Improving the Distribution of N-Grams in Phrase Tables Obtained by the Sampling-Based Method

Juan Luo[1](✉), Adrien Lardilleux[2], and Yves Lepage[1]

[1] IPS, Waseda University, 2-7 Hibikino, Wakamatsu-ku, Kitakyushu-shi, Fukuoka 808-0135, Japan
juan.luo@suou.waseda.jp, yves.lepage@waseda.jp
[2] Affinity Engine, 4 Rue Doaren Molac, 56610 Arradon, France
adrien.lardilleux@affinity-engine.fr

Abstract. We describe an approach to improve the performance of sampling-based sub-sentential alignment method on translation tasks by investigating the distribution of n-grams in the phrase tables. This approach consists in enforcing the alignment of n-grams. We compare the quality of phrase translation tables output by this approach and that of the state-of-the-art estimation approach in statistical machine translation tasks. We report significant improvements for this approach and show that merging phrase tables outperforms the state-of-the-art techniques.

Keywords: Sub-sentential alignment · Statistical machine translation

1 Introduction

Phrase tables play an important role in the process of building statistical machine translation systems. Their quality is crucial for the quality of translation outputs. The most widely used state-of-the-art tool to generate phrase tables is MGIZA++ [1], which trains the IBM models [2] and the HMM introduced in [3] in combination with the Moses toolkit [4]. Phrase tables are also used in other domains, e.g., bilingual terminology extraction [5], creation of lexicon entries [6].

A phrase table is a list of phrase pairs that are translations of each other with feature scores (see Table 1). It is normally constructed in two steps by using MGIZA++ and Moses toolkit. The first step consists of the using alignment tool MGIZA++ to generate source-to-target and target-to-source word alignments between two languages. The second step uses Moses to extract bilingual phrase pairs from alignments through heuristic combination of both directions and compute feature scores.

A. Lardilleux – The work was done while the author was at TLP Group, LIMSI-CNRS, France.

Z. Vetulani and J. Mariani (Eds.): LTC 2011, LNAI 8387, pp. 419–431, 2014.
DOI: 10.1007/978-3-319-08958-4_34

Table 1. Example of a phrase table.

Source language	Target language	Feature scores			
French	English	$\phi(f\|e)$	$lex(f\|e)$	$\phi(e\|f)$	$lex(e\|f)$
rapport	report	0.921	0.924	0.917	0.841
parlement européen	european parliament	0.811	0.187	0.897	0.773
mais	, but	0.087	0.719	0.044	0.190
activités et	activities and	0.615	0.502	0.889	0.613
le président,	president,	0.953	0.889	0.870	0.965
monsieur le président	mr president	0.918	0.979	0.947	0.874
l' union européenne	european union	0.836	0.661	0.851	0.886
la commission européenne	european commission	0.836	0.852	0.968	0.987

In this article, we investigate a different approach to the production of phrase tables: the sampling-based approach [7], available as a free open-source tool called Anymalign.[1] Being in line with the association alignment approach (see e.g. [8–10]), it is much simpler than the models implemented in MGIZA++, which are in line with the estimation approach (e.g. [11–14]).

In sampling-based alignment, only those sequences of words that appear exactly in the same sentences of the corpus are considered for alignment. The key idea is to produce more candidate words by artificially reducing the size of the input corpus, i.e., many subcorpora of small sizes are obtained by sampling and processed one after another. Indeed, the smaller a subcorpus, the less frequent its words, and the more likely they are to share the same distribution.

The subcorpus selection process is guided by a probability distribution that ensures a proper coverage of the input parallel corpus:

$$\mathrm{p}(k) = \frac{-1}{k\ \log(1 - k/n)} \quad \text{(to be normalized)}$$

where k denotes the size (number of sentences) of a subcorpus and n the size of the complete input corpus. This function is very close to $1/k^2$ and gives more credit to small subcorpora, which happen to be the most productive [7]. Once the size of a subcorpus has been chosen according to this distribution, its sentences are randomly selected from the complete input corpus according to a uniform distribution. Then, from each subcorpus, sequences of words that share the same distribution are extracted to constitute alignments along with the number of times they were aligned.[2]

Eventually, the list of alignments is turned into a full-fledged phrase table by calculating various features for each alignment. In the following, we use two translation probabilities and two lexical weights as proposed by [15], as well as the commonly used phrase penalty, for a total of five features.

