Procesamiento del Lenguaje Natural, Revista nº 49 septiembre de 2012, pp 93-100 recibido 05-05-12 revisado 05-06-12 aceptado 08-06-12

A Simple Approach to Use Bilingual Information Sources for Word Alignment

Una manera sencilla para usar fuentes de información bilingüe para el alineamiento de palabras

Miquel Esplà-Gomis, Felipe Sánchez-Martínez, Mikel L. Forcada
Dep. de Llenguatges i Sistemes Informàtics
Universitat d'Alacant, E-03071 Alacant, Spain
{mespla,fsanchez,mlf}@dlsi.ua.es

Resumen: En este artículo se describe un método nuevo y sencillo para utilizar fuentes de información bilingüe para el alineamiento de palabras en segmentos de texto paralelos. Este método puede ser utilizado *al vuelo*, ya que no requiere de entrenamiento. Además, puede ser utilizado con corpus comparables. Hemos comparado los resultados de nuestro método con los obtenidos por la herramienta GIZA++, ampliamente utilizada para el alineamiento de palabras, obteniendo unos resultados bastante similares.
Palabras clave: Alineamiento de palabras, fuentes de información bilingüe

Abstract: In this paper we present a new and simple method for using sources of bilingual information for word alignment between parallel segments of text. This method can be used *on the fly*, since it does not need to be trained. In addition, it can also be applied on comparable corpora. We compare our method to the state-of-the-art tool GIZA++, widely used for word alignment, and we obtain very similar results.
Keywords: Word alignment, sources of bilingual information

1 *Introduction*

In this paper we describe a method which uses sources of bilingual information (SBI) such as lexicons, translation memories, or machine translation, to align the words of a segment with those in its translation (parallel segments) without any training process. Our approach aligns the sub-segments in a pair of segments S and T by using any SBI available, and then aligns the words in S and T by using a heuristic method which does not require the availability of a parallel corpus. It is worth noting that many SBIs which could be used to align words with our method are currently freely available in the Internet: MT systems, such as Apertium[1] or Google Translate;[2] bilingual dictionaries, such as Dics.info;[3] or Word Refference[4] or translation memories, such as Linguee[5] or MyMemory.[6] This method is inspired on a previous approach (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2011) that was proposed to detect sub-segment alignments (SSAs) and help translators to edit the translation proposals produced by translation-memory-based computer-aided translation tools by suggesting the target words to change. A similar technique was also successfully applied to cross-lingual textual entailment detection (Esplà-Gomis, Sánchez-Martínez, and Forcada, 2012). Here, we propose to use these SSAs to obtain word alignment *on the fly*.

Related works. Many previous works tackle the problem of word alignment. The existing approaches may be divided in statistical approaches and heuristic approaches. One of the most remarkable works in the first group is the one by Brown et al. (1993), which describes a set of methods for word alignment based on the expectation-maximisation algorithm (Dempster, Laird, and Rubin, 1977), usually called *IBM models*. In this work, au-

[1] http://www.apertium.org
[2] http://translate.google.com
[3] http://www.dics.info
[4] http://www.wordreference.com
[5] http://www.linguee
[6] http://mymemory.translated.net

ISSN 1135-5948

thors propose five models, from a very simple one considering just one-to-one alignments between words, to more complex models which allow a word to be aligned with many words. Other authors (Vogel, Ney, and Tillmann, 1996; Dagan, Church, and Gale, 1993) propose using a hidden Markov model for word alignment. Both methods were combined and extended by Och and Ney (2003), who also developed the tool GIZA++, implementing all these methods.

