Development and Application of a Cross-language Document Comparability Metric

Fangzhong Su, Bogdan Babych
Centre for Translation Studies, University of Leeds
LS2 9JT, Leeds, UK
E-mail: smlfs@leeds.ac.uk, b.babych@leeds.ac.uk

Abstract

In this paper we present a metric that measures comparability of documents across different languages. The metric is developed within the FP7 ICT ACCURAT project, as a tool for aligning comparable corpora on the document level; further these aligned comparable documents are used for phrase alignment and extraction of translation equivalents, with the aim to extend phrase tables of statistical MT systems without the need to use parallel texts. The metric uses several features, such as lexical information, document structure, keywords and named entities, which are combined in an ensemble manner. We present the results by measuring the reliability and effectiveness of the metric, and demonstrate its application and the impact for the task of parallel phrase extraction from comparable corpora.

Keywords: comparable corpora, comparability metric, parallel phrase extraction

1. Introduction

In recent years, the application of cross-lingual comparable corpora (i.e., corpora in different languages that do not contain translated documents, but are within the same subject domain and contain similar text types) has attracted considerable attention in the NLP community. This is primarily driven by the difficulties in collecting large-scale parallel corpora, e.g., limited availability of high-quality human translations between many language pairs (especially for some under-resourced languages and under-represented subject domains). MT developers can collect comparable corpora from the Web with less effort, in comparison to the effort required for the creation of parallel resources.

Most of the applications focus on discovering translation equivalents from comparable corpora to support machine translation. For example, comparable corpora have been successfully used for the tasks of bilingual lexicon extraction (Rapp, 1995, Rapp, 1999, Yu and Tsujii, 2009, Morin et al., 2007, Prochasson and Fung, 2011, Li and Gaussier, 2010), parallel phrase extraction (Munteanu and Marcu, 2006), and parallel sentence extraction (Fung and Cheung, 2004a, Fung and Cheung, 2004b, Munteanu et al., 2004, Munteanu and Marcu, 2005, Smith et al., 2010).

However, successful detection of translation equivalents from comparable corpora very much depends on the quality of these corpora, specifically – on the degree of their closeness and successful alignment on different levels of text units (e.g., corpus, document, and sub-document units). Therefore, the goal of this work is to provide a comparability metric which can reliably identify comparable documents from raw corpora collected by crawling the web, and characterise the degree of their similarity, which enriches comparable corpora with the document alignment information, filters out documents that are not useful and eventually leads to extraction of good-quality translation equivalents from the corpora.

To achieve this goal, we need a qualitative definition and quantitative measures of *comparability*, which is the key concept for this task, applicable on the level of corpora, documents and sub-document units. However, so far there is no widely accepted definition of comparability, even though this concept is frequently used informally, to characterise the overlap in the subject domain or genre of the compared documents. Fung and Chueng (2004b) define different levels of corpus comparability as below. A *noisy parallel corpus* contains many parallel sentences that are not strictly aligned. A *comparable corpus* contains topic aligned documents that are not translations of each other. A *quasi-comparable corpus* contains documents that can talk about the same topic or not.

Similar to Fung and Chueng (2004b), we also analyse the degree of comparability for comparable corpus. For the purposes of our study, we can directly characterise comparability by how useful comparable corpora are for the task of detecting translation equivalents in them, and ultimately – how much finally to machine translation. In this work, we focus on document-level comparability, and use three broad categories for qualitative definition of comparability levels:

- *Parallel* documents are traditional parallel texts that are translation of each other, or with minor language-specific variations.
- *Strongly-comparable* documents are texts that are from the same source or independently-written in different languages, but talk about the same event or subject (e.g., linked articles in Wikipedia about the same topic). These documents can be aligned on the document level on the basis of their origin.
- *Weakly-comparable* documents are texts in the same narrow domain which describe different events, e.g., customer reviews about hotels and restaurants in London. These texts do not have an independent alignment across languages.

