HAL open science

Phrase-Based Language Model in Statistical Machine Translation

Achraf Ben Romdhane, Salma Jamoussi, Abdelmajid Ben Hamadou, Kamel Smaïli

To cite this version:

Achraf Ben Romdhane, Salma Jamoussi, Abdelmajid Ben Hamadou, Kamel Smaïli. Phrase-Based Language Model in Statistical Machine Translation. International Journal of Computational Linguistics and Applications, 2016. hal-01336485

HAL Id: hal-01336485

https://inria.hal.science/hal-01336485v1

Submitted on 23 Jun 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Phrase-Based Language Model in Statistical Machine Translation

Achraf Ben Romdhane[1], Salma Jamoussi[1],
Abdelmajid Ben Hamadou[1], and Kamel Smaïli[2]

[1] MIRACL Laboratory, University of Sfax, BP 3023 Sfax, TUNISIA,
{achraf.ramdhan, jamoussi, abdelmajid.benhamadou}@gmail.com,
[2] SMa^{r}T, LORIA, Campus scientifique, BP 139, 54500 Nancy, France,
Smaili@loria.fr

Abstract. As one of the most important modules in statistical machine translation (SMT), language model measures whether one translation hypothesis is more grammatically correct than other hypotheses. Currently the state-of-the-art SMT systems use standard word n-gram models, whereas the translation model is phrase-based. In this paper, the idea is to use a phrase-based language model. For that, target portion of the translation table are retrieved and used to rewrite the training corpus and to calculate a phrase n-gram language model. In this work, we perform experiments with two language models word-based (WBLM) and phrase-based (PBLM). The different SMT are trained with three optimization algorithms MERT, MIRA and PRO. Thus, the PBLM systems are compared to the baseline system in terms of BLUE and TER. The experimental results show that the use of a phrase-based language model in SMT can improve results and is especially able to reduce the error rate.

Keywords: Machine Translation, Phrases, Phrase based language model, Decoding optimization

1 Introduction

Machine translation systems have evolved since several decades from the use of a word to the use of a sequence of words (phrases) as basic units for translation. Currently, all the Statistical Machine Translation (SMT) systems are based on phrases. Succinctly, in the decoding step, the source sentence is segmented into phrases, each phrase is then translated into the target language and finally phrases are reordered [1]. At each step of the decoding phase, translation hypotheses are created and expanded until all words of the source sentence are covered. The expansion step produces a huge number of possible translations which are constrained by a cost estimation depending on many parameters such as the language model and the translation model probabilities. To achieve good translation quality, SMT researchers make a lot of effort in improving the translation model which moved from the original single-word-based to phrase-based

model [1], in order to better capture the context dependencies of words in the translation process. Despite the improvements made in language modelling [2], [3], the state-of-the-art SMT systems still use standard word n-gram models.

The idea, in this paper is to enhance the quality of SMT systems by improving their Language Models (LM). For that, we propose to use a phrase-based LM. This kind of models has already shown good performances in speech recognition tasks [4], [5], [6], [7] and hopefully it can improve the quality of SMT systems as well.

In SMT, few research work has been done for designing phrase-based language models. Many researchers attempt to improve the translation system with Continuous-Space Language Models (CSLM) [8], [9]. These works have shown that CSLM can improve the SMT when compared to back-off n-gram language models (BNLM). Schwenk in [8] uses a CSLM in reranking step of the n-best translation hypotheses. In this approach, the authors uses two differents language models. BNLM is used in decoding step to produce a n-best translation. Then, a CSLM is used to rerank those translations. In the same way, [10] proposes a method for converting CSLM into BNLM that will be used directly in SMT decoding.

Recently, [11] proposes a phrase-based language model for statistical machine translation. This model considers the sentence segmentation into phrases as a new variable to estimate sentence probability. For the sentence segmentation, all the word sequences are considered as phrases. This leads to huge phrases vocabulary and increases the number of possible segmentation solutions. In the same way, [12] propose a phrase-based LM using a limited set of sequences. This work, has shown that the Kneser-Ney phrase-based LM cannot outperform the Kneser-Ney word-based LM. This can be explained by the strict phrases vocabulary size constraint and the data sparsity problem. In [20], the authors showed that the phrase-based language model could improve the translation quality of SMT. In this work, the Interlingual Triggers based on Conditional Mutual Information (CMI) are used to extract a translation table. The target sequences of this table are used for segmenting the training target corpus phrases in order to learn a Phrase-Based Language Model (PBLM).

