Automatic Building and Using Parallel Resources for SMT from Comparable Corpora

Santanu Pal[1], Partha Pakray[2], Sudip Kumar Naskar[3]
[1]Universität Des Saarlandes, Saarbrücken, Germany
[2]Computer & Information Science,
Norwegian University of Science and Technology, Trondheim, Norway
[3]Department of Computer Science & Engineering,
Jadavpur University, Kolkata, India
[1]santanu.pal@uni-saarland.de,
[2]partha.pakray@idi.ntnu.no,
[3]sudip.naskar@cse.jdvu.ac.in

Abstract

Building parallel resources for corpus based machine translation, especially Statistical Machine Translation (SMT), from comparable corpora has recently received wide attention in the field Machine Translation research. In this paper, we propose an automatic approach for extraction of parallel fragments from comparable corpora. The comparable corpora are collected from Wikipedia documents and this approach exploits the multilingualism of Wikipedia. The automatic alignment process of parallel text fragments uses a textual entailment technique and Phrase Based SMT (PB-SMT) system. The parallel text fragments extracted thus are used as additional parallel translation examples to complement the training data for a PB-SMT system. The additional training data extracted from comparable corpora provided significant improvements in terms of translation quality over the baseline as measured by BLEU.

1 Introduction

Comparable corpora have recently attracted huge interest in natural language processing research. Comparable corpora are now considered as a rich resource for acquiring parallel resources such as parallel corpus or parallel text fragments,. Parallel text extracted from comparable corpora can take an important role in improving the quality of machine translation (MT) (Smith et al. 2010). Parallel text extracted from comparable corpora are typically added with the training corpus as additional training material which is expected to facilitate better performance of SMT systems specifically for low density language pairs.

In the present work, we try to extract English−Bengali parallel fragments of text from comparable corpora. We have collected document aligned corpus of English−Bengali document pairs from Wikipedia which provides a huge collection of documents in many different languages. For automatic alignment of parallel fragments we have used two-way textual entailment (TE) system and a baseline SMT system.

Textual entailment (TE), introduced by (Dagan and Glickman, 2004), is defined as a directional relationship between pairs of text expressions, denoted by the entailing *text* (T) and the entailed *hypothesis* (H). T entails H if the meaning of H can be inferred from the meaning of T. Textual Entailment has many applications in NLP tasks, such as summarization, information extraction, question answering,

Proceedings of the 3rd Workshop on Hybrid Approaches to Translation (HyTra) @ EACL 2014, pages 47–56,
Gothenburg, Sweden, April 27 2014. ©2014 Association for Computational Linguistics

information retrieval, machine translation, etc. In machine translation, textual entailment can be applied to MT evaluation (Pado et al., 2009). A number of research works have been carried out on cross-lingual Textual entailment using MT (Mehdad et al.,2010; Negri et al., 2010; Neogi et al., 2012). However, to the best of our knowledge, the work presented here is the first attempt towards employing textual entailment for the purpose of extracting parallel text fragments from comparable corpora which in turn are used to improve MT system.

Munteanu and Marcu (2006) suggested that comparable corpora tend to have parallel data at sub-sentential level. Hence, instead of finding sentence level parallel resource from comparable corpora, in the present work we mainly focus on finding parallel fragments of text.

We carried out the task of automatic alignment of parallel fragments using three steps: (i) mining comparable corpora form Wikipedia, (ii) sentence level alignment using two-way TE and a baseline Bengali−English SMT system, and finally (iii) clustering the parallel sentence aligned comparable corpora using textual entailment and then aligning parallel fragments of text by textual entailment and a baseline Bengali−English SMT system.

Although, we have collected document aligned comparable corpora, the documents in the corpus do not belong to any particular domain. Even with such a corpus we have been able to improve the performance of an existing machine translation system which was built on tourism domain data. This also signifies the contribution of this work towards domain adaptation of MT systems.

