Phrase-based Language Modelling for Statistical Machine Translation

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

[1]MIRACL Laboratory, ISIM Sfax, Pôle Technologique, TUNISIA
[2]SMART team LORIA Nancy, FRANCE

achraf.ramdhan@gmail.com jamoussi@gmail.com
abdelmajid.benhamadou@isimsf.rnu.tn smaili@loria.fr

Abstract

In this paper, we present our submitted MT system for the IWSLT2014 Evaluation Campaign. We participated in the English-French translation task. In this article we focus on one of the most important component of SMT: the language model. The idea is to use a phrase-based language model. For that, sequences from the source and the target language models are retrieved and used to calculate a phrase n-gram language model. These phrases are used to rewrite the parallel corpus which is then used to calculate a new translation model.

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, hypothesis are created and expanded until all words of the source sentence are covered. The expansion step produces a huge number of hypothesis which are constrained by the future cost estimation depending on 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 models to phrase-based-models [1], in order to better capture the context dependencies of the words in the translation process. In the other hand and despite the improvements made in language modelling [2], [3], the state-of-the-art SMT systems 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] and we hope that it can help in the improvement of the machine translation task. In SMT, the language model is calculated on the target language. Then, to get a phrase-based language model, the target language model should be rewritten in terms of sequence of words. To do this, we propose to extract the source phrases using triggers [4]. We then use the inter-lingual triggers to retrieve the corresponding target sequences [6], [7]. Both source and target phrases are used to rewrite the parallel corpus which is used to train the language and the translation models. In section 2, we give an overview of the source phrase extraction method. Then in section 3, we present the method which associates to each source sequence its equivalent sequences in the target language. A description of the used corpora and the results achieved are presented and discussed in section 5 and 6. We end with a conclusion which points out the strength of our method and gives some tracks about future work in our research group.

2. Source phrases extraction

We use the concept of triggers [4],[7],[8] to extract pertinent sequences from a corpus. A trigger is composed of a word and its best correlated triggered words estimated in terms of mutual information (MI) :

$$I(x,y) = P(x,y)\log\frac{P(x,y)}{P(x)P(y)} \qquad (1)$$

Where $P(x,y)$ is the joint probability and $P(x)$ and $P(y)$ are marginal probabilities. This allows to build a sequence of 2 words, to identify long phrases, an iterative process retrieves first, sequences of two words by grouping contiguous words which have a high MI then, in the second iteration, phrases of length 3 are identified, etc. To maintain a reasonable number of phrases, only the sequences which have a higher MI than the average MI of all sequences are kept for the forth coming steps. At the end of the process, we get a list of phrases which is used to rewrite the source corpus in terms of words and sequences. Examples of the retrieved phrases are given in table 1.

Since classical triggers allow to establish a triggering-triggered relationship between two events from the same language, Lavecchia et al. in [7] proposed to determine correlations between words coming from two different languages. These triggers called inter-lingual triggers. Each of them is

Phrases	**MI** $\times 10^{-5}$
parlement_européen	69.07
projet_européen	0.78
populaire_européen	0.22
politique_économique	0.17
commission_des_affaires_juridiques	0.039
commission_des_relations_économiquess	0.045
je_voudrais_vous_demander	0.032

Table 1: Examples of source phrases

composed of a triggering source event and its best correlated triggered target events.

3. Target phrases extraction

Once we have determined the list of the source sequences, we can then determine their corresponding sequences in the target side. For that, we used the method proposed by Lavecchia et al. [7] based on n-to-m inter-lingual trigger model. This method allows to associate 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, we choose one or more target sequences of length k $\pm\Delta$k without performing any word alignment. In our case, for the language pair English-French, we set Δk to 1 in a way that a sequence of two words will be associated with the target sequences of length one, two or three words. Thus, we select for each source phrase the first 30 most relevant target sequences that have the best MI. An example of the extracted phrases with their best corresponding target sequences is presented in table 2.

Source phrases	Target phrases	**MI** $\times 10^{-2}$
parlement_européen	**european parliament**	**2.3**
	the european parliament	2.01
	parliament	1.7
	europen	1.6
	the european	1.3
je_vous_remercie	**thank you**	**0.43**
	thank you very much	0.091
	thank you for your	0.067
	i thank you	0.063
	very much	0.054

Table 2: Example of inter-lingual phrases

4. How to process the parallel corpus?

The objective in this section is to show how to rewrite both source and target copora in terms of phrases. For each source phrase, we select all possible target phrases by using inter-lingual triggers. The target phrases are added, in a decreasing order of MI, to a dictionary of phrases. Then the target corpus is rewritten in terms of these phrases. In case of conflict, the algorithm will prefer the phrase with the highest MI. At this point, we get a bilingual training corpus written in terms of word and phrases. The achieved corpora are then used to train the language and translation models. Table 3 illustrates some examples of sentences of the obtained training corpora.

thank_you_very_much for_your_attention . je_vous_remercie de_votre_attention .
thank_you_very_much for_your_contributions and support . merci de_vos_contributions et de votre soutien .
i declare the_session_of_the_european_parliament adjourned . je déclare interrompue la_session_du_parlement_européen .
adjournment of_the_session interruption de_la_session
a_new deal for_the_new world une_nouvelle donne pour le_nouveau monde
it_is easier in certain_areas . c'_est plus facile dans certains_domaines .

