Improving Sampling-based Alignment Method for Statistical Machine Translation Tasks

Juan LUO Jing SUN Yves LEPAGE
Graduate School of Information, Production and Systems
Waseda University
{juanluoonly@suou,cecily.sun@akane,yves.lepage@aoni}.waseda.jp

Abstract

We describe an approach to improve the performance of the sampling-based multilingual alignment method implemented by Anymalign on translation tasks. The idea of the approach is to enforce the alignment of N-grams. We compare the quality of the phrase translation table output by our approach and that of MGIZA++ for statistical machine translation tasks. We improved the performance of Anymalign in the baseline system, but did not beat MGIZA++ as we expected.

1 Introduction

In machine translation, alignment plays an important role in the process of building a machine translation system. The quality of the alignment, which identifies the relations between words or phrases in the source language and those in the target language, is crucial for the final results and the quality of a machine translation system. Training various alignment models requires alignment tools, that is, aligners. Currently, the state-of-the-art tool is MGIZA++ [2].

In this paper, we investigate methods and techniques of a different approach to subsentential alignment, the sampling-based method, implemented in Anymalign [6], and we propose an improvement. Experimental results using the Europarl parallel corpus [3] are presented. The organization of the paper is as follows. Section 2 provides the basic concepts and techniques of the subsentential alignment method. Section 3 presents the proposed method of Anymalign1-N to improve sampling-based alignment for statistical machine translation tasks. Section 4 describes the results obtained from experiments using Europarl data. Finally, in section 5, conclusion and future work are presented.

2 Sampling-based Alignment Method

There are various methods and models being suggested and implemented to solve the problem of alignment. Our work will follow and focus on the sampling-based subsentential alignment method proposed in [6]. This approach is implemented in Anymalign as a free software.[1] The approach is much simpler than the estimative approach, implemented in MGIZA++. Also its ability to perform multilingual alignment simultaneously is worth drawing attention.

In the sampling-based alignment method, terms appearing exactly on the same lines is central. In small corpora, such terms tend to become hapaxes, that is, terms with one occurrence only. Hapaxes have been shown to safely align across languages [6].

A multilingual parallel corpus is, firstly, assimilated without boundary between languages to a "monolingual" corpus, which is referred to as an alingual corpus. Then, subcorpora of the alingual corpus are selected to extract sequences of words appearing exactly on the same lines and thus generate alignments, as well as counting the number of times they have been obtained. In order to ensure the coverage of the corpus as it is sampling-based, a probability distribution for the sampling into subcorpora is introduced:

$$p(k) = \frac{-1}{k\ log(1-k/n)}$$

Here k and n denote the size of subcorpora and the size in lines of the alingual corpus. k/n is the probability that a particular sentence is chosen and $(1-k/n)$ is the probability for a sentence not to be chosen.

In obtaining translation probabilities of multilingual alignment, we collect counts of alignments $C(s_1, ..., s_L)$. $C(s_i)$ is the sum of counts over all alignments. Therefore, the translation probability of a sequence of words s_i is:

$$P(s_1, ..., s_{i-1}, s_{i+1}, ..., s_L | s_i) = \frac{C(s_1, ..., s_L)}{C(s_i)}$$

The lexical weights [4] are adapted accordingly in the situation of multilingual alignment:

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

$$W(s_1, ..., s_{i-1}, s_{i+1}, ..., s_L|s_i) =$$

$$\prod_{w_i \in s_i} max_{w_j \in \cup_{i \neq j} s_j} D(w_j|w_i)$$

where D is the lexical translation probability distribution.

3 Anymalign1-N

3.1 Problem Definition

The sampling-based approach has been proven in [7] to excel in aligning unigrams, which makes it very good at multilingual lexicon induction. However, the generated phrase tables are not sufficient for performing machine translation tasks up to the level of MGIZA++. This comes from the fact that Anymalign does not align enough N-grams.

3.2 Alignment with N-grams

We propose here a method to force the sampling-based approach to align more N-grams.

Consider that we have a parallel input corpus, i.e., a pair of corresponding sentences, for instance, in French and English. Groups of characters that are separated by spaces in these sentences are considered as words. Those single words are referred to as unigrams. Two words and three words are called bigrams and trigrams respectively and longer sequences of words are simply called N-grams.

Theoretically, since the sampling-based alignment method is good at aligning unigrams, if we could make Anymalign to align bigrams, trigrams, or even N-grams as if they were unigrams, the approach would presumably show better performance in producing phrase translation tables and, hence, better performance in terms of machine translation tasks. This is done by replacing spaces in the sentences by underscore symbols and reduplicating words as many times as needed. In this way, bigrams, trigrams and N-grams appear as unigrams. Table 1 depicts the way of forcing N-grams into unigrams.

