Weighting scheme and pipeline design for ad hoc phrase tables produced with the sampling-based alignment method

Lishi Zhang Yves Lepage
Graduate School of Information, Production and Systems, Waseda University
{chocolishi@akane., yves.lepage@}waseda.jp

1 Introduction

Translation tables are the main bilingual data in phrase-based statistical machine translation (PB-SMT). They are compiled from a bilingual corpus. The standard process of producing phrase tables in the usual setting training-tuning-testing of building statistical machine translation systems can be said to be of the eager learning type. The data is compiled in advance and the result of compilation is applied to some unknown data. By opposition, building ad hoc phrase tables or on-demand phrase tables or generating phrase tables on-the-fly is of the lazy learning type, because it consists in extracting the ad hoc phrase tables from the bilingual corpus only after the sentences to be translated are known. Such ad hoc phrase tables are created for a specific purpose. Ideally, they should only contain necessary and specific phrase pairs to translate the test set. This way of doing has something in common with example-based machine translation (EBMT).

The weighted sampling-based alignment method is a method used to produce ad hoc phrase tables [3]. It relies on the use of the sampling-based alignment method, implemented in Anymalign[1] [2], which is an associative sub-sentential alignment method that has been proved to obtain better results than state-of-the-art methods (Giza++) on bilingual lexicon induction when evaluating against human-built dictionaries [1]. Anymalign samples a large number of sub-corpora randomly to compute co-occurrence distributions of bilingual pairs or sequences of words. By computing translation probabilities and lexical weightings from the co-occurrence distribution, a phrase table is obtained.

In this work, we first propose a scheme of weighting for the weighted sampling-based alignment method to estimate the relevance between sentences contained in the training corpus and the sentences to be translated by using the similarity between the training corpus and the sentences to translate. Secondly, we design a pipeline that fits the weighted sampling-based alignment method. To this end, we determine the optimal size of test sets, on the contrary to current practice in PB-SMT, we select a suitable tuning set by adapting it to the test set.

[1] https://anymalign.limsi.fr/

2 Sketch of the method and weighting scheme

Our method will use the standard sampling-based alignment method to produce phrase tables without major modification. In order to select necessary phrases only, we propose to bias the sampling in the sampling-based method when drawing sub-corpora, by giving a higher chance of being selected to those lines in the training corpus that share more with the test set. The weighted sampling-based alignment method will thus essentially rely on the definition of a function to assign a weight to each sentence pair in the training corpus. Each such weight reflects the potential of the line to produce phrase alignments that will be necessary to translate the test set.

2.1 Need for balanced phrase entries

As the phrase tables should be produced in order to translate a given test set, the phrase tables should ideally contain only entries in the source language that appear in the test set. However, in such a phrase table, there is a danger that the probabilities from the target phrases are wrongly estimated if unbalanced translation possibilities from target to source are not provided. This danger may originate from the fact that target phrases are not translated in a blanaced and general way if only source phrases from the source test are considered. Consequently, there is a necessity to adopt a two-step method that will first go from source to target, and then from target to source, so that the final phrase table contains more than just source phrases from the test set, but also phrases that are possible translations in the training corpus, of target phrases obtained from source phrases from the test set.

Our method will thus first extracts all (source language) N-grams from the test set. These N-grams are used in a first step to assign weights to lines in the training corpus. These weights will bias the sampling implemented in the sampling-based alignment method (Anymalign) so as to produce to produce a first ad hoc phrase table. Lines with a higher weight just have a higher chance of being used when drawing sub-corpora.

The first produced ad hoc phrase table will then be used to extract a new set of target language N-grams. This new set of N-grams is the set of all phrases contained on the target side of the first ad hoc phrase table that correspond to source phrases found in the test set. This new set of N-grams will be used to apply the same weighting scheme again, but on the target side, so as to give a weight to each line of the training corpus.

2.2 Weighting scheme

The weighting scheme we use to assign a weight to a line containing a sentence f_i in the training corpus is given in Formula 1. $\hat{L}$ is the N-gram representation of a sentence or a set of sentences, so that in particular, $\hat{T}$ is the N-gram representation of the test set (the set of all N-grams contained in the test set in the first step of the method, or the set of target phrases (N-grams) in the first ad hoc phrase table corresponding to source N-grams from the test set in the second step of the method). $|p|$ is the length of a phrase p. $|S|$ is the cardinality of set S (for $\hat{T} \cap \hat{f}_i$, $\hat{T}$ and $\hat{f}_i$).

$$\begin{aligned} \text{weight}(f_i) \;=\; & \frac{\sum_{p \in \hat{T} \cap \hat{f}_i} \text{Info}(p) \times |p|}{\sum_{p \in \hat{T}} \text{Info}(p) \times |p|} \times \\ & \frac{\left|\hat{T} \cap \hat{f}_i\right|}{\left|\hat{T}\right|} \times \frac{\left|\hat{T} \cap \hat{f}_i\right|}{\left|\hat{f}_i\right|} \end{aligned} \tag{1}$$

This weighting scheme takes three points into consideration: phrase coverage, phrase frequency and length of sentence. Table 1 gives a sketch of the desired weights that should be assigned to lines in the training corpus depending on the amount of N-grams they share with a set of N-grams, for the two following points: importance of phrases sentence length and phrase coverage.