Some heuristic approaches have also been proposed. Rapp (1999) proposes an approach based in the idea that groups of words which usually appear together in a language should also appear together in other languages. To obtain word alignments from this idea, the author uses a window of a given number of words to look for the most usual groups of words in each monolingual corpora. Then, coocurrence vectors are computed for the words appearing frequently together inside the window and word alignments are computed by comparing these coocurrence vectors. Fung and McKeown (1997) propose a similar method which introduces some SBIs. In this case, authors use bilingual dictionaries to obtain an initial alignment between *seed words* in a parallel text. To choose reliable seed words, they use only those words having a univocal translation in both directions and appearing with enough frequency to become useful references in both texts of the parallel corpus. Then, these initial alignments are used to align other words appearing around them in the parallel texts using a similar method to that used by Rapp (1999). Another family of heuristic methods for word alignment are based on cognates. Schulz et al. (2004) use word similarity between Spanish and Portuguese for word alignment. The most important limitation of this work is that it is only useful for closely-related languages. Other works (Al-Onaizan and Knight, 2002) try to overcome this problem by using transliteration to obtain the way in which a word in a language may be written in another language. In this case, Al-Onaizan and Knight (2002) use transliteration to find out the most likely way in which English proper nouns could be written in languages such as Arabic or Japanese in order to find their translations. Although statistical approaches have proved to obtain better results than heuristic ones, one of the advantages of heuristic approaches is that they can be used not only on parallel corpora, but in comparable corpora.

Novelty. In this work we propose a method for word alignment using previously existing bilingual resources. Although some works in the bibliography also use SBIs to perform alignment (Fung and McKeown, 1997), the main difference between this work and the previous approaches is that our method does not need any training process or bilingual corpus, i.e. it can be run *on the fly* on a pair of parallel segments. This kind of alignment method may be useful in some scenarios, as is the case of some computer-aided translation systems, to help users to detect which words should be post-edited in the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In addition, this method can be applied on comparable corpora to find partial alignments.

The paper is organized as follows: Section 2 describes the method used to collect the bilingual information and obtain the word alignment; Section 3 explains the experimental framework; Section 4 shows the results obtained for the different features combination proposed; finally, the paper ends with some concluding remarks.

2 Methodology

The method presented here uses the available sources of bilingual information (SBIs) to detect parallel sub-segments in a given pair of parallel text segments S and T written in different languages. Once sub-segments have been aligned, a simple heuristic method is used to extract the most likely word alignments from S to T and from T to S. Finally, both alignments are symmetrised to obtain the word alignments.

Sub-segment alignment. To obtain the sub-segment alignments, both segments S and T are segmented in all possible ways to obtain sub-segments of length $l \in [1, L]$, where L is a given maximum sub-segment length measured in words. Let σ be a sub-segment from S and τ a sub-segment from T. We consider that σ and τ are aligned if any of the available SBIs confirm that σ is a translation of τ, or vice versa.

Suppose the pair of parallel segments *S=Costarà temps solucionar el problema*, in Catalan, and *T=It will take time to solve the problem*, in English. We first obtain all the

possible sub-segments σ in S and τ in T and then use machine translation (MT) as a SBI by translating the sub-segments in both directions. We obtain the following set of SSAs:

temps	$\leftrightarrow$	*time*
problema	$\leftrightarrow$	*problem*
solucionar el	$\rightarrow$	*solve the*
solucionar el	$\leftarrow$	*to solve the*
el problema	$\leftrightarrow$	*the problem*

It is worth noting that multiple alignments for a sub-segment are possible, as in the case of the sub-segment *solucionar el* which is both aligned with *solve the* and *to solve the*. In those cases, all the sub-segment alignments available are used. Figure 1 shows a graphical representation of these alignments.

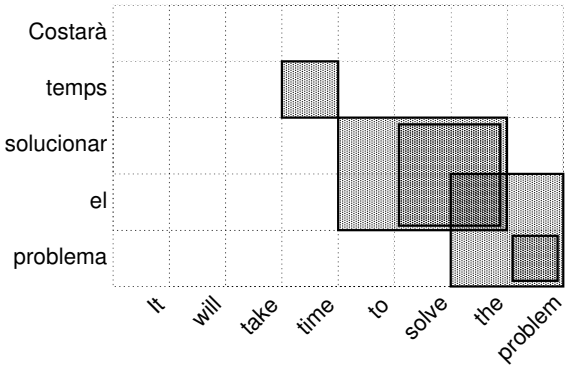

Figure 1: Sub-segment alignments.

