Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model

Daniel Ortiz-Martínez[1], Ismael García-Varea[2], and Francisco Casacuberta[1]

[1] Dpto. de Sist. Inf. y Comp., Univ. Politécnica de Valencia, 46071 Valencia, Spain
dortiz@dsic.upv.es fcn@dsic.upv.es
[2] Dpto. de Sist. Inf., Univ. de Castilla-La Mancha, 02071 Albacete, Spain
ivarea@info-ab.uclm.es

Abstract. We present a phrase-based statistical alignment model togheter with a set of different smoothing techniques to be applied when the best phrase-to-phrase alignment for a pair of sentences is to be computed. We follow a loglinear approach, which allows us to introduce different scoring functions to control specific aspects of phrase-level alignments. Experimental results for a well-known shared task on word alignment evaluation are reported, showing the great importance of smoothing in the generation of alignments. As a step forward, we also discuss the adaptation of the proposed model for its use in a CAT (Computer Assisted Translation) system.

1 Introduction

Statistical Machine translation (SMT) is an area of great interest in the NLP community that deals with the transformation of text or speech from a *source language* into a *target language*.

From a purely statistical point of view, the translation process can be formulated as follows: A source language string $\mathbf{f}$ is to be translated into a target language string $\mathbf{e}$. Every target string is regarded as a possible translation for the source language string with maximum a-posteriori probability $Pr(\mathbf{e}|\mathbf{f})$. According to Bayes' theorem, the target string $\hat{\mathbf{e}}$ that maximizes the product of both the target language model $Pr(\mathbf{e})$ and the string translation model $Pr(\mathbf{f}|\mathbf{e})$ must be chosen. The equation that models this process is:

$$\hat{e}_1^I = \arg\max_{\mathbf{e}} \left\{ Pr(\mathbf{e}) \cdot Pr(\mathbf{f}|\mathbf{e}) \right\} \quad (1)$$

State-of-the-art statistical translation systems follow a phrase-based approach, that is, the structural relations between source and target sentences are captured by means of phrases instead of isolated words.

In this paper we tackle the problem of generating alignments at phrase level by means of smoothed phrase-based statistical alignment models. As far as we know the problem of finding the best alignment at phrase level has not been extensively addressed in the literature. For example, in [1] three different techniques

for obtaining phrase-level alignments are compared, but there is no mention at all of the phrase-level aligment coverage problems that arise when real tasks and applications are used.

Different applications can benefit from the techniques proposed here, ranging from phrase-based SMT systems to machine-aided NLP tools, as for example CAT [2]. Under the CAT framework, we are given a source sentence and a prefix of the target sentence, and the goal is to obtain the best suffix that constitutes a complete translation. Then the first problem to be solved is how to align the given prefix with the corresponding portion of the source sentence. The techniques proposed here can be easily adapted to deal with this problem.

2 Phrase-based SMT

Different translation models have been proposed depending on how the relation between the source and the target languages is structured; that is, the way a target sentence is generated from a source sentence. This relation is summarized using the concept of *alignment*; that is, how the constituents (typically words or groups-of-words) of a pair of sentences are aligned with each other.

For the translation model, $Pr(\mathbf{f}|\mathbf{e})$, in Eq. (1), Phrase-based Translation (PBT) can be explained from a generative point of view as follows [3]:

1. The target sentence $\mathbf{e}$ is segmented into K phrases $(\tilde{e}_1^K)$.
2. Each target phrase $\tilde{e}_k$ is translated into a source phrase $\tilde{f}$.
3. Finally, the source phrases are reordered in order to compose the source sentence $\tilde{f}_1^K = \mathbf{f}$.

