A Gold Standard for English–Swedish Word Alignment

Maria Holmqvist and Lars Ahrenberg
Department of Computer and Information Science
Linköping University, Sweden
firstname.lastname@liu.se

Abstract

Word alignment gold standards are an important resource for developing and evaluating word alignment methods. In this paper we present a free English–Swedish word alignment gold standard consisting of texts from Europarl with manually verified word alignments. The gold standard contains two sets of word aligned sentences, a test set for the purpose of evaluation and a training set that can be used for supervised training. The guidelines used for English–Swedish alignment were created based on guidelines for other language pairs and with statistical machine translation as the targeted application. We also present results of intrinsic evaluation using our gold standard and discuss the relationship to extrinsic evaluation in a statistical machine translation system.

1 Introduction

Translated texts are rich sources of information about language differences and translation. A fundamental step in extracting translation information from parallel text is to perform word alignment and determine which words and phrases are translations of each other in the source and target texts. Word alignment forms the basis of (phrase-based) statistical machine translation (PBSMT) but alignments are also used in other data-driven approaches to machine translation to extract bilingual dictionaries and learn translation rules.

The task of identifying corresponding words in a parallel text is difficult and manual word alignment can be time-consuming. Unsupervised methods for automatic word alignment have dominated the machine translation field (Och and Ney, 2003), but an increasing amount of research is devoted to improving word alignment quality through supervised training (e.g., Ayan and Dorr, 2006; Blunsom and Cohn, 2006; Ittycheriah and Roukos, 2005). Supervised methods require a set of high quality alignments to train the parameters of a discriminative word alignment system. These alignments are often hand-made alignment gold standards. Gold standards are also an important resource for evaluation of word alignment accuracy.

In this paper, we present an English–Swedish word alignment gold standard. It consists of 1164 sentence pairs divided into a training set and a test set. The training set was produced to be used as training data for supervised word alignment (Holmqvist, 2010) and the test set was created for the purpose of word alignment evaluation. The test set alignments have confidence labels for ambiguous links in order to be able to calculate more fine-grained evaluation measures. The gold standard and alignment guidelines can be downloaded from http://www.ida.liu.se/~nlplab/ges . Alignments are stored in NAACL format (Mihalcea and Pedersen, 2003).

This paper is organized as follows. In Section 2 we review available gold standards for English–Swedish and compare them to our newly created resource. The selection of parallel texts is described in Section 3 and the guidelines for manual word alignment are motivated and exemplified in Section 4. We then review recent research on word alignment evaluation in Section 5. In Section 6 we use our gold standard reference alignment to compare intrinsic evaluation with extrinsic evaluation in a phrase-based statistical machine translation system. Finally, Section 7 contains conclusions and directions for future work.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 106–113

2 Related work

Gold standards consisting of parallel text with manually annotated word alignments exist for several language pairs including English–French (Och and Ney, 2003), Dutch–English (Macken, 2010) and English–Spanish (Lambert et al., 2005).

For some language pairs, parallel resources have been developed in the form of parallel treebanks. Parallel treebanks consist of parallel syntactic trees that have manual alignments between corresponding words and phrases as well as between subtrees. The added effort of verifying syntactic structure and aligning subtrees makes treebanks even more labor-intensive to produce than alignment gold standards. However, word alignments from large parallel treebanks such as the English-Arabic treebank from LDC are also used to train and evaluate word alignment systems, (e.g., Gao et al., 2010).

Currently, available resources for English–Swedish word alignment include two parallel treebanks, Smultron (Volk et al., 2009) and LinES (Ahrenberg, 2007). Smultron is a multi-lingual treebank consisting of 1500 sentences from three domains with subsentential alignments. The Smultron alignment guidelines are similar to our test data guidelines where two types of links are used, one for regular links and one for more fuzzy correspondences. LinES is an English–Swedish treebank containing 2400 sentences from four subcorpora. This treebank was primarily designed to investigate and measure the occurrence of translation shifts and the word alignments in LinES are sparse. Furthermore, LinES is not an open resource, but it can be queried through a web interface. Another resource of free parallel English–Swedish data is OPUS, an open source collection of multilingual parallel data with automatic sentence and word alignments (Tiedemann, 2009).

