A novel alignment model inspired on IBM Model 1

Jesús González-Rubio[1], Germán Sanchis-Trilles[1], Alfons Juan[1], and Francisco Casacuberta[1]

Departamento de Sistemas Informáticos y Computación
Universidad Politécnica de Valencia, Valencia, Spain.
{jegonzalez,gsanchis,ajuan,fcn}@dsic.upv.es

Abstract. We present an extension to IBM Model 1 for training word-to-word lexicon probabilities. This model takes into account a given fixed segmentation of the source and target sentences in the estimation of the statistical dictionary. Our experimentation on the Europarl corpus shows that a statistical consistent improvement in the translation quality can be achieved by including our proposed model as a new information source in a log-linear combination of models.

1 Statistical Machine Translation

The goal of Machine Translation is the translation of a text given in some source language into a target language. We are given a source language sentence $\mathbf{f} = f_1 \dots f_j \dots f_J$ which is to be translated into a target language sentence. Among all possible target language sentences, we will choose the sentence $\hat{\mathbf{e}} = e_1 \dots e_i \dots e_I$ which maximises the posterior probability. Such statement is formalised in the Fundamental Equation of Machine Translation:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\text{argmax}}\{Pr(\mathbf{e}|\mathbf{f})\} = \underset{\mathbf{e}}{\text{argmax}}\{Pr(\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e})\} \,. \tag{1}$$

The argmax operation denotes the search problem, i.e. the generation of the output sentence in the target language. The decomposition in Eq. (1) allows an independent modelling of the target *language model* $Pr(\mathbf{e})$ and the (inverse) *translation model* $Pr(\mathbf{f}|\mathbf{e})$[1], known as source-channel model [1]. This decomposition has a very intuitive interpretation: the translation model $Pr(\mathbf{f}|\mathbf{e})$ will capture the word relations between both input and output languages, whereas the language model $Pr(\mathbf{e})$ will ensure that the output sentence is a well-formed sentence belonging to the target language.

Many statistical translation models [2–5] try to model word-to-word correspondences between source and target words. Known as statistical alignment models, these models typically yield the following equation:

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{\mathbf{a}} \{Pr(\mathbf{a}|\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e}, \mathbf{a})\} \,. \tag{2}$$

The alignment model in Eq. (2) introduces a 'hidden' word alignment $\mathbf{a} = a_1^J$, which describes a mapping from a source position j to a target position a_j.

[1] We use $Pr(\cdot)$ to denote general probability distributions and $p(\cdot)$ to denote model-based probability distributions.

Word-based translation models were later on extended by phrase-based models [6–8], which have proved to provide a very efficient framework for machine translation. Phrase-based models compute the translation probability of a given *phrase*, i.e. sequence of words, and hence they introduce information about context. Statistical machine translation systems implementing these models have mostly outperformed single-word models such as IBM Model 1 [2], becoming predominant in the state-of-the-art [9] nowadays.

In order to combine the positive contributions of different approaches, statistical machine translation models can be merged using a log-linear combination [10]. In this framework, we have a set of M feature functions $h_m(\mathbf{e}|\mathbf{f}),\ m = 1, \ldots, M$. For each feature function, there exists a model parameter $\lambda_m,\ m = 1, \ldots, M$. The following decision rule is obtained:

$$\hat{\mathbf{e}} = \underset{\mathbf{e}}{\operatorname{argmax}} \frac{\exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}|\mathbf{f})]}{\sum_{\mathbf{e}'} \exp[\sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}'|\mathbf{f})]} = \underset{\mathbf{e}}{\operatorname{argmax}} \sum_{m=1}^{M} \lambda_m h_m(\mathbf{e}|\mathbf{f}) \,. \quad (3)$$

In this paper we present a novel word alignment model (Section 3) intended to overcome some of the problems inherent to IBM Model 1 (Section 2). We will show that an improvement in translation quality, on the Europarl corpus, can be achieved when using our proposed model as one more feature function in a log-linear machine translation model (Sections 4 and 5).

