Improving N-gram Distribution for Sampling-based Alignment by Extraction of Longer N-grams

Suchen Zhang, Juan Luo and Yves Lepage
Graduate School of IPS, Waseda University
{jimzhang@moegi.,juan.luo@suou., yves.lepage@}waseda.jp

Abstract

Translation tables are an essential component of statistical machine translation system. The sampling-based alignment method is a way of building translation tables. It has advantages in speed and accuracy, but lays slightly behind in translation evaluations compared to the standard alignment technique.

Previous research has proved that the sampling-based alignment method does not generate enough long N-gram alignments. This harms in translation evaluation. This paper investigates translation table obtained by the sampling-based alignment method in detail and introduces an improved distribution to allot time for different N-gram lengths. The new model helps in outputting more numerous longer N-grams. We report significant improvements in BLEU scores in 110 Europarl corpus language pairs.

1 Introduction

To build a statistical machine translation system using parallel language data, it typically requires 2 processes: training and tuning. The training process aligns phrases with different numerical features. The tuning process determines the parameters to weight the different features during decoding. There are various models and methods proposed to implement phrase alignment. The state-of-the-art alignment tool is a combination of MGIZA++ [1] with the Moses heuristic grow-diag-final [6].

Here, we introduce the sampling-based alignment approach. This approach [8] is different from the combination of MGIZA++/Moses. It is available as a free open-source tool[1] called Anymalign. It can be interrupted at any time, and is simpler than the models implemented in MGIZA++/Moses, which meets the trend of associative alignment introduced by recent works [10]. In Anymalign, only those sequences of words that appear exactly on the same lines of the corpus are considered to be a "perfect alignment". The core idea of sampling-based alignment method is to randomly select sentences from the original corpus as sub-corpora so as to generate numerous "perfect alignments". This process is repeated following a particular sampling so as to ensure the coverage of the entire corpus by the sampled sub-corpora [8].

In this paper, we firstly investigates the N-gram distribution of the MGIZA++ translation table after applying pruning [3]. We then use sampling-based method imitate this N-gram distribution. To this end, we introduce a model to allot time distribution in the sampling-based alignment method. Our goal is to output longer N-gram alignments. This paper improves over the work reported in [9]. The paper is organized as follows: Section 2 reviews related work. Section 3 describes the experiment settings. Section 4 explains our model and analyses the results.

2 Related Works

2.1 Enforcing Sampling-based Alignment N-grams

As it is shown by experiments reported in [7] and [4], the sampling-based alignment method excels in aligning unigrams. Practically, it is possible to enforce the processing of longer N-grams as unigrams by replacing space between some words with underscore symbols. Such a technique to process N-grams as if they were unigrams, has been adopted in [2]. There, *word packing* is implemented to obtain $n-m$ alignments based on co-occurrence frequencies.

Because the sampling-based alignment method can be interrupted at any time, it is possible to allot different time ranges to each of the n $\times$ m cells using word packing. In [9], a normal distribution model along the main diagonal has been proposed to generate longer N-gram alignments with the same length. It improves the quality in machine translation over a baseline using the standard sampling-based alignment method.

3 Sampling-based Alignment Experiments Setting

3.1 Language Data

We use the standard Europarl Corpus [5] as the data set for our experiments. Tuning set and test set are indepen-

[1] http://users.info.unicaen.fr/~alardill/anymalign/

dent and randomly selected. Total 110 language pairs are used and all the sentences are translation across all other languages in the same line, so that the experiment over the 110 language pairs can be considered similar in contexts. Details are shown below:

- Training: 347,614 sentences
- Tuning: 500 sentences
- Testing: 1,000 sentences

3.2 Experiment Settings

We first investigate the performance of the sampling-based alignment approach implemented by Anymalign in statistical machine translation tasks. A preliminary experiment compares Anymalign with the standard MGIZA++/Moses. Although Anymalign and MGIZA++ are both capable of parallel processing, for fair comparison, we run them as single processes individually in the relative experiments. We use Moses [6], MERT (Minimum Error Rate Training) to tune the parameters of translation tables [10], and the SRILM toolkit [12] to build a target language model. These two baseline systems were both evaluated using BLEU[11].