Our gold standard is a freely available resource designed for the purpose of improving word alignment for statistical machine translation. First of all, it has the advantage that it contains over 1000 sentences with full-text word alignments from a single domain. The Europarl domain was chosen since it is an open source corpus that is large enough for training an English–Swedish SMT system. By building translation systems from different alignments we can measure the impact of the alignment on translation quality and compare it to intrinsic measures of word alignment accuracy.

Furthermore, the alignment guidelines used for our gold standard work is based on a similar effort by Lambert et al. (2005) to produce a gold standard for English–Spanish word alignment for machine translation. Especially the test data in the gold-standard was created based on their findings on how to build reference alignments that will strengthen the correlation between word alignment accuracy and translation quality.

3 Text Selection

The parallel texts in the gold standard were taken from the English–Swedish part of the Europarl[1] corpus (Koehn, 2005) with texts collected between the years 1997-2003. The texts from the 4th quarter of 2000 were not included in our corpus since these texts are commonly used as test sets for machine translation evaluation.

The corpus was sentence aligned (Gale and Church, 1991) and sentences longer than 40 words were removed. This step removed 20% of the sentence pairs resulting in a parallel corpus with 704852 parallel segments and about 1,5 million words per language.

A random sample of 1200 sentence pairs from the first 20 000 sentences was divided into a training set of 1000 sentence pairs and a test set of 200 sentence pairs. About 3% of the sentence pairs were removed from the data because their sentence alignment was incorrect. Table 1 shows the final size and characteristics of the training and test corpora in terms of sentences, word tokens and word types.

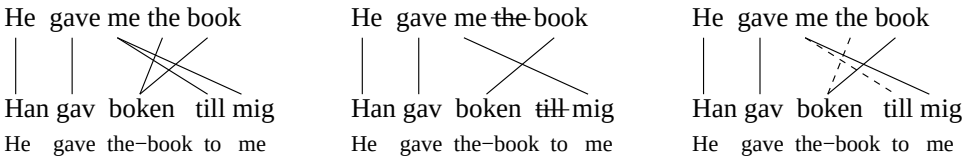
		English		Swedish	
Corpus	Size	Words	Types	Words	Types
Training	972	20340	3374	18343	4181
Test	192	4263	1332	3837	1395
Total	1164	24603	4706	22180	5576

Table 1: Corpus statistics for training and test data.

4 Manual Word Alignment

This section describes the manual word alignment process and presents guidelines for English–Swedish word alignment. Section 4.1 presents a range of factors that must be considered before settling on a set of word alignment guidelines and Section 4.2 and 4.3 presents the guiding principles for alignment of test and training data respectively.

[1] Europarl v. 2.0, http://www.statmt.org/europarl/archives.html.



(a) The Blinker guidelines (b) LinES guidelines (c) EPPS guidelines

Figure 1: Example of word alignment guidelines. Lines represent Sure links, dashed lines represent Possible links and strikethrough represents null links.

4.1 Alignment Guidelines

Word alignment is an ambiguous task for human annotators as there can be several plausible options for linking words in a given sentence pair. In order to produce a gold standard alignment that is consistent and yet fair to different alignment strategies, it is common to let several human annotators perform the alignment task independently based on a set of guidelines that explains the alignment strategy. A strategy for word alignment must contain decisions on many levels, including:

Size of alignment units. Alignments can be more or less oriented to lexical items instead of larger segments such as phrases. Phrase alignments are usually created by linking each word in the source phrase to each word in the target phrase.

Alignment purpose and coverage. The final purpose of the alignment will influence alignment strategy and affect the need for coverage. For example, some applications only require alignment of a pre-determined set of words (translation spotting) instead of full-text alignment (Véronis, 2000).

Confidence labels. A label can be attached to each word link to distinguish between *sure* and *possible* links.

Criteria for correspondence. Criteria for translation correspondence can be biased in favor of semantic or structural correspondence.