The rest of the paper is organized as follows. Section 2 describes the related work. Section 3 describes the mining process of the comparable corpora. The two-way TE system architecture is described in section 4. Section 5 describes the automatic alignment technique of parallel fragment of texts. Section 6 describes the tools and resources used for this work. The experiments and evaluation results are presented in section 7. Section 8 concludes and presents avenues for future work.

2 Related Work

Comparable corpora have been used in many research areas in NLP, especially in machine translation. Several earlier works have studied the use of comparable corpora in machine translation. However, most of these approaches (Fung and McKeown, 1997; Fung and Yee, 1998; Rapp, 1999; Chiao and Zweigenbaum, 2002; Dejean et al., 2002; Kaji, 2005; Otero, 2007; Saralegui et al., 2008; Gupta et al., 2013) are specifically focused on extracting word translations from comparable corpora. Most of the strategies follow a standard method based on the context vector similarity measure such as finding the target words that have the most similar distributions with a given source word. In most of the cases, a starting list contains the "seed expressions" and this list is required to build the context vectors of the words in both the languages. A bilingual dictionary can be used as a starting list. The bilingual list can also be prepared form parallel corpus using bilingual correlation method (Otero, 2007). Instead of a bilingual list, multilingual thesaurus could also be used for this purpose (Dejean, 2002).

Wikipedia is a multilingual encyclopedia available in different languages and it can be used as a source of comparable corpora. Otero et al. (2010) stored the entire Wikipedia for any two languages and transformed it into a new collection: CorpusPedia. Our work shows that only a small ad-hoc corpus containing Wikipedia articles could prove to be beneficial for existing MT systems.

In the NIST shared task on Recognizing Textual Entailment Challenge (RTE), several methods have been proposed to tackle the textual entailment problem. Most of these systems use some form of lexical matching, e.g., n-gram, word similarity, etc. and even simple word overlap. A number of systems represent the texts as parse trees (e.g., syntactic or dependency trees)

before the actual task. Some of the systems use semantic features (e.g., logical inference, Semantic Role Labelling) for solving the text and hypothesis entailment problem. MacCartney et al. (2006) proposed a new architecture for textual inference in which finding a good alignment is separated from evaluating entailment. Agichtein et al. (2008) presented a supervised machine learning approach to train a classifier over a variety of lexical, syntactic, and semantic metrics. Malakasiotis (2009) used string similarity measures applied to shallow abstractions of the input sentences and a Maximum Entropy classifier to learn how to combine the resulting features.

In the present work, we used the textual entailment system of Pakray et al. (2011) which performed well on various RTE tasks and datasets, as well as other NLP tasks like question answering, summarization, etc. We integrated a new module to by using reVerb [1] tool and optimized all the features produced by different modules.

The main objective of the present work is to investigate whether textual entailment can be used to establish alignments between text fragments in comparable corpora and whether the parallel text fragments extracted thus can improve MT system performance.

3 Mining Comparable Corpora

We collected comparable corpora from Wikipedia - online collaborative encyclopedia available in a wide variety of languages. English Wikipedia contains largest volume of data such as millions of articles; there are many language editions with at least 100,000 articles. Wikipedia links articles on the same topic in different languages using "interwiki" linking facility. Wikipedia is an enormously useful re-source for extracting parallel resources as the documents in different languages are already aligned. We first collect an English document from Wikipedia and then find the same document in Bengali if there exists any inter-language link. Extracted English−Bengali document pairs from Wikipedia are already comparable since they are written about the same entity. Although each English−Bengali document pairs are comparable and they discuss about the same topic, most of the times they are not exact translation of each other; as a result parallel fragments of text are rarely found in these document pairs. The bigger the size of the fragment may result less probable parallel version will be found in the target side. Nevertheless, there is always chance of getting parallel phrase, tokens or even sentences in comparable documents.

