Improving Phrase-Based Statistical Translation Through Combination of Word Alignments

Boxing Chen and Marcello Federico

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica
Via Sommarive 18, 38050 Povo (Trento), Italy
{boxing, federico}@itc.it

Abstract. This paper investigates the combination of word-alignments computed with the competitive linking algorithm and well-established IBM models. New training methods for phrase-based statistical translation are proposed, which have been evaluated on a popular traveling domain task, with English as target language, and Chinese, Japanese, Arabic and Italian as source languages. Experiments were performed with a highly competitive phrase-based translation system, which ranked at the top in the 2005 IWSLT evaluation campaign. By applying the proposed techniques, even under very different data-sparseness conditions, consistent improvements in BLEU and NIST scores were obtained on all considered language pairs.

1 Introduction

The recent years have seen a growing interest in Statistical Machine Translation (SMT). Besides its very competitive performance, a reason for its popularity is also the availability of public software to develop SMT components. A notable advance in this direction was the release of the GIZA++ tool [1], which implements the quite tricky word-alignment models introduced by IBM [2] in the early 90s, plus a few other models. Currently, most state-of-the-art SMT systems are trained on parallel texts aligned with GIZA++.

In general, phrase-based SMT [3] exploits IBM word-alignments computed in both directions, i.e. from source to target words and vice versa. Hence, a combination of the two alignments is taken, and phrase pairs are extracted from it. Up to now, this approach has proved to be successful over a range of tasks and language pairs.

Alternative word-alignment models have been recently proposed which are simpler and much faster to compute [4,5,6,7]. However, up to now, experimental comparisons between such models and the well established IBM models have only addressed the accuracy of the resulting word-alignments and not their impact on translation performance[8]. In fact, in several venues it has been argued whether alignment accuracy is indeed a good indicator of translation accuracy.

The original contribution of this work is the combined use of different word-alignment methods within a state-of-the-art phrase-based SMT system. More specifically, we focus on the comparison of translation performance of different

T. Salakoski et al. (Eds.): FinTAL 2006, LNAI 4139, pp. 356–367, 2006.

word alignments generated under a widely used IBM-model setting and with the *competitive linking algorithm* (CLA) proposed by [4]. Briefly, the CLA computes an association score between all possible word pairs within the parallel corpus, and then applies a greedy algorithm to compute the best word-alignment for each sentence pair. The algorithm works under the one-to-one assumption, i.e. each source word is aligned to one target word only, and vice versa.

Experiments were conducted on data from the BTEC corpus, which are distributed by the International Workshop on Spoken Language Translation - IWSLT [9,10]. In particular, translation into English from a variety of source languages was considered: Chinese, Arabic, Japanese, and Italian. For all language pairs, a standard training condition of 20K sentence pairs was assumed, which corresponds to the core tracks of the 2005 IWSLT Evaluation Campaign. For Italian and Chinese, training with larger amounts of data was also investigated, namely up to 60K and 160K sentence pairs, respectively.

This paper is organized as follows. Section 2 presents our phrase-based SMT framework, including IBM word-alignment settings, and the phrase-pair extraction method. Section 3 reviews the competitive linking algorithm and the adopted associative score. Section 4 and 5, respectively, present and discuss the experimental results. Section 6 is devoted to conclusions.

2 Phrase-Based SMT

In phrase-based translation, words are no longer the only units of translation, but they are complemented by strings of consecutive words, the phrases.

Our phrase-based system [11] is based on a log-linear model which extends the original IBM Model 4 [2] to phrases. The output translation for a given source sentence $\mathbf{f}$ is computed through a dynamic-programming beam-search algorithm [12] which maximizes the criterion:

$$\tilde{\mathbf{e}}^* = \arg\max_{\tilde{\mathbf{e}}} \max_{\mathbf{a}} \sum_{r=1}^{R} \lambda_r h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a}),$$