In this paper, we propose a new method for using a phrase-based language model in machine translation (SMT-PBLM) and we argue that this model can improve the SMT quality. This method is not limited on a small set of phrases. Indeed, we use all the phrases of translation table for rewriting the training corpora where the words are joined to constitute phrases. The new corpora is used to train the phrase-based language model. Then, we tried three optimization algorithms (MERT, MIRA and PRO) to tune our translation system. Finally, we developed an optimization method in order to better take into account the change of the structure of lexical units in our SMT-PBLM.

In Section 2, we present the phrase-based language model. Then, in section 3, we give an overview of the sentence extraction method and we present how to define a phrase-based training corpus in section 4. A description of the used corpora and the results achieved are presented and discussed in section 5. We

end with a conclusion which points out the strength of our method and gives some tracks about future work.

2 Phrase-Based Language Model

Language models play an important role in SMT, it measures whether one translation hypothesis is grammatically better than other hypotheses. A statistical language model assigns a probability $P(e_1^m)$ to string of words $e_1^m = w_1, w_2..., w_m$. In the case of n-gram model the probability is calculated as follows:

$$P(e_1^m) = \prod_{i=1}^{m} P(w_i|w_1^{i-1}) \tag{1}$$

after making the Markov assumption which retains the most recent $n-1$ words to predict the next word.

$$\prod_{i=1}^{m} P(w_i|w_1^{i-1}) \approx \prod_{i=1}^{m} P(w_i|w_{i-n+1}^{i-1}) \tag{2}$$

Currently, most of the state-of-the-art MT systems use word-based n-gram language model, considers a sentence as a suite of words, whereas phrase-based translation model considers a sentence as a suite of phrases. In this paper, we propose a phrase-based language model in order to use the same units for both language and translation models.

In phrase-based SMT, the target translation hypothesis is a suite of translation options. Each translation option represents a word or phrase. A phrase-based language model considers the $n-1$ previous phrases to predict the next phrase. Given a translation hypothesis $h = h_1h_2...h_n$, the phrase-based language model calculates the following probability:

$$P(h) = P(h_1h_2...h_n) = p(h_1)p(h_2|h_1)...p(h_n|h_1...h_{n-1}) \tag{3}$$

where h_i is a translation option (word or phrase). The phrase based LM can capture longer relationship easily e.g. a 5-gram phrase-based language model could take into account phrases up to 40 words (in the case of a 5-gram where each gram is composed of a phrase of 8 words). A word-based language model is actually a special case of a phrase-based language model, where each phrase is reduced to its smallest unit: a word.

3 Extracting Phrases

The goal is to select automatically a set of phrases that will be extracted and integrated into the training corpus and consequently into the language model. In [6], the authors determined the sequences which reduce the perplexity of a language model. This, is an iterative process that merges at each iteration, word

pairs into a single unit (phrase) such as "white house" which will be transformed into "white_house". At each iteration, the best phrases according to a selection criterion are retained and used into the training corpus. This algorithm has been applied to a monolingual corpus and at the end of this process the authors obtain a set of phrases and a training corpus re-written in terms of words and phrases.

Other work focuses on the extraction of sequences based on a bilingual corpus. There are two main methods. One is based on word-to-word alignment [1], thus the phrases are selected with a set of heuristics. Any contiguous source words must be a translation of any contiguous target words under the condition that words from both sides are mapped to each other. The second method extract phrases without word alignment. Lavecchia et al. in [13] proposed to determine correlations between words coming from two different languages. This method use the intra-lingual triggers to extract the source phrases then the inter-lingual triggers associates to each source phrase of n words a set of target sequences of variable size m. In fact, for each source phrase of k words, one or more target sequences of length $k \pm \Delta k$ are retained.

