. 1995. Identifying word translations in non-parallel texts. In *Proc. of the 33rd Annual Conf. of the Association for Computational Linguistics*, pages 321–322.

S. A. Rauf and H. Schwenk. 2009. On the use of comparable corpora to improve SMT performance. In *EACL*, pages 16–23, April.

L. Shi, C. Niu, M. Zhou, and J. Gao. 2006. A dom tree alignment model for mining parallel data from the web. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 489–496, Morristown, NJ, USA. Association for Computational Linguistics.

I. Skadina, A. Vasiljevs, R. Skadins, R. Gaizauskas, D. Tufis, and T. Gornostay. 2010. Analysis and evaluation of comparable corpora for under resourced areas of machine translation. In *Proceedings of the International Conference on Language Resources and Evaluation: Workshop on Building and Using Comparable Corpora (This volume)*, May.

R. Steinberger, B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, and D. Varga. 2006. The jrcacquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*.

G. Thurmair. 2006. Using corpus information to improve mt quality. In *Proceedings of the Workshop LR4Trans-III, LREC*.

J. Tiedemann. 2009. News from OPUS - a collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing (vol V)*, pages 237–248, Amsterdam/Philadelphia. John Benjamins.

D. Tufiş, R. Ion, A. Ceauşu, and D. Ştefănescu. 2006. Improved lexical alignment by combining multiple reified alignments. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL2006)*, pages 153–160, April.

D. Tufiş, K. Koeva, E. Erjavec, M. Gavrilidou, and C. Krstev. 2008. 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.). In *Proceedings of the Sixth International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL 2008)*, pages 145–152, September.

M. Utiyama and H. Isahara. 2003. Reliable measures for aligning japanese-english news articles and sentences. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 72–79, Morristown, NJ, USA. Association for Computational Linguistics.

J. Xu, R. Zens, and H. Ney. 2006. Partitioning parallel

documents using binary segmentation. In *Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL): Proceedings of the Workshop on Statistical Machine Translation*, pages 78–85.

B. Zhao and S. Vogel. 2002. Adaptive parallel sentences mining from web bilingual news collection. In *Proceedings of the 2002 IEEE International Conference on Data Mining*, page 745, Washington, DC, USA. IEEE Computer Society.

Y. Zhou, F. Huang, and H. Chen. 2008. Combining probability models and web mining models: a framework for proper name transliteration. *Information Technology and Management*, 9(2):91–103.