Untranslated items. Some word alignment guidelines include a special link type for untranslated words, a null link, while others let these words be unaligned.

We will illustrate different alignment strategies with the relatively simple sentence *He gave me the book – Han gav boken till mig* which has been aligned using three different guidelines: Blinker (Melamed, 1998), LinES (Ahrenberg, 2007) and the guidelines of Lambert et al. (2005) (henceforth referred to as EPPS).

The Blinker guidelines in Figure 1(a) aim to align as small segments as possible including as many words as necessary to achieve semantic correspondence. Blinker allows two types of links, regular links and null-links.

The same link types are used in the LinES guidelines in Figure 1(b) but in these guidelines one-to-one links are strongly preferred over many-to-many links and function words are null-linked if they lack a corresponding token in the other language.

The EPPS guidelines in Figure 1(c) incorporates both alignment strategies by labeling unambiguous correspondences as *sure* links while the function words without corresponding tokens are labeled as *possible* links. When sure and possible labels are used in an alignment reference, sure alignments should be more important than possible alignments during evaluation. The EPPS alignment would therefore be fair to systems that follow either Blinker or LinES alignment guidelines.

4.2 Test Data Alignment

Word aligned test data is used as a reference when evaluating the quality of computed alignments. The guidelines for aligning English–Swedish reference data are based on the EPPS guidelines which are adapted to the task of producing full-text reference word alignments for alignment evaluation and machine translation (Lambert et al., 2005).

The basic correspondence criterion for English–Swedish word alignment follows the definition in Lambert et al. (2005) that "the correspondence between two lexical units should involve on both sides as few words as possible but as many words as necessary, with the requirement that the linked words or groups bear the same meaning." Correspondences between multiword units are created by linking each word in the source segment to each word in the target segment. The EPPS guidelines adds a confidence label to each word link in the

reference where alignments labeled *sure* (S) are obligatory while alignments labeled *possible* (P) are acceptable alignments during evaluation. As shown in the previous section, confidence labels ensure that the reference alignment is reasonably fair to different alignment strategies.

The *alignment error rate* (AER) (Och and Ney, 2003) is a common evaluation measure for word alignment that takes the confidence label of the reference links into account when error rate is calculated (See section 5 for more details). Lambert et al. (2005) show that a large proportion of possible links in the reference will lead to an AER that favours high precision alignments. Since recall is just as, if not more, important than precision for statistical machine translation, the EPPS guidelines are designed to create reference alignments with a large proportion of sure links in order to increase the importance of alignment recall. Unlike the EPPS guidelines we also use explicit *null links* to mark words and phrases that have no translation in the other language.

The proper use of sure and possible links in our guidelines are illustrated by Figures 2 and 3. As a rule, two words or phrases correspond if they are semantically and structurally equivalent and alignments should be kept as fine-grained as possible. A word link is sure if the correspondence meets both semantic and syntactic criteria and possible if only one criterion is met, when a correspondence is uncertain or if a word has many alignment possibilities.

For example, Figure 2(a) contains a word-by-word translation of the noun phrase *the red car* annotated with S links. The noun phrase in Figure 2(b), however, lacks a Swedish lexical item corresponding to the definite article *the*, and since the definiteness instead is expressed with a Swedish definite suffix, the article is linked to the noun with a P link. In short, function words should be S linked to corresponding function words if possible. However, in cases where the syntactic function is expressed by a content word, a P link should be used between the function and content words. In Figure 2(c), for example, *om* has an attributive function and is P-linked to the English attribute *threshold*.

P links can also be used when two content words correspond on a structural level but not on the semantic level such as the words *worst* and *större* (*Eng.* larger) in Figure 3. These words correspond in the given sentence but they might not work well as translations of each other in another context.