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

[2] Contrary to the widely used terminology where it denotes a set of links between the source and target words of a sentence pair, we call "alignment" a (source, target) phrase pair, i.e., it corresponds to an entry in the so-called [phrase] translation tables.

One important feature of the sampling-based alignment method is that it is *anytime* in essence: the number of random subcorpora to be processed is not set in advance, so the alignment process can be interrupted at any moment. Contrary to many approaches, *quality* is not a matter of time, however *quantity* is: the longer the aligner runs (i.e. the more subcorpora processed), the more alignments produced, and the more reliable their associated translation probabilities.

Intuitively, since the sampling-based alignment process can be interrupted without sacrificing the quality of alignments, it should be possible to allot more processing time for n-grams of similar lengths in both languages and less time to very different lengths. For instance, a source bigram is much less likely to be aligned with a target 9-gram than with a bigram or a trigram. The experiments reported in this paper make use of the anytime feature of Anymalign and of the possibility of allotting time freely.

This article is organized as follows: Sect. 2 defines the problem. Section 3 proposes a variant in order to improve the translation performance. Section 4 describes the merge of two aligners' phrase tables. Section 5 provides the conclusion.

2 Description of the Problem

In order to measure the performance of the sampling-based alignment approach implemented in Anymalign in statistical machine translation tasks, we conducted a preliminary experiment and compared with the standard alignment setting: symmetric alignments obtained from MGIZA++. Although Anymalign and MGIZA++ are both capable of parallel processing, for fair comparison in time, we run them as single processes in all our experiments.

2.1 Experimental Setup

A sample of the French-English parts of the Europarl parallel corpus was used for training, tuning and testing. A detailed description of the data used in the experiments is given in Table 2. To perform the experiments, a standard statistical machine translation system was built for each different alignment setting, using the Moses decoder [4] MERT (Minimum Error Rate Training) [16] and the SRILM toolkit [17]. As for the evaluation of translations, the BLEU metric [18] was used.

2.2 Problem Definition

In a first setting, we evaluated the quality of translations output by the Moses decoder using the phrase table obtained by making MGIZA++'s alignments symmetric. In a second setting, this phrase table was simply replaced by that produced by Anymalign. Since Anymalign can be stopped at any time, for a fair comparison it was run for the same amount of time as MGIZA++: seven hours in total. The experimental results are shown in Table 3. In order to investigate

Table 2. Statistics on the French-English parallel corpus used for the training, development, and test sets.

		French	English
Train	Sentences	100,000	
	Word tokens	3,986,438	2,824,579
	Word types	42,919	32,588
Dev	Sentences	500	
	Word tokens	18,120	13,261
	Word types	2,300	1,941
Test	Sentences	1,000	
	Word tokens	38,936	27,965
	Word types	3,885	3,236

Table 3. Evaluation results on a statistical machine translation task using phrase tables obtained from MGIZA++ and Anymalign (baseline).

	BLEU
MGIZA++	27.42
Anymalign (baseline)	22.85

the differences between MGIZA++ and Anymalign phrase tables, we analyzed the distribution of n-grams of both aligners, The distributions are shown in Table 4 and Table 5.

In Anymalign's phrase table, the number of alignments is 8 times that of 1×1 n-grams in MGIZA++ phrase table, or twice the number of 1×2 n-grams or 2×1 n-grams in MGIZA++ phrase table. Along the diagonal (m × m n-grams) for $m > 2$, the number of alignments in Anymalign table is approximately hundred

Table 4. Distribution of phrase pairs in phrase tables (MGIZA++).