Word alignment from sub-segment alignments. The information provided by the SSAs can then be used for word alignment. We define the *alignment strength* A_{jk} between the j-th word in S and the k-th word in T as

$$A_{jk}(S,T,M) = \sum_{(\sigma,\tau)\in M} \frac{\text{cover}(j,k,\sigma,\tau)}{|\sigma|\cdot|\tau|}$$

where M is the set of SSAs detected for the pair of parallel segments S and T, $|x|$ is the length of segment x measured in words, and $\text{cover}(j,k,\sigma,\tau)$ equals 1 if σ covers the j-th word in S and τ the k-th word in T, and 0 otherwise. This way of computing the alignment strengths is based on the idea that SSAs apply *alignment pressures* on the words; so the larger the surface covered by the SSA, the weaker the word-alignment strength obtained. Following our example, the alignment strengths for the words covered by the SSAs are presented in Figure 2. The words *temps* and *time* are only covered by a SSA (*temps*,*time*), so the surface is 1 and the alignment strength is $A_{1,4} = 1$. However, words *the* and *el* are covered by three SSAs: (*solucionar el*, *solve the*), (*solucionar el*, *to solve the*), and (*el problema*, *the problem*). So the alignment strength is $A_{3,6} = 1/4 + 1/6 + 1/4 = 2/3$.

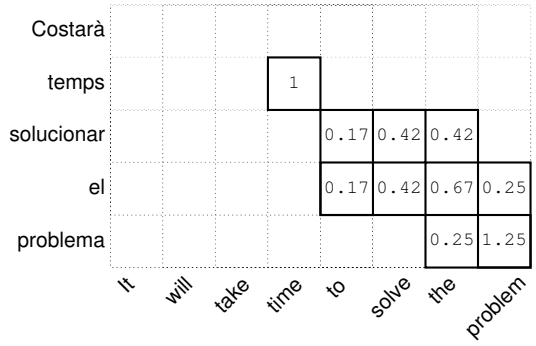

Figure 2: Alignment strengths.

The alignment strengths are then used to obtain word alignments. We simply align the j-th word in S with the k-th word in T if $A_{jk} > 0 \wedge A_{jk} \geq A_{jl}, \forall l \in [1, |T|]$, and vice versa. Note that one word in one of the segments can be aligned with multiple words in the other segment. Figures 3 and 4 show, respectively, the Catalan-to-English and the English-to-Catalan word alignments for the running example.

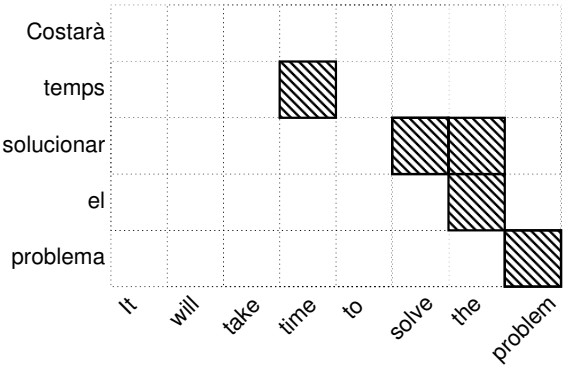

Figure 3: Resulting Catalan to English word alignment.

Figure 5 shows two possible symmetrised word alignments obtained by computing, in the first case, the intersection of the alignments shown in Figures 3 and 4, and, in the second case, the the widely-used *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003). It is worth noting that some words remain unaligned in Figure 5. This is a situation which can also be found in other state-of-the-art word alignment methods and, in this case, can be caused both by the symetrisation method, such as the word *to* in the alignment symetrysed through the intersection, or

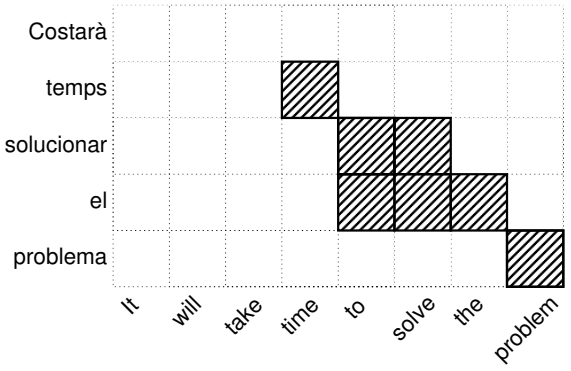

Figure 4: Resulting English to Catalan word alignment.