In PBT, it is assumed that the relations between the words of the source and target sentences can be explained by means of the hidden variable $\tilde{a}_1^K$, which contains all the decisions made during the generative story.

$$Pr(\mathbf{f}|\mathbf{e}) = \sum_{K,\tilde{a}_1^K} Pr(\tilde{f}_1^K, \tilde{a}_1^K|\tilde{e}_1^K) = \sum_{K,\tilde{a}_1^K} Pr(\tilde{a}_1^K|\tilde{e}_1^K) Pr(\tilde{f}_1^K|\tilde{a}_1^K, \tilde{e}_1^K) \qquad (2)$$

where each $\tilde{a}_k \in \{1 \dots K\}$ denotes the index of the target phrase $\tilde{e}$ that is aligned with the k-th source phrase $\tilde{f}_k$.

Different assumptions can be made from the previous equation. For example, in [3] all possible segmentations have the same probability, and in [4], it is also assumed that the alignments must be monotonic. In both cases the model parameters that have to be estimated are the translation probabilities between phrase pairs $(\{p(\tilde{f}|\tilde{e})\})$, which typically are estimated via relative frequencies as $p(\tilde{f}|\tilde{e}) = N(\tilde{f}, \tilde{e})/N(\tilde{e})$, where $N(\tilde{f}|\tilde{e})$ is the number of times that $\tilde{f}$ has been seen as a translation of $\tilde{e}$ within the training corpus.

According to Eq. (2), and following a maximum approximation, the problem stated in Eq. (1) can be reframed as:

$$\hat{\mathbf{e}} \approx \arg\max_{\mathbf{e},\mathbf{a}} \{p(\mathbf{e}) \cdot p(\mathbf{f}, \mathbf{a}|\mathbf{e})\} \qquad (3)$$

State-of-the-art statistical machine translation systems model $p(\mathbf{f}, \mathbf{a}|\mathbf{e})$ following a loglinear approach [5], that is:

$$p(\mathbf{f}, \mathbf{a}|\mathbf{e}) \propto \exp\left[\sum_i \lambda_i f_i(\mathbf{f}, \mathbf{e}, \mathbf{a})\right] \quad (4)$$

where each $f_i(\mathbf{f}, \mathbf{e}, \mathbf{a})$ is a feature function, and weights λ_i are optimized using a minimun error rate training (MERT) criteria [6] to optimize a particular quality metric (for example maximize the BLEU metric for translation quality, or minimize the *Alignment Error Rate* (AER) for alignment quality) on a development corpus.

3 Phrase-based alignments

The problem of finding the best alignment at phrase level has not been extensively addressed in the literature. A first attempt can be found in [1]. The concept of phrase-based alignment can be stated formally as follows:

Let $\mathbf{f} \equiv f_1, f_2, \ldots, f_J$ be a source sentence and $\mathbf{e} \equiv e_1, e_2, \ldots, e_I$ the corresponding target sentence in a bilingual corpus. A phrase-alignment between $\mathbf{f}$ and $\mathbf{e}$ is defined as a set $\mathcal{S}$ of ordered pairs included in $\mathcal{P}(\mathbf{f}) \times \mathcal{P}(\mathbf{e})$, where $\mathcal{P}(\mathbf{f})$ and $\mathcal{P}(\mathbf{e})$ are the set of all subsets of consecutive sequences of words, of $\mathbf{f}$ and $\mathbf{e}$, respectively. In addition, the ordered pairs contained in $\mathcal{S}$ have to include all the words of both the source and target sentences.

A phrase-based alignment of length K ($\tilde{A}_K$) of a sentence pair $(\mathbf{f}, \mathbf{e})$ is defined as a triple $\tilde{A}_K \equiv (\tilde{f}_1^K, \tilde{e}_1^K, \tilde{a}_1^K)$, where $\tilde{a}_1^K$ is a specific one-to-one mapping between the K segments/phrases of both sentences ($1 \leq K \leq \min(J, I)$).