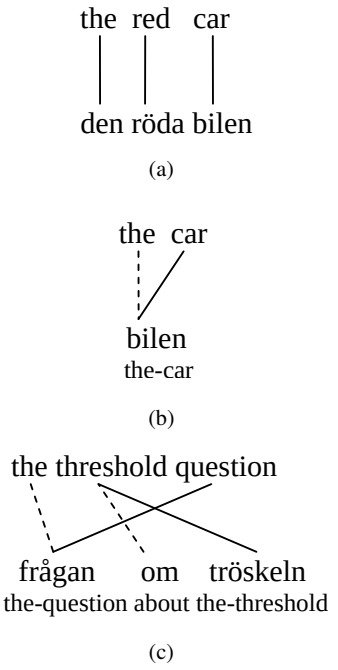


Figure 2: Noun phrase alignments.

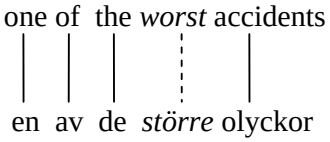


Figure 3: P-linked content words.

4.2.1 Inter-annotator Agreement

Two annotators, the authors of this paper, were given the task of drawing links between corresponding words and to give at least one link to all words in the sentence pair, using S, P and null links according to the guidelines. All links from both alignments were added to the final reference alignment. If annotators disagreed on a link or on the label of a link, that link was labeled P in the reference. If a word was null-linked by one annotator and linked to a word by the other annotator, the null link was removed and the word link was labeled P in the reference. The final reference alignment contains a large proportion of sure links (73% if not counting null-links). The proportion of each link type is shown in Table 2.

The alignment consistency between the two annotators was relatively high. Inter-annotator agreement shown in Table 3 was calculated as $\text{AGR} = 2 * I/(A1 + A2)$ where $A1$ and $A2$ is the sets of links created by the first and second annotator and I is the intersection of these sets. The agreement between A1 and A2 was 85.8% for labeled agreement and 91.3% for unlabeled agreement. Possible links were the link type on which annotators disagreed the most.

Reference	%S	%P	%Null	Links
A1	70.1	15.1	14.8	5527
A2	68.3	13.5	18.2	5086
Final	63.6	23.5	12.9	5254

Table 2: Alignment data.

Links	% Agreement
s	93.7
p	55.2
n	79.6
s+p (unlabeled)	87.1 (93.7)
s+p+n (unlabeled)	85.8 (91.3)

Table 3: Inter-annotator agreement per link type.

Ref	Alignment	F-score	AER
A2	A1	93.67	5.78
A1	A2	93.67	3.26
Final	A1	98.58	0.00
Final	A2	95.23	0.00

Table 4: Evaluation of each annotator.

We also evaluated each annotation with the final gold standard as reference (ignoring null-links). The result in Table 4 shows that these evaluations resulted in an alignment error rate of 0, which makes sense since both are perfectly valid human alignments. Table 4 also contains the result of evaluating each annotation with the other as reference which resulted in very low error rates.

4.3 Training Data Alignment

The training data guidelines were written to produce word alignments that cover as much of the text as possible while preserving the semantic and syntactic correspondence between aligned segments. The alignment followed the same principles as the test data alignment except that no confidence labels were used. A link is either a regular link or a null link. Words are the preferred unit of alignment and phrase alignments should be as small as possible. Training data alignment guidelines are similar to the Blinker guidelines although null links are used more often to avoid large phrase alignments. The training data was aligned by one annotator.

4.4 Alignment Annotation Tools

The manual alignments were produced using two tools, I*Link (Ahrenberg et al., 2003) and the Alpaco_sp editor[2] (Lambert et al., 2005). I*Link is a a tool for interactive word alignment that simplifies the manual alignment process by suggesting alignments which the user can choose to either accept or reject. The user can also override suggested alignments and add new ones. This method of alignment is relatively fast but I*link can not distinguish between possible and sure links, and it does not allow alignment of discontinuous segments such as when the two parts of a particle verb *låser upp* (*Eng*. unlock) are separated by a pronoun as in *låser* du *upp* – you *unlock*.

The Alpaco_sp alignment editor is a tool for manual word alignment with confidence labels. After aligning sure and continuous segments with I*Link, annotators used Alpaco to refine the alignments by adding possible links and links for discontinuous segments where appropriate.