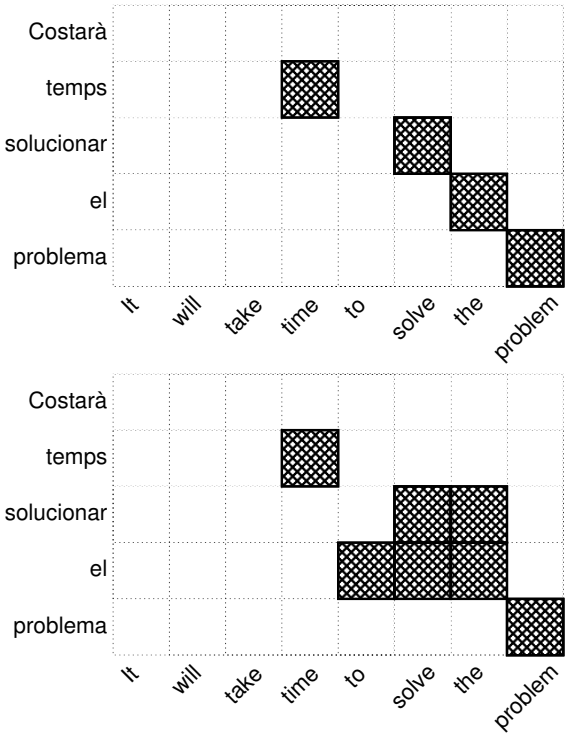

Figure 5: Two possible symmetrised word alignments, the first one using the intersection heuristic and the second one using the grow-diag-final-and heuristic.

by the lack of bilingual evidence relating the words, such as the words *Costarà*, *It*, *will*, and *take*. Depending on the needs of the task, more bilingual sources can be used in order to reduce the number of unaligned words. However, it is worth noting that unaligned words can also be caused by incorrect or excessively free translations, so keeping them unaligned may improve the overall alignment quality.

In addition, alignment strengths can be seen as a measure of the confidence on the relationships between the words. In future works, we plan to use the average alignment strength as a measure of the confidence on the SSAs. In this way, it could be possible to set a threshold to discard less-trusted SSAs. In the running example, the average alignment strength for the SSA (*solucionar el*, *to solve the*) is 0.37, whereas for the SSA (*el problema*, *the problem*) the average alignment strength is 0.60. Therefore, we see that (*el problema*, *the problem*) is a more reliable SSA than (*solucionar el*, *to solve the*).

3 Experimental setting

We evaluated the success of our system for word alignment using a *gold-standard* English–Spanish parallel corpus in which word alignments are annotated. We ran our method in both directions (Spanish to English and English to Spanish) and symmetrised the alignment obtained through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003) implemented in Moses (Koehn et al., 2007). We compared the performance of our system with that obtained by GIZA++ (Och and Ney, 2003) in different scenarios.

Test corpus. We used the test parallel corpus from the *tagged EPPS corpus* (Lambert et al., 2005) as a gold-standard parallel corpus.[7] It consists of 400 pairs of sentences from the English–Spanish Europarl (Koehn, 2005) parallel corpus and is provided with the corresponding gold-standard for word alignment. Two levels of confidence are defined for word alignments in this corpus, based on the judgement of the authors of the gold-standard: *sure* alignments and *possible* (less trusted) alignments.

Sources of bilingual information. We used three different MT systems as SBIs to translate the sub-segments from English into Spanish and vice versa:

- *Apertium*:[8] a free/open-source platform for the development of rule-based MT systems (Forcada et al., 2011). We used the English–Spanish MT system from the project's repository[9] (revision 34706).
- *Google Translate*:[10] an online MT system

[7] `http://gps-tsc.upc.es/veu/LR/epps_ensp_alignref.php3` [last visit: 2nd May 2012]

[8] `http://www.apertium.org` [last visit: 2nd May 2012]

[9] `https://apertium.svn.sourceforge.net/svnroot/apertium/trunk/apertium-en-es/` [last visit: 2nd May 2012]

[10] `http://translate.google.com` [last visit: 2nd May 2012]

by Google Inc (translations performed on 28th April 2012).

- *Microsoft Translator*:[11] an online MT system by Microsoft (translations performed on 27th April 2012).