2 IBM Model 1

IBM Model 1 [2], is a word alignment model which was originally developed to provide reasonable initial parameter estimates for more complex word alignment models, but it has subsequently found a host of additional uses, as segmenting long sentences for improved word alignment [11] or extracting parallel sentences from comparable corpora [12]. Furthermore, at the 2003 John Hopkins summer workshop on statistical machine translation, a large number of features were tested to discover which ones could improve a state-of-the-art translation system, and the only feature that produced a "truly significant improvement" was the IBM Model 1 score [13].

IBM Model 1 is defined as a particularly simple alignment model, where all word-to-word alignments have the same probability, i.e. $Pr(\mathbf{a}|\mathbf{e})$ is modelled using a uniform distribution (which [2] show yields Eq. (4)). Hence, word order does not affect alignment probabilities.

$$p(\mathbf{f}|\mathbf{e}) = \prod_{j=1}^{J} \left[\frac{1}{I+1} \sum_{i=0}^{I} p(f_j|e_i) \right] \,. \quad (4)$$

IBM Model 1 clearly has many shortcomings as a translation model due to its simplicity. The *distortion problem* and the fact that some words act as *garbage collectors* are some of them. The distortion problem is a structural limitation of the IBM Model 1 due to the fact that the position of any word in the target sentence is independent of the position of the corresponding word in the source sentence, or the positions of any other source language words or their translations. The other problem with IBM Model

1, as standardly trained, is that rare words in the source language tend to act as "garbage collectors" [14, 13], aligning too many words in the target sentence.

Our proposal attempts to reduce the shown problems of IBM Model 1 by including information about a given segmentation of the input and output sentences in the estimation process of the lexicon dictionary. Similar aims, but differently approached, are pursued by [15], which extends the word-to-word alignment approach allowing one-to-many alignments, or [16], that deals with problems related to the suboptimal performance of the standard training method for IBM Model 1.

3 Model Description

Our alignment model is an enhancement of the IBM Model 1, which takes into account a given segmentation of the input and output sentences to estimate a statistical dictionary. The aim of our model is to benefit those alignments which are coherent with a fixed given segmentation which is considered optimal. We expect to reduce the dispersion of the lexical probabilities, concentrating the probability mass in those words which are revealed by the segmentation as potential candidates to be a correct translation. In addition, our model also aims to reduce the "garbage words" problem of IBM Model 1, which tends to concentrate alignment points in some words, independently of the distance between source and target words.

We are given a source sentence $\mathbf{X}$ divided into K segments $\mathbf{X} = X_1 \dots X_k \dots X_K$, where each segment X_k is a sequence of Γ_k words $X_k = x_{k1} \dots x_{kk'} \dots x_{k\Gamma_k}$. This source sentence is to be translated into a target sentence $\mathbf{Y}$ which is divided into L segments $\mathbf{Y} = Y_1 \dots Y_l \dots Y_L$, where each segment Y_l is a sequence of Λ_l words $Y_l = y_{l1} \dots y_{ll'} \dots y_{l\Lambda_l}$. The segmentation of the source and target sentences is given as input for our model and remains fixed throughout all the process.

In order to take into account the segmentations of the input and output sentences, we modify the statistical alignment model in Eq. (2) as follows:

$$Pr(\mathbf{X}|\mathbf{Y}) = \sum_{\mathbf{c},\mathbf{b}} Pr(\mathbf{c}|\mathbf{Y}) Pr(\mathbf{X}, \mathbf{b}|\mathbf{Y}, \mathbf{c}) \ . \tag{5}$$

Instead of only considering one 'hidden' word alignment $\mathbf{a}$, as IBM Model 1 does, our proposal has two 'hidden' alignments. First, we introduce a segment alignment $\mathbf{c} = c_1 \dots c_k \dots c_K$, which describes a mapping from a source segment k to a target segment $l = c_k$. Once the segment alignment is determined, we include a word alignment $\mathbf{b} = b_1 \dots b_k \dots b_K, \ \forall k \ b_k = b_{k1} \dots b_{kk'} \dots b_{k\Gamma_k}$ which describes a mapping from the k'th word of source segment k to the l'th word of target segment l, with $l' = b_{kk'}$. Hence, alignment $\mathbf{c}$ maps a given source segment into a specific target segment, and then alignment $\mathbf{b}$ maps the words on the source segment into the words in the target segment.