5 Word Alignment Evaluation

Word alignment quality is evaluated either intrinsically by comparing alignments to a reference alignment or extrinsically by measuring the impact of the alignments in an application. In intrinsic evaluation, standard measures of precision (1), recall (2) and F-measure (3) are calculated by comparing a set of computed alignments A to a set of gold standard alignments G.

$$\text{Precision}(A,G) = \frac{|G \cap A|}{|A|} \quad (1)$$

$$\text{Recall}(A,G) = \frac{|G \cap A|}{|G|} \quad (2)$$

$$\text{F-measure}(P,R) = \frac{2PR}{P+R} \quad (3)$$

Different weights can be assigned to precision and recall when calculating F-score by varying

[2] http://gps-tsc.upc.es/veu/personal/lambert/scripts/alpaco_sp.tgz

α in the general formulation of the F-measure, shown in (4).

$$\text{F-measure}(A,G,\alpha) = \frac{1}{\frac{\alpha}{Precision(A,G)} + \frac{(1-\alpha)}{Recall(A,G)}} \quad (4)$$

Setting $\alpha = 0.5$ results in the standard balanced F-measure that gives equal weight to precision and recall. A lower α-constant will weight recall higher and a larger constant will favor high precision.

Alignment Error Rate (5) is a quality measure that uses the confidence labels in the gold standard (Och and Ney, 2003). It takes into account the fact that sure links (S) should be more important to get right than possible links (P) when calculating alignment accuracy. It is based on a different formulation of precision and recall, where recall errors only can be made if the computed alignment lacks a sure link (6) and precision errors only when a computed link is not even a possible link in the reference (7). Sure links are by definition also possible.

$$\text{AER}(A,P,S) = 1 - \frac{|S \cap A| + |P \cap A|}{|S| + |A|} \quad (5)$$

$$\text{Recall}(A,S) = \frac{|S \cap A|}{|S|} \quad (6)$$

$$\text{Precision}(A,P) = \frac{|P \cap A|}{|A|} \quad (7)$$

5.1 Word Alignment and SMT

Researchers in statistical MT want to improve word alignment in order to produce better translations. However, several studies have shown that improvements in terms of AER often fail to result in improved translation quality, (e.g., Ayan and Dorr, 2006; Fraser and Marcu, 2006). Translation quality can be measured in terms of the Bleu metric (Papineni et al., 2001). One reason for this lack of correlation between intrinsic and extrinsic evaluation measures is that AER favours high-precision alignments. Fraser and Marcu (2006) found that although precision is important for translation systems trained on small corpora, the importance of recall increases as the amount of data grows and alignment quality improves. In a standard PBSMT system, word alignments control which phrases are extracted as possible translations. A sparse, high-precision alignment is more ambiguous and phrase extraction heuristics will extract more alternative phrase translations. Especially for systems trained on small corpora the many alternative translations in the phrase table seem to be beneficial to translation quality (Lambert et al., 2009).

The connection between word alignment and phrase extraction suggests that other alignment characteristics than alignment precision and recall might be important for extracting the right phrases. For instance, correctly aligned discontinuous phrases such as German particle verbs can prevent the extraction of useful phrases from embedded words and removing these (correct) links improved translation quality (Vilar et al., 2006).

A better correlation between intrinsic measures of alignment quality and translation quality have been found by having a large proportion of S links in the reference (Lambert et al., 2005) or by only having S links (Fraser and Marcu, 2006). In addition, Fraser and Marcu, achieved better correlation for Arabic–English and French–English when using the general F-measure weighed in favor of recall as the intrinsic measure instead of AER.

6 Experiments on English–Swedish Europarl

In this section we use our gold standard to compare intrinsic alignment quality measures with translation quality for PBSMT systems built on the English–Swedish Europarl corpus. Our aim is to investigate how well alignment quality metrics and translation quality correlate for this corpus and how variables such as corpus size and translation direction affect the correlation.