Then, given a pair of sentences $(\mathbf{f}, \mathbf{e})$ and a phrase-based alignment model, we have to obtain the best phrase-alignment $\tilde{A}_K$ (or Viterbi phrase-alignment $V(\tilde{A}_K)$) between them. Assuming a phrase-alignment of length K, $V(\tilde{A}_K)$ can be computed as:

$$V(\tilde{A}_K) = \arg\max_{\tilde{A}_K} \left\{p(\tilde{f}_1^K, \tilde{a}_1^K|\tilde{e}_1^K)\right\} \quad (5)$$

where, following the assumptions of [3], $Pr(\tilde{f}_1^K, \tilde{a}_1^K|\tilde{e}_1^K)$ can be efficiently computed as:

$$p(\tilde{f}_1^K, \tilde{a}_1^K|\tilde{e}_1^K) = \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k}) \quad (6)$$

On the basis of Eq. (6), a very straightforward technique can be proposed for finding the best phrase-alignment of a sentence pair $(\mathbf{f}, \mathbf{e})$. This can be conceived as a sort of *constrained* translation. In this way, the search process only requires the use of a regular SMT system which filters its phrase-table in order to obtain those translations of $\mathbf{f}$ that are compatible with $\mathbf{e}$.

In spite of its simplicity, this technique has no practical interest when applied on regular tasks. Specifically, the technique is not applicable when the alignments

cannot be generated due to coverage problems of the phrase-based alignment model (i.e. one or more phrase pairs required to compose a given alignment have not been seen during the training process). This problem cannot be easily solved, since standard estimation tools such as THOT [7] and MOSES [8] do not guarantee the complete coverage of sentence pairs even if they are included in the training set; this is due to the great number of heuristic decisions involved in the estimation process.

One possible way to overcome the above-mentioned coverage problems requires the definition of an alternative technique that is able to consider every source phrase of $\mathbf{f}$ as a possible translation of every target phrase of $\mathbf{e}$. Such a technique requires the following two elements:

1. A general mechanism to assign probabilities to phrase pairs, no matter if they are contained in the phrase-table or not
2. A search algorithm that enables efficient exploration of the set of possible phrase-alignments for a sentence pair

The general mechanism for assigning probabilities to phrase pairs can be implemented by means of the application of smoothing techniques over the phrase-table. As shown in [9], well-known language model smoothing techniques can be imported into the PBT framework. As will be shown in section 4, the PBT smoothing techniques described in [9] can also be adapted to the generation of phrase-based alignments.

Regarding the search algorithm to be used, different search strategies can be adopted, as for example dynamic-programming-based or branch-and-bound algorithms. In this study, a *branch-and-bound* search strategy has been adopted. Our branch-and-bound search algorithm attempts to iteratively expand partial solutions, called hypotheses, until a complete phrase-alignment is found. The hypotheses are stored in a stack and ordered by their score. Since the number of possible alignments for a given sentence pair may become huge, it is necessary to apply heuristic prunings in order to reduce the search space. Such heuristic prunings include the limitation of the maximum number of hypotheses that can be stored in the stack and also the maximum number of different target phrases that can be linked to an unaligned source phrase when expanding a partial hypothesis.

3.1 A loglinear approach to phrase-to-phrase alignments

The score for a given alignment can be calculated according Eq (6). This scoring function has an important disadvantage. Specifically, it does not allow control of basic aspects of the phrase alignment, such as the lengths of the source and target phrases, and the reorderings of phrase alignments. This problem can be alleviated following the approach stated in Eq. (4), thus introducing different feature functions as scoring components in a log-linear fashion.