3.1 Model assumptions

Next, we describe the assumptions made in the derivation of our model. First, the second term on Eq. (5) is analysed, on Eq. (6) we assume that the alignment of a given segment

does not depend on the alignment of the previous segments, whereas on Eq. (7) we perform a similar assumption on the word level, i.e. the alignment of a given word does not depend on the previous word alignments.

$$Pr(\mathbf{X}, \mathbf{b}|\mathbf{Y}, \mathbf{c}) = \prod_{k=1}^{K} Pr(X_k, b_k|\mathbf{Y}, \mathbf{c}, X_1^{k-1}, b_1^{k-1}) \approx \prod_{k=1}^{K} p(X_k, b_k|\mathbf{Y}, c_k) \quad (6)$$

$$= \prod_{k=1}^{K} \prod_{k'=1}^{\Gamma_k} p(x_{kk'}, b_{kk'}|\mathbf{Y}, c_k, x_k{}_1^{k'-1}, b_k{}_1^{k'-1})$$

$$\approx \prod_{k=1}^{K} \prod_{k'=1}^{\Gamma_k} p(x_{kk'}, b_{kk'}|\mathbf{Y}, c_k) \; . \quad (7)$$

The same assumption done on Eq. (6) can be applied to the first term on Eq. (5), yielding

$$Pr(\mathbf{c}|\mathbf{Y}) \; = \; \prod_{k=1}^{K} Pr(c_k|\mathbf{Y}, c_1^{k-1}) \; \approx \; \prod_{k=1}^{K} p(c_k|\mathbf{Y}) \; . \quad (8)$$

Lastly, we will perform the same assumption as IBM Model 1, modelling the mappings between input and output positions in the alignments as uniform distributions.

3.2 Our model

The final formulation of our model is shown in Eq. (9) and Eq. (10):

$$p(X|Y) \; = \; \prod_{k=1}^{K} \left[\frac{1}{L+1} \sum_{l=0}^{L} p(X_k|Y_l) \right] \; . \quad (9)$$

$$p(X_k|Y_l) = \prod_{k'=1}^{\Gamma_k} \left[\frac{1}{\Lambda_l + 1} \sum_{l'=0}^{\Lambda_l} p(x_{kk'}|y_{ll'}) \right] \; . \quad (10)$$

Our model can be seen as a composition of two models: the first component (equation (9)) models the mapping between the segments of the input and output sentences ($\mathbf{c}$ alignment) while the second one (equation (10)), which is embedded into Eq. (9), models the alignment between the words of one source segment and the words in the corresponding target segment ($\mathbf{b}$ alignment). However, it is important to point out that both components are estimated jointly and build up our entire model.

As the standard IBM Model 1, the parameters of our model constitute a statistical word dictionary $p(x_{kk'}|y_{ll'})$.

We use the Expectation-Maximisation (EM) algorithm [17] to obtain the maximum-likelihood estimates of the parameters.

The parameter re-estimation process in the EM algorithm shows the differences between our model and IBM Model 1. IBM Model 1 obtains the expected value for an alignment with the following equation [2]:

$$a_{nji}^{(t)} = \frac{p(x_{nj}|y_{ni})^{(t)}}{\sum_{i'=0}^{I} p(x_{nj}|y_{ni'})^{(t)}} \; . \quad (11)$$

In our case, we took into account the segmentation of the input and output sentences to obtain the expected value for an alignment, yielding the following equation:

$$(c_{nkl} \cdot b_{kll'})^{(t)} = \frac{p(x_{nkk'}|y_{nll'})^{(t)}}{\sum_{l''=0}^{\Lambda_l} p(x_{nkk'}|y_{nll''})^{(t)}} \cdot \frac{p(X_k|Y_l)}{\sum_{l'''=0}^{L} p(X_k|Y_{l'''})} . \quad (12)$$

In the original IBM Model 1 (equation (11)) each word alignment has the same significance, no matter the positions of the words. In our formulation (equation (12)) the importance of each word alignment is weighted by the significance of the alignment of the segments the words belong to with respect to the rest of segment alignments. Hence, we benefit those alignments coherent with the given segmentation which is considered optimal.