where $\tilde{\mathbf{e}}$ represents a string of phrases in the target language, $\mathbf{a}$ an alignment from the words in $\mathbf{f}$ to the phrases in $\tilde{\mathbf{e}}$, and $h_r(\tilde{\mathbf{e}}, \mathbf{f}, \mathbf{a})$ $r = 1, \ldots, R$ are *feature functions* designed to model different aspects of the translation process. In particular, feature functions are defined around the following steps of the search algorithm, which progressively add phrases $\tilde{e}$ to the target string, by covering corresponding source phrases (see Figure 1): the *permutation model*, which sets the position of the first word of the next source phrase to cover; the *fertility model*, that establishes its length; the *lexicon model* which generates target translations $\tilde{e}$; the *language model*, which measures the fluency of $\tilde{e}$ with respect to its left context. Notice that according to our model target phrases might have fertility equal to zero, hence they do not translate any source word. Moreover, uncovered source positions can be associated to a special target word (*null*), according to specific fertility and permutation models.

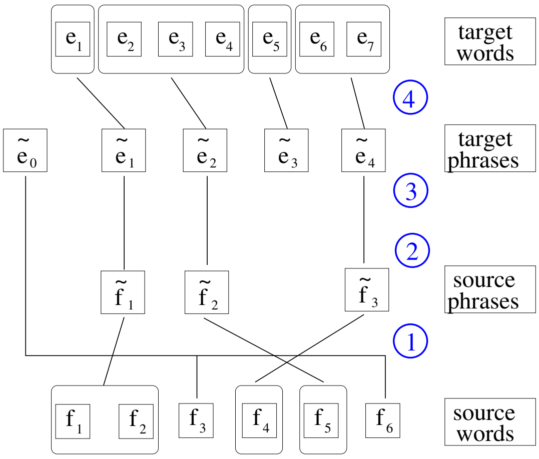


Fig. 1. Phrase-based SMT. Feature functions used in the translation process: (1) permutation model, (2) fertility model, (3) lexicon model, (4) language model.

In order to reduce the computation complexity of the search algorithm, constraints on phrase re-ordering are applied. In particular, if re-ordering is not permitted at all we have so-called monotone search, otherwise we have non-monotone search.

The resulting log-linear model has eight feature functions, whose parameters are either estimated from data or empirically fixed. In particular, fertility and lexicon models exploit relative frequencies computed on a sample of *phrase pairs* extracted from a parallel corpus. A detailed description of these features can be found in [13]. The scaling factors λ_i of the log-linear model are estimated on a development set, by applying a *minimum error training* procedure [14].

2.1 Phrase-Pair Extraction

Phrase pairs are collected from a parallel corpus containing sentence pairs $(\mathbf{f}, \mathbf{e})$ provided with some word alignment $\mathbf{c}$. For each sentence pair, all phrase-pairs are extracted corresponding to sub-intervals of the source and target positions, J and I, such that the alignment $\mathbf{c}$ links all positions of J into I and vice versa (links to the null word are disregarded). In the experiments, phrases were extracted with maximum length in the source and target set to 8.

In this work, we propose three methods to compute the alignment $\mathbf{c}$: the union of direct and inverse IBM alignments, the intersection of direct and inverse IBM alignments with expansion [15], and the competitive linking algorithm.

2.2 IBM Word-Alignment

IBM models use a many-to-one alignment scheme, i.e. each word in the source sentence is mapped to exactly one word in the target sentence. For the sake of phrase-extraction, alignments from source to target and from target to source are computed.

IBM alignments in both directions were computed through the GIZA++ toolkit [8].

3 Competitive Linking Algorithm

The competitive linking algorithm [4] works under the one-to-one assumption, i.e. each source word can be aligned to one target word only, and vice versa. An association score is computed for every possible translation pair, and a greedy algorithm is applied to select the best word-alignment. Alignment quality strongly depends on the association score. Several scores for word-pairs have been proposed in the literature, such as Mutual information, t-score, Dice coefficient, χ^2, log-likelihood ratio, etc. In this paper, we use a log-linear combination of two probabilities, as suggested in [6]: the first addresses the co-occurrence of word pairs, the other their position difference.