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**89,788**	44,941	10,700	2,388	486	133	52	148,488
	bigrams	61,007	**288,394**	86,978	20,372	5,142	1,163	344	463,400
	trigrams	19,235	149,971	**373,991**	105,449	27,534	7,414	1,857	685,451
	4-grams	5,070	47,848	193,677	**335,837**	106,467	31,011	9,261	729,171
	5-grams	1,209	13,984	73,068	193,260	**270,615**	98,895	32,349	683,380
	6-grams	332	3,856	24,333	87,244	177,554	**214,189**	88,700	596,208
	7-grams	113	1,103	7,768	33,278	91,355	157,653	**171,049**	462,319
	total	176,754	550,097	770,515	777,828	679,153	510,458	303,612	3,768,417

Table 5. Distribution of phrase pairs in phrase tables (Anymalign).

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams ···	total
Source	unigrams	**791,099**	105,961	9,139	1,125	233	72	37 ···	1,012,473
	bigrams	104,633	**21,602**	4,035	919	290	100	44 ···	226,176
	trigrams	10,665	4,361	**2,570**	1,163	553	240	96 ···	92,268
	4-grams	1,698	1,309	1,492	**1,782**	1,158	573	267 ···	61,562
	5-grams	378	526	905	1,476	**1,732**	1,206	642 ···	47,139
	6-grams	110	226	467	958	1,559	**1,694**	1,245 ···	40,174
	7-grams	40	86	238	536	1,054	1,588	**1,666** ···	35,753
	···	···	···	···	···	···	···	··· ···	···
	total	1,022,594	230,400	86,830	55,534	42,891	37,246	34,531 ···	1,371,865

times less than in MGIZA++ table. This confirms the results given in [19] that the sampling-based approach excels in aligning unigrams, which makes it better at multilingual lexicon induction than, e.g., MGIZA++. However, its phrase tables do not reach the performance of symmetric alignments from MGIZA++ on translation tasks. This basically comes from the fact that Anymalign does not align enough long n-grams.

3 Anymalign1-N

3.1 Phrase Subtables

To solve the above-mentioned problem, we propose a method to force the sampling-based approach to align more n-grams.

Consider that we have a parallel input corpus, i.e., a list of (source, target) sentence pairs, for instance, in French and English. Groups of characters that are separated by spaces in these sentences are considered as words. Single words are referred to as unigrams, and sequences of two and three words are called bigrams and trigrams, respectively.

Theoretically, since the sampling-based alignment method excels at aligning unigrams, we could improve it by making it align bigrams, trigrams, or even longer n-grams as if they were unigrams. We do this by replacing spaces between words by underscore symbols and reduplicating words as many times as needed, which allows to make bigrams, trigrams, and longer n-grams appear as unigrams. Table 6 depicts the way of forcing n-grams into unigrams. The same trick was used in a work by [20].

It is thus possible to use various parallel corpora, with different segmentation schemes in the source and target parts. We refer to a parallel corpus where source n-grams and target m-grams are assimilated to unigrams as a *unigramized n-m corpus*. These corpora are then used as input to Anymalign to produce phrase subtables, as shown in Table 7. Practically, we call Anymalign1-N the process of running Anymalign with all possible unigramized n-m corpora, with n and

Table 6. Transforming n-grams into unigrams by inserting underscores and reduplicating words for both the French part and English part of the input parallel corpus.

n	French	English
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 7. List of n-gram phrase subtables (TT) generated from the training corpus. These subtables will then be merged together into a single phrase table.

		Target					
		unigrams	bigrams	trigrams	4-grams	···	N-grams
Source	unigrams	TT1 × 1	TT1 × 2	TT1 × 3	TT1 × 4	···	TT1 × N
	bigrams	TT2 × 1	TT2 × 2	TT2 × 3	TT2 × 4	···	TT2 × N
	trigrams	TT3 × 1	TT3 × 2	TT3 × 3	TT3 × 4	···	TT3 × N
	4-grams	TT4 × 1	TT4 × 2	TT4 × 3	TT4 × 4	···	TT4 × N
	···	···	···	···	···	···	···
	N-grams	TTN × 1	TTN × 2	TTN × 3	TTN × 4	···	TTN × N

m both ranging from 1 to a given N. In total, this corresponds to N × N runs of Anymalign. All phrase translation subtables are finally merged together into one large phrase table, where translation probabilities are re-estimated given the complete set of alignments.