In this work we propose to use Moses toolkit to train phrase-based translation model from parallel corpora [14]. Then we collect all target phrases from translation table and use them to train the phrase-based language model. Afterward we merge the words of the target phrases and we replace these in the translation table by the new phrases without modifying the initial parameters ($P(e|f)$,$P(f|e)$,$lex(e|f)$,$lex(f|e)$) of the translation table (see 1). This is done due to the fact that the translation process would propose them such as translations which will be used in the target language model. Table 1 illustrates some examples of phrases in the new translation table.

Table 1. Some examples of phrases in the intial and new translation tables where $P(e|f)$ is direct phrase translation probability, $lex(e|f)$ is direct Lexical weighting, $P(f|e)$ is an inverse phrase translation probability and $lex(f|e)$ is an inverse lexical weighting.

	source (f)		target (e)		$P(e\|f)$	$lex(e\|f)$	$P(f\|e)$	$lex(f\|e)$
	un petit peu plus	|||	a little bit more	|||	0.13	0.044	0.588	0.0206
Initial	un petit peu plus	|||	a little more	|||	0.0447	0.02261	0.176	0.070
Translation	voudrais dire	|||	'd like to say	|||	0.2	0.0089	0.181	0.0029
Table	voudrais dire	|||	want to say	|||	0.105	0.0060	0.363	0.017
	voudrais dire	|||	would like to say	|||	0.5	0.0043	0.181	0.0023
	voudrais dire	|||	would like to tell	|||	1	0.0045	0.090	0.0012
	un petit peu plus	|||	a_little_bit_more	|||	0.13	0.044	0.588	0.0206
New	un petit peu plus	|||	a_little_more	|||	0.0447	0.02261	0.176	0.070
Translation	voudrais dire	|||	'd_like_to_say	|||	0.2	0.0089	0.181	0.0029
Table	voudrais dire	|||	want_to_say	|||	0.105	0.0060	0.363	0.017
	voudrais dire	|||	would_like_to_say	|||	0.5	0.0043	0.181	0.0023
	voudrais dire	|||	would_like_to_tell	|||	1	0.0045	0.090	0.0012

4 Rewriting the training corpus

It is a crucial issue to rewrite adequately the training corpus in order to train a phrase-based language model. Due to the high number of phrases in the translation table, it is necessary to take into account the issue of the overlapping phrases in the segmentation phase of the same sentence. For example in the sentence "*a lot of sociologists actually are quite disappointed*" and the following list of sequences ($lot_of, a_lot, a_lot_of, are_quite$, $actually_are$) what are the phrases to use for merging words in the above sentence? $actually_are$ or are_quite ? At this point, there are two possible solutions. The first solution consists in finding the segmentation that minimizes the likelihood or the perplexity of the sentence. This ensures a good quality language model in terms of perplexity but increases the problem of OOV words because the minimization of the perplexity in the experiments we achieved, reduces the number of phrases used. In fact, only those that guarantee an improvement of the perplexity are selected. The second solution do not segment in order to avoid the errors of wrong segmentation. What we propose here is to increase the size of the training corpus by inserting all the possible segmentations. Indeed, we can produce all the possible hypotheses directed by using the right part phrases of the translation table. For example, for the sentence "*a lot of sociologists actually are quite disappointed*", there are four possible segmentations with the list of sequences given in the above example:

a lot_of sociologists actually_are quite disappointed
a lot_of sociologists actually are_quite disappointed
a_lot of sociologists actually are_quite disappointed
a_lot of sociologists actually_are quite disappointed
a_lot_of sociologists actually_are quite disappointed
a_lot_of sociologists actually are_quite disappointed

This method ensured that all the sentences of the translation table will be integrated in the training corpus of the language model. Thus, we get a bilingual training corpus where the target corpus is written in terms of words and phrases containing some redundancy of sentences. This redundancy would be useful for sequences that have low number of occurrences. The obtained corpus is then used to train the phrase-based language model.

