The INESC-ID Machine Translation System for the IWSLT 2010

Wang Ling, Tiago Luís, João Graça, Luísa Coheur and Isabel Trancoso

L^2F Spoken Language Systems Lab
INESC-ID Lisboa
{wang.ling,tiago.luis,joao.graca,luisa.coheur,imt}@l2f.inesc-id.pt

Abstract

In this paper we describe the Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento (INESC-ID) system that participated in the IWSLT 2010 evaluation campaign. Our main goal for this evaluation was to employ several state-of-the-art methods applied to phrase-based machine translation in order to improve the translation quality. Aside from the IBM M4 alignment model, two constrained alignment models were tested, which produced better overall results. These results were further improved by using weighted alignment matrixes during phrase extraction, rather than the single best alignment. Finally, we tested several filters that ruled out phrase pairs based on puntuation. Our system was evaluated on the BTEC and DIALOG tasks, having achieved a better overall ranking in the DIALOG task.

1. Introduction

This paper describes the machine translation system employed by Instituto de Engenharia de Sistemas e Computadores Investigação e Desenvolvimento (INESC-ID) for the International Workshop on Spoken Language Translation (IWSLT) 2010. In this year's evaluation we participated in the BTEC and DIALOG tasks for both language directions. Our system is a phrase-based statistical machine translation system. Our work focuses on the building of better phrase-tables, which are created from the phrase extraction process. We improve this step by using better alignments models than the IBM M4 alignment model, using the posterior distribution over alignments instead of the single best alignment. Moreover, we filtered phrase pairs from being extracted based on punctuation and phrase length differences.

This paper is organized as follows: in Section 2 we will present the methods we used to improve the phrase extraction algorithm and Section 3 will describe the corpus used and data preparation. In Section 4, we will report the experimental results and, in Section 5, we will conclude the paper.

2. Phrase Extraction

The most common phrase extraction algorithm [1] uses word alignment information to constraint the possible phrases that can be extracted. Given a word alignment, all phrase pairs consistent with that word alignment are extracted from the parallel sentence (a phrase pair is consistent with a word alignment if all words in one language contained in the phrase pair are aligned only to words in the other language which are also contained in the phrase pair). That is to say, all phrase pairs that include at least one aligned point, but do not contradict an alignment by including an aligned word in one language without its translation in the other language, are extracted. So, on the one hand if there are too many incorrect alignment points forming a cluster, the correct phrases cannot be extracted without the spurious words, leading to missing words/phrases from the phrase table. In addition, unaligned words act as wild cards that can be aligned to every word in the neighborhood, thus increasing the size of the phrase table. Another undesirable effect of unaligned words is that they will only appear (in the phrase table) in the context of the surrounding words. Moreover, the spurious phrase pairs will change both the phrase probability and the lexical weight feature. The work by [2] conclude that the factor with most impact was the degradation of the translation probabilities due to noisy phrase pairs.

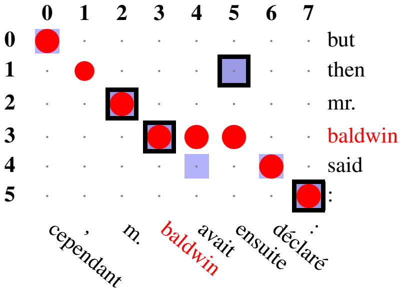
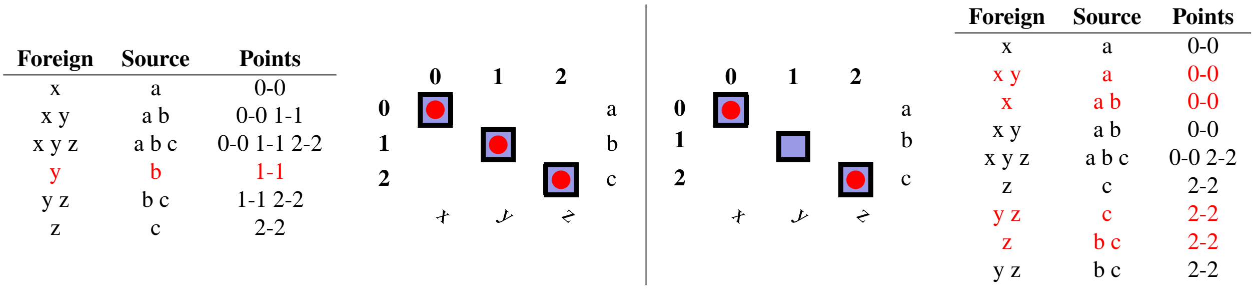


Figure 2: Example of a word alignment suffering from the garbage collector effect.