Although Anymalign is capable of directly producing alignments of sequences of words, we use it with a simple filter[3] so that it only produces (typographic) unigrams in output, i.e., n-grams and m-grams assimilated to unigrams in the input corpus. This choice was made because it is useless to produce alignments of sequences of words, since we are only interested in *phrases* in the subsequent machine translation tasks. Those phrases are already contained in our (typographic) unigrams: all we need to do to get the original segmentation is to remove underscores from the alignments.

3.2 Equal Time Configuration

The same experimental process (i.e., replacing the phrase table) as in the preliminary experiment was carried out on Anymalign1-N with equal time distribution, i.e., uniformly distributed time among subtables. For a fair comparison, the same amount of time was given: seven hours in total. The results are given in

[3] Option `-N 1` in the program.

Table 8. Anymalign1-4 with equal time for each $n \times m$ n-grams alignments.

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**171,077**	118,848	39,253	13,327	0	0	0	342,505
	bigrams	119,953	**142,721**	67,872	24,908	0	0	0	355,454
	trigrams	45,154	75,607	**86,181**	42,748	0	0	0	249,690
	4-grams	15,514	30,146	54,017	**60,101**	0	0	0	159,778
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	351,698	367,322	247,323	141,084	0	0	0	1,107,427

Table 12. On the whole, MGIZA++ significantly outperforms Anymalign1-N, by more than 4 BLEU points. However, the proposed approach, Anymalign1-N, produces better results than Anymalign in its basic version, with the best increase with Anymalign1-4 (+1.4 BP).

The comparison of Table 8 and 4 shows that Anymalign1-N delivers too many alignments outside of the diagonal ($m \times m$ n-grams) and still not enough along the diagonal. Consequently, this number of alignments should be lowered. A way of doing so is by giving less time for alignments outside of the diagonal.

3.3 Time Distribution among Subtables

To this end, we distribute the total alignment time among phrase subtables according to the standard normal distribution:

$$\phi(n, m) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(n-m)^2}$$

The alignment time allotted to the subtable between source n-grams and target m-grams will thus be proportional to $\phi(n, m)$.

In a third evaluation, we compare this new setting (with a total amount of processing time of 7 h) with MGIZA++, Anymalign in its standard use, and Anymalign1-N with equal time distribution (Table 12). There is an increase in BLEU scores for almost all Anymalign1-N, from Anymalign1-3 to Anymalign1-10, when compared with equal time distribution. The greatest increase in BLEU is obtained for Anymalign1-10 (almost +2 BP). Anymalign1-4 shows the best translation quality among all other settings, but gets a less significant improvement (+0.2 BP).

Again, we investigated the number of entries in Anymalign1-N run with this normal time distribution. We compare the number of entries in Anymalign1-4 with equal time distribution (Table 8) and standard normal time distribution (Table 9). The number of phrase pairs on the diagonal roughly doubled when using standard normal time distribution. We can see a significant increase in the number of phrase pairs of similar lengths, while the number of phrase pairs with

Table 9. Anymalign1-4 with standard normal time distribution.

		Target							
		unigrams	bigrams	trigrams	4-grams	5-grams	6-grams	7-grams	total
Source	unigrams	**255,443**	132,779	13,803	469	0	0	0	402,494
	bigrams	134,458	**217,500**	75,441	8,612	0	0	0	436,011
	trigrams	15,025	86,973	**142,091**	48,568	0	0	0	292,657
	4-grams	635	10,516	61,741	**98,961**	0	0	0	171,853
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	405,561	447,768	293,076	156,610	0	0	0	1,303,015

different lengths tends to decrease slightly. This means that the standard normal time distribution allowed us to produce much more numerous useful alignments (a priori, phrase pairs with similar lengths), while maintaining the noise (phrase pairs with different lengths) to a low level, which is a neat advantage over the original method.

