Analogy-based machine translation using secability

Tatsuya Kimura, Jin Matsuoka, Yusuke Nishikawa and Yves Lepage
Graduate School of Information, Production and Systems, Waseda University, Japan
{tatsuya-kimura@ruri.,jinmatsuoka@akane.,y_nishikawa@asagi.,yves.lepage@}waseda.jp

***Abstract*—The problem of reordering remains the main problem in machine translation. Computing structures of sentences and the alignment of substructures is a way that has been proposed to solve this problem. We use secability to compute structures and show its effectiveness in an example-based machine translation.**

***Index Terms*—Example-based machine translation; proportional analogy; secability; alignment; translation table**

I. Introduction

Lepage and Denoual (2005) have proposed an analogy-based framework for translation which is an instance of example-based machine translation (EBMT) [1]. But the translation of long sentences remains difficult. The difficulty comes from the difference in sentence structure in different languages. These differences become larger when sentences become longer. Consequently for this technique, the computation of the structure of sentences is an important issue. In this paper we propose a method that does not use part-of-speech information. It is an unsupervised learning method. We use the notion of 'secability' to compute the structure of sentences. This consists in cutting sentence according to the degree to which words or phrases bind together. During the translation table generation process, we select those alignments that maximize lexical weights. These alignments are stored in translation tables with their translation probabilities and lexical weights. They are exploited in the analogy-based system to build analogies. We conduct some preliminary experiments and test the hypothesis that secability can helps to translate in an effective way in conjunction with the analogy-based framework.

II. Translation Probabilities and Lexical Weights

The translation probability for a bilingual alignment is computed as follows. The transition probability of a phrase $\bar{f} = (f_1 f_2 \dots f_i)$ given a phrase $\bar{e} = (e_1 e_2 \dots e_j)$ is the number of times they appear together in the sentences which are translation of one another, derided by the total number of occurrences of $\bar{e}$ in the bilingual corpus. See Equation (1).
For an example of calculation of lexical weights. In this scheme, each of the French words f_i is aligned with some English words e_j with the word translation probability $w(f_i|e_j)$ [3].

$$w(\bar{f}|\bar{e}) = \frac{C(\bar{f},\bar{e})}{C(\bar{e})} \tag{1}$$

$$lex(\bar{f}|\bar{e}) = \sqrt[n]{\prod_{i=1}^{n} \frac{1}{|\{j|(i,j) \in a\}|} \sum_{\forall (i,j) \in a} w(f_i|e_j)} \tag{2}$$

III. Secability

Secability means the quality of being divisible. For machine translation, secability was proposed by Chenon [2]. Based on secability, it is possible to build a structure for sentences. Equation (3) shows how to compute the secability value between two words $e2$ and $e3$ based on the probabilities of bigrams $e_i e_{i+1}$. The probabilities we use in our actual computations are smoothed according to Equation (4) For a given sentence, secability values are computed between each word of the sentence. The secability value with the highest value indicates the weakest place in the sentence, i.e., the place where the sentence can be most easily divided into two parts. This division process is repeated recursively in both parts of the sentences.

$$sec(e_2 e_3) = \frac{p(e_1 e_2) \cdot p(e_2 e_3) \cdot p(e_3 e_4)}{p(e_1 e_2 e_3) \cdot p(e_2 e_3 e_4)} \tag{3}$$

$$p(e_i e_{i+1}) = \frac{C(e_i e_{i+1}) + \delta}{N + \delta \times V} \tag{4}$$

$$\delta \ = \ 1 \ \ (Laplace's \ \ law)$$

where $C(e_i e_{i+1})$ is the number of occurrences of the bigram $(e_i e_{i+1})$ in the data, N is the total number of n-grams and V is the actual number of n-grams.

IV. Phrase-to-phrase alignment using Secability

The division of sentences can be applied to all sentences of a bicorpus, for the source sentences and the target sentences. This division is performed independently in each language on the contrary to, e.g., inversion transduction grammars [5]. In addition, and independently, it is possible to compute the correspondence between source and target words using word-to-word alignment tools Anymalign [4]. A partial sentential alignment using word alignments is shown graphically in Fig. 1, by drawing lines from some of the source words to some of the target words. Based on the tree structure obtained by secability and the partial alignment between words in corresponding sentences, we create the correspondences between subtrees across the two languages. Fig. 2 shows the tree structure in English (above) and French (below) with their partial word-to-word correspondences (indicated by lines) and substructure correspondences (indicated by boxes). Based on such substructure correspondences, we extract subsentential alignments by projection of the substructures. In this way, for the example in Fig. 3, we get the correspondences between "le fruit ", and "the fruit ", between "mange le fruit "and "eats the fruit ", between "mange le fruit "and "the fruit "and between "

le fruit ”and ”eats the fruit ”. For long sentences, this technique allows to get correspondences between long pieces, hence a translation table that contains long entries can be produced.