The first probability is defined as follows. Given two words f and e, with joint frequency n_{ef} and marginal frequencies n_f and n_e, we compute the probability that f and e co-occur just by chance with the hyper-geometric distribution

$$P_{cooc}(f,e) = \frac{\binom{n}{n_{ef}}\binom{n-n_f}{n_e-n_{ef}}}{\binom{n}{n_e}}$$

where n indicates the number of sentence pairs in the training corpus. For each word, only one occurrence per sentence is taken into account, as suggested in [4].

The probability considers the chance of observing a certain position difference between two randomly drawn positions inside two sentences of equal lengths. Hence, assuming the source and target sentences have lengths m and l, respectively, the normalized position difference between words f_j and e_i is computed by:

$$dist(j,i) = \left| j - i \cdot \frac{m}{l} \right|$$

Probabilities of observing any distance values for two randomly drawn positions were pre-computed for a fixed length $L = 50$ and tabulated as follows:

$$P_{pos}(dist) = \begin{cases} 7/L & \text{if } dist \leq 3 \\ 4/L & \text{if } 3 < dist \leq 5 \\ 1 - 11/L & \text{if } 5 < dist \end{cases}$$

The two probabilities are log-linearly combined with empirically determined weights:

$$S(f_j, e_i) = -\log P_{cooc}(f_j, e_i) + 4 \log P_{pos}(dist(j,i))$$

Notice that the negative logarithm is taken for the first score, as a small probability corresponds to a strong association score.

Table 1. Statistics of training, development and testing data used for the IWSLT 2005 supplied data condition. For Italian-English a comparable set was collected.

		IWSLT 2005				Italian-English	
		Chinese	Arabic	Japanese	English	Italian	English
Train Data	Sentences	20,000				20,000	
	Running words	173K	171K	159K	181K	149K	155K
	Vocabulary	8,536	9,251	18,150	7,348	9,611	6,885
Dev. Data	Sentences	500			500 × 16	100	100 × 16
	Running words	3,860	3,538	3,359	64,884	788	14,001
Test Data	Sentences	506			506 × 16	506	506 × 16
	Running words	3,514	3,531	3,259	65,616	3,574	65,615

Computing alignments of the training data with the CLA requires $\mathcal{O}(n\ m\ l)$ operations for the scoring function, and $\mathcal{O}(n\ m\ l\ \log m\ l)$ operations to align the corpus, where m and l indicate the lengths of the longest source and target sentences.

4 Training Modalities

Four training modalities for our phrase-based SMT system have been investigated which either change the way word-alignments are estimated or the way phrase-pairs are generated.

IBM Union

It represents the baseline modality: direct and inverse word alignments are computed by means of IBM models and successively phrase-pairs are extracted from the union of the two alignments.

IBM Intersection

Starting from the intersection alignemnt **c** of the direct and inverse IBM word alignments, additional links (i,j) are iteratively added to **c** if they satisfy the following criteria: a) links (i,j) only occur in the direct or inverse IBM alignment; b) they already have a neighbouring link in **c** or both of the words f_j and e_i are not aligned in **c**. Phrase-pairs are then extracted from the new alignment.

CLA

Word alignments are computed with the competitive linking algorithm and phrase-pairs are extracted from them.

Inter+CLA

Phrase-pairs obtained from the previous two methods (IBM Intersection and CLA) are joined.

Table 2. Examples of English sentences in the BTEC task

I'd like to take a sightseeing tour.
Do you have any of these?
Do you have travel accident insurance?
Take this baggage to the JAL counter, please.
How do you eat this?

Table 3. Statistics of extended BTEC data

Training Data	Chinese	English
Sentences	160,000	
Running words	1,106K	1,154K
Vocabulary	15,222	13,043
Training Data	Italian	English
Sentences	60,000	
Running words	463K	480K
Vocabulary	15,775	10,828

5 Experiments

5.1 Translation Tasks and Data

Experiments were carried out on the Basic Traveling Expression Corpus (BTEC) [16]. BTEC is a multilingual speech corpus that contains translation pairs taken from phrase books for tourists. We conducted experiments on four language pairs: Chinese-English, Japanese-English, Arabic-English and Italian-English. For the first three language pairs, we used data sets distributed for the IWSLT 2005 Evaluation Campaign[1], corresponding to the so-called *supplied data* evaluation condition. For Italian we used an equivalent test-suite kindly made available by the C-STAR Consortium[2], which will be distributed for IWSLT 2006. For each source sentence of the development and test sets, 16 references are available. Detailed statistics of training, development, testing data are reported in Table 1. A few examples of English sentences occurring in the test set are shown in Table 2.