We designed a crawler to collect comparable corpora for English−Bengali document pairs. Based on an initial seed keyword list, the crawler first visits each English page of Wikipedia, saves the raw text (in HTML format), and then follows the cross-lingual link for each English page and collects the corresponding Bengali document. In this way, we collect English−Bengali comparable documents in the tourism domain. We retain only the textual information and all the other details are discarded. We extract English and Bengali sentences from each document. The extracted sentences from each English document are not parallel with the corresponding Bengali document. Moreover, Bengali documents are contained limited information compare to the English document. We align sentences of English−Bengali from these comparable corpora through a baseline PB-SMT system. A Bengali-English baseline PB-SMT system has been developed which was trained on English−Bengali tourism domain corpus. We translated Bengali sentences into English. The translated sentence is then examined for entailment in the English comparable document by using two-way TE system proposed in section 4. If it is more than 50% entailed with the target document then the target sentence is directly fetched form the comparable English document and the source-target sentence pair are saved in a list. In this way, we extract parallel sentences from comparable corpora. These parallel sentences except those are 100% entailed may

[1] http://reverb.cs.washington.edu/

not be completely parallel but they are comparable. So, we created a parallel fragment list which is proposed in section 5.

4 Two-way Textual Entailment System

A two-way automatic textual entailment (TE) recognition system that uses lexical, syntactic and semantic features has been described in this section. The system architecture has been shown in Figure 1. The TE system has used the Support Vector Machine (SVM) technique that uses thirty-one features for training purpose. In lexical module there are eighteen features and eleven features from syntactic module, one feature by using reVerb and one feature from semantic module.

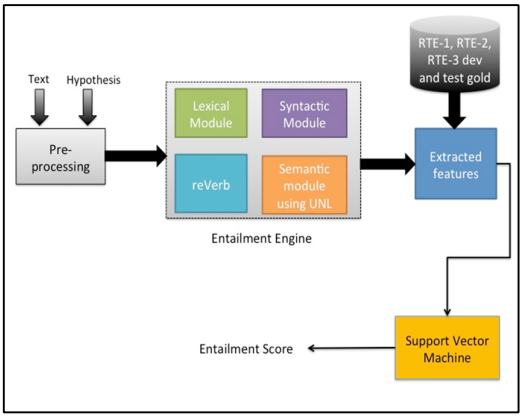

Fig.1 Two way TE architecture

4.1 Lexical Module

In this module six lexical comparisons and seventeen lexical distance comparisons between text and hypothesis has used.

Six lexical comparisons are WordNet (Fellbaum, 1998) based unigram match, bigram match, longest common sub-sequence, skip-gram, stemming and named entity matching. We have calculated weight from each of these six comparisons in equation (1).

$$weight = \frac{\sum number-of-common-tokens-between-text-and-hypothesis}{\sum number-of-tokens-in-hypothesis} \quad (1)$$

The API for WordNet Searching (JAWS) [2] provides Java applications with the ability to retrieve data from the WordNet 2.1 database.

For Named entity detection we have used Text Tokenization Toolkit (LT-TTT2)[3] (Grover et. al., 1999). The LT-TTT2 named entity component has been used.

For lexical distance measure, we have used features of Vector Space Measures (Euclidean distance, Block distance, Minkowsky distance, Cosine similarity, Matching Coefficient), Set-based Similarities (Dice, Jaccard, Overlap, Harmonic), Edit Distance Measures (Levenshtein distance, Smith-Waterman distance, Jaro Distance). Lexical distance measurement has used the libraries SimMetrics[4], SimPack[5] and SecondString[6]. SimMetrics is a Similarity Metric Library, e.g., from edit distance (Levenshtein, Gotoh, Jaro etc) to other metrics, (e.g Soundex, Chapman).