In this paper, we present a comparability metric to automatically measure the comparability level of cross-lingual comparable documents. Several information are taken into account for the metric design, including lexical information, document structure, keywords and named entities. These features are then combined by a simple average weighted strategy in the metric. The experimental results show that the proposed metric can effectively predict the comparability degree of comparable documents. Moreover, we also investigate the applicability of the metric by measuring its impact to the task of parallel phrase extraction from comparable corpora. It turns out that, the metric can be applied to automatically select higher comparable documents from raw comparable corpus, leading to more number of parallel phrases extracted from the resulting good-quality comparable documents.

The remainder of this paper is organized as follows. Section 2 discusses previous work. Section 3 introduces our comparability metrics. Section 4 presents the experimental results and evaluation. Section 5 describes the application of the metrics, followed by conclusion and future work in Section 6.

2. Related Work

Although there has been a considerable amount of literature tackling the detection of translation equivalents (e.g., bilingual lexicon, parallel phrases and sentences) from comparable corpora, there are only few papers which analyse the characteristics of corpus comparability. In this section, we will introduce some representative work which are closer to ours.

Some studies (Sharoff, 2007; Maia, 2003; McEnery and Xiao, 2007) analyse comparability by assessing corpus composition, such as structural criteria (e.g., format and size), and linguistic criteria (e.g., topic, domain, and genre). Kilgarriff and Rose (1998) determine the similarity of monolingual corpora based on the Chi-square statistic on the top-n most frequent words extracted from the compared corpora. Resnik and Smith (2003) identify similar webpages by analysing their HTML document structure. Saralegi et al. (2008) measure the degree of comparability of comparable corpora (English and Basque) based on topic distribution and publication dates of documents.

Munteanu and Marcu (2005; 2006) select more comparable document pairs in a cross-lingual information retrieval (CLIR) based manner by using a bilingual dictionary and article information of publication date. The top-n retrieved document pairs then serve as input for the tasks of parallel sentence and sub-sentence extraction. Smith et al. (2010) extract comparable corpora from Wikipedia and use "interwiki" links to identify aligned comparable document pairs for the task of parallel sentence extraction. Li and Gaussier (2010) propose a comparability metric to select more comparable texts from other external sources into the original corpora for bilingual lexicon extraction. The comparability is determined by measuring the proportion of overlapping words between two documents by looking up a bilingual dictionary.

3. Methodology

To measure the comparability of two documents in different languages (A and B), we need to translate or map lexical items from the text in language A into language B, so that we can compare them within the same language B. Usually this mapping is done by using bilingual dictionaries (Li and Gausier, 2010; Prassasson and Fung, 2011) or existing machine translation tools.

Using bilingual dictionaries is often problematic, since they are not always available, especially for under-resourced languages. Another problem is that texts generated by mapping using bilingual dictionaries are very different from naturally written texts, or texts created by human or machine translation. There the translation is done word for word, and words which do not occur in the dictionary are omitted. The word order in the translated texts directly mirrors the structure of the source language, so important information about grammar, morphology, syntactic structure and named entities is lost. To overcome these problem, in this work we use the existing MT tool (Microsoft Bing Translator – in the form of open API interface[1]) to generate document translation, which is still not as good as human translation, but is significantly better than bilingual dictionary based word-for-word mapping.

If the language pair contains a well-resourced and an under-resourced language, (e.g., English-Lithuanian), we usually translate the documents into the better-resourced language (English). This allows us to apply various available NLP tools (e.g., POS tagging, word stemming and lemmatization, and named entity recognition) on the side of the well-resourced languages and gives additional useful information and features for comparability metric.

