Translation Paraphrases in Phrase-Based Machine Translation

Francisco Guzmán and Leonardo Garrido

Center for Intelligent Systems
ITESM Campus Monterrey, Mexico
guzmanhe@gmail.com, leonardo.garrido@itesm.mx

Abstract. In this paper we present an analysis of a phrase-based machine translation methodology that integrates paraphrases obtained from an intermediary language (French) for translations between Spanish and English. The purpose of the research presented in this document is to find out how much extra information (i.e. improvements in translation quality) can be found when using Translation Paraphrases (TPs). In this document we present an extensive statistical analysis to support conclusions.

1 Introduction

Statistical methods have proven to be very effective when addressing linguistic problems, specially when dealing with Machine Translation [1]. There have been several attempts to improve the performance of such systems. Non-syntactic phrase-based translation systems[2] certainly outperform word-based systems[3].

Nevertheless, Statistical Machine Translation (STMT) effectiveness is limited to situations where large amounts of data are available. Such a condition, limits the performance of SMT systems over"low density" language pairs [4]. Scarce training data, often leads to a low coverage problem, that is, a low amount of learned translations for a language pair.

There are several efforts trying to improve translation quality of STMT systems. Many state-of-the-art systems involve the introduction of syntactic information to phrase-based machine translations [5,6,7,8,9].

On the other hand, we find several efforts which do not use syntactic information. One main topic of discussion is the usage of paraphrases. For example Callison [4] improves translation quality by giving alternatives to broaden coverage of a phrase-based machine translation system through the use of paraphrases. They use paraphrases in cases when a phrase is not found in their phrase-tables. Other effort is conducted by Guzman and Garrido [10] who obtain what they call "translation paraphrases" from pivoting through an intermediary language.

In this paper we analyze their methodology to assess whether the inclusion of Translation Paraphrases (TP) in a STMT system are useful to improve translation quality, in comparison to systems that do not include such features.

This paper is organized as follows: In Sec.2 we explain the methodology followed throughout our experimentation. In Sec.3 explain thoroughly the results

A. Gelbukh (Ed.): CICLing 2008, LNCS 4919, pp. 388–398, 2008.

and their statistical analysis. In Sec.4, we discuss the implications of results and we propose further research directions regarding translation paraphrases.

1.1 Translation Paraphrases

The strategy proposed by [10] to tackle the coverage problem is to extend phrase-tables that are used for phrase-based STMT with translation paraphrases learned from a third language. Figure 1 exemplifies this point. In their scope, translation paraphrases are the mechanism of preserving meaning through translation. While bridging through a third language, translation paraphrases are to give more flexible interpretations of source texts, as well as to reinforce translations that are more likely to be good translations regardless of the translation process.

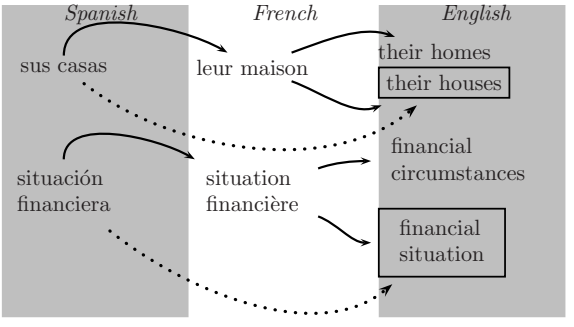


Fig. 1. Example of a translation paraphrase: When translating from Spanish to English with a Spanish-English trained phrase-table, we only get "their homes" English phrase as an alternative to "sus casas" Spanish phrase. However, if using translation paraphrases issued from French, we get "their homes" and "their houses" alternatives.

2 Experimental Setup

2.1 System Training

Every STMT system needs to be trained over a pair of aligned corpora. Aligned corpora are collections of documents, for which each line in one document has its counterpart in other language, which has been translated by a human (Fig. 2). Using the information in these documents, a STMT systems constructs a model that estimates the likelihood of a phrase in one language to be translated into another.

After the model training, we end up with a phrase-table, which is a collection of correspondences between phrases in both languages with their corresponding probability of being translated into each other.

