Two Ways to Use a Noisy Parallel News corpus for improving Statistical Machine Translation

Souhir Gahbiche-Braham **Hélène Bonneau-Maynard** **François Yvon**

Université Paris-Sud 11
LIMSI-CNRS
91403 Orsay, France
{souhir,hbm,yvon}@limsi.fr

Abstract

In this paper, we present two methods to use a noisy parallel news corpus to improve statistical machine translation (SMT) systems. Taking full advantage of the characteristics of our corpus and of existing resources, we use a bootstrapping strategy, whereby an existing SMT engine is used both to detect parallel sentences in comparable data and to provide an adaptation corpus for translation models. MT experiments demonstrate the benefits of various combinations of these strategies.

1 Introduction

In Statistical Machine Translation (SMT), systems are created from *parallel corpora* consisting of a set of source language texts aligned with its translation in the target language. Such corpora however only exist (at least are publicly documented and available) for a limited number of domains, genres, registers, and language pairs. In fact, there are a few language pairs for which parallel corpora can be accessed, except for very narrow domains such as political debates or international regulatory texts. Another very valuable resource for SMT studies, especially for under-resource languages, are *comparable corpora*, made of pairs of monolingual corpora that contain texts of similar genres, from similar periods, and/or about similar topics.

The potential of comparable corpora has long been established as a useful source from which to extract bilingual word dictionaries (see eg. (Rapp, 1995; Fung and Yee, 1998)) or to learn multilingual terms (see e.g. (Langé, 1995; Smadja et al., 1996)).

More recently, the relative corpus has caused the usefulness of comparable corpora be reevaluated as a potential source of parallel fragments, be they paragraphs, sentences, phrases, terms, chunks, or isolated words. This tendency is illustrated by the work of e.g. (Resnik and Smith, 2003; Munteanu and Marcu, 2005), which combines Information Retrieval techniques (to identify parallel documents) and sentence similarity detection to detect parallel sentences.

There are many other ways to improve SMT models with comparable or monolingual data. For instance, the work reported in (Schwenk, 2008) draws inspiration from recent advances in unsupervised training of acoustic models for speech recognition and proposes to use self-training on in-domain data to adapt and improve a baseline system trained mostly with out-of-domain data.

As discussed e.g. in (Fung and Cheung, 2004), comparable corpora are of various nature: there exists a continuum between truly parallel and completely unrelated texts. Algorithms for exploiting comparable corpora should thus be tailored to the peculiarities of the data on which they are applied.

In this paper, we report on experiments aimed at using a noisy parallel corpus made out of news stories in French and Arabic in two different ways: first, to extract new, in-domain, parallel sentences; second, to adapt our translation and language models. This approach is made possible due to the specificities of our corpus. In fact, our work is part of a project aiming at developing a platform for processing multimedia news documents (texts, interviews, images and videos) in Arabic, so as to streamline the

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 44–51,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

work of a major international news agency. As part as the standard daily work flow, a significant portion of the French news are translated (or adapted) in Arabic by journalists. Having access to one full year of the French and Arabic corpus (consisting, to date, of approximately one million stories (150 million words)), we have in our hands an ideal comparable resource to perform large scale experiments.

These experiments aim at comparing various ways to build an accurate machine translation system for the news domain using (i) a baseline system trained mostly with out-of-domain data (ii) the comparable dataset. As will be discussed, given the very large number of parallel news in the data, our best option seems to reconstruct an in-domain training corpus of automatically detected parallel sentences.

The rest of this paper is organized as follows. In Section 2, we relate our work to some existing approaches for using comparable corpora. Section 3 presents our methodology for extracting parallel sentences, while our phrase-table adaptation strategies are described in Section 4. In Section 5, we describe our experiments and contrast the results obtained with several adaptation strategies. Finally, Section 6 concludes the paper.

2 Related work

From a bird's eye view, attempts to use comparable corpora in SMT fall into two main categories: first, approaches aimed at extracting parallel fragments; second, approaches aimed at adapting existing resources to a new domain.

2.1 Extracting parallel fragments

Most attempts at automatically extracting parallel fragments use a two step process (see (Tillmann and Xu, 2009) for a counter-example): a set of candidate parallel texts is first identified; within this short list of possibly paired texts, parallel sentences are then identified based on some similarity score.