3.1 Features

For the comparability metric, we extract the following features from each of the compared documents:

- **Lexical features**: Lemmatized bag-of-word representation of each document after stop-word filtering. Obviously, the proportion of overlapped lexical information in two documents is the key factor in measuring their comparability. We apply cosine similarity measure to the lexical feature vectors and obtain the lexical similarity score (denoted by W_L) for each pair of comparable documents.
- **Structure similarity**: We approximate it by the number of content words (adjectives, adverbs, nouns, verbs and proper nouns) and the number of sentences in each document, denoted by C_D and S_D respectively. The intuition is that, if two documents are parallel or strongly-comparable, their document length should be

[1]http://code.google.com/p/microsoft-translator-java-api/

similar. Thus, the structure similarity (denoted by W_S) of two documents D1 and D2 is defined as below.

$$W_S=0.5*(C_{D1}/C_{D2})+ 0.5*(S_{D1}/S_{D2}),$$

suppose that $C_{D1}<=C_{D2}$, and $S_{D1}<=S_{D2}$ (switch C_{D1} and C_{D2} if $C_{D1}>C_{D2}$, and S_{D1} and S_{D2}, if $S_{D1}>S_{D2}$).

- **Keyword features**: Top-20 words (ranked by TFIDF weight) of each document. The idea is that TFIDF measure can help select more informative words (keywords) from documents (Frank et al., 1999, Liu et al., 2009). If any two documents share more keywords, they should be more comparable. Cosine similarity measure is applied to capture the keyword similarity (denoted by W_K) of each document pair.
- **Named Entity features**: Named entities identified in each document. If more named entities co-occur in two documents, these documents are very likely to talk about the same event or subject and thus should be more comparable. We apply Stanford NER toolkit[2] to extract named entities from the texts (Finkel et al., 2005). Again, cosine similarity is then applied to measure the closeness between named entity vectors (denoted by W_N) in each compared document pair.

3.2 Ensemble combination

After obtaining the four individual comparability scores (W_L, W_S, W_K, and W_N) for lexical feature, structure feature, keywords and named entities, we apply a weighted average strategy to combine these different types of scores in the comparability metric. Specifically, in the metric, each individual score is associated with a constant weight, indicating the relative confidence (or importance) of the corresponding type of score. Thus, the overall comparability score (denoted by *SC*) of a document pair is computed as below:

$$SC=\alpha*W_L+\beta*W_S+\gamma*W_K+\delta*W_N$$

where α, β, γ, and $\delta \in [0, 1]$, and $\alpha+\beta+\gamma+\delta=1$[3]. Therefore, *SC* should be a value between 0 and 1, and larger *SC* value indicates higher comparability level.

4. Experiment and evaluation

4.1 Data source

To investigate the reliability of the proposed comparability metric, we use the initial comparable corpora (ICC) collected in FP7 ICT ACCURAT project (http://www.accurat-project.eu/) for experiments. ICC contains cross-lingual comparable corpora for under-resourced languages and narrow subject domains. It was manually annotated by project partners at the document level (document pairs) for comparability levels (parallel, strongly- comparable, weakly-comparable) based on the definition described in section 1. Hence, we use ICC as gold standard, and perform the experiments on seven language pairs: German-English (DE-EN), Greek-English (EL-EN), Estonian-English (ET-EN), Lithuanian-English (LT-EN), Latvian-English (LV-EN), Romanian-English (RO-EN), and Slovenian-English (SL-EN).

4.2 Experimental results

We adopt a simple method for evaluation. For each language pair, we compute the average scores for all the document pairs in the same comparability level, and compare them to the gold standard labels. In addition, in order to better reveal the relation between the comparability scores automatically obtained from the proposed metrics and the gold standard comparability levels, we also measure the Pearson correlation between them. More specifically, we calibrate the comparability levels into a numerical scale: Parallel=3, strongly-comparable =2, and weakly-comparable =1. The correlation is then computed between these calibrated values and the corresponding average comparability scores.