3.4 Pruning Phrase Tables

Inspired by the work of Johnson et al. [21], we applied the technique of pruning on phrase tables of Anymalign (standard normal time distribution).

In [21], Fishers exact significance test is used to eliminate a substantial number of phrase pairs. The significance of the association between a (source, target) phrase pair is evaluated and their probability of co-occurrence in the corpus is calculated. A two by two contingency table for the phrase pair $(\tilde{s}, \tilde{t})$ is shown in Table 10.

The hypergeometric distribution is used to compute the observed probability of joint occurrence $C(\tilde{s}, \tilde{t})$, with $\tilde{s}$ a source phrase and $\tilde{t}$ a target phrase:

$$p_h(C(\tilde{s}, \tilde{t})) = \frac{\binom{C(\tilde{s})}{C(\tilde{s},\tilde{t})}\binom{N-C(\tilde{s})}{C(\tilde{t})-C(\tilde{s},\tilde{t})}}{\binom{N}{C(\tilde{t})}} \tag{1}$$

Table 10. 2×2 contingency table for $\tilde{s}$ and $\tilde{t}$

$C(\tilde{s}, \tilde{t})$	$C(\tilde{s}) - C(\tilde{s}, \tilde{t})$	$C(\tilde{s})$
$C(\tilde{t}) - C(\tilde{s}, \tilde{t})$	$\mathrm{N} - C(\tilde{s}) - C(\tilde{t}) + C(\tilde{s}, \tilde{t})$	$\mathrm{N} - C(\tilde{s})$
$C(\tilde{t})$	$\mathrm{N} - C(\tilde{t})$	N

Here, N is the number of sentences in the input parallel corpus. The p-value is calculated as:

$$\text{p-value}(C(\tilde{s},\tilde{t})) = \sum_{k=C(\tilde{s},\tilde{t})}^{\infty} p_h(k) \tag{2}$$

Any phrase pair with a p-value greater than a given threshold will be filtered out.

In a fourth evaluation, we compare with the previous settings. We used $\alpha+\varepsilon$ and $\alpha-\varepsilon$ filters. The proportion of phrase pairs filtered out from the phrase tables is shown in Fig. 1. In both cases, the number of phrase pairs discarded from phrase tables varies according to N: it amounts to around 87 % for Anymalign1-1, but only to about half of the phrase pairs for Anymalign1-3 to Anymalign 1-10.

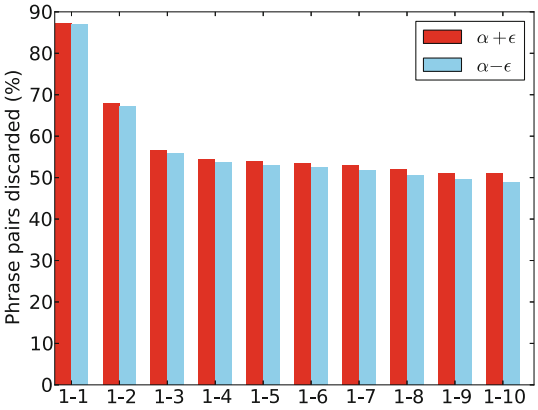

Fig. 1. Proportion of entries discarded in phrase tables of Anymalign1-N.

Evaluation results are given in Table 12. The phrase table size reduction results in slight but consistent improvements in translation quality. Among all Anymalign1-N, Anymalign1-4 once again gets the highest BLEU score of 25.11 ($\alpha+\varepsilon$ filter) and 25.14 ($\alpha-\varepsilon$ filter). This allowed us to achieve a slight improvement (+0.7 BLEU points) over Anymalign1-4 with standard normal time distribution and a significant improvement (+2.2 BLEU points) over Anymalign baseline.