Our alignment and translation experiments were performed on two corpora of different size, a small corpus containing 100K sentence pairs and a large corpus of 700K sentence pairs. We used the Giza++ word alignment system (Och and Ney, 2003) to create four alignments for each corpus with varying precision and recall. The four alignments were produced using different heuristics to create a single alignment from the source-to-target and target-to-source alignments produced by Giza++. The alignment with highest precision takes the *intersection* (I) of links from the two alignments, the alignment with highest recall takes the *union* (U), and heuristics *grow-diag* (GD) and *grow-diag-final* (GDF) create alignments from the intersection and add links from the union to increase alignment recall.

	small						large					
					Bleu%						Bleu%	
Align	P	R	F	AER	en-sv	sv-en	P	R	F	AER	en-sv	sv-en
I	95.6	56.4	71.0	16.3	22.9	28.3	96.5	60.8	74.6	12.9	23.7	30.0
GD	83.2	73.7	78.2	15.0	23.1	28.5	85.3	76.9	80.9	12.6	24.7	30.7
GDF	73.7	77.4	75.5	19.5	22.8	28.3	77.6	79.5	78.6	16.2	24.7	30.6
U	69.9	78.4	73.9	21.7	22.4	27.9	75.0	80.7	77.7	17.5	24.9	30.3

Figure 4: Intrinsic and extrinsic evaluation of Swedish–English word alignment.

Alignments were evaluated against the 192 sentences in the gold standard test set. To investigate the correlation between intrinsic quality measures and machine translation quality on our corpus, we built standard phrase-based SMT systems using Moses (Koehn et al., 2007), one for each alignment and translation direction, resulting in eight systems for each corpus size.

System parameters were tuned using a development set of 1000 sentence pairs. Each system was evaluated on a test set of 2000 sentences and translation quality was measured in Bleu. Table 4 contains alignment quality scores precision, recall, F-measure and AER and Bleu scores for each translation system.

The table shows that different alignments generally have a small effect on Bleu score. The change is 0.6-0.7 Bleu points for the small systems and 0.7-1.2 Bleu points for the large systems.

The alignment heuristic with the best AER (*grow-diag*) produces the best translation in most systems, but the correlation between AER and Bleu is not strong for all conditions and alignment heuristics. Table 5 shows the correlation between $1-AER$ and Bleu measured by the Pearson product-moment correlation coefficient, r. This correlation is quite strong for the small dataset ($r = 0.92$ and $r = 0.84$), but negative for the larger dataset ($r = -0.59$ and $r = -0.01$). This is consistent with earlier findings that high-precision alignments which are favored by the AER measure tend to result in better translation quality when systems are trained on smaller corpora.

Higher correlation have been reported between F-score and Bleu. Fraser and Marcu (2006) found the highest correlation by adapting the precision/recall weights in the F-measure to different corpora sizes and language pairs.

To find the optimal weights of precision and recall for our data set we set α in (4) to different values in the range $0.1, .., 0.9$. Table 5 shows the α that results in the best correlation with Bleu for each system. For the small dataset, the best correlation was found with a constant of 0.6 and for the large data set the best constant was 0.1 and 0.5 respectively. This also supports the hypothesis that precision is more important to systems trained on small corpora while recall is more important for systems trained on large corpora.

There are also differences in the optimal balance between precision and recall between the two translation directions for the system trained on the large corpus. Translation from English to Swedish seems to benefit from higher alignment recall, while the quality of Swedish to English translation depends more equally on precision and recall.

		Correlation r			
Corpus		Best α	F_α	$F_{0.5}$	1-AER
small	en-sv	0.6	0.91	0.45	0.92
	sv-en	0.6	0.87	0.48	0.84
large	en-sv	0.1	0.99	0.80	-0.59
	sv-en	0.5	0.99	0.99	-0.01

Table 5: Correlation between measures of word alignment accuracy and Bleu.

7 Conclusion

We have presented a freely available gold standard for English–Swedish word alignment which can be used to train and evaluate word alignment systems. We described the alignment guidelines for manual annotation that we developed for Swedish–English word alignment which were based on previous research in producing gold standards for other languages for the purpose of statistical machine translation.

In addition, we showed how the gold standard reference can be used to evaluate different word alignment methods and compared it to an external evaluation in a statistical machine translation system. We measured the correlation between alignment quality metrics and translation quality.