The work reported in (Zhao and Vogel, 2002) concentrates on finding parallel sentences in a set of comparable stories pairs in Chinese/English. Sentence similarity derives from a probabilistic alignment model for documents, which enables to recognize parallel sentences based on their length ratio, as well as on the IBM 1 model score of their word-to-word alignment. To account for various levels of parallelism, the model allows some sentences in the source or target language to remain unaligned.

The work of (Resnik and Smith, 2003) considers mining a much larger "corpora" consisting of documents collected on the Internet. Matched documents and sentences are primarily detected based on surface and/or formal similarity of the web addresses or of the page internal structure.

This line of work is developed notably in (Munteanu and Marcu, 2005): candidate parallel texts are found using Cross-Lingual Information Retrieval (CLIR) techniques; sentence similarity is indirectly computed using a logistic regression model aimed at detecting parallel sentences. This formalism allows to enrich baseline features such as the length ratio, the word-to-word (IBM 1) alignment scores with supplementary scores aimed at rewarding sentences containing identical words, etc. More recently, (Smith et al., 2010) reported significant improvements mining parallel Wikipedia articles using more sophisticated indicators of sentence parallelism, incorporating a richer set of features and cross-sentence dependencies within a Conditional Random Fields (CRFs) model. For lack of find enough parallel sentences, (Munteanu and Marcu, 2006; Kumano and Tokunaga, 2007) consider the more difficult issue of mining parallel *phrases*.

In (Abdul-Rauf and Schwenk, 2009), the authors, rather than computing a similarity score between a source and a target sentence, propose to use an existing translation engine to process the source side of the corpus, thus enabling sentence comparison to be performed in the target language, using the edit distance or variants thereof (WER or TER). This approach is generalized to much larger collections in (Uszkoreit et al., 2010), which draw advantage of working in one language to adopt efficient parallelism detection techniques (Broder, 2000).

2.2 Comparable corpora for adaptation

Another very productive use of comparable corpora is to *adapt* or *specialize* existing resources (dictionaries, translation models, language models) to specific domains and/or genres. We will only focus here on adapting the translation model; a review of the literature on language model adaptation is in (Bellagarda, 2001) and the references cited therein.

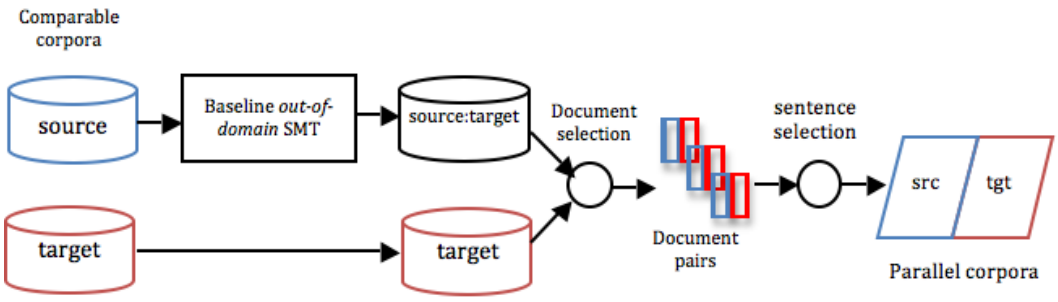

Figure 1: Extraction of parallel corpora

The work in (Snover et al., 2008) is a first step towards augmenting the translation model with new translation rules: these rules associate, with a tiny probability, every phrase in a source document with the most frequent target phrases found in a comparable corpus specifically built for this document.

The study in (Schwenk, 2008) considers *self-training*, which allows to adapt an existing system to new domains using monolingual (source) data. The idea is to automatically translate the source side of an in-domain corpus using a reference translation system. Then, according to some confidence score, the best translations are selected to form an *adaptation corpus*, which can serve to retrain the translation model. The authors of (Cettolo et al., 2010) follow similar goals with different means: here, the baseline translation model is used to obtain a phrase alignment between source and target sentences in a comparable corpus. These phrase alignments are further refined, before new phrases *not in the original phrase-table*, can be collected.