In our experiments we trained the three combinations of language pairs in the set {English, French, Spanish }. Thus, we obtained three phrase tables: English-French, English-Spanish and French-Spanish. For the purposes of our experiments,

12 there needs firstly to be clarity between all of the groups of this house and then between this house and the commission .	**12** la clarté doit tout d'abord régner entre tous les groupes de cette assemblée et ensuite entre le parlement et la commission .
13 we should not find ourselves late in the day in the unfortunate position where the one or other institution creates an unnecessary fracture in institutional relationships .	**13** nous ne devons pas nous retrouver en fin de compte dans la position malheureuse où l'une ou l'autre institution crée une rupture inutile dans les relations institutionnelles .

Fig. 2. Example of an aligned corpora, extracted from Europarl corpus

we trained over aligned corpora containing 10k, 20k, 40k, 80k and 100k sentence-pairs (k-size) issued from the European Parliament Proceedings Corpus (Europarl) [11] from year 2001. For the model training, we used Giza++ [12].

2.2 Phrase Table Consolidation

In their paper Guzman and Garrido[10] describe a methodology for creating TPs from a trilingual aligned corpus. They combine the phrase-tables issued from training English-French and French-Spanish language pairs to obtain a English-Spanish Translation Paraphrases phrase-table using the following equation:

$$p_{\text{otp}}(e|s) = \sum_{f} p(e|f)p(f|s) \quad . \tag{1}$$

That is, the marginalized probability of translating a Spanish phrase to a French phrase and then translating that phrase to an English phrase. For instance, let us call the phrase-table containing these new probabilities, the Only-TP phrase table.

Furthermore, Guzman and Garrido describe a method for combining the Only-TP phrase-table with a phrase-table trained directly from English-Spanish (a Non-TP phrase-table) using the following model:

$$p_{\text{mix}}(e|s) = \alpha\, p_{\text{otp}}(e|s) + (1-\alpha)\, p_{\text{ntp}}(e|s) \quad . \tag{2}$$

In our experiments we trained Non-TP and Only-TP phrase-tables for each of the k-sizes and afterwards we combined them to produce mixed phrase-tables by using (2) while varying alpha from 0.1 to 0.9. After this stage, we ended up with 55 phrase tables (eleven for each k-size). For clarity, see Fig.3.

Having a phrase-table, half of the training to produce a STMT system is done. The second half is to fine-tune the decoder, which speaking generally is the piece of software that uses the information in the phrase-tables to produce a translation. With a phrase-table we can build a rough non-tuned STMT system that might be able of performing low quality translations. Therefore the second part of the training phase has to do with tuning the parameters of the decoder for an optimal output.

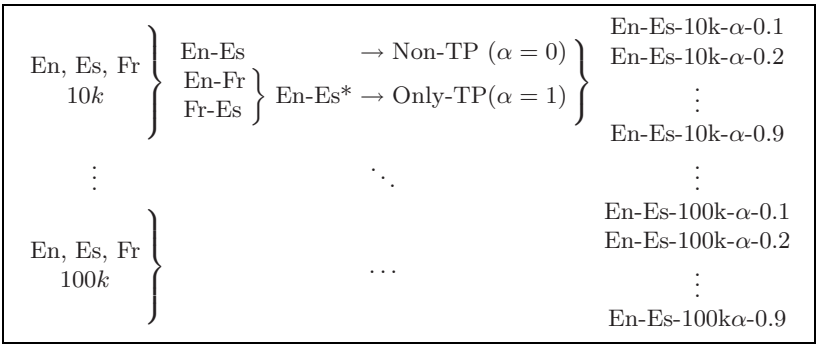


Fig. 3. Outline of the experimental procedure to merge phrase-tables. First we train the En-Es, Fr-Es and En-Fr models to obtain their corresponding phrase-tables. Then we obtain the translation paraphrase-table from merging En-Fr and Fr-Es phrase-tables. Finally, we merge the En-Es* translation paraphrase-table with the En-Es phrase-table at different levels to obtain the mixed phrase-tables.

2.3 MERT Training

The Minimum Error Rate Training (MERT) [13] is the process with which we tune the factors of the log-linear model described in Och and Ney in[14]. Roughly, the process consists in testing combinations of parameters and determine which combination give the best output. This is done by translating an specific document and then evaluating the translation quality. In other words, we train the decoder's parameters for it to be an "expert" in translating a given set of documents.

In this training phase we used a random subset of 100 lines randomly extracted from the documents of Europarl Corpus of 2002. We ran the MERT over each of the 55 phrase-tables, to ensure that each one was configured to perform at its best.