Metrics. We computed the precision (P) and recall (R) for the alignments obtained both by our approach and by the baseline:

$$P = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{WA}|}$$

$$R = 100\% \cdot \frac{|\text{WA} \cap \text{GS}|}{|\text{GS}|}$$

where WA is the set of alignments obtained and GS is the set of alignments in the gold standard. Then, we combined both measures to obtain the F-measure:

$$F = \frac{2 \cdot P \cdot R}{P + R}$$

These three metrics were computed, only for the sure alignments and also for both sure and possible alignments.

Baseline. We compared the performance of our word-alignment method to that of GIZA++ (Och and Ney, 2003), a toolkit for word alignment which implements different statistical alignment strategies. We run GIZA++ in both directions (source to target and target to source) and then we combine both sets of alignments through the *grow-diag-final-and* heuristic (Koehn, Och, and Marcu, 2003).

GIZA++ is widely used for word-alignment in statistical MT. In this scenario, it is usually trained on the parallel corpus to be aligned. However, it is also possible to use pre-trained models to align new pairs of segments, in order to avoid training a new alignment model for each new alignment task. As our system is aimed at performing word alignment on the fly, we consider that the most adequate scenario to compare our approach with GIZA++ is using pre-trained alignment models to align the test corpus. Therefore, for a better comparison of our method to state-of-the-art techniques, we define two baselines. In the first one, henceforth *basic-GIZA++ baseline*, we train and run GIZA++ on the test corpus. In the second one, henceforth *pre-trained-GIZA++ baseline*, we train GIZA++ on a larger parallel corpus and use the resulting models to align the test corpus. To train the alignment models for the pre-trained-GIZA++ baseline, we used the parallel corpus from the News Commentary corpus distributed for the machine translation task in the Workshop on Machine Translation 2011.[12] This corpus was lowercased, tokenized and cleaned to keep only those parallel segments containing up to 40 words. After this process, we obtained a corpus of 126,419 pairs of segments.

[11] `http://www.microsofttranslator.com` [last visit: 2nd May 2012]

segs.	sure			sure ∪ possible		
	P	**R**	**F**	**P**	**R**	**F**
100	57.1	59.9	58.5	64.7	47.4	54.7
200	57.5	61.2	59.3	64.9	47.5	54.9
300	59.7	63.6	61.6	67.8	50.1	57.7
400	59.9	64.2	62.0	68.2	50.5	58.0

Table 2: Precision (P), recall (R), and F-measure (F) obtained by the basic-GIZA++ baseline for sure alignments, and for all sure and possible alignments when aligning the gold-standard corpus in portions of 100, 200, 300, and 400 pairs of segments (segs).

4 Results and discussion

Table 1 shows the results obtained by our system and both baselines based on GIZA++: the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

As can be seen, the method proposed in this paper obtains F-measures very similar to those obtained by both GIZA++-based baseline approaches. Another important detail is that our method obtains better precision in alignment than the two baselines proposed, although the results on recall obtained by the basic-GIZA++ baseline are better than ours.

Table 2 presents the results obtained by the basic-GIZA++ baseline when using portions of the test corpus with a different number of pairs of segments. The results presented in this table are useful to understand that, although the basic-GIZA++ yields slightly better results than the other approaches in Table 1, it clearly depends on the size of the parallel corpus to align. Of course, using this approach is not possible when trying to align a pair of segments on the fly, and obtains lower results when trying to align a very small set of parallel segments.

[12] `http://www.statmt.org/wmt11/translation-task.html`

Alignment kind	SBI-based approach			basic-GIZA++			pre-trained-GIZA++		
	P	R	F	P	R	F	P	R	F
sure	68.5%	57.6%	62.6%	59.9%	64.2%	62.0%	61.5%	55.8%	58.5%
sure ∪ possible	75.7%	43.9%	55.6%	68.2%	50.5%	58.0%	67.3%	42.2%	51.8%

Table 1: Precision (P), recall (R), and F-measure (F) obtained for the sure alignments, and also for all sure and possible alignments when aligning the gold-standard corpus. The results included correspond to our SBI-based approach and to both the basic-GIZA++ baseline and the pre-trained-GIZA++ baseline.