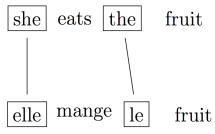

Fig. 1. A partial alignment from some of the English words to some of the French words

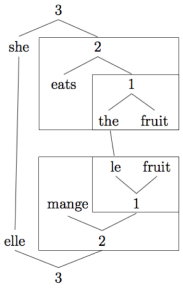

Fig. 2. Two secability trees in English (above) and French (below) with their partial word-for-word correspondences

V. Analogy-based translation using secability

We use the notion of analogy proposed in [1]. In all generality, an analogy A : B :: C : D means that “ A is to B as C is to D ”. For example, If we want to translate the following sentence.

”she eats the hamburger”

We segment the sentence by using secability. Then, the sentence is divided into the following three parts :

1) ”the hamburger”
2) ”eats the hamburger”
3) ”she eats the hamburger”

We translate in order of shorter substructures in the sentence. By doing so, it is possible to translate long sentences by combining shorter sequences.

the fruit : eats the fruit :: X = the hamburger : eats the hamburger

le fruit : mange le fruit :: le hamburger : Y = mange le hamburger

We add intermediate translation results to the translation table.

eats the fruit : she eats the fruit :: X = eats the hamburger : she eats the hamburger

mange le fruit : elle mange le fruit :: mange le hamburger : Y = elle mange le hamburger

VI. Experiments

We use the Europarl corpus [6]. Table I gives statistics about the data. We use Spanish and Portuguese (usually highest BLEU score) and French and Finnish (usually lowest BLEU score). English and French are used for comparison.

TABLE I
STATISTICS ABOUT TRAINING SET AND TEST SET

		French	English	Finnish	French	Portuguese	Spanish
Train	Sentences	347,614		347,614		347,614	
	Words	10,959,243	9,945,400	7,180,028	10,959,243	10,302,370	10,472,185
Test	Sentences	100		100		100	
	Words	2,880	2,638	1,838	2,846	2,709	2,747
	Sent. length: avg ± stdev	30 ± 10	26 ± 9	19 ± 7	29 ± 10	27 ± 9	28 ± 9

TABLE II
OF ENTRIES IN TRANSLATION TABLES USING SECABILITY

	fr-en	fr-fi	pt-es
# of entries	1,280,483	1,045,670	1,337,194
Length of Entries: avg ± stdev	10.04 ± 14.96	11.40 ± 17.40	10.47 ± 15.56

We performed two experiments. In the first one, the best translation candidates are selected using translation probabilities. In the second one, lexical weights are used instead of translation probabilities.

TABLE III
EVALUATION RESULTS IN EBMT BY ANALOGY

	fr-en		en-fr		fi-fr		fr-fi		pt-es		es-pt	
	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER	BLEU	TER
Translation probability	**13.5**	**0.73**	**10.2**	**0.71**	0.4	0.88	**1.0**	**1.24**	**23.7**	**0.58**	**20.9**	**0.61**
Lexical weight	7.6	0.97	7.3	0.85	**0.9**	**0.85**	0.8	1.49	12.0	0.90	11.7	0.81

VII. Conclusion

When comparing results of translation by analogy, better results are obtained with translation probabilities than with lexical weights. The reason may be the following:

- Lexical weights are less sensitive to word co-occurrence because they are the product of translation probabilities of words.
- In example-based machine translation of longer units, word co-occurrences are important. This is better reflected by transition probabilities.

References

[1] Y. Lepage and E. Denoual. Purest ever example-based machine translation: Detailed presentation and assessment. *Machine Translation*, 19(3):251— 282, 2005b.

[2] C. Chenon. Vers une meilleure utilisabilité des mémoires de traduction, fondée sur un alignement sous-phrastique *PhD thesis*, Université Joseph Fourier-Grenoble 1, 2006.

[3] P. Koehn, F.J. Och, and D. Marcu. Statistical phrase-based translation. In *Proceedings of the Human Language Technology and North American As- sociation for Computational Linguistics Conference (HLT/NAACL)*, pages 127—133, Edmonton, Alberta, 2003.

[4] A. Lardilleux and Y. Lepage. Sampling-based multilingual alignment. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214—218, 2009.

[5] D. Wu. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3):377—403, 1997.

[6] P. 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.