5 Experiments

This section describes the performance of the proposed language model in a machine translation task. We use the IWSLT2010 test data for the French to English translation system [15]. Table 2 provides relevant statistics about the data used. The language models (word and phrase based) have been trained with SRILM toolkit [16]. The word alignment of the parallel corpora is generated by GIZA++ Toolkit [14] in both directions. Afterwards, the alignments are combined using the grow-diag-final-and heuristic to obtain symetric word

Table 2. Details about corpora used in this work.

Corpus	Sentences	Tokens (fr-en)	Unique Tokens (fr-en)
Train	176857	3.4M-3.1M	75K-58K
Dev (2010)	887	20K-20K	3.8K-3K
Test (2010)	1664	34K-32K	4.7K-3.8K

alignment model [1]. For decoding we used Moses toolkit [17] and finally an optimization algorithm is applied to estimate the optimal value of each weight on the development data set. Eight SMT features are used namely: Bidirectional phrase translation probability $(p(e|f), p(f|e))$, Bidirectional lexical probability $(lex(e|f), lex(f|e))$, Phrase penalty, Word penalty, Distortion, language model (word or phrase based). It should be noted that the system parameters were trained on the development corpus dev2010 and the test is performed on the corpus tst2010. In this evaluation, we compare the performance of the different language models in a translation task in terms of BLEU [21], TER [24] and (TER-BLEU)/2. BLEU is an n-gram precision metric, i.e. higher values are better, while TER is an error rate, i.e. lower values are better and (TER-BLEU)/2 aims at simultaneously maximizing BLEU and minimizing TER. Our SMT-PBLM is compared to a baseline SMT. For the baseline system, we evaluate 3 language models (5, 4 and 3-grams standard word-based language models). For the SMT-PBLM, we use 5, 4, 3 and 2 grams. In this work, we observed that the optimization algorithms are not well suited to the use of phrase-based language model. That is why we used, obviously the most widely algorithm MERT[3], but also we tried two other optimizations algorithms PRO[4] and MIRA[5]. Finally, we were not satisfied by the results, we developed an optimization method inspired from the greedy search algorithm that will be explained in section (5.3).

5.1 MERT optimization

An optimization procedure in SMT attempts to improve translation quality by searching the weights minimizing a given error measure, or equivalently maximizing a given translation metric. MERT, proposed by [14] obeys to this concept. Table 3 shows the results obtained with SMT-WBLM and SMT-PBLM where the optimization task is realized with MERT algorithm.

In terms of BLUE we find that the word-based language model is more efficient than the phrase-based language model. Table 3 shows that it could be acceptale to use just a bigram of phrase-based language model since the results are similar to 4-grams PBLM. This could be explained by the fact that, rarely succession of 5 or 4-grams of phrase have been found on the test corpus. In terms

[3] MERT: Minimum Error Rate Training
[4] PRO: Pairwise Ranking Optimization
[5] MIRA: Margin-Infused Relaxed Algorithm

Table 3. Results for the French → English translation task with MERT optimization algorithm.

System	Dev BLEU	Test BLEU	Test TER	Test (TER-BLEU)/2
SMT-WBLM LM(5-grams)	26.45	**33.23**	49.68	8.225
SMT-WBLM LM(4-grams)	26.41	33.18	49.49	**8.155**
SMT-WBLM LM(3-grams)	26.40	33.09	49.93	8.42
SMT-PBLM LM_{phrase}(5-grams)	25.60	32.69	49.13	8.22
SMT-PBLM LM_{phrase}(4-grams)	25.34	32.63	**49.07**	8.22
SMT-PBLM LM_{phrase}(3-grams)	25.32	32.58	49.31	8.365
SMT-PBLM LM_{phrase}(2-grams)	25.35	32.64	49.22	8.29

of TER, SMT-PBLM gives better results than the best SMT-WBLM. The improvement is of 0.42 in comparison to SMT-WBLM (4-grams) and of 0.55 in comparison to SMT-WBLM (5-grams). This means that it is easier for a human being to correct a translation achieved by SMT-PBLM than with SMT-WBLM. Overall, the combined score (TER-BLEU)/2 is substantially improved.