In the experiments, we randomly extract a small subset from ICC as development set to tune the weights for the parameter α, β, γ, and δ, and the rest of ICC is used as test set. We tried out several different combinations of weights for the four different types of features on the development set and empirically set $\alpha = 0.5$, $\beta = \gamma = 0.2$, and $\delta = 0.1$, as this setting performs best among all the tested combinations. The weight assignment is actually based on the assumption that, lexical feature can best characterize the comparability given the good translation quality provided by the powerful MT system, while keyword and named entity features are also better indicators of comparability than the simple document length information. The results which list number of tested document pairs in each comparability level and the average comparability score of these document pairs (in bold), are presented in Table 1 and Chart 1 below.

Overall, from the average cosine scores for each comparability level presented in Table 1 and Chart 1, we can see that the scores obtained from the comparability metric can reliably reflect the comparability levels across different languages. This is because the average scores for higher comparable levels are always significantly larger than that of lower comparable levels, namely SC(parallel) > SC(strongly- comparable) > SC(weakly-comparable). In addition, from Table 1, we can also see that the correlation scores are very close to 1 for all the 7 language pairs, indicating that there is strong correlation between the comparability scores obtained from the metrics and the corresponding comparability levels. These results thus confirm that, on the level of average scores for the document collection, the comparability level predicted by our metric corresponds to the independently defined levels of comparability, based on the origin of the collected texts.

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

[3] We define that the scale of the four weight parameters between 0 and 1, which sum up to 1. This will guarantee the final comparability score (***SC***) in the scale of 0 and 1, which is better for threshold setting in the applications.

Language pair	Overall number of document pairs	Parallel	Strongly-comparable	Weakly-comparable	Pearson Correlation
DE-EN	1286	531 **0.912**	715 **0.622**	40 **0.326**	**0.99998**
EL-EN	834	85 **0.841**	400 **0.635**	349 **0.250**	**0.98505**
ET-EN	1648	182 **0.765**	987 **0.547**	479 **0.310**	**0.99971**
LT-EN	1177	347 **0.755**	509 **0.613**	321 **0.308**	**0.97855**
LV-EN	1252	184 **0.770**	558 **0.627**	510 **0.236**	**0.96588**
RO-EN	130	20 **0.782**	42 **0.614**	68 **0.311**	**0.98658**
SL-EN	1795	532 **0.779**	302 **0.582**	961 **0.373**	**0.99985**

Table 1: Number of document pairs (top) and average comparability scores (bottom, **bold**) for different comparability

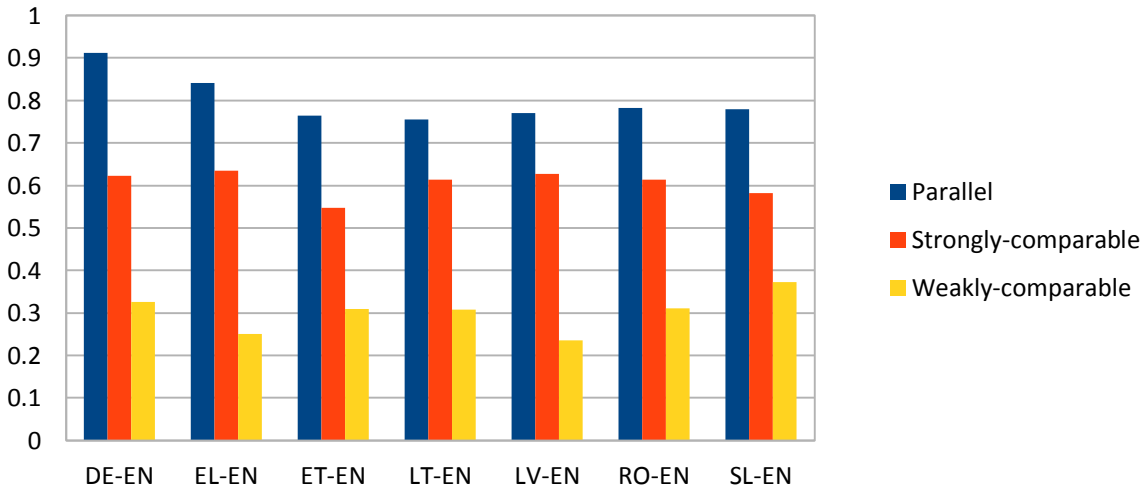

Chart 1: Average comparability scores for each of the comparability levels in ICC.