These results confirm that the approach proposed here can obtain alignments of a quality comparable to that obtained by the state-of-the-art GIZA++ tool, at least when trying to align small corpora, without needing any training process. These results set a bridge between the work of Esplà, Sánchez-Martínez, and Forcada (2011) and Esplà-Gomis, Sánchez-Martínez, and Forcada (2011), allowing to use SBI-based word alignment to help users to modify the translation proposals of a computer-aided translation system. It is worth noting that the weakness of our method is the recall, which may be improved by combining other SBIs.

5 Concluding remarks

In this work we have presented a new and simple approach for word alignment based on SBIs. This method can use any bilingual source of sub-sentential bilingual knowledge to align words in a pair of parallel segments on the fly. In this way, this process can be run without any training, which is useful in some scenarios, as is the case of computer-aided translation tools, in which word alignment can be used to guide translators when modifying the translation proposals (Kranias and Samiotou, 2004; Esplà, Sánchez-Martínez, and Forcada, 2011). In the experiments performed, our approach obtained results similar to those obtained by the state-of-the-art word-alignment GIZA++ tool. It is worth noting that the method proposed in this paper is a naïve approach which could be extended to obtain better results. Currently, we are evaluating new possibilities to improve the results obtained, such as using stemming or adding other SBIs available on-line.

In addition, we are developing a machine-learning-based approach which uses the ideas presented in this paper to perform word alignment in a more elaborate way, in order to improve the results obtained by the current approach. In this work we simply rely on the idea of *alignment pressures* to obtain the alignment strengths. However, it is possible to fit a maximum-entropy function, using a set of features obtained from the sub-segment alignments in order to obtain better alignment strengths. Although fitting the function would require a training process, once it is performed it could be applied to any new pair of segments on the fly. Another possible improvement may be to set weights for the different SBIs used for alignment, in order to promote those sources which are more reliable.

Acknowledgements: Work supported by the Spanish government through project TIN2009-14009-C02-01 and by Universitat d'Alacant through project GRE11-20. Google Translate service was provided by the *University Research Program for Google Translate*.

References

Al-Onaizan, Y. and K. Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 400–408, Philadelphia, Pennsylvania.

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

Dagan, I., K.W. Church, and W.A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*, pages 1–8, Columbus, USA.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. volume 39 of *Series B*. Blackwell Publishing, pages 1–38.

Esplà, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using word alignments to assist computer-aided translation users by marking which target-side words to change

or keep unedited. In *Proceedings of the 15th Conference of the European Association for Machine Translation*, pages 81–88, Leuven, Belgium.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2011. Using machine translation in computer-aided translation to suggest the target-side words to change. In *Proceedings of the 13th Machine Translation Summit*, pages 172–179, Xiamen, China.

Esplà-Gomis, M., F. Sánchez-Martínez, and M.L. Forcada. 2012. UAlacant: using online machine translation for cross-lingual textual entailment. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics*, pages 472–476, Montreal, Quebeq, Canada.

Forcada, M.L., M. Ginestí-Rosell, J. Nordfalk, J. O'Regan, S. Ortiz-Rojas, J. Pérez-Ortiz, F. Sánchez-Martínez, G. Ramírez-Sánchez, and F. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25(2):127–144.

Fung, P. and K. McKeown. 1997. Finding terminology translations from non-parallel corpora. pages 192–202.

Koehn, P. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X*, pages 79–86, Phuket, Thailand.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL*, pages 177–180, Prague, Czech Republic.

Koehn, P., F.J. Och, and D. 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*, pages 48–54, Edmonton, Canada.

Kranias, L. and A. Samiotou. 2004. Automatic translation memory fuzzy match post-editing: A step beyond traditional TM/MT integration. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 331–334, Lisbon, Portugal.

Lambert, P., A. De Gispert, R. Banchs, and J. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39(4):267–285.

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

Rapp, R. 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*, pages 519–526, College Park, USA.

Schulz, S., K. Markó, E. Sbrissia, P. Nohama, and U. Hahn. 2004. Cognate mapping: a heuristic strategy for the semi-supervised acquisition of a Spanish lexicon from a Portuguese seed lexicon. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Vogel, S., H. Ney, and C. Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proceedings of the 16th International Conference on Computational Linguistics*, pages 836–841, Copenhagen, Denmark.