Our results support the findings for other language pairs that recall plays a more important role for MT systems trained on large corpora, while precision is more important for systems trained on smaller corpora. However, in the translation direction Swedish–English translation quality was not as dependent on alignment recall. We believe this observation warrants further investigation using a larger sample of alignments.

We also plan to investigate the relationship between alignment and translation by measuring other characteristics of the alignment which may affect translation quality, such as aligned word types and the number of discontinuous links.

References

Lars Ahrenberg, Magnus Merkel, and Michael Petterstedt. 2003. Interactive word alignment for language engineering. In *Proceedings of EACL 2003*, pages 49–52, Budapest, Hungary.

Lars Ahrenberg. 2007. LinES: An English-Swedish parallel treebank. In *Proceedings of Nodalida 2007*, pages 270–273, Tartu, Estonia.

Necip Fazil Ayan and Bonnie J. Dorr. 2006. A maximum entropy approach to combining word alignments. In *Proceedings of HLT-NAACL 2006*, pages 96–103, Morristown, NJ, USA.

Phil Blunsom and Trevor Cohn. 2006. Discriminative word alignment with conditional random fields. In *Proceedings of COLING-ACL 2006*, pages 65–72, Sydney, Australia.

Alexander Fraser and Daniel Marcu. 2006. Semi-supervised training for statistical word alignment. In *Proceedings of COLING-ACL 2006*, pages 769–776, Sydney, Australia.

William A. Gale and Kenneth W. Church. 1991. A program for aligning sentences in bilingual corpora. In *Proceedings of ACL 1991*, pages 177–184, Berkeley, California, USA.

Qin Gao, Nguyen Bach, and Stephan Vogel. 2010. A semi-supervised word alignment algorithm with partial manual alignments. In *Proceedings of WMT and MetricsMATR*, pages 1–10, Uppsala, Sweden.

Maria Holmqvist. 2010. Heuristic word alignment with parallel phrases. In *Proceedings of LREC 2010*, Valletta, Malta.

Abraham Ittycheriah and Salim Roukos. 2005. A maximum entropy word aligner for Arabic–English machine translation. In *Proceedings of HLT-EMNLP 2005*, pages 89–96, Vancouver, Canada.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL 2007, demo session*, Prague, Czech Republic.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, Phuket, Thailand.

Patrik Lambert, Adrià de Gispert, Rafael Banchs, and José B. Mariño. 2005. Guidelines for word alignment evaluation and manual alignment. *Language Resources and Evaluation*, 39:267–285.

Patrik Lambert, Yanjun Ma, Sylwia Ozdowska, and Andy Way. 2009. Tracking relevant alignment characteristics for machine translation. In *Proceedings of MT Summit XII*, Ottawa, Canada.

Lieve Macken. 2010. An annotation scheme and gold standard for Dutch-English word alignment. In *Proceedings of LREC 2010*, Valletta, Malta.

I. Dan Melamed. 1998. Annotation style guide for the Blinker project, version 1.0. Technical report, University of Pennsylvania.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *HLT-NAACL 2003 Workshop: Building and Using Parallel Texts*, pages 1–10, Edmonton, Canada.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of ACL 2001*, pages 311–318, Philadelphia, Pennsylvania, USA.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In *RANLP*, vol. V, pages 237–248. Borovets, Bulgaria.

David Vilar, Maja Popović, and Hermann Ney. 2006. AER: Do we need to "improve" our alignments? In *Proceedings of IWSLT 2006*, pages 205–212, Kyoto, Japan, November.

Martin Volk, Torsten Marek, and Yvonne Samuelsson. 2009. SMULTRON (version 2.0) - The Stockholm MULtilingual parallel TReebank. An English-German-Swedish parallel Treebank with sub-sentential alignments.

Jean Véronis. 2000. Evaluation of parallel text alignment systems: the ARCADE project. In *Parallel text processing: Alignment and use of translation corpora*, pages 369–388. Dordrecht: Kluwer Academic Publishers.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955