We propose the following set of feature functions:

– $f_1(\mathbf{f}, \mathbf{a}, \mathbf{e}) = \prod_{k=1}^{K} p(\tilde{e}_{\tilde{a}_k}|\tilde{f}_k)$: direct phrase model probability

- $f_2(\mathbf{f}, \mathbf{a}, \mathbf{e}) = \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{e}_{\tilde{a}_k})$: inverse phrase model probability
- $f_3(\mathbf{f}, \mathbf{a}, \mathbf{e}) = \prod_{k=1}^{K} p(|\tilde{e}_k|)$: target phrase length model. This component can be modeled by means of a uniform distribution (penalizes the length of the segmentation) or a geometric distribution (penalizes the length of the source phrases)
- $f_4(\mathbf{f}, \mathbf{a}, \mathbf{e}) = \prod_{k=1}^{K} p(\tilde{f}_k|\tilde{f}_{k-1})$: distortion model. This component is typically modeled by means of a geometric distribution (penalizes the reorderings)
- $f_5(\mathbf{f}, \mathbf{a}, \mathbf{e}) = \prod_{k=1}^{K} p(|\tilde{f}_k| \mid |\tilde{e}_{\tilde{a}_k}|)$: source phrase length model given the length of the target phrase. This component can be modeled by means of different distributions: uniform (does not take into account the relationship between the length of source and target phrase), Poisson or geometric

The corresponding weights $\lambda_i, i \in \{1, 2, \ldots, 5\}$ can be computed by means of MERT training.

Regarding the probability distribution used to model feature functions f_3, f_4, and f_5, we tested all possible combinations of uniform, geometric, and Poisson distributions in the experiments that we describe in section 5.

3.2 Application to CAT

The technique presented above for generating complete phrase-alignments can be easily adapted for the generation of partial alignments. As was mentioned in section 1, a good example in which the generation of partial alignments is required is the Computer Assisted Translation (CAT) framework. Under this framework, we are given a source sentence $\mathbf{f}$, and a prefix of the target sentence, which we will call $\mathbf{p}$, and the goal is to obtain the best suffix of $\mathbf{p}$ that constitutes a complete translation of $\mathbf{f}$. The generation of the suffix in CAT can be seen as a two-stage process. First we partially align the prefix $\mathbf{p}$ with only a part of $\mathbf{f}$, and second, we translate the unaligned portion of $\mathbf{f}$ (if any). The formalism presented at the beginning of this section requires few modifications to allow the generation of partial alignments. Specifically, given $\mathbf{f}$ and $\mathbf{p}$ we have to obtain the set $\mathcal{S}'$ of ordered pairs that contains all the words of $\mathbf{p}$ and only a subset of the words of $\mathbf{f}$.

4 Smoothing techniques

As was mentioned in section 3, the application of smoothing techniques is crucial in the generation of phrase-alignments. Although smoothing is an important issue in language modeling and other areas of statistical NLP (see for example [10] for more details), it has not received much attention from the SMT community. However, most of the well-known language model smoothing techniques can be imported to the SMT field and specifically to the PBT framework, as it is shown in [9].

In spite of the fact that PBT and the generation of phrase-alignments are similar tasks, it should be noted that the two problems differ in a key aspect.

While in PBT the probabilities of unseen events are not important (since the decoder only proposes phrase translations that are contained in the model, see [9]), in the generation of phrase alignments, assigning probabilities to unseen events is one of the most important problems that has to be solved.

In the rest of this section, we describe the smoothing techniques that we have implemented. They are very similar to those proposed in [9], although in our case we have strongly focused on the appropriate treatment of unseen events.

4.1 Statistical estimators

Training data can be exploited in different ways to estimate statistical models. Regarding the phrase-based models, the standard estimation technique is based on the relative frequencies of the phrase pairs (see section 2). Taking this standard estimation technique as a starting point, a number of alternative estimation techniques can be derived.

Phrase-based model estimators We have implemented the following estimation techniques for phrase-based models:

- Maximum-likelihood estimation (ML)
- Good-Turing estimation (GT)
- Absolute-discount estimation (AD)
- Kneser-Ney smoothing (KN)
- Simple discount (SD)

As was mentioned above, ML estimation uses the concept of relative frequency as a probability estimate. Once the counts of the phrase pairs have been obtained, three different well-known estimation techniques can be applied, namely, GT estimation and two estimation techniques based on the subtraction of a fixed quantity from all non-zero counts: AD estimation and KN estimation (see [9] for more details). In addition, we have implemented a very simple estimation technique (labeled as SD) which works in a similar way to AD estimation but it subtracts a fixed probability mass instead of a fixed count.