2.4 Translation

Upon the conclusion of our systems' training, we wanted to test the performance of each configuration against a controlled testing set of 30 samples. Each of those samples contained 50 lines of text, which were randomly extracted from the Europarl test set [11], containing documents from October 2000 to December 2000. This random sampling process was done in order to diminish the effects of the clustering of translation difficulty. An equivalent process was performed by [15] in their "broad sampling" where they followed a deterministic rule to form samples containing lines of text from different parts of the corpus.

For translating the samples, we used the moses decoder [16] with the parameters issued from the MERT training.

In order to evaluate the phrasal translation quality, we used the BLEU metric [17] (which is one of the standard measurements of quality of translation) with a single source of reference translation. Although recent studies suggest that BLEU's correlation with human judgments is not as strong as previously thought

[18], doing manual evaluation implies having infrastructure and resources (human judges, evaluation framework, etc.) which we do not currently possess.

3 Results and Discussion

The results of our experiments are summarized in Tab.1: At a first glance it is difficult to perform an analysis by just looking at these results. The first piece of information that out stands is the maxima for each group of training sentence pairs (k-group). From the table we can see that as the number of pairs increases, the maxima moves from an α of 0.6 at 10k to 0.5 at 20k and 40k; and to 0.4 at 80k and 100k. This suggests that as we increase the training data size, translation paraphrases become less handy.

To better analyze the information gathered throughout our experiments we performed an statistical analysis for each group.

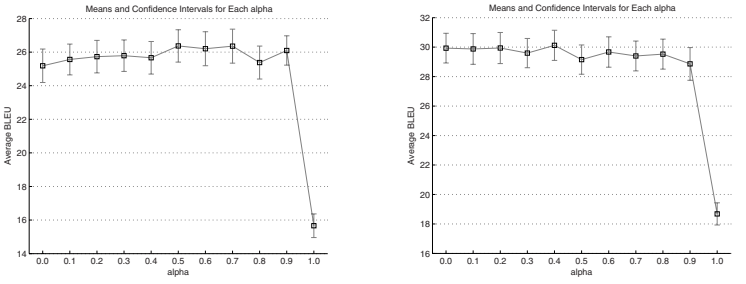
Table 1. Experimental results presented by alpha and number of training sentence pairs. For each registry we have the average BLEU of the 30 translation problems ($\bar{x}$) and their standard error($\sigma_{\bar{x}}$).

	10k		20k		40k		80k		100k	
α	$\bar{x}$	$\sigma_{\bar{x}}$	$\bar{x}$	$\sigma_{\bar{x}}$	$\bar{x}$	$\sigma_{\bar{x}}$	$\bar{x}$	$\sigma_{\bar{x}}$	$\bar{x}$	$\sigma_{\bar{x}}$
0.0	24.01	0.43	25.19	0.49	27.58	0.44	29.93	0.49	30.60	0.48
0.1	24.22	0.46	25.56	0.45	28.08	0.48	29.88	0.51	30.51	0.51
0.2	24.39	0.46	25.73	0.47	28.17	0.47	29.94	0.52	30.40	0.48
0.3	24.36	0.46	25.79	0.46	28.04	0.46	29.60	0.49	30.69	0.49
0.4	24.24	0.46	25.66	0.47	28.26	0.46	**30.12**	0.50	**30.72**	0.48
0.5	24.23	0.44	**26.37**	0.47	**28.69**	0.45	29.15	0.49	30.43	0.50
0.6	**24.61**	0.46	26.20	0.49	28.23	0.45	29.67	0.50	30.31	0.48
0.7	24.10	0.47	26.36	0.50	28.34	0.46	29.40	0.50	30.63	0.47
0.8	24.18	0.46	25.38	0.48	28.23	0.44	29.52	0.50	30.29	0.47
0.9	23.94	0.43	26.10	0.43	27.44	0.45	28.86	0.54	29.51	0.44
1.0	13.74	0.35	15.66	0.34	17.31	0.37	18.68	0.37	19.39	0.37

3.1 Confidence Intervals

Since our research question was to find if we obtained any improvements in translation quality by the usage of TPs, a good starting point was to look at the confidence intervals for each mean, to determine if the systems belong to the same population (that is, they share a common population mean and thus, no improvement has been made).