4 Sampling-based Alignment Experiments Results and Analysis

4.1 Cell-by-Cell Comparison of Anymalign and MGIZA++/Moses Translation Tables

To investigate the difference between Anymalign and MGIZA++/Moses translation tables, we force Anymalign to output the same N-gram distribution[2] as MGIZA++/Moses in all 110 Europarl language pairs. We compare the two tables in each language pair for each N-gram cell. For the phrase pairs in each cell, we further investigate the intersection: how many entries have the exact same source phrase $\widehat{s}$ and target phrase $\widehat{t}$; for the translation probabilities and lexical weights, we calculate the difference between translation probabilities to see the distribution of the variance between Anymalign and MGIZA++.

According to [4], among all the Europarl language pairs, uni-gram to uni-gram alignments were used more than 65% during decoding process, while less than 10% were used for the n-gram length longer than 3. To save time and try not to sacrifice BLEU score significantly, we decide to generate the same number of MGIZA++/Moses limited to the n-gram length up to 5. This is because that for the n-gram maximum length which is greater than 5, the alignments only count for less than 10% in numbers of entries but cost more than 30 hours of time for Anymalign, while only less than 5% are used during decoding.

[2]The -A option of Anymalign allows to control the number of entries output.

language pair	overlap ratio	alignments	overlap
pt-es	25%	1,561,573	401,342
language pair	overlap ratio	alignments	overlap
nl-fi	13%	500,839	66,139

Table 1: The language pairs with the most/least alignment numbers and the most/least overlap

(a) Overlap in N-gram & ratio to MGIZA++ TT % (pt-es)

		Target				
		unigram	bigram	3-gram	4-gram	5-gram
Source	unigram	19173 27.4%	4882 13.1%	404 4.7%	37 2.2%	5 0.0%
	bigram	10630 16.8%	107420 37.9%	21420 19.4%	1145 6.4%	67 2.1%
	trigram	1103 6.6%	18199 17.7%	110771 37.6%	21206 19.9%	1557 8.4%
	4-gram	84 2.7%	1125 5.9%	10180 14.8%	41769 26.4%	11559 18.4%
	5-gram	4 0.0%	83 2.6%	651 5.6%	3439 11.5%	14429 21.9%

(b) Overlap in N-gram & ratio to MGIZA++ TT % (nl-fi)

		Target				
		unigram	bigram	trigram	4-gram	5-gram
Source	unigram	11673 17.6%	1184 7.4%	77 3.8%	6 20%	0 -
	bigram	10543 15.2%	17821 18.8%	1768 9.7%	112 4.3%	3 0.0%
	trigram	4678 15.3%	6693 11.7%	4314 11.7%	489 6.2%	48 3.7%
	4-gram	991 10.9%	1630 7.8%	1316 6.8%	925 7.5%	116 4.6%
	5-gram	84 4.5%	433 6.1%	453 5.5%	401 5.8%	381 8.3%

We analysed the sum of overlap for 110 language pairs and found that the language pairs with the most alignment numbers and the most overlap in total is pt-es, and the least one nl-fi (see Table 1). It is obvious that the language pair with the most/least n-gram alignments attained the most/least overlap between Anymalign and MGIZA++/Moses and vice versa. Thus the overlap ratio for all 110 languages ranges from 13% to 25%.

We inspect the translation table N-grams cell by cell. We consider the two previous language pairs: pt-es and nl-fi for comparison, then output the overlap between Anymalign and MGIZA++/Moses. The ratio figures under the overlaps are divided by the original number in the translation table. The figures are shown in Table 1(a) and Table 1(b). Among all the cells in pt-es, more than 37% of the alignments overlap for bigram-bigram and trigram-trigram and more than 27% for unigram-unigram, while none of the cells in nl-fi has the overlap ratio more than 20%. As a consequence, pt-es got the highest BLEU score, while nl-fi is one of the lowest among 110 language BLEU matrix. This is an evidence to prove that more longer n-gram alignments (mainly bigrams, trigrams, 4-grams) will result in better translation evaluation scores (see Table 4).