To perform experiments under different data sparseness conditions, additional parallel texts available through the C-STAR Consortium were used as well. These extend the Italian-English and Chinese-English texts up to 60K and 160K sentence pairs, respectively. Statistics of the extended data are reported in Table 3. In Figure 2 vocabulary size is plotted for each language against increasing amounts of training data. Notice, that the different vocabulary-growth curves of Italian and Chinese are mainly due to different strategies used to create the Italian-English and Chinese-English corpora, rather than to intrinsic properties of the two languages.

[1] http://www.is.cs.cmu.edu/iwslt2005/

[2] httt://www.c-star.org

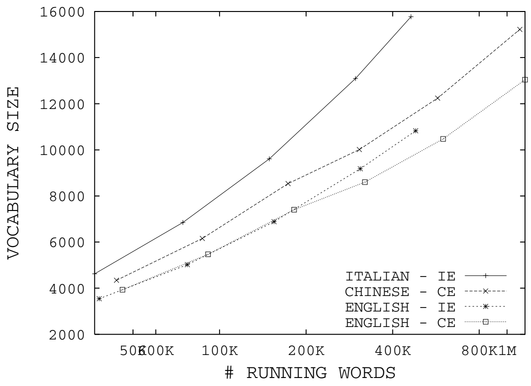


Fig. 2. Vocabulary growth in the extended BTEC data

Table 4. BLEU% scores and NIST scores under different training conditions

Language	Chinese		Japanese		Arabic		Italian	
	BLEU	NIST	BLEU	NIST	BLEU	NIST	BLEU	NIST
IBM Union	38.88	7.411	42.52	7.731	58.23	8.880	62.20	9.846
CLA	39.41	7.457	45.96	7.770	57.26	8.977	62.38	9.822
IBM Inter.	41.26	7.387	46.59	7.778	59.05	8.925	63.18	9.842
Inter+CLA	41.93	7.492	47.76	7.858	59.79	9.191	63.92	9.853

Before the experiments, some pre-processing was applied to the texts. Arabic, Chinese and Japanese characters were converted into a full ASCII encoding. Even if Chinese texts were provided with a manual segmentation at the word level, they were re-segmented with an in-house tool, trained from the original segmentation. We found that this permits the smoothing of inconsistencies in the manual segmentation. All texts were finally tokenized and put in lower case.

The search algorithm was configured similarly for all language pairs. Non-monotone search was applied for all languages, with less re-ordering allowed for Italian than for all other source languages.

Translation performance is here reported in terms of BLEU [17] score and NIST[3] score (case insensitive with punctuation).

5.2 Experimental Results

First experiments evaluated the different training modalities on all four language pairs. All experiments used the same amount of training data, i.e. 20K sentence pairs. Results are reported in Table 4.

The comparison between phrase-based training with IBM union alignments and CLA alignments shows that it is hard to say which alignment performs absolutely better. IBM union alignment works better on Arabic-English, but