In the Fig. 4 we display the confidence intervals for 20k and 80k to a level of 95%. As one can see, for both graphs it might seem clear that, excluding the rightmost system (α=1.0), the others belong to the same group. Nevertheless, this conclusion wouldn't be as valid if we had different groups. For instance, if for 20k we only analyze the confidence intervals for α=$\{0.0,0.5,1.0\}$ we wouldn't



(a) **(b)**

Fig. 4. Confidence Intervals for (a)20k and (b) 80k k-sizes

draw the same conclusions since their means do not fall into each other's confidence intervals, suggesting that their means are different. Therefore we needed to perform other analysis to obtain sound conclusions. Withal, what we do can conclude is that a system trained with only translation paraphrases (α=1.0) perform worse than any other system.

3.2 One Way Analysis of Variance (ANOVA)

The ANOVA test is useful when dealing with several groups for which we want test if they belong to a single population (meaning that they share the same mean). Although, it serves only to test the null hypothesis of all means being equal, and does not tells us anything about differences between individual groups, it is relevant that we reject the experiment's null hypothesis (all means are equal) so we can decouple individual groups to do pairwise comparison under a least significant differences scheme (LSD).

In Tab. 2 we show the results of the ANOVA tests for every k-group. As we can see the p-values are very low (basically zero), allowing us to reject the experiment's null hypothesis for each of the k-groups. This is not news, because from plotting confidence intervals we could see that the mean performance of the systems with α=1.0 was very different from the others. But as we said before, rejecting the experiment's null hypothesis allows us to perform pairwise comparisons.

3.3 Unplanned Pairwise Comparisons

Since we had many groups of α (α-groups) to compare, we decided to analyze only three α-groups: Non-TP (α=0.0), Best-TP (the α-group with best performance within a k-group) and Only-TP (α=1.0). The method for comparisons that we used was an approximate randomization test for the paired sample comparison of means. We used this test because for every system, we ran them over the same set of problems, and therefore the different samples were not independent. Besides, being a computer intensive method, we needed not to care

Table 2. ANOVA results for each k-group

k-size	Source	SS	df	MS	F	pvalue
	Groups	3011	10	301.1	51	0.00
10k	Error	1883.3	319	5.9		
	Total	4894.3	329			
	Groups	2866	10	286.6	44.87	0.00
20k	Error	2037.7	319	6.39		
	Total	4903.7	329			
	Groups	3213.4	10	321.34	53.13	0.00
40k	Error	1929.3	319	6.05		
	Total	5142.7	329			
	Groups	3297.9	10	329.79	45.32	0.00
80k	Error	2321.1	319	7.28		
	Total	5619	329			
	Groups	3343.5	10	334.35	50.05	0.00
100k	Error	2131	319	6.68		
	Total	5474.5	329			

about parametric distributions' requirements for validity. The only assumptions we made is that our samples are random and representative.

3.4 The Approximate Randomization Test

This test allow us to test the null hypothesis that the means of α-groups are equal. Having two samples of individuals S_A and S_B formed by the BLEU metrics for the translation of each problem in the test set, the test statistic for a pairwise comparison of two means is given by the expression:

$$\theta = \sum_{i=1}^{N}(a_i - b_i)/N \text{ for } a_i \in S_A \text{ and } b_i \in S_B \tag{3}$$

where N is the number of translation problems given to each system (30).

In a randomization test, we randomly shuffle the sign of individual differences to get a sampling distribution of θ^* which is a pseudo-statistic.

Running 9999 iterations of this randomization test for the pairs (Non-TP,Best-TP), (Non-TP,Only-TP) and (Best-TP,Only-TP) for k=80k, we get the $\theta*$ distributions displayed in Fig.5.

From these distributions we can take the probability of $P(\theta^* \geq \theta)$ which is displayed in the Tab. 3.

Using these results, we can run the following hypothesis: $H_0 : \theta = 0$ using the alternative hypothesis of $H_1 : \theta > 0$ for each one, using the probability just obtained as the p-value.