Lexical distributions A good way to tackle the problem of unseen events is the use of probability distributions that decompose phrases into words. Two different techniques are mentioned in [9] for this purpose: the noisy-or and an alternative technique which is based on alignment matrices. In our work we have applied another technique which consists in obtaining the IBM 1 model probability as defined in [11] for phrase pairs instead of sentence pairs (this distribution will be referred to as LEX).

4.2 Combining estimators

The statistical estimators described in the previous subsection can be combined in the hope of producing better models. In our work we have chosen three different techniques for combining estimators:

– Linear interpolation
– Backing-off
– Log-linear interpolation

The linear interpolation technique consists of making a linear combination of different estimators, ensuring that the weights of such combination determine a probability function. We have implemented linear combinations of two estimators. One of them is a phrase-based model estimator and the second one is the lexical distribution described in section 4.1. This combination scheme has been specially chosen to deal with unseen events.

The backing-off combination technique consults different models in order depending on their specificity. Again, we have implemented backoff models which combine two different estimators in the same way as has been described for the case of linear interpolation. In this particular case, only GT and SD estimation were implemented.

Finally, phrase-based model estimators and lower order distributions can also be combined by means of log-linear interpolation. In this case, the procedure consists of adding new components to the initial log-linear model described in section 3.1. Again, the main goal of the combination is to achieve good treatment of unseen events. For this purpose, lexical distributions in both directions are incorporated into the log-linear model as score components. In this case, only GT estimation was implemented.

5 Experimental Results

Different experiments were carried out in order to assess the proposed phrase-to-phrase alignment smoothing techniques.

5.1 Corpora and evaluation

The experiments consisted of obtaining phrase-to-phrase alignments between pairs of sentences following the different smoothing techniques described in the previous section. Specifically, a test set containing several sentence pairs to be aligned was used. The test set was taken from the shared tasks in word alignments developed in HLT/NAACL 2003 [12]. This shared task involved four different language pairs, but we only used English-French in our experiments.

A subset of the Canadian Hansards corpus was used in the English-French task. The English-French corpus is composed of 447 English-French test sentences and about a million training sentences.

We were interested in evaluating the quality of the phrase-to-phrase alignments obtained with the different phrase alignment smoothing techniques that we proposed. Unfortunately, there does not exist a gold standard for phrase alignments, so we needed to refine the obtained phrase alignments to word alignments in order to compare them with other existing word alignment techniques.

Taking these considerations into account, we proceeded as follows: Given a pair of sentences to be aligned we first aligned them at phrase level, obtaining

a phrase-to-phrase alignment. Afterwards, we obtained a word-to-word IBM1 alignment for each pair of aligned phrases. Finally, these "intra-phrase" word alignments were joined, resulting in a word level alignment for the whole sentence. We could thus make a fair comparison of the proposed smoothing techniques with the ones presented in the HLT/NAACL 2003 shared task.

To evaluate the quality of the final alignments obtained, different measures were taken into account: *Precision*, *Recall*, *F-measure*, and *Alignment Error Rate*. Given an alignment A and a reference alignment G (both A and G can be split into two subsets A_S,A_P and G_S, G_P, respectively representing *Sure* and *Probable* alignments) *Precision* (P_S,P_P), *Recall* (R_S, R_P), *F-measure* (F_S, F_P) and *Alignment Error Rate* (AER) were computed (see [12] for more details).