The distribution of phrase pairs in pruned phrase tables are shown in Table 11(a) with $\alpha+\varepsilon$ filter and Table 11(b) with $\alpha-\varepsilon$ filter. The largest difference when compared with the non-pruned phrase table of Anymalign 1-4 with standard normal time distribution (Table 9) is visible in the cell corresponding to 1-to-1 entries. As a consequence, the largest number of entries are now 2-to-2 phrase pairs, which account for around 19 % of the total number of phrase pairs in both cases.

Table 11. Anymalign1-4 with standard normal time distribution (after pruning).

(a) $\alpha + \varepsilon$

		Target							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source	1-grams	**60,297**	59,099	8,819	328	0	0	0	128,543
	2-grams	58,232	**110,415**	51,557	6,954	0	0	0	227,158
	3-grams	9,777	58,604	**69,431**	28,046	0	0	0	165,858
	4-grams	474	8,586	31,209	**31,666**	0	0	0	71,935
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	128,780	236,704	161,016	66,994	0	0	0	593,494

(b) $\alpha - \varepsilon$

		Target							
		1-grams	2-grams	3-grams	4-grams	5-grams	6-grams	7-grams	total
Source	1-grams	**63,252**	59,621	8,844	332	0	0	0	132,049
	2-grams	58,595	**115,664**	52,062	7,006	0	0	0	233,327
	3-grams	9,801	58,850	**69,749**	28,287	0	0	0	166,687
	4-grams	477	8,701	31,391	**31,848**	0	0	0	72,417
	5-grams	0	0	0	0	0	0	0	0
	6-grams	0	0	0	0	0	0	0	0
	7-grams	0	0	0	0	0	0	0	0
	total	132,125	242,836	162,046	67,473	0	0	0	604,480

4 Merging Phrase Tables

In order to check exactly how different the phrase table of MGIZA++ and that of Anymalign are, we performed an additional set of experiments in which MGIZA++'s phrase table is merged with that of Anymalign baseline. As for the feature scores in the phrase tables for the intersection part of both aligners, we adopted parameters either from MGIZA++ or from Anymalign for evaluation.

Evaluation results on machine translation tasks with merged phrasse tables are given in Table 12. This setting outperforms MGIZA++ on BLEU scores.

Table 12. Evaluation results.

MGIZA++	**27.42**									
Anymaslign (baseline)	**22.85**									
	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10
Anymalign1-N (equal time)	19.84	24.06	24.03	**24.23**	23.76	23.49	23.71	22.53	22.96	21.82
Anymalign1-N (std.norm.)	19.84	24.04	24.41	**24.42**	24.36	24.03	24.05	23.66	24.02	23.61
Anymalign1-N (std.n., $\alpha + \varepsilon$)	19.53	24.25	24.13	**25.11**	24.57	24.59	24.19	24.46	24.61	24.59
Anymalign1-N (std.n., $\alpha - \varepsilon$)	19.76	24.10	24.70	**25.14**	24.57	24.47	24.16	24.18	24.58	24.40
Merge (MGIZA++ para.)	**27.54**									
Merge (Anymalign para.)	**27.47**									

The phrase table with Anymalign parameters for the intersection part is slightly behind the phrase table with MGIZA++ parameters. This may indicate that the feature scores in Anymalign phrase table need to be revised.

5 Conclusions and Future Work

In this article, we proposed a method to improve the performance of the sampling-based sub-sentential alignment method on statistical machine translation tasks.

We analyzed the strengths and weaknesses of this method according to the distribution of phrase pairs in phrase tables, and pointed to directions for proposed work building on its strengths. By introducing a method to enforce the alignment of n-grams, called Anymalign1-N, we increased the phrase coverage. A gain of 1.3 BLEU point over Anymalign baseline was observed. In order to balance the distribution of n-grams in phrase tables, a standard normal time distribution has been introduced. Within the same amount of processing time, the number of n-m phrase pairs of similar lengths increased substantially, which led to additional improvements in translation quality with Anymalign1-N (+0.2 BLEU point). The translation quality was further improved by pruning phrase tables. In total, the experiments proposed in this article allowed us to achieve a significant improvement of more than 2.2 BLEU point over Anymalign baseline. Finally, merging Anymalign's phrase table with that of MGIZA++ allowed to outperform MGIZA++ alone.