The approaches developed below borrow from both traditions: given (i) the supposed high degree of parallelism in our data and (ii) the size of the available comparable data, we are in a position to apply any of the above described technique. This is all the easier to do as all stories are timestamped, which enables to easily spot candidate parallel texts. In both cases, we will apply a bootstrapping strategy using as baseline a system trained with out-of-domain data.

3 Extracting Parallel Corpora

This section presents our approach for extracting a parallel corpus from a comparable in-domain corpora so as to adapt a SMT system to a specific domain. Our methodology assumes that both a baseline *out-of-domain* translation system and a comparable *in-domain* corpus are available, two requirements that are often met in practice.

As shown in Figure 1, our approach for extracting an *in-domain* parallel corpus from the *in-domain* comparable corpus consists in 3 steps and closely follows (Abdul-Rauf and Schwenk, 2009):
translation: translating the source side of the comparable corpora;
document pairs selection : selecting, in the comparable corpus, documents that are similar to the translated output;
sentence pairs selection : selecting parallel sentences among the selected documents.

The main intuition is that computing document similarities in one language enables to use simple and effective comparison procedures, instead of having to define *ad hoc* similarities measures based on complex underlying alignment models.

The **translation** step consists here in translating the source (Arabic) side of the comparable corpus using a baseline *out-of-domain* system, which has been trained on parallel *out-of-domain* data.

The **document selection** step consists in trying to match the automatic translations (source:target) with the original documents in the target language. For each (source:target) document, a similarity score with all the target documents is computed. We contend here with a simple association score, namely the Dice coefficient, computed as the number of words in common in both documents, normalized by the length of the (source:target) document.

A priori knowledge, such as the publication dates

of the documents, are used to limit the number of document pairs to be compared. For each source document, the target document that has the best score is then selected as a potential parallel document. The resulting pairs of documents are then filtered depending on a threshold T_d, so as to avoid false matches (in the experiments described below, the threshold has been set so as to favor precision over recall).

At the end of this step, a set of similar source and target document pairs has been selected. These pairs may consist in documents that are exact translations of each other. In most cases, the documents are noisy translation and only a subset of their sentences are mutual translation.

The **sentence selection** step then consists in performing a sentence level alignment of each pair of documents to select a set of parallel sentences. Sentence alignment is then performed with the hunalign sentence alignment tool (Varga et al., 2005), which also provides alignment confidence measures. As for the document selection step, only sentence pairs that obtain an alignment score greater than a predefined threshold T_s are selected, where T_s is again chosen to favor prevision of alignments of recall. From these, $1:1$ alignments are retained, yielding a small, adapted, parallel corpus. This method is quite different from (Munteanu and Marcu, 2005)'s work where the sentence selection step is done by a Maximum Entropy classifier.

4 Domain Adaptation

In the course of mining our comparable corpus, we have produced a translation into French for all the source language news stories. This means that we have three parallel corpora at our disposal:

- The **baseline training corpus**, which is large (a hundred million words), delivering a reasonable translation performance quality of translation, but *out-of-domain*;
- The **extracted *in-domain* corpus**, which is much smaller, and potentially noisy;
- The **translated *in-domain* corpus**, which is of medium-size, and much worse in quality than the others.

Considering these three corpora, different adaptation methods of the translation models are explored. The first approach is to concatenate the **baseline** and **in-domain** training data (either **extracted** or **translated**) to train a new translation model. Given the difference in size between the two corpus, this approach may introduce a bias in the translation model in favor of *out-of-domain*.

The second approach is to train separate translation models with **baseline** on the one hand, and with *in-domain* on the other data and to weight their combination with MERT (Och, 2003). This alleviates the former problem but increases the number of features that need to be trained, running the risk to make MERT less stable.

A last approach is also considered, which consists in using only the **in-domain** data to train the translation model. In that case, the question is the small size of the *in-domain* data.

The comparative experiments on the three approaches, using the three corpora are described in next section.

5 Experiments and results

5.1 Context and data

The experiments have been carried out in the context of the Cap Digital SAMAR[1] project which aims at developping a platform for processing multimedia news in Arabic. Every day, about 250 news in Arabic, 800 in French and in English[2] are produced and accumulated on our disks. News collected from December 2009 to December 2010 constitute the comparable corpora, containing a set of 75,975 news for the Arabic part and 288,934 news for the French part (about 1M sentences for Arabic and 5M sentences for French).