The entries common to Anymalign and MGIZA++/Moses may have different feature scores (translation probabilities and lexical weights). We compute the difference and plot their distribution in Figure 1 and Figure 2. The distribution in y axis is the percentage of the counts of each difference with two digits. Average and variance deviation values also added on the graphs. For uni-gram to uni-gram entries, Anymalign's translation probabilities is very close to the ones generated by MGIZA++/Moses, while the variance is bigger for nl-it. Among 110 language pairs in Europarl, the average values are always less than 0, which means the feature score from Anymalign translation table is always sightly over the one from MGIZA++/Moses. The slight difference exists because sampling-based alignment method estimates translation probability by the $C(\widehat{s},\widehat{t})/C(\widehat{t})$ in "perfect alignment" set, while MGIZA++/Moses find the counts in the whole alignment sets. The average and variance deviation is bigger for the language pair nl-fi and it is always like that when Finnish act as target language. Linguistically we know in advance that Finnish has more hapaxes (the word on appear once in the corpus), so that when Finnish serves as a target language, Anymalign will extract more $\widehat{t}$ alignments. For this the reason, in Figure 2, the right side of x axis after zero shows lower features for Anymalign than for MGIZA/Moses.

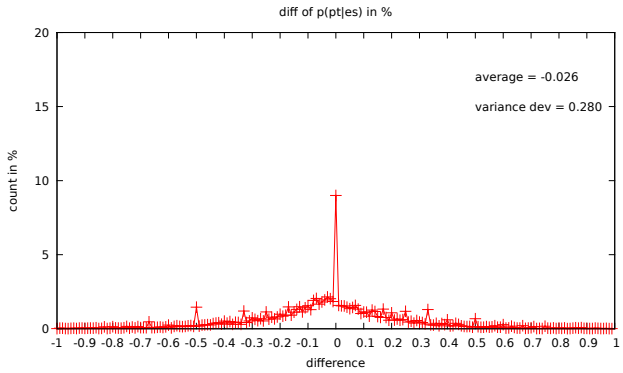

Figure 1: difference of translation probability $p(pt \mid es)$ unigram to unigram

In this experiment where we force Anymalign to output the same number of N-grams as MGIZA++/Moses, twice the time is needed for Anymalign. But the BLEU scores lays behind those of MGIZA++/Moses. In the next section, we introduce a time distribution model to focus more on the N-grams that are the most useful for better translation performance.

4.2 Multivariate Normal Distribution Model Experiments

In the previous subsection, we reported on mimicking MGIZA++/Moses translation table N-grams distributions. In this section, we apply a multivariate normal distribution model to simulate MGIZA++ N-gram distribution by allotting a given time range for each N-gram cell. More time will be allotted to the cells which contribute more to decoding, such as unigram to unigram and bi-gram to bi-gram.

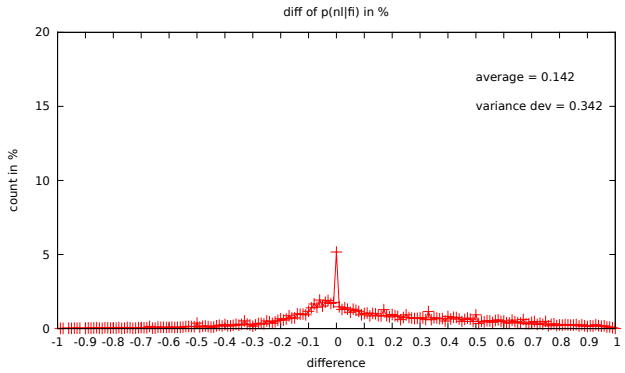

Figure 2: difference of translation probability $p(nl \mid fi)$ unigram to unigram

$$f(n,m) = \frac{1}{2\pi\sigma_n\sigma_m\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\right) \times$$

$$\exp\left(\frac{(n-\mu_n)^2}{\sigma_n^2} + \frac{(m-\mu_m)^2}{\sigma_m^2} - \frac{2\rho(n-\mu_n)(m-\mu_m)}{\sigma_n\sigma_m}\right)$$

since the translation table matrix only has two dimensions: source n-grams and target m-grams, thus bi-variate normal distribution $f(n,m)$ will be applied. In this equation, n and m refers to the source and target N-gram index of the cells. For bivariate normal distribution equation, means and variances are

$$\mu = \begin{pmatrix}\mu_n \\ \mu_m\end{pmatrix}, \quad \Sigma = \begin{pmatrix}\sigma_n^2 & \rho\sigma_n\sigma_m \\ \rho\sigma_n\sigma_m & \sigma_m^2\end{pmatrix}.$$

in this equation, $\sigma_n > 0$ and $\sigma_m > 0$. The parameters will be determined by simulating the distribution of the translation table generated from MGIZA++. For instance, in the case of fr-en, $\mu_f = 2.8$, $\mu_e = 2.6$, $\sigma_f = 1.4$, $\sigma_e = 1.3$, $\rho = 0.1$ To assess the fitness of the distribution in actual Moses translation tables, we use the classical notion of error rate and compute the following value:

$$\Delta_{ER} = \sqrt{\sum_{n,m}^{N}(A_{(n,m)} - M_{(n,m)})^2}$$

where A and M refer to Anymalign and MGIZA++/Moses, and n and m refer to the source and target N-gram length. N represents the maximum length for n-grams. The smaller the error rate, the better the model simulates MGIZA++/Moses N-gram distribution. The parameters for fr-en is generated by selecting the minimum of error rate of 0.041.

We applied the multivariate normal distribution model and run Anymalign for 7 hours. We also manually set the

Max N-gram length	N≤ 2	N≤ 3	N≤ 4	N≤ 5	N≤ 6
standard model	25.30	**25.48**	24.94	24.35	24.02
multivariate model	24.09	26.81	**27.02**	26.58	26.44

Table 2: BLEU score of Anymalign using multivariate normal distribution model for different length of N-gram

	da	de	el	en	es	fi	fr	it	nl	pt	sv
da	-	20.14	21.97	31.06	26.76	14.02	24.31	21.95	23.25	24.40	29.12
de	24.08	-	21.26	27.42	25.54	12.25	21.25	21.97	26.26	24.27	21.08
el	23.53	18.94	-	30.86	31.72	13.07	27.12	25.38	22.51	29.34	22.76
en	29.05	19.29	27.17	-	34.61	15.10	30.50	26.78	24.35	31.73	29.37
es	24.63	19.03	27.81	34.59	-	12.91	35.46	30.96	23.54	36.17	23.73
fi	20.88	15.24	18.41	22.86	22.40	-	19.78	17.46	17.83	20.01	18.83
fr	21.61	16.73	24.19	29.40	33.13	10.31	-	30.26	21.10	31.22	19.30
it	22.89	17.33	25.45	30.42	34.94	11.21	33.69	-	23.10	32.84	20.11
nl	24.82	21.34	20.98	27.85	24.19	11.36	23.02	20.27	-	23.52	21.29
pt	24.44	19.29	26.87	32.74	38.80	12.02	34.40	30.79	22.60	-	23.01
sv	33.18	20.24	23.18	33.09	29.49	15.04	25.93	23.05	23.70	26.81	-

Table 3: BLEU score matrix of MGIZA++/Moses baseline

maximum N-gram length for aligning from 2 to 7 so as to see when Anymalign performs best. The BLEU score of Anymalign model both reached the peak when the maximum length of N-gram is 4, and for that the multivariate normal distribution model increased 6% than the previous model. Table 2 shows that the multivariate normal distribution provide better time allotting model than former work which output more longer N-grams with better translation quality.

Although evaluation of BLEU of Anymalign were still slightly behind MGIZA++/Moses, we can see their performances become real close within 5% or even outperformed especially for Spanish, French and Italian (see Table 3 and Table 5 in bold figures).

5 Conclusion

In this paper, we presented a comparison between sampling-based alignment tool and the state-of-the-art alignment tool and found more hints to improve sampling-based alignment. By improving the standard normal distribution on the main diagonal, the improved multivariate normal distribution model allotted more time to the cells that contribute more to translation quality. Our proposed method outputs significantly better results than the unmodified sampling-based alignment method and reaches the level of MGIZA++/Moses in some of the language pairs like fr-it. Since the sampling-based alignment method is much faster and generates less configuration, it is worth considering for its simplicity. In the future work, we would like to modify the way of computing feature scores so as to further increase translation quality.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number C 23500187.

	da	de	el	en	es	fi	fr	it	nl	pt	sv
da	-	16.13	15.76	25.44	19.98	10.27	20.15	17.21	18.80	18.58	26.01
de	17.88	-	15.34	20.03	18.48	8.44	18.38	15.93	20.18	17.05	15.07
el	16.66	14.20	-	25.21	25.18	8.83	24.26	21.87	16.87	23.58	16.01
en	21.90	15.52	20.65	-	25.84	10.10	22.70	22.06	20.06	23.45	25.84
es	16.95	13.86	20.62	26.38	-	8.64	31.28	27.87	17.42	31.93	16.82
fi	14.61	11.20	12.53	17.24	14.84	-	12.73	13.56	12.77	13.75	13.46
fr	16.79	14.02	19.09	24.82	27.39	7.94	-	27.04	17.23	27.83	15.86
it	16.18	13.73	19.46	24.49	30.43	8.50	30.62	-	16.73	28.48	15.38
nl	17.58	17.42	14.60	22.07	18.74	7.97	19.01	16.66	-	17.10	15.55
pt	16.51	13.80	20.33	24.94	32.84	8.68	31.78	27.87	16.87	-	15.76
sv	27.23	15.45	16.99	26.68	20.52	9.97	20.84	17.28	18.02	18.79	-