Figure 1 shows the phrase tables extracted from two word alignments for the same sentence pair. These alignments only differ in one point: *y*-*b*. However, the nonexistence of this point in the second word alignment, results in the removal of the phrase *y*-*b* in the second phrase table. Hence we would not be able to translate *y* as *b* except in the contexts shown in that table. Figure 2 shows an example of a word alignment where a rare source word is aligned to a group of target words, an effect known as garbage collector,



Foreign	Source	Points
x	a	0-0
x y	a b	0-0 1-1
x y z	a b c	0-0 1-1 2-2
y	b	1-1
y z	b c	1-1 2-2
z	c	2-2

Foreign	Source	Points
x	a	0-0
x y	a	0-0
x	a b	0-0
x y	a b	0-0
x y z	a b c	0-0 2-2
z	c	2-2
y z	c	2-2
z	b c	2-2
y z	b c	2-2

Figure 1: Machine Translation phrase extraction from word alignments example.

which leads that the word *baldwin* cannot be extracted without the incorrect surrounding context. This will make the pair *baldwin, baldwin* unavailable outside the given context. These led us to the work described in the following sections.

2.1. Constrained Alignments

Rather than using the IBM M4 alignment model, we use the posterior regularization framework with symmetric constraints [3], which produces better overall results. This constraint takes into account that if a certain unit a is aligned to unit b in the source to target alignment model, b should also be aligned to a in the target to source alignment model. We also made some tests with bijective constrains [3], which had better overall results compared to the IBM M4 model, but that were not as good as the symmetric constraints. In both of these alignments we use a threshold of 0.4 for accepting an alignment point and we train them using the conditions described in [3]. For each model, we initialize the translation tables with the results produced by IBM M1 with 5 iterations. The "HMM" model was run for 5 iterations, while the "BHMM" and "SHMM" were run for 2 iterations. Both "BHMM" and "SHMM" require two parameters: the constraint set slack and convergence stopping criteria. For both, we use the value 0.001 (we refer the reader from the original paper for an explanation of the meaning of these parameters).

2.2. Weighted Alignment Matrixes

We use information about the posterior distributions in the alignments, rather than the single best alignment to obtain better results. Given the posterior distribution for an alignment link, we use the soft union heuristic (the average of each link) to obtain a symmetrized alignment with link posteriors. Given these alignments links, we calculate the phrase translation probability and the link probability using the approach proposed for weighted alignment matrixes, as described in [4]. We only accept a phrase if its phrase posterior probability is above a particular threshold. For both the BTEC and DIALOG corpora we use a threshold of 0.1. We set the values based on the results of the original paper and leave the tuning of this particular threshold as future work, as lowering it does not always yields better results.

2.3. Phrase Pair Filtering

For each phrase pair that is extracted from a sentence pair, we apply an acceptor to decide whether that phrase pair is accepted or not. We build special acceptors that deal with punctuation. The idea is that punctuation normally translates one to one. Moreover, we observed that spurious punctuation tend to appear in the translation due to incorrect phrases. To this end we tried three different acceptors: *no-punc*, *no-terminal-punc* and *no-terminal-punc-unless-last*. The first rejects all phrases that contain punctuation; the second rejects all the phrases that contain a terminal punctuation; the last acceptor rejects phrases that has a punctuation anywhere except in the end of the phrase.

3. Corpus

This sections presents a description of the corpus used to train our translation system, as well as the pre- and posprocessing techniques used.

3.1. Data

In the BTEC task we used the training set, with approximately 20K sentence pairs, provided by the IWSLT 2010 to train our translation system. The provided development set was divided into development and test sets. We used the file devset1_CSTAR03 for the development set and the devset2_IWSLT04 as test set. The data we used in the DIALOG task was analogous to the BTEC task. We used the provided training set, with approximately 30K sentence pairs, and divided the development corpus into development and test sets. For the en-cn language pair, we used the devset1_CSTAR03 and devset2_IWSLT04, as in BTEC. For the cn-en language pair, we used the DIALOG.devset as development set and devset3_IWSLT05 as test set.