4.2 Syntactic Module

The syntactic module compares the dependency relations in both hypothesis and text. The system extracts syntactic structures from the text-hypothesis pairs using Combinatory Categorial Grammar (C&C CCG) Parser [7] and Stanford Parser [8] and compares the corresponding structures to determine if the entailment relation is established. Two different systems have been implemented one system used Stanford Parser output and another system used C&C CCG Parser. The system accepts pairs of text snippets (text and hypothesis) at the input and gives score for each comparison. Some of the important comparisons on the dependency structures of the text and the hypothesis are Subject-subject comparison, WordNet Based Subject-Verb

[2] http://lyle.smu.edu/~tspell/jaws/index.html
[3] http://www.ltg.ed.ac.uk/software/lt-ttt2
[4] http://sourceforge.net/projects/simmetrics/
[5] https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/index.html
[6] http://sourceforge.net/projects/secondstring/
[7] http://svn.ask.it.usyd.edu.au/trac/candc/wiki
[8] http://nlp.stanford.edu/software/lex-parser.shtml

Comparison, Subject-Subject Comparison, Object-Verb Comparison, WordNet Based Object-Verb Comparison, Cross Subject-Object Comparison Number Comparison, Noun Comparison, Prepositional Phrase Comparison, Determiner Comparison and other relation Comparison.

4.3 reVerb Module

ReVerb [9] is a tool, which extracts binary relationships from English sentences. The extraction format is in Table 1.

Extraction Format	arg1 rel arg2
Example	A person is playing a guitar
reVerb Extracts	arg1= {A person} rel = {is playing} arg2 = {a guitar}

Table **1**: Example by reVerb Tool

The system parsed the text and the hypothesis by reverb tool. Each of the relations compares between text and hypothesis and calculates a score for each pair.

4.4 Semantic Module

The semantic module based on the Universal Networking Language (UNL) (Uchida and Zhu, 2001). The UNL can express information or knowledge in semantic network form with hyper-nodes. The UNL is like a natural language for computers to represent and process human knowledge. There are two modules in UNL system - En-converter and De-converter module. The process of representing natural language sentences in UNL graphs is called En-converting and the process of generating natural language sentences out of UNL graphs is called De-converting. An En-Converter is a language independent parser, which provides a framework for morphological, syntactic, and semantic analysis synchronously. The En-Converter is based on a word dictionary and a set of enconversion grammar rules. It analyses sentences according to the en-conversion rules. A De-Converter is a language independent generator, which provides a framework for syntactic and morphological generation synchronously.

An example UNL relation for a sentence "*Pfizer is accused of murdering 11 children*" is shown in Table 2.

[S:00]
{org:en} *Pfizer is accused of murdering 11 children* {/org}
{unl}
obj(accuse(icl>do,equ>charge,cob>abstract_thing,agt>person,obj>person).@entry
.@present,pfizer.@topic)
qua:01(child(icl>juvenile>thing).@pl,11)
obj:01(murder(icl>kill>do,agt>thing,obj>living_thing).@entry,child(icl>juvenile
>thing).@pl)
cob(accuse(icl>do,equ>charge,cob>abstract_thing,agt>person,obj>person).@entr
y.@present,:01)
{/unl}
[/S]

Table **2**: Example of UNL

The system converts the text and the hypothesis into UNL relations by En-Converter. Then it compares the UNL relations in both the text and the hypothesis and gives a score for each comparison.

4.5 Feature Extraction Module

The features are listed in Table 3:

Name of Features	**No of features**
Lexical Module	18
Syntactic Module	11
reVerb Module	1
Semantic Module	1

Table **3**: Features for SVM

4.6 Support Vector Machines (SVM)

Support Vector Machines (SVMs) [10] are supervised learning models used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for

[9] http://reverb.cs.washington.edu/

[10] http://en.wikipedia.org/wiki/Support_vector_machine

each given input, which of two possible classes form the output, making it a non-probabilistic binary linear classifier.

The SVM based our Textual Entailment system has used the following data sets: RTE-1 development and RTE-1 annotated test set, RTE-2 development set and RTE-2 annotated test set, RTE-3 development set and RTE-3 annotated test set to deal with the two-way classification task. The system has used the LIBSVM -- A Library for Support Vector Machines[11] for the classifier to learn from this data set.