4 Experimental setup

In our experimentation we include scores derived from our model into a log-linear combination, as another feature functions, with the purpose of improving the translation quality of the log-linear model.

We perform our experiments on the second version of the Europarl corpus [18], which is built from the proceedings of the European Parliament. This corpus is divided into three separate sets: one for training, one for development and one for test and was the corpus used in the 2006 Workshop on Machine Translation (WMT) of the ACL [19]. We focused on the German–English (De–En), French–English (Fr–En) and Spanish–English (Es–En) subcorpora of the Europarl corpus, as done in the 2006 WMT of the ACL.

		De	En	Es	En	Fr	En
Training	Sentences	751K		731K		688K	
	Run. words	15.3M	16.1M	15.7M	15.2M	15.6M	13.8M
	Avg. len.	20.3	21.4	21.5	20.8	22.7	20.1
	Voc.	195K	66K	103K	64K	80K	62K
Development	Sentences	2000		2000		2000	
	Run. words	55K	59K	61K	59K	67K	59K
	Avg. len.	27.6	29.3	30.3	29.3	33.6	29.3
	OoV	432	125	208	127	144	138
Test	Sentences	2000		2000		2000	
	Run. words	54K	58K	60K	58K	66K	58K
	Avg. len.	27.1	29.0	30.2	29.0	33.1	29.3
	OoV	377	127	207	125	139	133

Table 1. Statistics of the Europarl corpus for each of the subcorpora. OoV stands for "Out of Vocabulary" words, K for thousands of elements and M for millions of elements.

Since the original corpus is not sentence-aligned, different corpora are obtained while building the parallel bilingual corpora. The statistics of these corpora are displayed in Table 1. The language models used in our experimentation were computed

with the SRILM [20] toolkit, using 5-grams and applying interpolation with the Kneser-Ney discount. The perplexity of the various subsets of the corpora, according to these language models, are shown in Table 2.

It seems important to point out the fact that the average sentence length in the training sets is much shorter than in the other sets is because in the cited workshop the training sets were restricted to sentences with a maximum length of 40 words, whereas the rest of sets did not have this restriction.

	German	English	Spanish	French
Development	148.6	89.9	89.0	66.5
Test	149.8	88.9	90.6	66.7

Table 2. Perplexity of the various corpora subsets with 5-grams.

Since the translations in the corpus have been written by a big number of different human translators, a same sentence may be translated in several different ways, all of them correct. This fact increases the difficulty of the corpus, and can be seen in the number of different pairs that constitute the training set, which is very similar to the total number of pairs, and also worsens the problem of "garbage collector" words, which our model attempts to reduce. An example is the English sentence "*We shall now proceed to vote.*": it appears translated into Spanish both as "*Se procede a la votación.*", which is quite a faithful translation, and "*El debate queda cerrado.*", which means "*the debate is now closed.*". Although these two Spanish sentences are clearly different, one can easily imagine a scenario where both translations would fit.

To train our models, we previously need a segmentation of the corpus (see Section 3). There are a number of algorithms to segment a corpus [21, 22, 11]. In our case, the segmentation was obtained following the technique described in [23]. First, a phrase-based model trained on a training set is used to translate the training set itself. Then, the alignment inherent to the translation of each sentence pair of the training set is used to segment this sentence pair. The resulting segmented corpora is used by our model as input.

The evaluation has been carried out using the WER and BLEU measures, following previous works in statistical machine translation and for comparison purposes. The WER criterion is similar to the edit distance used in Speech Recognition. It computes the minimum number of editions (substitutions, insertions and deletions) needed to convert the translated sentence into the sentence considered ground truth. The BLEU measure [24] computes the precision of unigrams, bigrams, trigrams and 4-grams with respect to the reference translation with a penalty for too short sentences.