For the test corpus, in terms of BLEU the difference is of 0.54, while for the development it is of 0.85. This could be explained by the fact that the MERT algorithm is not able to handle adequatly the parameters of the SMT-PBLM. In fact phrases are composed of words which have been concatenated. Consequently, they are considered such as words and not as sequence of words.

Table 4. Weight of each parameter obtained by MERT

System	WP	PP	D	LM	TM			
SMT-WBLM	-0.37	0.05	0.11	0.13	0.08	0.075	0.09	0.06
SMT-PBLM	-0.21	0.31	0.10	0.10	0.10	0.008	0.07	0.07

Table 4 shows the weight of each parameter obtained by MERT. For SMT-PBLM, we notice that the weight of the phrase penalty (PP) is much greater than the weight of the word penalty (WP). This means that the decoder promotes the translations that have a large number of sequences and small number of words. Unlike the baseline system that increases the number of words and minimize the number of sequences in the translation. Consequently, we need an other algorithm of optimization which can correctly handle the phrases of our language model. In the following, we will test other algorithms of optimizations. In Table 5, we presents some translations achieved by both MT systems, we can observe that globally our system achieves well translated sentences, and sometimes better and more understandable than the baseline system. Furthermore, we can see that the last two sentences produced by our SMT-PBLM are closer to the reference sentencs than those given by SMT-WBLM.

Table 5. Few examples of translations based on the phrase-base language model

SRC	et c' est ce qui s' est passé à la fin de ces trois mois .
REF	and that 's what happened at the end of that three month period .
SMT-WBLM	and this is what happened to the end of these three months .
SMT-PBLM	and that_'s what_happened at the_end of_these three months_.
SRC	qu' est-ce qui définit une histoire ?
REF	what defines a story ?
SMT-WBLM	what is it that defines a story ?
SMT-PBLM	what that_defines a story_?
SRC	le deuxième facteur concerne les services que nous utilisons .
REF	the second factor is the services we use .
SMT-WBLM	the second aspect is about the services that we use .
SMT-PBLM	the_second_aspect is the_services_we_use .
SRC	ces dernières années , nous avons commencé à apprendre des choses sur le bonheur des deux " moi " .
REF	so in recent years , we have begun to learn about the happiness of the two selves .
SMT-WBLM	in the last few years , we 've started to learn things about happiness of the two " me . "
SMT-PBLM	in_recent years ,_we started learn_things_about happiness of_the two "_me ._"

5.2 MIRA and PRO optimization

Alternative discriminative parameters training algorithms for SMT have been proposed in the last few years, such as using a margin infused relaxed algorithm (MIRA) [19] and pairwise ranking optimization (PRO) [18]. PRO generates lists of k-best candidate translations for each sentence, and tunes the weight vector for those candidates. This method seeks the weight vector which classifies pairs of candidate translations into correctly ordered and incorrectly ordered, based on the scoring function. Tables 6 and 7 show the obtained results by the baseline and the SMT-PBLM where the optimization task is realized respectively with PRO and MIRA. From Table 6 we see that MIRA algorithm allows us to

Table 6. French → English translation result with MIRA optimization algorithm.

	Dev	Test		
System	BLEU	BLEU	TER	(TER-BLEU)/2
SMT-WBLM	26.30	**33.16**	49.86	8.35
SMT-PBLM	25.19	32.80	**49.18**	**8.19**

improve the results of SMT-PBLM. Indeed, we achieve a BLEU of 32.80 with MIRA, when we obtained a BLEU of 32.69 with MERT. An improvement of the results with MIRA has been achieved for BLEU and TER. The improved results are presented in bold style in Table 6. This shows that it is possible to

Table 7. French → English translation result with PRO optimization algorithm.