The specificity of this comparable corpus is that many Arabic stories are known to be translation of news that were first written in French. The translations may not be entirely faithful: when translating a story, the journalist is in fact free to rearrange the structure, and to some extend, the content of a document (see example Figure 2).

In our experiments, the *in-domain* comparable corpus then consists in a set of Arabic and French

[1] http://www.samar.fr

[2] The English news have not been used in this study.

Arabic:
واضاف نحن في حماس لا مانع لدينا من استئناف المفاوضات غير المباشرة حول الصفقة من النقطة التي انتهت اليها والتي حاول ان يفشلها نتانياهو.
And he added, we in Hamas don't have a problem to resume indirect negotiations about the deal from the point at which it ended and at which Netanyahu tried to fail.

French: Le porte-parole a réaffirmé que le Hamas était prêt à reprendre les tractations au point où elles s'étaient arrêtées.
The spokesman reaffirmed that Hamas was ready to resume negotiations at the point where they stopped.

Figure 2: An example of incorrect/inexact translation in a pair of similar documents.

documents which are parallel, partly parallel, or not parallel at all, with no explicit link between Arabic and French parts.

5.2 Baseline translation system

The baseline *out-of-domain* translation system was trained on a corpus of 7.6 million of parallel sentences (see Table 1), that was harvested from publicly available sources on the web: the United Nations (UN) document database, the website of the World Health Organization (WHO) and the Project Syndicate Web site. The "UN" data constitutes by far the largest portion of this corpus, from which only the Project Syndicate documents can be considered as appropriate for the task at hand.

A 4-gram backoff French language model was built on 2.4 billion words of running texts, taken from the parallel data, as well as notably the Gigaword French corpus.

	ar		fr	
Corpus	**#tokens**	**voc**	**#tokens**	**voc**
baseline	162M	369K	186M	307K
extracted	3.6M	72K	4.0M	74K
translated	20.8M	217 K	22.1M	181K

Table 1: Corpus statistics: total number of tokens in the French and Arabic sides, Arabic and French vocabulary size. Numbers are given on the preprocessed data.

Arabic is a rich and morphologically complex language, and therefore data preprocessing is necessary to deal with data scarcity. All Arabic data were preprocessed by first transliterating the Arabic text with the BAMA (Buckwalter, 2002) transliteration tool. Then, the Arabic data are segmented into sentences. A CRF-based sentence segmenter for Arabic was built with the Wapiti[3] (Lavergne et al., 2010) package. A morphological analysis of the Arabic text is then done using the Arabic morphological analyzer and disambiguation tool MADA (Nizar Habash and Roth, 2009), with the MADA-D2 since it seems to be the most efficient scheme for large data (Habash and Sadat, 2006).

The preprocessed Arabic and French data were aligned using MGiza++[4] (Gao and Vogel, 2008). The Moses toolkit (Koehn et al., 2007) is then used to make the alignments symmetric using the *grow-diag-final-and* heuristic and to extract phrases with maximum length of 7 words. A distortion model lexically conditioned on both the Arabic phrases and French phrases is then trained. Feature weights were set by running MERT (Och, 2003) on the development set.

5.3 Extraction of the *in-domain* parallel corpus

We follow the method described in Section 3: Arabic documents are first translated into French using the baseline SMT system. For the document selection step each translated (ar:fr) document is compared only to the French documents of the same day. The thresholds for document selection and sentence selection were respectively set to 0.5 and 0.7. For a pair of similar documents, the average percentage of selected sentences is about 43%.

The document selection step allows to select documents containing around 35% of the total number of sentences from the initial Arabic part of the comparable corpus, a percentage that goes down to 15% after the sentence alignment step. The resulting *in-domain* parallel corpus thus consists in a set of 156K pairs of parallel sentences. Data collected during the last month of the period was isolated from the resulting corpus, and was used to randomly extract a development and a test set of approximately 1,000

[3] http://wapiti.limsi.fr

[4] http://geek.kyloo.net/software/doku.php/mgiza:overview

Reference:	Le ministre russe des Affaires étrangères, Sergueï Lavrov *a prévenu* mercredi [...]
Baseline:	*Pronostiquait* Ministre des affaires étrangères russe, Sergei Lavrov mercredi [...]
Extracted:	Le ministre russe des Affaires étrangères, Sergueï Lavrov *a averti* mercredi [...]
Reference:	Le porte-parole de Mme Clinton, *Philip Crowley*, a toutefois reconnu [...]
Baseline:	Pour *ukun FILIP Cruau* porte-parole de Clinton a reconnu ...
Extracted:	Mais *Philip Crowley*, le porte-parole de Mme Clinton a reconnu [...]