To test the statistical significance of the results, we have followed the methods described in [25] and [26]. Zhang and Vogel present a bootstrap [27] based algorithm that computes a confidence interval, based on bootstrap percentiles, for the discrepancy between the two machine translation systems (X and Y) under study. This algorithm extracts N bootstrap samples $T_1 \dots T_i \dots T_N$ from the translated test set. If system X scored x_i on T_i and system Y scored y_i, then the discrepancy score between system X and Y on T_i is $\delta_i = x_i - y_i$. From the N discrepancy scores, we find the 2.5th percentile and the 97.5th percentile, which is the 95% confidence interval for the discrepancy be-

tween the systems. Bisani and Ney present a similar method where instead of returning and interval they compute the *Paired Probability of Improvement* (PPoI) which is the relative number of times system X outperforms system Y and vice versa.

5 Experiments

For each language pair, we trained two of our alignment models on the corresponding segmented training set, one model for each translation direction. These will be called, hereafter, our direct and inverse extended lexicalised models.

We used the Moses toolkit [28] to train the phrase-based models from the training subcorpora of Europarl and the parameters of the log-linear models were optimised using the development subcorpora via the MERT procedure [29], using BLEU as the measure to be optimised.

The standard Moses translation model includes five translation scores for each phrase pair in the phrase table [30]: two phrase translation scores (direct and inverse), based on counting the co-occurrences of each phrase pair and normalising the counts, two lexical weights, whose purpose is to assert the lexical soundness of each bilingual phrase pair, and a constant value called phrase penalty.

Similarly, we can obtain two lexical probabilities given by the likelihood of the phrase pair $[X_k, Y_l]$ according to our direct and inverse extended lexicalised models (equation (10)).

	Monotonic				Non Monotonic			
	Baseline		Extended		Baseline		Extended	
Language Pair	WER	BLEU	WER	BLEU	WER	BLEU	WER	BLEU
Es-En	58.25	31.01	**57.87**	**31.27**	57.67	31.56	**57.35**	**31.99**
En-Es	59.50	30.16	**59.26**	**30.52**	58.37	31.26	**58.23**	**31.54**
De-En	66.82	25.00	**66.71**	**25.01**	65.45	26.21	**65.06**	**26.49**
En-De	72.45	18.04	**71.71**	**18.42**	71.57	18.81	**71.33**	**18.92**
Fr-En	57.67	30.83	**57.59**	**30.99**	57.34	31.46	**57.08**	**31.71**
En-Fr	60.50	32.31	**60.41**	**32.37**	59.17	33.34	**58.76**	**33.75**

Table 3. BLEU and WER translation results for test set. Baseline stands for the standard Moses log-linear model, Extended for the standard Moses log-linear combination plus the two (direct and inverse) scores of our models, Monotonic for monotonic decoding and Non Monotonic for non monotonic decoding.

Table 3 shows the translation quality for the test set as measured by BLEU and WER. *Baseline* stands for the standard Moses log-linear translation model, whereas the *Extended* combination is obtained by including the direct and inverse scores of our extended lexicalised models into the *Baseline* system. Results are shown for both *monotonic* and *non monotonic* decoding. In this context, *monotonic* implies that both the segmentation of the training set and the final translation of the test set were performed monotonically. In contrast, *non monotonic* implies that both the segmentation and the translation were performed using the standard lexicalised reordering implemented into Moses.

The inclusion of our lexicalised models is reflected in an improvement of the translation quality, as measured by WER and BLEU scores, both in the monotonic and the non monotonic cases. Our interpretation for this fact is that the model presented here incorporates further information into the log-linear combination of models, which is evidenced by a slight, but systematic, improvement in the translation quality over all the language pairs.