3.3 Phrase Translation Tables

In the process of building a statistical machine translation system, it is essential to generate phrase translation tables for the machine translation tasks. The approach to produce a translation table with N-grams alignment using the sampling-based method, that is, Anymalign, is as follows: the two subparts (source and target) of a parallel corpus are processed separately to make them into bigram texts, trigram texts, and so on, and enforced into unigrams as described above. These corpora are then processed to produce phrase translation tables, as shown in Table 2. All phrase translation tables obtained are then merged into one big translation table for the purpose of better suiting the machine translation tasks.

Table 2: Merging all N-gram translation tables (TT) generated from training the source and the target corpora into one translation table.

		Target			
		unigrams	bigrams	trigrams	N-grams
Source	unigrams	TT1-1	TT1-2	TT1-3	TT1-N
	bigrams	TT2-1	TT2-2	TT2-3	TT2-N
	trigrams	TT3-1	TT3-2	TT3-3	TT3-N
	N-grams	TTN-1	TTN-2	TTN-3	TTN-N

4 Experiments

We present in this section the experimental results on the quality of the phrase translation tables obtained from MGIZA++, off-the-shelf Anymalign and our method (Anymalign with N-grams).

The input French-English parallel corpus from Europarl parallel corpus was used for training, tuning and testing. The detailed description of the corpora used in the experiments is given in Table 3. To perform the experiments, a standard statistical machine translation system was built using the Moses decoder [5], the SRILM toolkit [12] and MGIZA++, which is a multi-threaded version of GIZA++ [9].

For the evaluation of translations, four automatic evaluation metrics were used: mWER [8], BLEU [10], NIST [1], and TER [11].

The quality of the phrase translation table obtained from training MGIZA++ was evaluated in a first experiment (baseline). In order to evaluate the quality of Anymalign translation tables for the machine translation tasks, the phrase table obtained with MGIZA++ was replaced by that of Anymalign, which was trained in a second experiment using the Moses standard statistical machine translation system. The same process was carried out for our approach (Anymalign1-N) to evaluate its trans-

Table 3: Summary of French-English corpora for training set, development set, and test set.

		French	English
Train	sentences	100,000	100,000
	words	3,986,438	2,824,579
	words/sentence	38	27
Dev	sentences	500	500
	words	18,120	13,261
	words/sentence	36	26
Test	sentences	1,000	1,000
	words	38,936	27,965
	words/sentence	37	27

Table 1: Transforming N-grams into unigrams by inserting underscores between words for both the French part and English part of the corpus.

	French part	English part
1	le debat est clos .	the debate is closed .
2	le_debat debat_est est_clos clos_.	the_debate debate_is is_closed closed_.
3	le_debat_est debat_est_clos est_clos_.	the_debate_is debate_is_closed is_closed_.
4	le_debat_est_clos debat_est_clos_.	the_debate_is_closed debate_is_closed_.
5	le_debat_est_clos_.	the_debate_is_closed_.

Table 4: Evaluation results on Europarl French-English corpus.

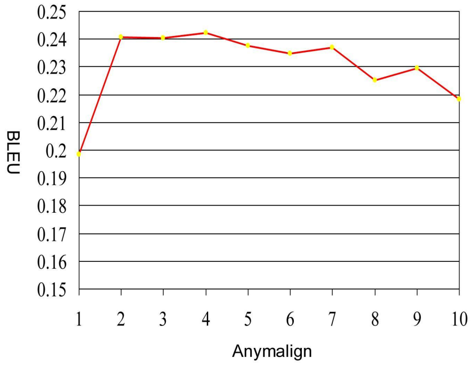
	mWER	BLEU	NIST	TER
MGIZA++	**0.5714**	**0.2742**	**6.6747**	**0.6170**
Anymalign1-10	0.6475	0.2182	5.8534	0.6886
Anymalign1-9	0.6279	0.2296	6.0261	0.6722
Anymalign1-8	0.6353	0.2253	5.9777	0.6794
Anymalign1-7	0.6157	0.2371	6.2107	0.6559
Anymalign1-6	0.6193	0.2349	6.1574	0.6634
Anymalign1-5	0.6099	0.2376	6.2331	0.6551
Anymalign1-4	0.6142	**0.2423**	6.2087	0.6583
Anymalign1-3	**0.6075**	0.2403	**6.3009**	**0.6507**
Anymalign1-2	0.6121	0.2406	6.2789	0.6536
Anymalign	0.6818	0.1984	5.6353	0.7188

lation quality in a third experiment. In order to be fair and comparable to the results produced by Moses/MGIZA++, we set the same amount of running time for Anymalign in the second and third experiments as that of MGIZA++. This is possible because Anymalign can be interrupted manually. The evaluation results of all experiments are shown in Table 4. On the whole, MGIZA++ outperforms Anymalign. Our approach Anymalign1-N gets much better results than Anymalign in its basic version.