	Dev	Test		
System	BLEU	BLEU	TER	(TER-BLEU)/2
SMT-WBLM	25.62	**33.92**	**48.06**	**7.07**
SMT-PBLM	24.81	32.49	48.41	7.96

improve the SMT-PBLM by using the adequate optimization algorithm. From Table 7, we can see that PRO on the baseline system has greatly improved results compared to MERT and MIRA algorithms. But, for the SMT-PBLM we notice a diminution of 0.2 points BLEU compared to MERT and 0.4 points BLEU compared to MIRA. This decrease is explained by the fact that PRO applies a clustering method on the n-best distinct hypothesis (in our experiment $n = 100$) to extract the features weight. This constraint disadvantages SMT-PBLM because in the n-best we have several sentences which are different in terms of segments but exactly the same in terms of words. Table 8 shows an example of 4-best hypothesis which are different in terms of sequences but are identical in terms of words.

Table 8. Example of 4-best hypothesis.

today everybody_speaks of_happiness .
today everybody_speaks of_happiness_.
today everybody speaks of_happiness .
today everybody_speaks of happiness_.

Using the PRO algorithm for this type of n-best can distort the classification task in PRO. Indeed, it can classify the same sentence (in terms of words) in two different classes. In our experiments, we found that only 30% of the n-best are really distinct in terms of words.

To solve this problem, we apply PRO algorithm on the n-best distinct hypothesis in terms of sequences and words. Table 9 shows the obtained results where we provide to PRO n-best distinct sentences in terms of phrases and words.

Table 9. Obtained results when updating the n-best list.

	Dev	Test		
System	BLEU	BLEU	TER	(TER-BLEU)/2
PBLM-SMT (distinct phrases)	24.81	32.49	**48.41**	7.96
PBLM-SMT (distinct phrases and words)	25.17	**33.00**	48.91	**7.955**

The use of the n-best distinct words and sequences shows an improvement of

0.51 points BLEU in test set while, for the development it is of 0.36 points BLEU.

5.3 Greedy search optimization

With these experiments, we claim that the existing algorithms of parameter optimization are not suitable for the model we propose. In order to find the best weight for the SMT-PBLM, we applied the greedy search algorithm. In this algorithm, we use the configuration of the MIRA algorithm as a starting point. Then iteratively we vary the weights one by one by keeping the other weights fixed. At each iteration we calculate the BLEU score on the development set. The weights that give the best BLEU score are used for the decoding. After a set of tests, the highest obtained score BLEU on the development corpus is 25.86 and 32.93 points BLEU in the test corpus. Table 10 shows the best weight obtained by greedy search. Table 10 indicates that the greedy algorithm retains

Table 10. Weights of each parameter obtained by our greedy search method.

System	WP	PP	D	LM	TM			
PBLM (MIRA)	-0.22	0.22	0.12	0.11	0.10	0.05	0.13	0.001
PBLM (greedy search)	-0.29	0.22	0.15	0.13	0.08	0.05	0.11	0.08

the same weight obtained by MIRA for Phrase Penalty parameter therefore MIRA estimate properly this parameter. While we observe a drop for weights of the Phrase Penalty and translation model and an increase for weights of the Distortion and language model. The greedy search algorithm optimization enables us to improve the SMT-PBLM score of 0.13 points BLEU compared to MIRA and 0.24 points BLEU compared to MERT optimization.

6 Conclusion

In this paper, we presented a new method that uses a phrase-based language model for statistical machine translation task. This method has two major advantages. First, it can capture longer dependencies between words. Second, it uses the same phrases as those used in the translation table and more exactly those of the target part of each entry. In addition, our PBLM based method uses all target phrases of the translation table. The extensive experiments on French-to-English translation show that the phrase-based language model can improve the quality of the SMT task. Indeed, the SMT-PBLM translation quality is better than the baseline model in terms of TER which means that our model allows less work in post-edition operation. Thus, we showed that the state of the art optimization algorithms are not suitable for concatenated phrases used in our model. For this reason, we used a new optimization method based on a greedy search strategy to optimize the weights. This method allowed us to improve the

results, we started with a BLEU of 32.80 and we improve it till 32.93 which is still slightly under the baseline which achieves a score of 33.16 which actually is just 0,23% better than ours. In future work, we plan to propose an extension of the optimization algorithms to better treat concatenated sequences. We also look at combining WBLM with PBLM in order to better estimate the grammatical quality of the target side and improve accordingly the quality of the translation system.