5.2 Alignment quality results

As described in [12], two different sets of evaluations were conducted:

- NULL alignments: given a word alignment $\mathbf{a}$ for a pair of sentences $(\mathbf{f}, \mathbf{e})$, if a word f_j ($j \in \{1 \ldots |\mathbf{f}|\}$) is not aligned with any e_i ($i \in \{1 \ldots |\mathbf{f}|\}$), or viceversa, that word is aligned with the NULL word.
- NO-NULL alignments: NULL alignments are removed, from the test set and from the obtained alignments.

In Table 1 the alignment quality results using different phrase-to-phrase alignment smoothing techniques are presented, for NO-NULL and NULL alignments. It is worth mentioning that the figures for *Sure* alignments are identical for NO-NULL and NULL alignments. In the table the first row shows the baseline, which consists of the results obtained using a maximum likelihood estimation (ML) without smoothing. The rest of the rows corresponds to different estimation techniques combined with linear interpolation except in those cases where a back-off (BO) or a log-linear interpolation (LL) were used.

For the NO-NULL alignment experiment, significant improvements in all alignment quality measures were obtained for all the smoothing techniques compared with the baseline. The baseline system results were worse due to the great number of times in which the segmentation of a sentence pair could not be completed due to coverage problems (in our experiments, 86.5% of the test pairs presented this problem); in such situations, the baseline system aligned all the source words of the source sentence with all the target words of the target sentence. Finally, it is worth pointing out that all those experiments that included the LEX distribution outperformed the others due to improved assignment of probabilities to unseen events.

With respect to the probability distribution used to model feature functions f_3 and f_5, we show the results corresponding to the use of a uniform distribution for f_3 and a geometric distribution for f_5, since such choices led to better results. As was mentioned in section 3.1, the use of a uniform distribution for f_3 penalizes the length of the segmentation and the use of a geometric distribution for f_5 makes it possible to establish a relationship between the length of source and target phrases (the use of a Poisson distribution also worked well).

	NO-NULL & NULL			NO-NULL				NULL			
Smooth. tech.	P_S	R_S	F_S	P_P	R_P	F_P	AER	P_P	R_P	F_P	AER
ML	64.39	76.62	69.98	77.49	28.31	41.47	20.04	55.10	29.38	38.32	36.42
GT	71.58	79.59	75.38	87.80	27.02	41.32	14.82	52.45	28.84	37.22	39.11
AD	69.11	77.64	73.13	84.02	26.56	40.36	17.12	51.10	28.10	36.26	40.18
KN	68.62	77.91	72.97	83.71	26.66	40.44	17.23	51.49	28.19	36.44	39.83
ML+LEX	72.56	83.31	77.57	89.67	28.37	43.10	12.03	55.39	30.09	39.00	35.80
GT+LEX	**72.64**	83.18	77.56	89.42	28.24	42.93	12.23	55.07	29.98	38.82	36.07
AD+LEX	71.92	81.95	76.61	90.03	27.80	42.48	12.55	54.25	29.58	38.29	37.10
KN+LEX	71.31	82.12	76.34	89.93	28.01	42.72	12.46	54.80	29.76	38.58	36.60
GT+LEX+BO	71.74	85.98	78.22	91.55	**29.64**	**44.78**	**09.77**	58.78	31.37	40.91	32.49
SD+LEX+BO	72.07	**86.16**	**78.44**	**91.52**	29.57	44.70	**09.77**	**59.09**	**31.45**	**41.05**	**32.18**
GT+LEX+LL	71.37	84.72	77.48	89.82	29.10	43.96	11.21	57.43	30.80	40.09	33.78

Table 1. Comparative alignment quality results (in %) using different smoothing techniques for NO-NULL and NULL alignments

It is also worth mentioning that despite the fact that phrase alignment techniques proposed here are not specifically designed to obtain word alignments, all the results are competitive with those presented in [12]. In the table, the best results for each column are highlighted showing that GT+LEX+BO and SD+LEX+BO obtained the best results.