Method	Fr-En	Cn-En	En-Cn
Moses IBM M4	61.05	40.29	44.68
Moses HMM	59.87	38.54	45.32
HMM	59.93	38.49	45.47
BHMM	62.45	38.17	44.43
SHMM	62.46	41.42	44.99

Table 1: BLEU using the default Moses extraction algorithm (one, kohen, moses) for the different alignment models for three different scenarions.

3.2. Pre-processing

Considering the English and French corpora, we simply tokenize and lowercase the sentences with the scripts provided by Moses. For Chinese, we also replace the punctuation (“。”, “，”, “？”, “！”) with the respective latin punctuation, which is needed since the punctuation based filters are not implemented to work with Chinese punctuation. As for Chinese segmentation, we leave the segmentation of Chinese characters as the one given in the corpus, as most Chinese segmentation tools are trained using external resources. For the DIALOG task, since the input comes from an ASR, and thus, does not have punctuation, we used the SRILM toolkit to punctuate the text using a n-gram model built with the same training corpus used to create the translation model.

3.3. Post-processing

First of all, texts are de-tokenized using the detokenizer provided by Moses. Since the evaluation is case sensitive, we use a maximum entropy-based capitalization system [5] to recase the output of the MT system. Finally, the punctuation in the Chinese text is also converted back.

4. Results

For all experiments we use the Moses decoder (`http://www.statmt.org/moses/`), and before decoding the test set, we tune the weights of the phrase table using Minimum Error Rate Training (MERT). The results are evaluated with the BLEU metric.

Table 1 compares the different word alignments using the default phrase extraction. As a sanity check, we compare our implementation of the default phrase extraction with the one provided by moses (“HMM” vs “Moses HMM”), which yielded very close results. The small difference in values is due to an implementation detail difference in the alignments used, when calculating the lexical weighting of phrase pairs that were generated from multiple alignments. When this happens, we need to choose which alignment to use to compute the lexical weighting. In the case of Moses, it picks the most frequent alignment, while we select the alignment from the first phrase pair that is selected. Comparing the results for the different alignment models we see that the constrained-based alignments have better overall results than the regular HMM and IBM M4. This is not specially surprising since this was observed before, specially under small data conditions [3], as the the ones used here.

Method	Fr-En	Cn-En	En-Cn
HMM	59.93	38.49	45.47
BHMM	62.45	38.17	44.43
SHMM	62.46	41.42	44.99
HMM-post	61.74	39.48	46.11
BHMM-post	62.74	40.69	45.23
SHMM-post	63.07	42.15	45.00

Table 2: Different weighting of each phrase using the score from weighted alignment matrix.

Table 2 shows the differences between using the default phrase extraction and using information about the alignment posterior. We note that for every scenario using the posteriors improves the translation quality over using a single alignment. The default phrase extraction suffers from the garbage collector effect, which does not allow a phrase pair to exist outside the context where it was seen in the training corpus. The constrained alignment models partially solve this problem by correctly dealing with the garbage collector effect. This is further improved since a word pair can be extracted even when it is not consistent with the existing alignments.

Finally, we build special acceptors that deal with punctuation. The results are shown in Table 3. One of those acceptors is the *no-punct*, that rejects all phrases that contains punctuation. This is the more radical approach and produces worst results. The reason is that some types of punctuation, commas for instance, are used in different contexts from one language to another. For instance, the sentence *“s'il vous plaît, cherchez mon non encore une fois”*, does not have the comma in the translation *“Please look for my name again”*. By discarding these phrases, these commas could never be translated, hence decreasing performance. The second heuristic *no-terminal-punct* rejects all phrases that contain a terminal punctuation. This heuristic produces small improvement for French to English and English to Chinese, but not for Chinese to English. This happens since Chinese has a lot of particles such as “吗”, “呢”, “啊” and “吧”, specially in spoken text, which are characters that are not aligned with any words in English. Since we do not use null translations, these must be aligned with something in the phrase table. In the case of “吗” and “呢”, because they are question particles, they tend to appear before question marks, so they are generally aligned like “吗？” to “?”, which would work like a null translation for the particle, but because of our heuristic, we would not allow the resulting phrase pair to be extracted. Thus, when the character “吗” is translated, the chosen translation is picked from phrase pairs that are created from incorrect alignments for the particle.