	BLEU				WER			
	Monotonic		Non Monotonic		Monotonic		Non Monotonic	
Lang. Pair	Improvement	PPoI	Improvement	PPoI	Improvement	PPoI	Improvement	PPoI
Es–En	**0.26±0.23**	**0.98**	**0.43±0.24**	**1.00**	**-0.38±0.21**	**1.00**	**-0.31±0.23**	**0.99**
En–Es	**0.36±0.26**	**0.99**	**0.28±0.23**	**0.99**	**-0.22±0.22**	**0.97**	-0.16±0.22	**0.85**
De–En	-0.03±0.18	0.35	**0.27±0.27**	**0.97**	-0.10±0.23	**0.85**	**-0.36±0.28**	**0.99**
En–De	**0.38±0.21**	**1.00**	0.09±0.25	0.79	**-0.72±0.25**	**1.00**	-0.27±0.28	**0.94**
Fr–En	**0.18±0.18**	**0.98**	**0.23±0.20**	**0.99**	-0.07±0.17	**0.82**	**-0.25±0.21**	**0.99**
En–Fr	0.05±0.23	0.73	**0.43±0.27**	**1.00**	-0.10±0.28	0.66	**-0.41±0.27**	**1.00**

Table 4. Average improvements with their confidence intervals at 95% and Paired Probabilities of Improvement (PPoI) of the Extended model with respect to the Baseline model, for both BLEU and WER measures. Bold improvements are statistically significant, and bold PPoIs reflect a real superiority of the Extended model.

Table 4 shows the average improvements with their confidence intervals, at a confidence level of 95%, of the Extended models with respect to the Baseline models for each of the language pairs considered and considering both the monotonic and non monotonic cases, following the technique described in [25]. Table 4 also displays the PPoI of the Extended system versus the Baseline system, according to [26].

Most of the results for non monotonic decoding show an improvement with confidence intervals that do not overlap with zero, so we can claim that the Extended model is statistically better than the Baseline model [25] for almost all the language pairs when using non monotonic decoding, and even in those cases where the improvement in the translation quality is not statistically significant the PPoI ranges between 0.8 and 1.0 so we can be confident that results reflect a real superiority of the Extended model [26]. On the other hand, when performing monotonic decoding, differences are statistically significant in less cases, and PPoI is, in general, lower than in the non monotonic case. This is due to the fact that, in our model, there is a correlation between the quality of the given segmentation of the corpus and the quality of the statistical dictionary estimated by our model. As the quality of the non monotonic segmentation is better than the quality of the monotonic one [23], our statistical dictionary is better estimated for the non monotonic case.

For both monotonic and non monotonic, translation quality results of the Extended model improve the Baseline model. However, a statistical dictionary allowing a significant improvement over the Baseline system was obtained only when the quality of the segmentation of the corpus was improved. This is specially interesting, given that the segmentation used is defined in [23] as *approximated* segmentation, and hence further improvements cannot be discarded if the segmentation is improved as well.

6 Conclusions

In this work a novel alignment model has been introduced, which enhances IBM Model 1 by including information about a fixed given segmentation of the input and output sentences in the estimation process of the statistical dictionary. This model has been used in combination with other models to improve the translation quality as measured by BLEU and WER on the Europarl corpus. Results obtained, when our model is incorporated as a new feature function in a log-linear combination, systematically improve baseline BLEU and WER scores. In addition most of these improvements are statistically significan or reflect a real superiority of the Extended model.

Our proposal is a first step towards a hybrid word and phrase based alignment model. Future work includes further research on the correlation of the quality of the statistical dictionary with the quality of the segmentation by trying out different segmentations. Within this line, the final aim is to calculate the statistical dictionary and simultaneously estimate the best segmentation of the corpus, instead of using a given one.

Acknowledgements

This work has been partially supported by the Spanish MEC under FPU scholarships AP2006-00691 and AP2005-4023, and grant Consolider Ingenio 2010 CSD2007-00018 and by the EC (FEDER) and the Spanish MEC under grant TIN2006-15694-CO2-01.