[3] http://www.nist.gov/speech/tests/mt/

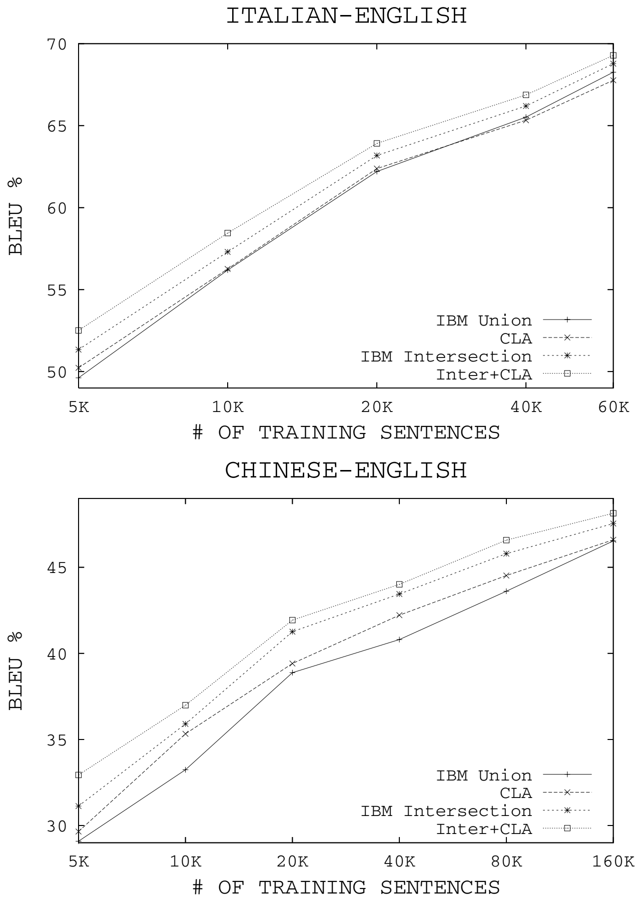


Fig. 3. Performance of training modalities against increasing amounts of training data

CLA obtains better results on the other three language-pairs. In particular, CLA alignment performs much better on Japanese-English, with a relative improvements in BLEU score around 8.1% (from 42.52 to 45.96). It is worth remarking that CLA alignments can be computed much more efficiently than IBM alignments.

IBM Intersection alignments always give better results in terms of BLEU score than union and CLA alignments. Differences in terms of NIST scores are however not so evident.

By applying the Inter+CLA training modality – i.e. concatenation of phrase-pairs from the IBM intersection alignments and CLA alignments – an improvement against IBM Intersection is observed with all four language-pairs. Relative BLEU improvements range from 1% (Italian-English) to 2% (Japanese-English). Improvements of NIST score are also consistent across all language pairs but less marked. Unfortunately, the testing samples are too small for statistically

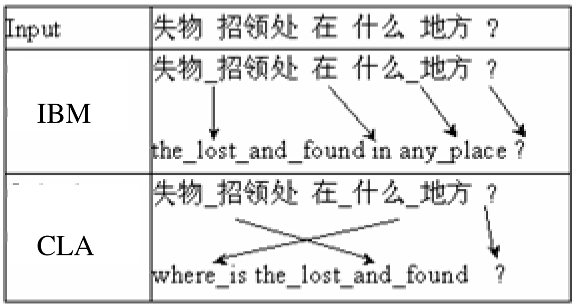

Input
失物 招领处 在 什么 地方 ?
IBM
失物_招领处 在 什么_地方 ?
the_lost_and_found in any_place ?
CLA
失物_招领处 在_什么_地方 ?
where_is the_lost_and_found ?

Fig. 4. MT output after training with IBM and CLA word alignments

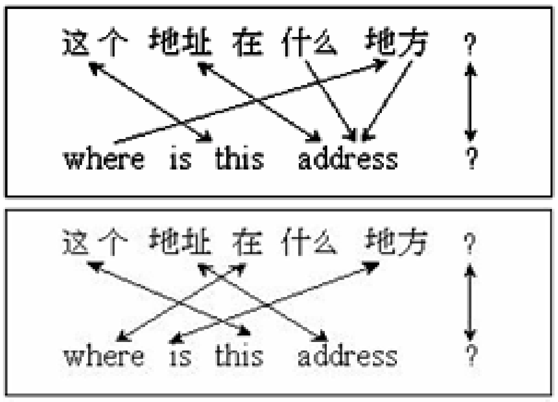


Fig. 5. Word alignments computed with IBM-models (top) and competitive linking algorithm (bottom)

assessing the reported BLEU score differences. However, a simple sign test[4] on the BLEU scores of the four tasks, with the assumption of less or equal performance, tells that improvements of the Inter+CLA method against each other method are significant at level $\alpha = 0.0625$.