2.2.1 Sentence length

We choose to give more weight to shorter sentences because longer sentences may contain more noise. This is done by using the factor $1/\left|\hat{f}_i\right|$. Now a short sentence is useful only if it is made out of important N-grams.

amount of phrases shared	sentence length	desired weight
none	short	low
none	long	low
only one	short	medium
only one	long	medium
small	short	medium or high
small	long	low or medium
large	short	high
large	long	medium or high

Table 1: Desired values of weight for a line according to the amount of phrases in common with a representation of the test set

2.2.2 Importance of N-grams

The frequency of a phrase or an N-gram is reflected in Formula 1 by using its self-information. A line which contains low frequency phrases should be given a higher weight.

$$\text{Info}(p) = -\log \text{freq}(p) \tag{2}$$

For translation purposes, we want to favor longer phrases. Informativeness being defined as the product of self-information and length, is more suited to favor in a blanced way phrases that are infrequent and long at the same time.

$$\text{Informativeness}(p) = \text{Info}(p) \times |p| \tag{3}$$

2.2.3 Phrase coverage

Sentences selected by higher weights should do not contain to much noise (i.e., they are not so long), and should contain important N-grams. I.e., such sentences should somehow focus on important shared N-grams. On the contrary, the more a sentence contains N-grams from the representation of the test set, the better this sentence. To that end, the formula consistently takes the ratio of $\left|\hat{T} \cap \hat{f}_i\right|$ over $\left|\hat{T}\right|$ for N-grams, while giving weight to the ratio of shared N-grams $\left|\hat{T} \cap \hat{f}_i\right|$ relatively to $\left|\hat{T}\right|$ as well as relatively to $\left|\hat{f}_i\right|$.

2.3 Restraining to lines covering the test set

We apply the weighting scheme defined above to the entire training corpus, in each of the two steps form source to target and target to source. So as to further focus the phrase table production step on the test set (i.e., its N-gram representation), we determine the set of lines with higher weights that cover the entire test set. To do so, we sort the lines of the training corpus by order of decreasing weights. Only the lines

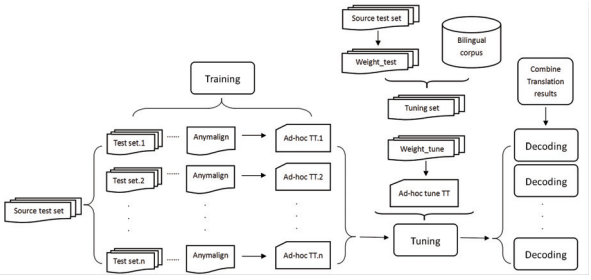

Figure 1: Pipeline for building a PB-SMT system with ad hoc phrase tables

which cover the N-gram representation of the test set are used in the proper production of phrase tables. This results in a faster processing time as well.

3 Pipeline for the weighted sampling-based alignment method

The pipeline for building a PB-SMT system equipped with ad hoc phrase table differs in two places from a usual PB-SMT pipeline where the test set is not known in advance. The training, i.e., the production of the ad hoc phrase table differs in that we use the test set. In our method, the weighting scheme and the production of phrase tables is performed in two directions: from source to target and from target to source. This results in an ad hoc phrase table that can be used in further tuning and decoding.

3.1 Splitting large test sets

In the case of test set of a large test set, e.g., several thousand lines against a training corpus of several hundreds of thousand lines, almost all lines from the training corpus are needed to cover the test set. Using ad hoc phrase tables is not justified in that case. The production of ad hoc phrase tables is only efficient for very small test sets. However, for experiment purposes, because we want to compare with standard settings in which the test set is usually large, we propose to cut a large test set into small parts and run our pipeline on each of the parts. The evaluation in translation accuracy will be performed on the all translation results.

3.2 Biased tuning

In standard settings, tuning or development is required and is a time-consuming process. It makes sense because the translation tables extracted from the training corpus are large enough and the test set is unknown. In our case, as the test set is known in advance and the ad hoc phrase tables are specific and small, the purpose of tuning becomes questionable. Still, it makes sense to extract a tuning set which is biased towards the test set by extracting 500 lines from the training corpus with the highest weights. Because this tuning set is biased towards the test set, we call this biased tuning.

The pipeline to evaluate the efficiency of the ad hoc phrase tables produced by our proposal on a large test set is sketched in Figure 1.

4 Experiments

4.1 Experiments settings and data used

We use the newest version of sampling-based alignment method, Anymalign to produce ad hoc phrase table which is within a PB-SMT system built by using the Moses toolkit[2].