References

1. Brown, P., Cocke, J., Pietra, S.D., Pietra, V.D., Jelinek, F., Lafferty, J., Mercer, R., Roossin, R.: A statistical approach to machine translation. Computational Linguistics **16** (1990)
2. Brown, P., Pietra, S.D., Pietra, V.D., Mercer, R.: The mathematics of machine translation. In: Computational Linguistics. Volume 19. (1993) 263–311
3. Vogel, S., Ney, H., Tillmann, C.: Hmm-based word alignment in statistical translation. In: Computational linguistics. (1996) 836–841
4. Tillmann, C., Vogel, S., Ney, H., Zubiaga, A.: A dp based search using monotone alignments in statistical translation. In: Computational Linguistics. (1997) 289–296
5. Niessen, S., Vogel, S., Ney, H., Tillmann, C.: A dp based search algorithm for statistical machine translation. In: Computational linguistics. (1998) 960–967
6. Tomas, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Proc. of the Machine Translation Summit VIII, Santiago de Compostela, Spain (2001) 357–361
7. Marcu, D., Wong, W.: Joint probability model for statistical machine translation. In: EMNLP, Pennsylvania, USA (2002)
8. Zens, R., Och, F., Ney, H.: Phrase-based statistical machine translation. In: Advances in artificial intelligence. Volume 2479. (2002) 18–32
9. Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., Schroeder, J.: (meta-)evaluation of machine translation. In: Proc. of the Workshop on Statistical Machine Translation. (2007)
10. Och, F., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proce. of the ACL. (2001) 295–302
11. Nevado, F., Casacuberta, F., Vidal, E.: Parallel corpora segmentation by using anchor words. In: Proc. of the of EACL 2003 workshop on EAMT, Budapest, Hungary (2003)

12. Munteanu, D., Fraser, A., Marcu, D.: Improved machine translation performance via parallel sentence extraction from comparable corpora. In: Proc. of the HLT. (2004) 265–272
13. Och, F., Gildea, D., Khudanpur, S., Sarkar, A., Yamada, K., Fraser, A., Kumar, S., Shen, L., Smith, D., Eng, K., Jain, V., Jin, Z., Radev, D.: A smorgasbord of features for statistical machine translation. In: HLT-NAACL. (2004) 161–168
14. Brown, P., Pietra, S.D., Pietra, V.D., Goldsmith, M., Hajic, J., Mercer, R., Mohanty, S.: But dictionaries are data too. In: Proc. of the HLT. (1993) 202–205
15. Och, F., Tillmann, C., Ney, H.: Improved alignment models for statistical machine translation. In: Proc. of the EMNLP-VLC, University of Maryland (1999) 20–2
16. Moore, R.: Improving ibm word-alignment model 1. In: Proc. of the ACL. (2004) 518
17. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society. **39**(1) (1977) 1–38
18. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: MT Summit. (2005)
19. Koehn, P., Monz, C.: Manual and automatic evaluation of machine translation between european languages. In: Proc. of the NAACL, New York City (2006) 102–121
20. Stolcke, A.: Srilm - an extensible language modeling toolkit. In: Proc. of the International Conference on Spoken Language Processing. Volume 2. (2002) 901–904
21. Michel, S., Plamondon, P.: Bilingual sentence alignment: Balancing robustness and accuracy (1996)
22. Wu, D.: Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. Computational Linguistics **23**(3) (1997) 377–403
23. Sanchis-Trilles, G., Casacuberta, F.: Increasing translation speed in phrase-based models via suboptimal segmentation. In: Proc. of PRIS, Barcelona (Spain) (2008)
24. Papineni, K., Kishore, A., Roukos, S., Ward, T., Zhu, W.: Bleu: A method for automatic evaluation of machine translation. In: Technical Report RC22176 (W0109-022). (2001)
25. Zhang, Y., Vogel, S.: Measuring confidence intervals for the machine translation evaluation metrics. In: Proc. of the TMI. (2004)
26. Bisani, M., Ney, H.: Bootstrap estimates for confidence intervals in asr performance evaluation. In: IEEE Conference on Acoustics, Speech, and Signal Processing. Volume 1. (2004)
27. Efron, B., Tibshirani, R.: An Introduction to Bootstrap. Chapman and Hall, New York (1993)
28. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantine, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. Proc. of the ACL (2007)
29. Och, F.: Minimum error rate training for statistical machine translation. In: Proc. of the ACL, Sapporo, Japan (2003)
30. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proc. of the NAACL-HLT. Volume 1. (2003) 48–54