5. Application

The comparability metric is useful for collecting high-quality comparable corpora, as it can help filter out low comparable or non-comparable document pairs from the raw crawled corpora. But is it also useful for other NLP tasks, such as translation equivalent detection from comparable corpora? In this section, we investigate its impact to the task of parallel phrase extraction from comparable corpora.

The algorithm of parallel phrase extraction used in our experiments, which develops the ideas of the algorithm presented in (Munteanu and Marcu, 2006), uses the lexical overlap and the structural matching measures (Ion, 2012). It first splits the source and target documents into phrases (Step 1). Then it computes parallelism score for each possible pair of phrases by using a bilingual dictionary (base dictionary) generated from GIZA++ (Och and Ney, 2003), and retains all the phrase pairs with score larger than a predefined threshold (Step 2). GIZA++ is then applied on the retained phrase pairs to detect new dictionaries entries, which are then added in the base dictionary (Step 3). Using the augmented dictionary, the algorithm iteratively executes Step 2 and Step 3 for several times (empirically set at 5) and outputs the detected phrase pairs.

For the experiment of parallel phrase extraction, we use another dataset called *USFD* corpora, which is much

Language pair	0.1<=SC<0.3 (ave 0.2)	0.3<=SC<0.5 (ave 0.4)	SC>=0.5 (ave 0.6)	Pearson correlation
DE-EN	927	1707	2788	**0.9956**
EL-EN	340	575	920	**0.9940**
ET-EN	312	528	897	**0.9887**
LT-EN	297	466	903	**0.9689**
LV-EN	433	924	1950	**0.9799**
RO-EN	2682	5130	8773	**0.9936**
SL-EN	519	1061	2504	**0.9673**

Table 2: Number of extracted parallel phrases for different intervals in USFD

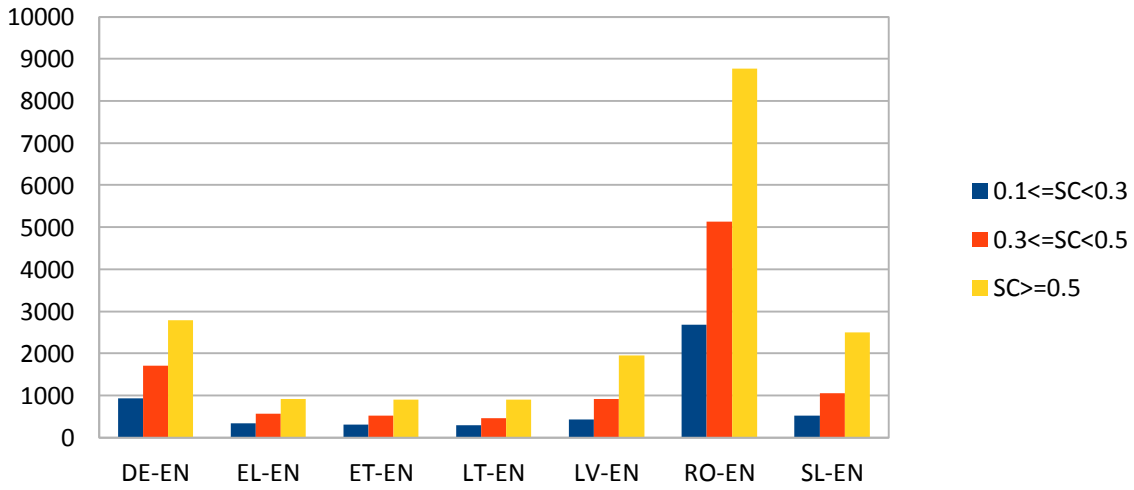

Chart 2: Number of extracted parallel phrases for different intervals for different comparability scores in USFD