Regarding the results for the NULL alignment experiment, there were small relative improvements in 5 out of 9 smoothing techniques compared with the baseline. The differences between these results and those for NO-NULL alignment experiment are due to the fact that the baseline generated a lot of alignments in which all words were aligned with all words due to coverage problems. In those situations, the IBM1 alignment model tended to align less words with the NULL word than when it was applied over intra-phrase alignments derived from successful segmentations of sentence pairs. If we compare column P_P of both experiments, a significant reduction in *precision* is obtained in the case of the NULL alignment experiment. This makes our results less competitive than those presented in [12] for the NULL alignment experiment.

According to these results, more research is needed in order to improve the intra-phrase word alignments. One possible solution is to use higher order word alignment models, for example HMM or IBM4 models.

6 Conclusions

We have presented a phrase-based statistical alignment model which can be used to obtain phrase-to-phrase alignments for pairs of sentences.

The proposed phrase-based statistical alignment model combines different smoothing techniques to overcome the coverage problems that the standard phrase-based models present.

The proposed system follows a loglinear approach which makes it possible to include different score components specifically designed to improve the phrase alignments.

Experimental results for a well-known shared task on word alignment evaluation have been reported. The results show the great impact of the smoothing techniques on alignment quality. As a step forward, we have also discussed the adaptation of the proposed model for its use in a CAT system.

Acknowledgments. This work has been partially supported by the Spanish research programme Consolider Ingenio 2010: MIPRCV (CSD2007-00018) and the EC (FEDER), the Spanish MEC under grant TIN2006-15694-CO2-01, the i3media project (CDTI 2007-1012) and the Spanish JCCM under grant PBI08-0210-7127.

References

1. García-Varea, I., Ortiz, D., Nevado, F., Gómez, P.A., Casacuberta, F.: Automatic segmentation of bilingual corpora: A comparison of different techniques. In: Proc. of the 2nd IbPRIA. Volume 3523 of LNCS., Estoril (Portugal) (June 2005) 614–621
2. Barrachina, S., Bender, O., Casacuberta, F., Civera, J., Cubel, E., Khadivi, S., Ney, A.L.H., Tomás, J., Vidal, E.: Statistical approaches to computer-assisted translation. Computational Linguistics (2008) In press
3. Zens, R., Och, F.J., Ney, H.: Phrase-based statistical machine translation. In: Advances in artificial intelligence. 25. Annual German Conference on AI. Volume 2479 of LNCS. Springer Verlag (September 2002) 18–32
4. Tomás, J., Casacuberta, F.: Monotone statistical translation using word groups. In: Proc. of the MT Summit VIII, Santiago de Compostela, Spain (2001) 357–361
5. Och, F.J., Ney, H.: Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In: Proc. of the 40th ACL, Philadelphia, PA (July 2002) 295–302
6. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proc. of the 41th ACL, Sapporo, Japan (July 2003) 160–167
7. Ortiz, D., García-Varea, I., Casacuberta, F.: Thot: a toolkit to train phrase-based statistical translation models. In: Proc. of the Machine Translation Summit X, Phuket, Thailand (September 2005) 141–148
8. 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., Constantin, A., Herbst, E.: Moses: Open source toolkit for statistical machine translation. In: ACL, Prague, Czech Republic (June 2007) 177–180
9. Foster, G., Kuhn, R., Johnson, H.: Phrasetable smoothing for statistical machine translation. In: Proc. of the EMNLP, Sydney, Australia, ACL (July 2006) 53–61
10. Manning, C.D., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts 02142 (2001)
11. 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**(2) (1993) 263–311
12. Mihalcea, R., Pedersen, T.: An evaluation exercise for word alignment. In Mihalcea, R., Pedersen, T., eds.: HLT-NAACL 2003 Workshop: Building and Using Parallel Texts: Data Driven Machine Translation and Beyond, Edmonton, Alberta, Canada, ACL (May 31 2003) 1–10