Figure 3: Comparative translations using the **baseline** translation and the **extracted** translation systems of two sentences: *"Russian Minister of Foreign Affairs, Sergueï Lavrov, informed Wednesday [...]"* and *"The spokesman for Mrs. Clinton, Philip Crowley, however, acknowledged [...]"*.

lines each. These 2,160 sentences were manually checked to evaluate the precision of the approach, and we found that 97.2% of the sentences were correctly paired. Table 1 compares the main characteristics of the three corpora used for training.

5.4 Translation Results

Translation results obtained on the test set are reported in terms of BLEU scores in Table 2, along with the corresponding phrase table sizes. The different adaptation approaches described in Section 4 were experimented with both **extracted** and **translated** corpora as adaptation corpus (see Section 3). As expected, adapting the translation model to the news domain is very effective. Compared to the baseline system, all adapted systems obtain much better results (from 2 to 6 BLEU points). The **extracted** system outperforms the baseline system by 5 BLEU points, even though the training set is much smaller (3.6M compared to 162M tokens). This result indirectly validates the precision of our methodology.

SMT System	#Phrase pairs	BLEU
baseline	312.4M	24.0
extracted	10.9M	29.2
baseline+extracted (1 table)	321.6M	29.0
baseline+extracted (2 tables)	312.4M + 9.9M	**30.1**
translated	39M	26.7
extracted+translated (2 tables)	9.9M + 39M	28.2

Table 2: Arabic to French translation BLEU scores on a test set of 1000 sentences

Concatenating the baseline and extracted data to train a single translation model does not improve the smaller **extracted** system, thus maybe reflecting the fact that the large *out-of-domain* corpus overwhelms the contribution of the *in-domain* data. However, a log-linear combination of the corresponding phrase tables brings a small improvement (0.8 BLEU point).

Another interesting result comes from the performance of the system trained only on the **translated** corpus. *Without using any filtering of the automatic translations*, this artificial dataset enables to build another system which outperforms the baseline system by 2.5 BLEU points. This is another illustration of the greater importance of having matched domain data, even of a poorer quality, than good parallel *out-of-domain* sentences (Cettolo et al., 2010).

In the last experiment, all the available *in-domain* data (**extracted** and **translated**) are used in conjunction, with a separate phrase-table trained on each corpus. However, this did not enable to match the results of the **extracted** system, a paradoxical result that remains to be analyzed more carefully. Filtering automatic translations may be an issue.

A rapid observation of the translations provided by both the baseline system and the **extracted** system shows that the produced output are quite different. Figure 3 displays two typical examples: the first one illustrates the different styles in Arabic ("News" style often put subject *"Le ministre russe des affaires étrangères"* before verb *" a prévenu"* or *"a averti"* — which are semantically equivalent — whereas "UN" style is more classical, with the verb *"Pronostiquait"* followed by the subject *"ministre russe des Affaires étrangères"*). The second one shows how adaptation fixes the transla-

tion of words (here *"Philip Crowley"*) that were not (correctly) translated by the baseline system (*"ukun FILIP Cruau"*).

6 Conclusion

We have presented an empirical study of various methodologies for (i) extracting a parallel corpus from a comparable corpus (the so-called "Noisy Corpus") and (ii) using in-domain data to adapt a baseline SMT system. Experimental results, obtained using a large 150 million word Arabic/French comparable corpus, allow to jointly validate the extraction of the in-domain parallel corpus and the proposed adaptation methods. The best adapted system, trained on a combination of the baseline and the extracted data, improves the baseline by 6 BLEU points. Preliminary experiments with self-training also demonstrate the potential of this technique.

As a follow-up, we intend to investigate the evolution of the translation results as a function of the precision/recall quality of the extracted corpus, and of the quality of the automatically translated data. We have also only focused here on the adaptation of the translation model. We expect to achieve further gains when combining these techniques with LM adaptation techniques.