Table 3: Examples of sentences of the training corpora

5. Resources Used in IWSLT 2014

Training the translation and language models is constrained to data supplied by the organizers. For this campaign, we only participated in the English-French translation task.
Among the parallel data provided, we use WIT^3 [9] and EUROPARL [10]. As usual, we clean the raw data before performing any model training. This includes the lowercasing conversion and removing of long sentences. After the preprocessing operation, we get a parallel corpus of 1 767 644 sentences. The English side has a total of 35 million words (117006 unique tokens). The French side has a total of 38 millions words (141150 unique tokens).
A 5-gram language model has been trained with SRILM toolkit [11]. The word alignment of the parallel corpora is generated using GIZA++ Toolkit [12] in both directions. Afterwards, the alignments are combined using the grow-diag-final-and heuristic to obtain symetric word alignment model [1]. For decoding we used Moses toolkit [13] and the standard MERT to tune the weights of our features on the 100-best translation assumptions of the development set. Eight default features are used:

- Bidirectional phrase translation probability $(p(e|f), p(f|e))$
- Bidirectional lexical probability $(lex(e|f), lex(f|e))$
- Phrase penalty
- Word penalty
- Distortion model
- 5-gram language model

6. Experiments

6.1. The retrieved phrases

In this task, we set the maximum size of a phrase to 8 words, this is due to the fact that in previous experiments [14] phrases with more than 8 words do not contribute effectively in the improvement of the machine translation quality. The method described in 2 is applied in a way that at each iteration, we retrieve phrases of different lengths depending on the size S of the source phrase. To control that, we keep only target phrases of T words with $T = S \pm \Delta S$. For instance, in the first iteration, only sequences of T words (with $T \in \{1, 2, 3\}$) are kept.

We extracted from the French part of the training corpus, a set of 23064 phrases. Then, for each source phrase of S words, we kept the 30 best potential translations of size T. These sequences are included in the translation table and used to rewrite the training corpus. In this way, the target corpus is composed of single words and phrases of at maximum of 8 words. Consequently, training a 5-gram language model will 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).

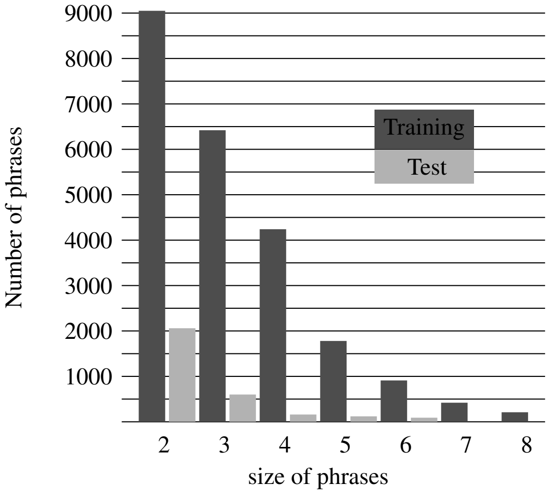

Figure 1: Histogram of the phrases number according to their size.

Figure 1 plots the histogram of the number of phrases contained in the training and the test corpora, according to their size. We can notice that the majority of phrases used are composed of two or three words which represents more than 60% of the extracted phrases. This histogram shows also that the number of phrases which occur in the test corpus is very low and does not exceed 12% of the whole extracted phrases.

6.2. Test data

The test data has to be written in the same way as the training corpus, for that two solutions are possible:

- Use the test corpora written in terms of words then we defragment our sequences belonging to the target part of the translation table.

- Rewrite the test data in terms of words and phrases. For this, the source sequences could be sorted according to their sizes or on their MI values. Then, for each sentence we explore the list of sequences in a decreasing manner. It worths mentioning that sorting sequences according to their size promotes the use of large size sequences while sorting sequence on their MI promotes the use of sequences short.

It should be noted that the system parameters were trained on the development corpus which combines the dev2010, tst2010 and tst2012. However we have chosen to report results on the tst2011, tst2013 and tst2014. Reported results are case-insensitive BLEU [15]. In addition, we performed tests on translation systems based on a training corpus written in terms of words and sequences:

- Sys1: uses a test corpus written in terms of words and sequences.

- Sys2: uses a test corpus written only in terms of words.

Table 4 illustrates the results obtained by different experiments on both development and test corpora.