larger than ICC, and was collected in ACCURAT project by the partner at the University of Sheffield. We test the performance of the extraction algorithm on the same seven language pairs. We first apply our comparability metric on *USFD* dataset to assign comparability scores for all the document pairs in *USFD*. For each language pair, we set three different intervals based on the comparability score (SC). They are (1) 0.1<=SC<0.3, (2) 0.3<=SC<0.5, and (3) SC>=0.5[4]. We then randomly select 500 document pairs from each interval, which serve as the data source for parallel phrase extraction. The experiment focuses on counting the number of extracted parallel phrases with parallelism score>=0.4 [5], and computes the average number of extracted phrases per 100000 words (the sum of words in a language pair) for each interval. In addition, the Pearson correlation measure is also applied to measure the correlation between the average comparability score[6] of each interval and the number of extracted parallel phrases. The results are presented in Chart 2 and Table 2.

From Chart 2 and Table 2, we can see that for all the 7 language pairs, based on the average number of extracted aligned phrases, clearly we have interval (3)>(2)>(1). In other word, higher comparability level always leads to significantly more number of aligned phrases extracted from the comparable documents.

Pearson's R correlation between the average numeric value of the comparability score and the number of extracted equivalents is very close to 1 for all language pairs, which indicates that the metric results are in line

[4] There is not special strategy about the selection of boundary for the intervals. We can set other intervals for experiment as well, such as SC>=0.6, as long as that there is enough amount of comparable document pairs in the corresponding intervals for evaluation.

[5] From the evaluation of the parallel phrase extraction performance on a small testing dataset, automatically extracted parallel phrase pairs with parallelism score>=0.4 are shown to be more reliable.

[6] The average comparability scores computed from the 500 randomly selected document pairs for the three intervals are about 0.2, 0.4, and 0.6, respectively.

with the performance of equivalent extraction algorithm. Therefore, in order to extract more parallel phrases from comparable documents, the comparability metric can be applied beforehand to select more comparable document pairs with higher comparability degree, where it is more likely to successfully extract a greater number of translation equivalents.

6. Conclusion and future work

We propose a comparability metric which incorporates lexical information, document structure, keywords, and named entities in an ensemble combination manner. The reliability and effectiveness of the metric have been confirmed by experiments, as the results show that it can be used to construct good-quality comparable corpora from the raw web crawling results. We also investigated the applicability of the metric by measuring its impact on parallel phrase extraction from comparable corpora. It turns out that higher comparability scores always lead to significantly more parallel phrases extracted from comparable documents.

However, in the metric, the text translation process is expensive, as it relies on the availability of the powerful MT systems (e.g., Bing translator or Google translator). Thus, in the future work, we will train alternative MT systems for text translation by using the available SMT toolkits (e.g., Moses) on large scale parallel corpora, such as Europarl[7] and JRC-Acquis[8]. We will also include a comprehensive evaluation of the proposed metric to capture its impact on the quality of machine translation systems with phrase tables derived from comparable corpora.

7. Acknowledgements

We thank our project partner RACAI for providing us the toolkit of parallel phase extraction, and the anonymous reviewers for valuable comments. This work is supported by the EU funded ACCURAT project (FP7-ICT-248347) at the Centre for Translation Studies, University of Leeds.

8. References

Babych, B., Sharoff, S., and Hartley, A. (2008). Generalising lexical translation strategies for MT using comparable corpora. *In proceedings of LREC 2008*, Marrakech, Morocco.

Chiao, Y-C., Zweigenbaum, P. (2002). Looking for candidate translational equivalents in specialized, comparable corpora. *In Proceedings of COLING 2002*, Taipei, Taiwan.

Finkel, J., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by Gibbs sampling. *In proceedings of ACL 2005*, University of Michigan, Ann Arbor, USA.