We also tested other acceptors, but results were not as good as the ones presented before. One of those acceptors consists in adding an acceptor that discards phrase pairs

SHMM-post	Fr-En	Cn-En	En-Cn
base	63.07	42.15	45.00
no-punct	62.75	40.87	45.75
no-terminal-punct	63.41	41.44	46.68
no-terminal-punct-unless-last	62.62	41.24	46.86

Table 3: Experiments with different features and acceptors for phrase extraction.

whose difference in the source and target phrase length are higher than a given threshold. The intuition behind this acceptor is that translations are mostly word by word, specially between English and French. Hence, if there is a huge difference between phrases, this is probably due to the unaligned words, and will lead to a lot of spurious phrases. However, this acceptor does not produced the expected results, mainly due to the translation of some English words like *“could”*, that is translated to the French sentence *“pourriez-vous , s’ il vous plaît ,”*, which is obviously not a good translation, since the correct translation of the French sentence is *“could you, please,”*. However, due to different writing styles used in both languages, this translation occurs very often, and when phrase pairs with length bigger than 4 are cut from the translation table, we could not perform such a translation.

4.1. Automatic Evaluation Results

We observe from the preliminary automatic evaluation results that our main problem in this evaluation was the fact we only used BLEU as our tuning and evaluation metric, rather than a combination of metrics, which limited the quality of our results. In fact, in many instances our system yielded better BLEU scores in the evaluation, but the overall score was worse. It is also important notice that the tuning process we used does not use any stabilization methods described in [6], which states that there is a high variance between different runs of the tuning process for the same translation model. Thus, there is also the possibility that we obtained a bad set of weights for one of our translation models. In the DIALOG task, we performed specially well in the IWSLT09 testset, probably due to the fact that this set bears more similarity to the data set we used for tuning. In the BTEC task, we were ranked worse. We think that one of the reasons was, as stated above, that we obtained a worse set of weights during the tuning procedure.

5. Conclusions

We have described the INESC-ID system for the BTEC and DIALOG tasks of the IWSLT 2010 evaluation campaign. By combining the constrained-based alignments, the alignment posterior weighting and the punctuation based acceptors we produced significant improvements in the results relatively to the baseline system.

As future work, we intend to deeply analyze the results attained in this work, in order to improve them. We wish to try other types of constrained alignments, as well as combining the existing ones. The acceptor heuristics, which can potentially reduce the size of the phrase table and improve the translation quality, did not always produce better results for a number of reasons, some of which have already been found. Thus, we will work towards solving these problems. Finally, we would also like to apply the phrase extraction algorithm to hierarchical and syntax machine translation and possibly combining these results.

6. Acknowledgements

This work was partially supported by FCT (INESC-ID multiannual funding) through the PIDDAC Program funds, and also through projects CMU-PT/HuMach/0039/2008 and CMU-PT/0005/2007. The PhD thesis of Tiago Luís is supported by FCT grant SFRH/BD/62151/2009. The PhD thesis of Wang Ling is supported by FCT grant SFRH/BD/51157/2010.

7. References

[1] P. Koehn, F. J. Och, and D. Marcu, “Statistical Phrase-based Translation,” in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology (NAACL)*. Morristown, NJ, USA: Association for Computational Linguistics, 2003, pp. 48–54.

[2] A. Lopez and P. Resnik, “Word-based Alignment, Phrase-based Translation: What´s the Link?” in *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA): Visions for the Future of Machine Translation*, Boston, MA, 2006, pp. 90–99.

[3] J. Graça, K. Ganchev, and B. Taskar, “Learning Tractable Word Alignment Models with Complex Constraints,” *Comput. Linguist.*, vol. 36, pp. 481–504.

[4] Y. Liu, T. Xia, X. Xiao, and Q. Liu, “Weighted Alignment Matrices for Statistical Machine Translation,” in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*. Association for Computational Linguistics, 2009, pp. 1017–1026.

[5] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, “Recovering Capitalization and Punctuation Marks for Automatic Speech Recognition: Case Study for Portuguese Broadcast News,” *Speech Communication*, vol. 50, no. 10, pp. 847–862, 2008.

[6] G. Foster and R. Kuhn, “Stabilizing Minimum Error Rate Training,” in *Proceedings of the Fourth Workshop on Statistical Machine Translation*. Athens, Greece: Association for Computational Linguistics, March 2009, pp. 242–249.