As we can see, there is strong evidence that suggests that the means of Best-TP and Non-TP groups are greater than the mean of the Only-TP group. But

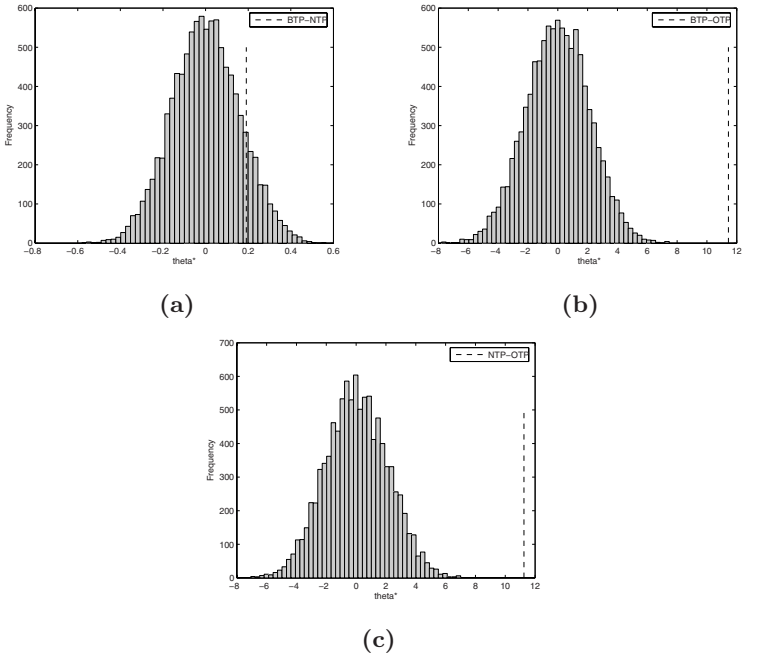


Fig. 5. Sampling distributions for the paired-sample comparisons of: (a) Best-TP vs Non-TP, (b) Best-TP vs Only-TP and (c) Non-TP vs Only-TP

Table 3. Probabilities of $p(\theta^* > \theta)$ for 80k

comparison	$p(\theta^* > \theta)$
Best-TP vs Non-TP	0.1112
Best-TP vs Only-TP	0.0001
Non-TP vs Only-TP	0.0001

there is not enough evidence that can ensure that the means of Non-TP and Best-TP groups are not equal at a significance level of 5%. For this last comparison, we should then keep the null hypothesis.

3.5 Summarized Comparisons

In the Tab.4 we show the p values for every pair and every k-size.

From this table, we observe that for k-groups 10k, 20k and 40k the Best-TP systems perform better than the Non-TP systems. Nevertheless as the size of training data increases (80k ,100k) such improvements are not significant. Therefore we can say that mixed TP systems lead to significant improvements in translation quality when training resources are scarce.

Table 4. Summarized results for pairwise comparisons presented by comparison and k-group

	Non-TP vs Best-TP		Best-TP vs Only-TP		Non-TP vs Only-TP	
size-k	θ	pvalue	θ	pvalue	θ	pvalue
10k	0.6003	0.0001	10.87	0.0001	10.268	0.0001
20k	1.1817	0.0000	10.71	0.0001	9.52	0.0001
40k	1.1	0.0000	11.38	0.0001	10.27	0.0001
80k	0.1910	0.1112	11.443	0.0001	11.252	0.0001
100k	0.121	0.1116	11.325	0.0001	11.204	0.0001

3.6 Best Practical Bound

In figure 6, we observe the average BLEU for groups Non-TP and Best-TP as well as the best practical bound (BPB). The BPB is the average of the best scores for each of the translation problems on the experiment at every k-group. The BPB allows us to detect ceiling effects (very hard problems that might obscure results).

This graph helps us to notice that at low k-values the Best-TP is close to the BPB but the differences with the Non-TP are significant so we can conclude that the Best-TP is the system that performs the best under almost every problem. Nevertheless, as k-size increases Non-TP and Best-TP become closer to each other, but distant from the BPB. This suggests that both could be performing better. Therefore we can conclude that no ceiling effect was observed and thus our results hold valid.