We perform experiments on the French–English language pair, using data from Europarl parallel corpus[3], v3. The training corpus is made of 347,614 sentences. The tuning set contains 500 sentences for experiments with GIZA++ and Anymalign in its basic usage. We use biased tuning in the other cases (see Section 3.2).

We prepare 3 test sets of 300 lines each. The first one is extracted from the same version of the Europarl corpus and is, of course, distinct from the training set. Two other sets are extracted from Europarl corpus v7. One of these sets is similar to the first test set. It is obtained by taking the 300 lines from Europarl v7 with the highest weights, relatively to the first test set. The other one is unrelated. It is obtained by sampling from Europarl corpus v7. We checked that no line in these two sets is found in the first test set, nor in the training set.

According to the design of the pipeline in Section 3, the test sets are divided into 10 and 15 parts, i.e., each part contains 30 lines or 20 lines.

4.2 Evaluation criteria

We evaluate our proposed technique using relevant measures:

- translation accuracy as measured by BLEU;
- sampling time, i.e., the time spent in biased sampling for sampling based alignment method (in minutes). This time is included in the total training time;
- tuning time (in minutes).

[2] http://www.statmt.org/moses/
[3] http://www.statmt.org/europarl/

Production of phrase tables			Type of test set					
	Ad hoc phrase tables		Original			Similar	Unrelated	
	# of parts	biaised	Processing time (h and min)			BLEU score		
	× # of lines	tuning	Training	Sampling	Tuning			
MGIZA++			5h13		2h31	34.25	34.28	34.25
Anymalign, basic usage			1h06	1h00	2h08	25.80	26.01	25.94
Anymalign	1 × 300	yes	19	15	1h44	30.43	30.52	30.41
”	10 × 30	no	46	15	1h56	32.31	32.23	31.99
”	”	yes	45	15	54	32.40	32.44	32.10
”	15 × 20	no	1h15	15	2h02	32.42	32.33	32.36
”	”	yes	1h14	15	49	32.61	32.79	32.58

Table 2: Sampling, training and tuning times and BLEU scores in different configurations. Except for MGIZA++, where this does not apply, the training time includes the sampling time. The confidence intervals of BLEU scores ranged from ±1.48 to ±1.55 in all cases

For comparison, we also build a standard baseline PB-SMT using MGIZA++, MOSES, MERT and SRILM.

4.3 Experiment results

The results of our experiments are reported in Table 2. In comparison to the basic use of the sampling-based alignment method, i.e., the simple use of Anymalign to produce phrase tables directly, the best configuration of our systems, the one with ad hoc phrase tables produced on 15 parts of 20 lines, and with biased tuning, shows an improvement of almost 7 BLEU points. Now, it is remarkable that such an increase in BLEU is also observed on the similar and unrelated test sets. However our best results still miss the state-of-the-art results obtained using MGIZA++ by a little bit more than half a confidence interval ($34.25 - 32.61 = 1.64 \simeq 1.55$).

As for time, the use of ad hoc phrase tables allows to divide the training time by more than 4 and the tuning time by more than 3 in comparison with state-of-the-art techniques. Biased tuning is twice as fast as standard tuning. As a whole, the times required to build a PB-SMT system with ad hoc phrase tables and biased tuning is twice as fast as the state-or-the art procedure ($1h14+49 = 3h03 \times 2 \leq 5h13+2h31 = 7h44$). Splitting the test set into parts results in a large overhead because the whole training corpus has to be read several times in order to calculate weights for each of the parts. This overhead is visible in Table 2 by comparing 1 set of 300 lines, i.e., without splitting, to the results with 10 or 15 parts.

5 Conclusion

In this paper, we proposed a weighting scheme for the production of ad hoc phrase tables, for use in conjunction with the sampling-based alignment method, that takes a balanced account of the test set on one side and the similarity between the training set and the test set on the other side. We designed an adequate SMT pipeline to evaluate the efficiency of our proposed production of ad hoc phrase tables.

We have shown large improvements in translation accuracy as measured by BLEU, in comparison with the basic usage of the sampling-based alignment method, while reducing the time to produce phrase tables and the time for tuning, but still lying behind a standard baseline PB-SMT system. Our proposal leads to much shorter times of development compared with the usual state-of-the-art PB-SMT pipeline.

Acknowledgments

This paper is part of the outcome of research performed under a Waseda University Grant for Special Research Project (Project number: 2015A-063).

References

[1] Adrien Lardilleux, Yves Lepage, and François Yvon. The contribution of low frequencies to multilingual sub-sentential alignment: a differential associative approach. *International Journal of Advanced Intelligence*, 3(2):189–217, July 2011.

[2] Adrien Lardilleux, François Yvon, and Yves Lepage. Generalizing sampling-based multilingual alignment. *Machine translation*, 27(1):1–23, 2013.

[3] Junjun Lee and Yves Lepage. Fast production of ad hoc translation tables using the sampling-based method. In *Proceedings of the 18th Annual Meeting of the Association for Natural Language Processing (ANLP 2012)*, pages 809–812, 2012.