References

1. Koehn, P., Och, F. J., Marcu, D.: Statistical phrase-based translation. IN: Proceedings of HLT-NAACL 2003, pp. 127–133, (2003)
2. Sarikaya, R., Deng, Y.: Joint Morphological-Lexical Language Modeling for Machine Translation. In: Proceedings of NAACL HLT 2007, pp. 145–148, (2007)
3. Khalilov, M.: Improving target language modeling techniques for statistical machine translation. IN: Proceedings of the Doctoral Consortium at the 8th EUROLAN Summer School, pp. 39–45, (2007)
4. Zitouni, I., Smaili, K., Haton, J.-P.: Statistical language modeling based on variable-length sequences. IN: Computer Speech and Language, pp. 27–41, (2003)
5. Deligne, S., Bimbot, F.: Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams. IN: Proceedings ICASSP, pp. 169–172, (1995)
6. Kuo, Hong-Kwang J., Reichl, W.: Phrase-based language models for speech recognition. IN: Sixth European Conference on Speech Communication and Technology (EUROSPEECH'99), (1999)
7. Lamel, L., Mori, R. D.: Speech Recognition Of European Languages. IN: Proceedings of the IEEE ASR Workshop, Snowbird, pp. 51–54, (1995)
8. Schwenk, H.: Continuous space language models. IN: Computer Speech and Language, pp. 492–518, (2007)
9. Le, H.S., Oparin, I., Allauzen, A., Gauvain, J.L., Yvon, F.: Structured output layer neural network language mode. IN: Proceedings of ICASSP, pp. 5524–5527, (2011)
10. Wang, R., Utiyama, M., Goto, I., Sumita, E., Zhao, H., Lu, B.-L.: Converting Continuous-Space Language Models into N-Gram Language Models for Statistical Machine Translation. IN: EMNLP, ACL, pp. 845–850, (2013)
11. Xu, J., Chen, G.: Phrase based language model for statistical machine translation. IN: CoRR, (2015)
12. Jiajun, Z., Shujie, L., Mu, L., Ming, Z., Chengqing, Z.: Beyond Word-based Language Model in Statistical Machine Translation. IN: CoRR, Vol. abs/1502.01446, (2015)
13. Lavecchia, C., Langlois, D., Smaili, K.: Discovering phrases in machine translation by simulated annealing. IN: INTERSPEECH, ISCA, pp. 2354–2357, (2008)
14. Och, F. J., Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. IN: Computational Linguistics, pp. 19–51, Vol. 29, (2003)
15. Michael, P., Marcello, F., Sebastian, S.: Overview of the IWSLT 2010 Evaluation Campaign. IN: Proceedings of the seventh International Workshop on Spoken Language Translation (IWSLT), Paris, France, pp. 3–27, (2010)
16. Stolcke, A.: SRILM: An Extensible Language Modeling Toolkit. IN: Proceedings of the 7th International Conference on Spoken Language Processing, pp. 901–904, (2002)

17. 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: The Association for Computer Linguistics, pp. 177–180, (2007)
18. Hopkins, M., May, J.: Tuning as Ranking. IN: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1352–1362, (2011)
19. Taskar, B., Chatalbashev, V., Koller, D., Guestrin, C.: Learning structured prediction models: a large margin approach. IN: Proceedings of the 22nd international conference on Machine learning , ICML '05, pp. 896–903, (2005)
20. Nasri, C., Latiri, C., Smaili, K.: Statistical Machine Translation Improvements based on Phrase Selection. IN: Recent Advances in Natural Language Processing (RANLP), Bulgaria, (2015)
21. Papineni, K., Roukos, S., Ward, T., Zhu, W.: A method for automatic evaluation of machine translation. IN: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Philadelphia, pp. 311–318, (2002)
22. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. IN: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, pp. 65–72 , (2005)
23. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. IN: Proceedings of Human Language Technology Conference, (2003)
24. Snover, M., Dorr, B., Schwartz, R., Micciulla, L., Makhoul, J.: A Study of Translation Edit Rate with Targeted Human Annotation. IN: Proceedings of AMTA, pp. 223–231, (2006)