System	Dev	tst11	tst13	tst14
baseline	28.91	36.84	-	-
Sys1	26.51	33.52	-	-
Sys2	28.27	35.48	30.91	26.97

Table 4: Results for the English → French MT task

On the development and the test corpus tst11, the use of a corpus written in terms of words (Sys2) is better than the one where the test data is rewritten in terms of phrases (Sys1). That's why, we decided to submit Sys2 as our primary SMT system. The small number of sequences used in our translation system and compared to the table of the baseline system is probably the reason which make our results worse than the baseline. Another explanation is related to the weak number of phrases contained in the test corpus, only 12% for tst13. Some translation examples are shown in Table 5.

7. Conclusions

In this paper, we evaluate our translation system on the data of IWSLT 2014 for English-French. Our contribution focuses on the use of a phrase-based language model and a translation model based on the phrases used in the language model. In order to train a phrase-based language model, we identified common source phrases by an iterative process.

Source	very often when i meet someone and they learn this about me there 's a certain kind of awkwardness .
Baseline	très souvent , lorsque je rencontre quelqu' un , et ils apprennent sur moi il y a une certaine gêne .
Sys2	très souvent quand je_rencontre quelqu' un , et ils_apprennent ce sur moi il y_a un_certain type_de gêne .
Reference	très souvent , quand je rencontre quelqu' un et qu' ils découvrent que je suis comme a , il y a un certain malaise .

Source	when we look at the population growth in terms of cars , it becomes even clearer .
Baseline	lorsque nous examinons la croissance de la population en termes de voitures , il devient encore plus clair .
Sys2	lorsque nous_examinons la_croissance_démographique en_termes de voitures , il devient encore plus clair .
Reference	quand nous regardons l' accroissement de la population en termes de voitures , ça devient même plus clair .

Table 5: Translation example from the tst11 set, comparing the baseline and the submitted system (Sys2) given a reference translation.

Then, we retrieved their potential translations by using interlingual triggers. These phrases are included in the translation table and used to rewrite the training corpus. The new corpus obtained is used to train the translation and language models. We evaluated the translation quality with the Bleu metric. The results showed that the state-of-the-art SMT system is better than our system. But, our results are encouraging and we plan to add some other features to the phrase based language model to improve the overall quality of our SMT system.

8. References

[1] P. Koehn, F. J. Och, and D. Marcu, “Statistical phrase-based translation”, Proceedings of HLT-NAACL 2003, 2003, pp. 127-133.

[2] R. Sarikaya, Y. Deng, “Joint Morphological-Lexical Language Modeling for Machine Translation"', In Proceedings of NAACL HLT 2007, Companion Volume, 2007, pp. 145-148.

[3] M. Khalilov, “Improving target language modeling techniques for statistical machine translation“, Proceedings of the Doctoral Consortium at the 8th EUROLAN Summer School, 2007, pp. 39-45.

[4] I. Zitouni, K. Smaïli, and J.-P. Haton, “Statistical language modeling based on variable-length sequences”, Computer Speech and Language, vol. 17, 2003, pp. 27-41.

[5] S. Deligne, F. Bimbot, ”Language modeling by variable length sequences: Theoretical formulation and evaluation of multigrams“, In Proceedings ICASSP, 1995, pp. 169-172.

[6] W. Kim and S. Khudanpur, ”Lexical triggers and latent semantic analysis for cross-lingual language model adaptation“, ACM Transactions on Asian Language Information Processing (TALIP), 2004, pp. 94-112.

[7] C. Lavecchia, D. Langlois, K. Smaïli, “Discovering phrases in machine translation by simulated annealing”, INTERSPEECH, ISCA, 2008, pp. 2354-2357.

[8] C. Tillmann, H. Ney, ”Word Triggers and the EM Algorithm“, In Proceedings of the Workshop Computational Natural Language Learning (CoNLL), 1997, pp. 117-124.

[9] M. Cettolo, C. Girardi and M. Federico, “WIT3: Web Inventory of Transcribed and Translated Talks”, In Proceedings of EAMT,2012, pp. 261-268.

[10] P. Koehn, “Europarl: A multilingual corpus for evaluation of machine translation”, In MT Summit, Thailand, 2005.

[11] A. Stolcke, “SRILM: An Extensible Language Modeling Toolkit,” in Proceedings of the 7th International Conference on Spoken Language Processing, 2002, pp. 901-904.

[12] F. J. Och, and H. Ney, “A Systematic Comparison of Various Statistical Alignment Models”, Computational Linguistics, vol. 29, no. 1,2003, pp. 19-51.

[13] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst, “Moses: Open source toolkit for statisti- cal machine translation,” in Annual Meeting of the Association for Computational Linguistics (ACL) demonstration session, Prague Republic, 2007, pp. 177-180.

[14] C. Nasri, K. Smaïli, and C. latiri Training Phrase-Based SMT without Explicit Word Alignment, 15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING), Nepal, 2014.

[15] K. Papineni, S. Roukos, T. Ward, and W. Zhu, “BLEU:a method for automatic evaluation of machine translation”, in Proceedings of ACL02, 2002, pp. 311-318.