5 Alignment of Parallel fragments using proposed TE system

We have extracted parallel fragment from the parallel sentence aligned comparable resource list as well as the training data. Initially, we make cluster on the English side of this list with the help of two-way TE method. More than 50% entailed sentences have been considered to take a part of the same cluster. The TE system divides the complete set of comparable resources list into some smaller sets of cluster. Each cluster contains at least two English sentences. Each English cluster is corresponding to the set comparable Bengali sentences. So in this way we have developed a number of English Bengali parallel clusters. We intersect between the both English and Bengali sentences which are belonging to the same clusters.

We try to align the English and Bengali fragments extracted from a parallel sentence aligned comparable resource list. If both sides contain only one fragment then the alignment is trivial, and we add such fragment pairs to seed another parallel fragment corpus that contains examples having only one token in both side. Otherwise, we establish alignments between the English and Bengali fragments using translation. If both the English and Bengali side contains *n* number of fragments, and the alignments of *n*-1 fragments can be established through translation or by means of already existing alignments, then the n^{th} alignment is trivial.

These parallel fragments of text, extracted from the comparable corpora are added with the tourism domain training corpus to enhance the performance of the baseline PB-SMT system.

6 Tools and Resources

A sentence-aligned English–Bengali parallel corpus contains 23,492 parallel sentences from the travel and tourism domain has been used in the present work. The corpus has been collected from the consortium-mode project "Development of English to Indian Languages Machine Translation (EILMT) System[12]". The Stanford Parser[13] and CRF chunker[14] (Xuan-Hieu Phan, 2006) have been used for parsing and chunking in the source side of the parallel corpus, respectively.

The experiments were carried out using the standard log-linear PB-SMT model as our baseline system: GIZA++ implementation of IBM word alignment model 4, phrase-extraction heuristics described in (Koehn et al., 2003), minimum-error-rate training (Och, 2003) on a held-out development set, target language model trained using SRILM toolkit (Stolcke, 2002) with Kneser-Ney smoothing (Kneser and Ney, 1995) and the Moses decoder (Koehn et al., 2007) have been used in the present study.

7 Experiments and Results

We randomly identified 500 sentences each for the development set and the test set from the initial parallel corpus. The rest is considered as the training corpus. The training corpus was filtered with the maximum allowable sentence length of 100 words and sentence length ratio of 1:2 (either way). Finally the training corpus

[11] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[12] The EILMT project is funded by the Department of Electronics and Information Technology (DEITY), Ministry of Communications and Information Technology (MCIT), Government of India.

[13] http://nlp.stanford.edu/software/lex-parser.shtml

[14] http://crfchunker.sourceforge.net/

contained 22,492 sentences. In addition to the target side of the parallel corpus, we used a monolingual Bengali corpus containing 488,026 words from the tourism domain for building the target language model. Experiments were carried out with different n-gram settings for the language model and the maximum phrase length and it was found that a 4-gram language model and a maximum phrase length of 7 produce the optimum baseline result on both the development and the test set. We carried out the rest of the experiments using these settings.

The collected comparable corpus consisted of 5582 English–Bengali document pairs. It is evident from Table 4 that English documents are more informative than the Bengali documents as the number of sentences in English documents is much higher than those in the Bengali documents. When the Bengali fragments of texts were passed to the Bengali–English translation module some of them could not be translated into English and also, some of them could be translated only partially. Therefore, some of the tokens were translated while some were not. Some of those partially translated text fragments were aligned through textual entailment; however, most of them were discarded. As can be seen from Table 4, 9,117 sentences were entailed in the English side, of which the system was able to establish cross-lingual entailment for 2,361 English–Bengali sentence pairs.

	No. of English sentence	**No. of Bengali sentence**
Extraction from Comparable corpora	579037	169978
more than 50% Entailed English Sentences	9117	-
more than 50% Entailed (sentence aligned comparable)	2361	2361
parallel fragment of texts from sentence aligned comparable list	3937	3937

Table **4**: Statistics of the sentence aligned comparable list and the aligned parallel text fragments.