A second series of experiments investigated the behavior of the training modalities against increasing amounts of data. These experiments are limited to the Chinese-English and Italian-English tasks.

Results are plotted in Figure 3. In the Chinese-English task, the superiority of the CLA over the IBM union modality consistently remains, independently from the data-sparseness condition. In the Italian-English task, IBM union modality and CLA perform very similar, CLA alignments work slightly better than IBM ones under the highest data-sparseness conditions.

Consistent conclusion can be also drawn for the combined training method Inter+CLA. For both language pairs, combined method outperform the IBM intersection modality in all considered data-sparseness conditions.

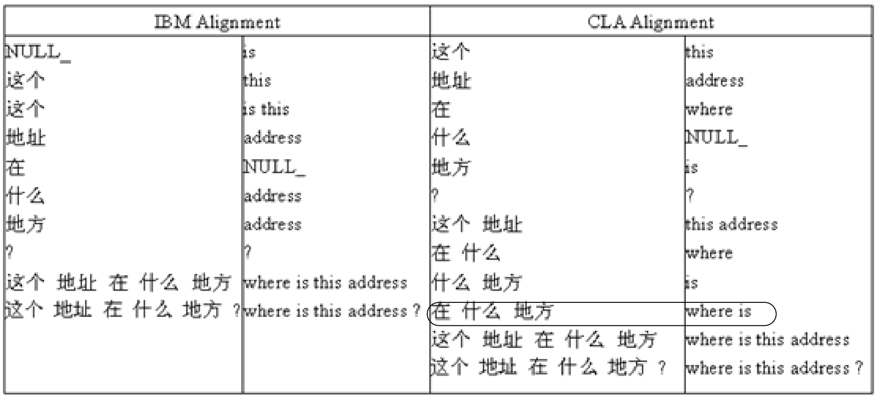
[4] http://home.clara.net/sisa/binomial.htm

6 Discussion

In order to better interpret the experimental results, a qualitative analysis of two very different word alignments can be informative, namely, CLA and IBM union alignments. A good starting point is given in Figure 4, which shows a Chinese sentence for which the system trained with CLA alignments performs better than the system training with IBM union alignments.

The problem with the IBM-model trained system is that it missed the translation of the last three Chinese words with the words *where is*. An inspection of the phrase table used by the decoder reveals that such translation is missing. By further looking into the training data we found that this translation pair could have been learned only from one sentence pair. This translation example is shown in Figure 5, together with the alignment computed with the CLA and the direct and inverse alignments computed with the IBM models. The union of the direct (arrows upward) and inverse (arrow downward) alignments is obtained by disregarding the direction of the links. Clearly, the one-to-one CLA alignment has a lower density than the IBM union alignment. According to our phrase-extraction methods, an alignment with fewer links often permits the generation of more phrase-pairs. This is indeed happens in the example shown in Figure 6, which also shows that the phrase-pair useful for the translation example in Figure 4 is indeed found in the CLA alignment.

The above example and some further manual inspections of alignments suggest the following general considerations. CLA alignments show in general lower recall and higher precision than IBM union alignments. (Formally, recall and precision of an automatic alignments should be measured by comparing all word-to-word links against some reference alignment.) From the point of view of phrase-extraction, a lower recall – i.e. number of links – can indeed result in a larger number of generated phrase pairs.



IBM Alignment		CLA Alignment	
NULL_	is	这个	this
这个	this	地址	address
这个	is this	在	where
地址	address	什么	NULL_
在	NULL_	地方	is
什么	address	?	?
地方	address	这个 地址	this address
?	?	在 什么	where
这个 地址 在 什么 地方	where is this address	什么 地方	is
这个 地址 在 什么 地方 ?	where is this address ?	在 什么 地方	where is
		这个 地址 在 什么 地方	where is this address
		这个 地址 在 什么 地方 ?	where is this address ?

Fig. 6. Phrase-pairs extracted from the IBM and CLA alignments in Figure 5, respectively. The useful translation pair for *where is* is pointed out.