In the future, we intend to modify the feature scores computed by Anymalign in order to make it better suited to statistical machine translation tasks.

Acknowledgments. Part of the research presented in this paper has been done under a Japanese grant-in-aid (Kakenhi C, 23500187: Improvement of alignments and release of multilingual syntactic patterns for statistical and example-based machine translation).

References

1. Gao, Q., Vogel, S.: Parallel implementations of word alignment tool. In: Software Engineering. Testing, and Quality Assurance for Natural Language Processing, Columbus, Ohio, pp. 49–57 (2008)
2. Brown, P., Pietra, S.D., Pietra, V.D., Mercer, R.: The mathematics of statistical machine translation: parameter estimation. Comput. Linguist. **19**(2), 263–311 (1993)
3. Vogel, S., Ney, H., Tillman, C.: HMM-based word alignment in statistical translation. In: Proceedings of the 16th International Conference on Computational Linguistics, Copenhagen, Denmark, pp. 836–841 (1996)
4. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, pp. 177–180 (2007)

5. Ideue, M., Yamamoto, K., Utiyama, M., Sumita, E.: A comparison of unsupervised bilingual term extraction methods using phrase-tables. In: Proceedings of MT Summit XIII, Xiamen, China, pp. 346–351 (2011)
6. Thurmair, G., Aleksic, V.: Creating term and lexicon entries from phrase tables. In: Proceedings of the 16th Annual Conference of the European Association for Machine Translation, Trento, Italy, pp. 253–260 (2012)
7. Lardilleux, A., Lepage, Y.: Sampling-based multilingual alignment. In: Proceedings of International Conference on Recent Advances in Natural Language Processing, Borovets, Bulgaria, pp. 214–218 (2009)
8. Gale, W., Church, K.: Identifying word correspondences in parallel texts. In: Proceedings of the 4th DARPA Workshop on Speech and Natural Language, California, pp. 152–157 (1991)
9. Melamed, D.: Models of translational equivalence among words. Comput. Linguist. **26**(2), 221–249 (2000)
10. Moore, R.: Association-based bilingual word alignment. In: Proceedings of the ACL Workshop on Building and Using Parallel Text, Ann Arbor, pp. 1–8 (2005)
11. Brown, P., Lai, J., Mercer, R.: Aligning sentences in parallel corpora. In: Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, California, pp. 169–176 (1991)
12. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Comput. Linguist. **29**(1), 19–51 (2003)
13. Liang, P., Taskar, B., Klein, D.: Alignment by agreement. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, New York, pp. 104–111 (2006)
14. Dyer, C., Chahuneau V., Smith, N. A.: A simple, fast, and effective reparameterization of IBM model 2. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Atlanta, pp. 644–648 (2013)
15. Koehn, P., Och, F.J., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Edmonton, pp. 48–54 (2003)
16. Och, F.J.: Minimum error rate training in statistical machine translation. In: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Sapporo, Japan, pp. 160–167 (2003)
17. Stolcke, A.: SRILM-an extensible language modeling toolkit. In: Proceedings of the 7th International Conference on Spoken Language Processing, vol. II, pp. 901–904 (2002)
18. Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 311–318 (2002)
19. Lardilleux, A., Chevelu, J., Lepage, Y., Putois, G., Gosme, J.: Lexicons or phrase tables? an investigation in sampling-based multilingual alignment. In: Proceedings of the 3rd Workshop on Example-based Machine Translation, Dublin, Ireland, pp. 45–52 (2009)

20. Henríquez Q, A.C., Costa-jussà, R.M., Daudaravicius, V., Banchs, E. R., Mariño, B. J.: Using collocation segmentation to augment the phrase table. In: Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR, Uppsala, Sweden, pp. 98–102 (2010)
21. Johnson, J.H., Martin, J., Foster, G., Kuhn, R.: Improving translation quality by discarding most of the phrasetable. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Czech Republic, pp. 967–975 (2007)