Frank, E., Paynter, G., and Witten, I. (1999). Domain-specific keyphrase extraction. *In proceedings of IJCAI 1999*, Stockholm, Sweden.

Fung, P., Cheung, P. (2004a). Mining very non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and EM. *In Proceedings of EMNLP 2004*, Barcelona, Spain.

Fung, P., Cheung, P. (2004b). Multi-level bootstrapping for extracting parallel sentences from a quasi comparable corpus. *In proceedings of COLING 2004*, Geneva, Switzerland.

Ion, R. (2012). PEXACC: a parallel data mining algorithm from comparable corpora. *In proceedings of LREC 2012*, Istanbul, Turkey.

Kilgarriff, A., Rose, T. (1998). Measures for corpus similarity and homogeneity. *In proceedings of EMNLP 1998*, Granada, Spain.

Li, B., Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. *In Proceeding of COLING 2010*, Beijing, China.

Liu, F., Pennell, D., Liu, F., and Liu, Y. (2009). Unsupervised approaches for automatic keyword extraction using meeting transcripts. *In proceedings of NAACL 2009*, Boulder, Colorado, USA.

Maia, B. (2003). What are comparable corpora? *In proceedings of the Corpus Linguistics workshop on Multilingual Corpora: Linguistic requirements and technical perspectives, 2003*, Lancaster, U.K.

McEnery, A., Xiao, Z. (2007). Parallel and comparable corpora? *In Incorporating Corpora: Translation and the Linguist. Translating Europe. Multilingual Matters*, Clevedon, UK.

Morin, M., Daille, B., Takeuchi, K., and Kageura, K. (2007). Bilingual terminology mining — using brain, not brawn comparable corpora. *In proceedings of ACL 2007*, Prague, Czech Republic.

Munteanu, D., Fraser, A. and Marcu. D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. *In proceedings of HLT-NAACL 2004*, Boston, USA.

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

Munteanu, D., Marcu, D. (2006). Extracting parallel sub-sentential fragments from non-parallel corpora. *In Proceedings of COLING/ACL 2006*, Sydney, Australia.

Och, F., Ney, H. (2003). A systematic comparison of various statistical alignment models. *In Computational linguistics, Volume 29, No.1, pp. 19-51, 2003.*

Prochasson, E., Fung, P. (2011). Rare word translation extraction from aligned comparable documents., *In Proceedings of ACL-HLT 2011*, Portland, USA.

Rapp, R. (1995). Identifying word translation in non-parallel texts. *In proceedings of ACL 1995*, Cambridge, Massachusetts, USA.

Rapp, R. (1999). Automatic identification of word translations from unrelated English and German corpora. *In proceedings of ACL 1999*, College Park, Maryland, USA.

Resnik, P., Smith, N. (2003). The web as a parallel corpus. *In computational Linguistics, 29(3):349–380.*

Saralegi, X., Vicente, I., and Gurrutxaga, A. (2008). Automatic extraction of bilingual terms from comparable corpora in a popular science domain. *In proceedings of the Workshop on Comparable Corpora, LREC 2008*, Marrakech, Morocco.

Sharoff, S. (2007). Classifying Web corpora into domain and genre using automatic feature identification. *In proceedings of 3rd Web as Corpus Workshop,*

[7] *Available at www.statmt.org/**europarl**/*

[8] Available at *http://langtech.jrc.it/JRC-Acquis.html*

Louvain-la-Neuve, Belgium.
Sharoff, S., Babych, B., and Hartley, A. (2006). Using comparable corpora to solve problems difficult for human translators. *In proceedings of ACL 2006*, Syndey, Australia.
Smith, J. Quirk, C., and Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. *In proceedings of NAACL 2010*, Los Angeles, USA.
Yu, K., Tsujii, J. (2009). Extracting bilingual dictionary from comparable corpora with dependency heterogeneity. *In proceedings of HLT-NAACL 2009*, Boulder, Colorado, USA.