CLA alignments can also induce phrase-pairs with a higher degree of non-monotonicity or, in other words, with a larger position mismatch between source and target phrases. This property could explain the better performance of CLA training in the Chinese-English task but the similar and lower performance in the Italian-English task. Translation between Italian and English seems to imply in fact much less word re-ordering than for Japanese and Chinese. Phrase-pairs extracted from CLA alignments are hence of little help. In other words, phrase-tables extracted just from CLA alignments seem less effective for language pairs with little word reordering.

Remarkably, a more consistent behavior emerges from the application of the combined training modalities. For all language pairs and data-sparseness conditions, it seems that merging information from the two types of alignments is always beneficial. More interestingly, the positive contribution is only slightly reduced when larger amounts of training data are used (see Figure 3).

7 Conclusions

We have presented novel training techniques based on the competitive linking algorithm which consistently improved performance of a phrase-based SMT system trained with conventional IBM word alignments. Extensive experiments were performed on a tourism domain including four language directions: Arabic-to-English, Chinese-to-English, Japanese-to-English, and Italian-to-English. Results showed that combining phrase-pairs extracted from IBM alignments with phrase-pairs extracted from CLA alignments gives consistent improvements in performance on all language pairs and under different data-sparseness conditions.

Acknowledgment

The authors would like to thank Laurens van der Werff and Nicola Bertoldi for useful suggestions and discussions. This work has been funded by the European Union under the integrated project TC-STAR - Technology and Corpora for Speech-to-Speech Translation - (IST-2002-FP6-506738, http://www.tc-star.org).

References

1. Och, F.J., Ney, H.: Improved statistical alignment models. In: Proceeding of the 38th Annual Meeting of the Association for Computational Linguistics, Hongkong, China (2000) 440–447
2. Brown, P.F., Della Pietra, S.A., Della Pietra, V.J., Mercer, R.L.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics **19** (1993) 263–312
3. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of HLT/NAACL 2003, Edmonton, Canada (2003) 127–133
4. Melamed, I.D.: Models of translational equivalence among words. Computational Linguistics **26** (2000) 221–249

5. Cherry, C., Lin, D.: A probability model to improve word alignment. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan (2003) 88–95
6. Kraif, O., Chen, B.: Combining clues for lexical level aligning using the null hypothesis approach. In: Proceedings of International Conference on Computational Linguistics (COLING), Geneva, Switzerland (2004) 1261–1264
7. Moore, R.C.: Association-based bilingual word alignment. In: Proceedings of ACL Workshop on Building and Using Parallel Texts, Ann Arbor, MI (2005) 1–8
8. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
9. Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., Tsujii, J.: Overview of the iwslt04 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Kyoto, Japan (2004) 1–12
10. Eck, M., Hori, C.: Overview of the iwslt 2005 evaluation campaign. In: Proceedings of the International Workshop on Spoken Language Translation (IWSLT), Pittsburgh, PA (2005) 11–32
11. Federico, M., Bertoldi, N.: A word-to-phrase statistical translation model. ACM Transactions on Speech and Language Processing **2** (2005) 1–24
12. Tillmann, C., Ney, H.: Word reordering and a dynamic programming beam search algorithm for statistical machine translation. Computational Linguistics **29** (2003) 97–133
13. Chen, B., Cattoni, R., Bertoldi, N., Cettolo, M., Federico, M.: The itc-irst smt system for iwslt 2005. In: Proceedings of the International Workshop on Spoken Language Translation - IWSLT, Pittsburgh, USA (2005) 98–104
14. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL), Sapporo, Japan (2003) 160–167
15. Och, F.J., Tillman, C., Ney, H.: Improved alignment models for statistical machine translation. In: Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, University of Maryland, MDPA (1999) 20–28
16. Kikui, G., Sumita, E., Takezawa, T., Yamamoto, S.: Creating corpora for speech-to-speech translation. In: Proceedings of the 4th European Conference on Speech Communication and Technology. Volume 2., Madrid, Spain (1995) 1249–1252
17. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. Research Report RC22176, IBM Research Division, Thomas J. Watson Research Center (2001)