This work was partly supported by the FUI/SAMAR project funded by the *Cap Digital* competitivity cluster.

References

Sadaf Abdul-Rauf and Holger Schwenk. 2009. On the use of comparable corpora to improve smt performance. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 16–23, Morristown, NJ, USA. Association for Computational Linguistics.

Jérôme R. Bellagarda. 2001. An overview of statistical language model adaptation. In *Proceedings of the ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition*, pages 165–174, Sophia Antipolis, France.

Andrei Z. Broder. 2000. Identifying and filtering near-duplicate documents. In *Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching*, COM '00, pages 1–10, London, UK. Springer-Verlag.

Tim Buckwalter. 2002. *Buckwalter Arabic Morphological Analyzer*. Linguistic Data Consortium. (LDC2002L49).

Mauro Cettolo, Marcello Federico, and Nicola Bertoldi. 2010. Mining Parallel Fragments from Comparable Texts. In Marcello Federico, Ian Lane, Michael Paul, and François Yvon, editors, *Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT)*, pages 227–234.

Pascale Fung and Percy Cheung. 2004. Multilevel bootstrapping for extracting parallel sentences from a quasi parallel corpus. In *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP 04)*, pages 1051–1057.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 414–420. Association for Computational Linguistics.

Qin Gao and Stephan Vogel. 2008. Parallel implementations of word alignment tool. In *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, SETQA-NLP '08, pages 49–57.

Nizar Habash and Fatiha Sadat. 2006. Arabic preprocessing schemes for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 49–52, Morristown, NJ, USA. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Morristown, NJ, USA. Association for Computational Linguistics.

Tadashi Kumano and Hideki Tanaka Takenobu Tokunaga. 2007. Extracting phrasal alignments from comparable corpora by using joint probability smt model. In Andy Way and Barbara Gawronska, editors, *Proceedings of the 11th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI'07)*, Skövde, Sweden.

Jean-Marc Langé. 1995. Modèles statistiques pour l'extraction de lexiques bilingues. *Traitement Automatique des Langues*, 36(1-2):133–155.

Thomas Lavergne, Olivier Cappé, and François Yvon. 2010. Practical very large scale crfs. In *Proceedings of the 48th Annual Meeting of the Association for*

Computational Linguistics, pages 504–513, Uppsala, Sweden.

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

Dragos Stefan Munteanu 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*, ACL-44, pages 81–88, Stroudsburg, PA, USA. Association for Computational Linguistics.

Owen Rambow Nizar Habash and Ryan Roth. 2009. Mada+tokan: A toolkit for arabic tokenization, diacritization, morphological disambiguation, pos tagging, stemming and lemmatization. In Khalid Choukri and Bente Maegaard, editors, *Proceedings of the Second International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, April. The MEDAR Consortium.

Franz Josef Och. 2003. Minimum error rate training in statistical machine translation. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 160–167, Sapporo, Japan.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, ACL '95, pages 320–322, Stroudsburg, PA, USA. Association for Computational Linguistics.

Philip Resnik and Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics*, 29:349–380.

Holger Schwenk. 2008. Investigations on large-scale lightly-supervised training for statistical machine translation. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 182–189, Hawaii, USA.

Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics*, 22:1–38, March.

Jason R. Smith, 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*, HLT '10, pages 403–411, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '08, pages 857–866, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christoph Tillmann and Jian-ming Xu. 2009. A simple sentence-level extraction algorithm for comparable data. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 93–96, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jakob Uszkoreit, Jay M. Ponte, Ashok C. Popat, and Moshe Dubiner. 2010. Large scale parallel document mining for machine translation. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1101–1109, Stroudsburg, PA, USA. Association for Computational Linguistics.

Dániel Varga, László Németh, Péter Halácsy, András Kornai, Viktor Trón, and Viktor Nagy. 2005. Parallel corpora for medium density languages. In *Proceedings of the RANLP*, pages 590–596.

Bing Zhao and Stephan Vogel. 2002. Adaptive parallel sentence mining from Web bilingual news collection. In *Proceedings of the International Conference on Data Mining*, pages 745–748. IEEE Computer Society.