Finally, the textual entailment based alignment procedure was able to align 3937 parallel fragments as reported in Table 4. Manual inspection of the parallel list revealed that most of the aligned texts were of good quality.

We carried out evaluation of the MT quality using four automatic MT evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), NIST (Doddington, 2002) and TER (Snover et al., 2006). Table 5 shows the performance of the PB-SMT systems built on the initial training corpus and the larger training corpus containing parallel text fragments extracted from the comparable corpora. Treating the parallel text fragments extracted from the comparable corpora as additional training material results in significant improvement in terms of BLEU (1.73 points, 15.84% relative) over the baseline system. Similar improvements are also obtained for the other metrics. The low evaluation scores could be attributed to the fact that Bengali is a morphologically rich language and has a relatively free phrase order; besides there were only one set of reference translations for the testset.

Experiments	**BLEU**	**NIST**	**METEOR**	**TER**
Baseline	10.92	4.16	0.3073	75.34
Baseline + parallel fragments of texts as additional training material	12.65	4.32	0.3144	73.00

Table **5**: Evaluation results

8 Conclusion and Future Work

In this paper, we have successfully extracted English–Bengali parallel fragments of text from comparable corpora using textual entailment techniques. The parallel text fragments extracted thus were able to bring significant improvements in the performance of an existing machine translation system. For low density language pairs, this approach can help to improve the state-of-art machine translation quality. A manual inspection on a subset of the output revealed that the additional training material

extracted from comparable corpora effectively resulted in better lexical choice and less OOV words than the baseline output. As the collected parallel text does not belong to any particular domain, this work also signifies that out of domain data is also useful to enhance the performance of a domain specific MT system. This aspect of the work would be useful for domain adaptation in MT. As future work, we would like to carry out experiments on larger datasets.

Acknowledgments

The research leading to these results has received funding from the EU project EXPERT –the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013<tel:2007-2013>/ under REA grant agreement no. [317471]. We acknowledge the support from Department of Computer and Information Science, Norwegian University of Science and Technology and also support from ABCDE fellowship programme 2012-1013.

References

Banerjee, Satanjeev and Alon Lavie. 2005. METEOR: 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, Ann Arbor, Michigan, pages 65–72.

Chiao, Yun-Chuang and Pierre Zweigenbaum. 2002. Looking for candidate translational equivalents in specialized, comparable corpora. In Proceedings of the 19th international conference on Computational linguistics, Volume 2, Association for Computational Linguistics, pages 1-5.

Dagan, Ido and Oren Glickman. 2004. Probabilistic textual entailment: generic applied modeling of language variability, In PASCAL Workshop on Learning Methods for Text Understanding and Mining, Grenoble, France.

De Marneffe, Marie-Catherine, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. 2006. Learning to distinguish valid textual entailments. In B. Magnini and I. Dagan (eds.), Proceedings of the Second PASCAL Recognizing Textual Entailment Challenge. Venice: Springer, pages 74–79.

Déjean, Hervé, Éric Gaussier, and Fatia Sadat. 2002. Bilingual terminology extraction: an approach based on a multilingual thesaurus applicable to comparable corpora. In Proceedings of the 19th International Conference on Computational Linguistics COLING, Pages 218-224.

Doddington, George. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In Proceedings of the second international conference on Human Language Technology Research . Morgan Kaufmann Publishers Inc, pages. 138-145.

Fung, Pascale and Kathleen McKeown. 1997. Finding terminology translations from non-parallel corpora. In Proceedings of the 5th Annual Workshop on Very Large Corpora, pages 192-202.

Fung, Pascale and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In Proceedings of the 17th international conference on Computational linguistics-Volume 1, Association for Computational Linguistics, pages 414-420.