Table 4: BLEU score matrix of Anymalign with same # to MGIZA++

	da	de	el	en	es	fi	**fr**	**it**	nl	pt	sv
da	-	17.80	18.74	28.17	22.79	11.18	**23.47**	20.17	21.12	21.85	27.27
de	20.54	-	17.54	22.52	20.43	9.46	**20.97**	18.35	22.81	19.60	17.02
el	19.73	16.48	-	28.43	27.71	10.34	**27.66**	**24.93**	19.45	26.39	18.40
en	24.69	17.09	23.83	-	28.90	11.52	24.58	25.17	22.69	26.74	24.14
es	19.51	15.91	23.39	29.18	-	9.79	**34.47**	**30.21**	19.69	**34.36**	18.76
fi	18.04	12.90	14.70	19.76	17.64	-	13.71	15.92	14.24	16.18	15.66
fr	18.59	15.27	20.80	27.02	30.90	8.50	-	**28.67**	18.76	**30.10**	17.62
it	18.56	15.24	22.55	26.90	32.91	10.01	**33.74**	-	19.26	**31.25**	17.81
nl	20.04	19.47	17.24	24.54	21.39	9.16	**22.27**	**19.01**	-	20.05	17.33
pt	19.20	15.98	23.27	27.64	34.72	9.73	**34.16**	**29.76**	19.47	-	18.04
sv	29.74	17.74	19.96	30.03	24.16	11.56	24.23	20.81	20.99	22.30	-

Table 5: BLEU score matrix of Anymalign using multivariate normal distribution model

謝辞

本研究はJSPS科研費 基盤C 23500187の助成を受けたものです。

References

[1] Qin Gao and Stephan Vogel. Parallel implementations of word alignment tool. In Association for Computational Linguistics, editor, *Software Engineering, Testing, and Quality Assurance for Natural Language Processing*, pages 49–57, Columbus, Ohio, 2008.

[2] A. Carlos Henríquez Q., R. Marta Costa-jussà, Vidas Daudaravicius, E. Rafael Banchs, and B. José Mariño. Using collocation segmentation to augment the phrase table. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and Metrics MATR (WMT 2010)*, pages 98–102, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[3] John Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. Improving translation quality by discarding most of the phrasetable. *Association for Computational Linguistics*, pages 967–975, jun 2007.

[4] Yves Lepage Juan Luo. Comparison of association and estimation approaches to alignment in word-to-word translation. In *Proceedings of the 10th Symposium on Natural Language Processing (SNLP-2013)*, pages 150–159, Phuket, Thailand, 2013.

[5] Philipp Koehn. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of the tenth Machine Translation Summit (MT Summit X)*, pages 79–86, Phuket, Thailand, 2005.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL 2007)*, pages 177–180, Prague, Czech Republic, 2007.

[7] Adrien Lardilleux and Yves Lepage. Hapax Legomena : their Contribution in Number and Efficiency to Word Alignment. *Lecture notes in computer science*, 5603:440–450, 2009.

[8] Adrien Lardilleux and Yves Lepage. Sampling-based multilingual alignment. In *Proceeding of the International Conference on Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria, 2009.

[9] Juan Luo, Adrien Lardilleux, Yves Lepage, et al. Improving sampling-based alignment by investigating the distribution of n-grams in phrase translation tables. In *Proceedings of the 25th Pacific Asia Conference on Language Information and Computing (PACLIC 25)*, pages 150–159, Singapore City, Singapore, 2011.

[10] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51, 2003.

[11] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318, Philadelphia, 2002.

[12] A. Stolcke. SRILM-an extensible language modeling toolkit. In *Proceeding of the Seventh International Conference on Spoken Language Processing (ICSLP 2002)*, volume 2, pages 901–904, Denver, Colorado, 2002.