A detailed description of the performance of Anymalign1-N on a statistical machine translation task is shown in Figure 1. The BLEU score shows a very significant increase from the unigram phrase translation table to the bigram phrase table: from 0.1984 to 0.2406. Anymalign1-4 gets the highest BLEU score of 0.2423. The score begins to decline from Anymalign1-5 and continues until Anymalign1-10. Overall, Anymalign1-4 shows the best performance in the statistical machine translation task on the Europarl French-English corpus.

Table 5 shows the number of N-gram entries in phrase translation tables of MGIZA++, Anymalign, and Anymalign1-N. The greatest number of N-gram entries in the MGIZA++ phrase tables is observed for tetragrams with 729,171 entries. The number of tetragram entries of Anymalign1-4 is the greatest among all Anymalign 4-gram entries. It suggests that the number of tetragrams has an important impact on the translation quality in the statistical machine translation tasks.

Figure 1: Translation quality in BLEU for different N of Anymalign1-N.



5 Conclusion

In this paper, we presented a method to significantly improve the translation quality of the sampling-based subsentential alignment approach: Anymalign is forced to align N-grams as if they were unigrams. A baseline statistical machine translation system was built to compare the translation performance of two aligners: MGIZA++ and Anymalign. While it still lies behind MGIZA++ for statistical machine translation of the Europarl French-English corpus, Anymalign1-N, the method presented here, obtains significantly better results as we expected and Anymalign1-4 shows the best performance. In the future we will focus on increasing the size of tetragrams of Anymalign phrase tables to improve the translation quality for statistical machine translation tasks.

References

[1] George Doddington. Automatic evaluation of machine translation quality using N-gram co-occurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, pages 138–145, San Diego, March 2002. Morgan Kaufmann Publishers Inc.

Table 5: Number of entries in phrase translation tables.

	unigram	bigram	trigram	tetragram	5-gram	6-gram	7-gram	8-gram	9-gram	10-gram	total
MGIZA++	148,488	463,400	685,451	**729,171**	683,380	596,208	462,319	0	0	0	3,768,417
Anymalign	819,569	0	0	0	0	0	0	0	0	0	819,569
Anymalign1-2	681,871	664,380	0	0	0	0	0	0	0	0	1,346,251
Anymalign1-3	465,607	496,817	311,481	0	0	0	0	0	0	0	1,273,905
Anymalign1-4	342,505	355,454	249,690	**159,778**	0	0	0	0	0	0	1,107,427
Anymalign1-5	258,745	266,976	185,854	134,187	86,993	0	0	0	0	0	932,755
Anymalign1-6	203,294	205,752	147,046	103,541	75,616	41,847	0	0	0	0	777,096
Anymalign1-7	165,742	167,771	116,552	86,339	62,179	35,712	20,670	0	0	0	654,965
Anymalign1-8	137,698	136,776	94,250	68,114	49,148	31,755	19,567	10,809	0	0	548,117
Anymalign1-9	119,074	114,740	79,044	55,992	42,212	27,090	15,062	8,843	6,493	0	468,550
Anymalign1-10	95,686	96,636	66,008	47,604	37,465	23,260	13,603	8,577	6,028	5,142	400,009

[2] 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, June 2007.

[3] 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, September 2005. URL `http://www.mt-archive.info/MTS-2005-Koehn.pdf`.

[4] Philipp Koehn, Franz J. Och, and Daniel Marcu. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 48–54, Edmonton, may 2003. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology-new/N/N03/N03-1017.pdf`.

[5] 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, June 2007. URL `http://www.aclweb.org/anthology/P/P07/P07-2045.pdf`.

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

[7] Adrien Lardilleux, Jonathan Chevelu, Yves Lepage, Ghislain Putois, and Julien Gosme. Lexicons or phrase tables? An investigation in sampling-based multilingual alignment. In Mikel Forcada and Andy Way, editors, *Proceedings of the third workshop on example-based machine translation*, pages 45–52, Dublin, Ireland, nov 2009. URL `http://www.computing.dcu.ie/~mforcada/ebmt3/proceedings/EBMT3-paper6.pdf`.

[8] Sonja Nießen, Franz Josef Och, Gregor Leusch, and Hermann Ney. An evaluation tool for machine translation: Fast evaluation for machine translation research. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC)*, pages 39–45, Athens, May 2000.

[9] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29, pages 19–51, March 2003. URL `http://acl.ldc.upenn.edu/J/J03/J03-1002.pdf`.

[10] 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, July 2002. URL `http://www.aclweb.org/anthology/P02-1040.pdf`.

[11] Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *Proceedings of Association for Machine Translation in the Americas (AMTA 2006)*, pages 223–231, Cambridge, Massachusetts, August 2006.

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