Gupta, Rajdeep, Santanu Pal, and Sivaji Bandyopadhyay. 2013. Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora. In proceedings of 6th workshop of Building and Using Comparable Corpora (BUCC), ACL, Sofia, Bulgaria, Pages 69-76.

Kneser, Reinhard and Hermann Ney. 1995. Improved backing-off for n-gram language modeling. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, volume I. pages 181-184.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico,Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ond rej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions. Association for Computational Linguistics, pages 177-180.

Koehn, Philipp, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In

Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1, Association for Computational Linguistics, pages 48-54.

Mehdad, Yashar, Matteo Negri, and Marcello Federico. 2010. Towards Cross-Lingual Textual entailment. In Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL-HLT 2010. LA, USA.

Munteanu, Dragos Stefan and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Association for Computational Linguistics, pages 81-88.

Negri, Matteo, and Yashar Mehdad. 2010. Creating a Bilingual Entailment Corpus through Translations with Mechanical Turk: $100 for a 10-day Rush. In Proceedings of the NAACL-HLT 2010, Creating Speech and Text Language Data With Amazon's Mechanical Turk Workshop. LA, USA.

Neogi, Snehasis, Partha Pakray, Sivaji Bandyopadhyay, and Alexander Gelbukh. 2012. JU_CSE_NLP: Language Independent Cross-lingual Textual Entailment System. (*SEM) First Joint Conference on Lexical and Computational Semantics, Collocated with NAACL-HLT 2012, Montreal, Canada.

Och, F. Josef. 2003. Minimum error rate training in statistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1, Association for Computational Linguistics, pages 160-167.

Och, F. Josef and Herman Ney. 2000. Giza++: Training of statistical translation models.

Otero, P. Gamallo. 2007. Learning bilingual lexicons from comparable english and spanish corpora. Proceedings of MT Summit xI, pages 191-198.

Otero, P. Gamallo and Isaac González López. 2010. Wikipedia as multilingual source of comparable corpora. In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC, pages 21-25.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, pages 311-318.

Prodromos Malakasiotis. 2009. "AUEB at TAC 2009", In TAC 2009 Workshop, National Institute of Standards and Technology Gaithersburg, Maryland USA.

Rapp, Reinhard. 1999. Automatic identification of word translations from unrelated English and German corpora. In Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, pages 519-526.

Saralegui, X., San Vicente, I., and Gurrutxaga, A. 2008. Automatic generation of bilingual lexicons from comparable corpora in a popular science domain. In LREC 2008 workshop on building and using comparable corpora.

Pado, Sebastian, Michel Galley, Dan Jurafsky, and Christopher D. Manning. 2009. Textual entailment features for machine translation evaluation. In Proceedings of the EACL Workshop on Statistical Machine Translation, Athens, Greece, pages 37–41.

Smith, R. Jason, Chris Quirk, and Kristina Toutanova. 2010. Extracting parallel sentences from comparable corpora using document level alignment. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, pages 403-411.

Snover, Matthew, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. Proceedings of Association for Machine Translation in the Americas, Cambridge, Massachusetts, USA, pages 223–231.

Pakray, Partha, Snehasis Neogi, Pinaki Bhaskar, Soujanya Poria, Sivaji Bandyopadhyay, and Alexander Gelbukh. 2011. A Textual Entailment System using Anaphora Resolution. System Report, Text Analysis Conference Recognizing Textual Entailment Track (TAC RTE) Notebook, November 14-15, 2011, National Institute of

Standards and Technology, Gaithersburg, Maryland USA

Stolcke, Andreas. 2002. SRILM-an extensible language modeling toolkit. In Proceedings of the international conference on spoken language processing, Volume 2, pages 901-904.

Wang, Rui and Günter Neumann. 2007. Recognizing Textual Entailment Using Sentence Similarity based on Dependency Tree Skeletons. In Proceedings of the third PASCAL Recognising Textual Entailment Challenge.

Xuan-Hieu Phan. 2006. CRFChunker: CRF English Phrase Chunker , http://crfchunker.sourceforge.net/.