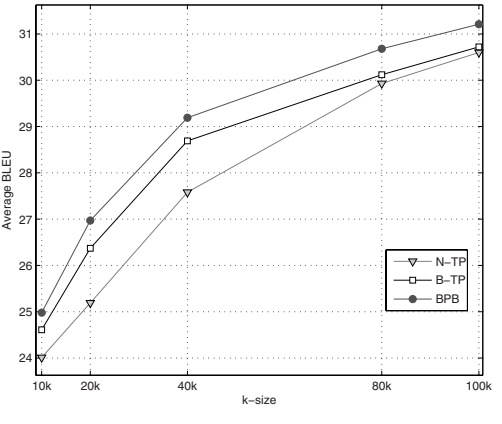


Fig. 6. Average BLEU vs. k-size for Best-TP, Non-TP and the Best Practical Bound (BPB)

4 Conclusions and Future Work

In this paper we analyzed the results obtained from using the translation paraphrases proposed by [10]. From our experiments we can draw the following conclusions:

1. As we increase the size of training corpora, we observe that the best translations are found at lower alphas, suggesting that for large training corpora, TPs have a lower impact.
2. For small training sizes, there is evidence that suggests that there is a significant improvement in translation quality by the utilization of TPs but at larger levels, there is no statistical evidence that suggest that a system's performance is affected by TPs. Therefore we can conclude that TPs bring significant improvements when dealing with scarce data.
3. TPs by themselves produce poor translations. Therefore they should not be used alone, but merged into phrase-tables of Non-TP systems.
4. There was no evidence that showed that we ran into a ceiling effect.

To assess the feasibility of using TPs as a translation aid, we need to test their translation quality improvements when dealing with scarce data. That is to test the improvements from merging small-corpus-trained Non-TP phrase-tables with large-corpus-trained Only-TP phrase-tables to verify whether the translation quality is bound (or not) to the size of the Non-TP phrase-tables.

Other experiments that are to be done is to assess the benefits of using TPs when addressing out-of-domain translation problems. So far we have been working under the same context: Europarl. It would be interesting to test the performance of TPs when dealing with translation problems from other sources. This could shed some light over the possibility of using TPs as a resource for out-of-domain translations.

Finally, we suspect that TPs' effectiveness is bound to the intermediate language used. In this study we used French as an intermediate between English and Spanish, because it seemed somewhat intuitive (French is related to Spanish and English). Nevertheless this assumption might not be optimal. Therefore an exploratory study to assess which intermediate language performs better for a given pair of languages, using information from contrastive linguistics, will be of great interest.

References

1. Brown, P.F., et al.: The mathematics of statistical machine translation: parameter estimation. Computational Linguistics 19, 263–311 (1993)
2. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Canada (2003)
3. Zens, R., Ney, H.: Improvements in phrase-based statistical machine translation. In: Proceedings of the Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL), Boston, MA, pp. 257–264 (2004)

4. Callison-Burch, C., Koehn, P., Osborne, M.: Improved statistical machine translation using paraphrases. In: Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Association for Computational Linguistics, Morristown, NJ, USA, pp. 17–24 (2006)
5. Langlais, P., Gotti, F.: Phrase-based smt with shallow tree-phrases. In: Proceedings of the Workshop on Statistical Machine Translation, Association for Computational Linguistics, New York City, pp. 39–46 (2006)
6. Giménez, J., Màrquez, L.: Combining linguistic data views for phrase-based SMT. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 145–148 (2005)
7. Alexandra Birch, M.O., Koehn, P.: Ccg supertags in factored statistical machine translation. In: ACL Workshop on Statistical Machine Translation (2007)
8. Hassan, K.S.H., Way, A.: Supertagged phrase-based statistical machine translation. In: 45th Annual Meeting of the Association for Comp. Linguistics (2007)
9. Vilar, J.M., Vidal, E.: A recursive statistical translation model. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts, Ann Arbor, Michigan, Association for Computational Linguistics, pp. 199–207 (2005)
10. Guzman, F., Garrido, L.: Using translation paraphrases from trilingual corpora to improve phrase-based statistical machine translation: A preliminary report. In: MICAI (2007)
11. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. MT Summit 2005 (2005)
12. Och, F., Ney, H.: Statistical machine translation. In: EAMT Workshop, Ljubljana, Slovenia, pp. 39–46 (2000)
13. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proc. of the Association for Computational Linguistics, Sapporo, Japan (2003)
14. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 295–302 (2002)
15. Koehn, P.: Statistical significance tests for machine translation evaluation. In: EMNLP (2004)
16. Koehn, P., et al.: Moses: Open source toolkit for statistical machine translation. In: Annual Meeting of the Association for Computational Linguistics (ACL), Prague, Czech Republic (2007)
17. Papineni, K., et al.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the Association of Computational Linguistics, pp. 311–318 (2002)
18. Chris Callison-Burch, M.O., Koehn, P.: Re-evaluating the role of bleu in machine translation research. In: EACL (2006)