Marker-Based Chunking in Eleven European Languages for Analogy-Based Translation

Kota Takeya and Yves Lepage(✉)

IPS, Waseda University, Hibikino 2-7, Kitakyushu, Fukuoka 808-0135, Japan
kota-takeya@toki.waseda.jp, yves.lepage@waseda.jp

Abstract. An example-based machine translation (EBMT) system based on proportional analogies requires numerous proportional analogies between linguistic units to work properly. Consequently, long sentences cannot be handled directly in such a framework. Cutting sentences into chunks would be a solution. Using different markers, we count the number of proportional analogies between chunks in 11 European languages. As expected, the number of proportional analogies between chunks found is very high. These results, and preliminary experiments in translation, are promising for the EBMT system that we intend to build.

1 Introduction

The work reported in this paper is part of the aim at building an EBMT system based on proportional analogies similar to the one described in [1]. For such an EBMT system to work well, the more numerous the proportional analogies, the better the translation outputs are expected to be. The translation method which we introduce in Sect. 2 can work on small sentences, but cannot handle long sentences, like the ones in the Europarl corpus. For long sentences, translating chunk by chunk could be a solution. As the number of analogies is the crucial point, this paper inspects ways of cutting sentences into chunks using different markers and examines the number of proportional analogies between them in 11 European languages.

The paper is organized as follows. Section 2 explains the notion of proportional analogy and shows how to translate using proportional analogy. Section 3 describes the basic notion of marker-based chunking used in the reported experiments. Section 4 presents the data for the experiments and the experimental protocol. Section 5 gives the results of the experiments. A conclusion summarizes this work in Sect. 6.

2 Proportional Analogy

2.1 Examples and Formalization

Proportional analogy is a general relationship between four objects, A, B, C and D, that states that 'A is to B as C is to D'. Its standard notation is $A : B :: C : D$.

This paper is part of the outcome of research performed under a Waseda University Grant for Special Research Project (project number: 2010A-906).

Z. Vetulani and J. Mariani (Eds.): LTC 2011, LNAI 8387, pp. 432–444, 2014.
DOI: 10.1007/978-3-319-08958-4_35

The following are proportional analogies between words (1), chunks (2) and sentences (3):

$$\text{relate} : \text{unrelated} :: \text{modulate} : \text{unmodulated} \tag{1}$$

$$\text{a key} : \text{the key} :: \text{a first visit} : \text{the first visit} \tag{2}$$

$$\text{I like music.} : \text{Do you go to lives?} : \text{I like jazz music.} : \text{Do you go to jazz lives?} \tag{3}$$

A formalization has been proposed in [2]. This formalization reduces to counting number of occurrences of symbols and computating edit distances. Precisely:

$$A : B :: C : D \Rightarrow \begin{cases} |A|_a - |B|_a = |C|_a - |D|_a, \ \forall a \\ \delta(A,B) = \delta(C,D) \\ \delta(A,C) = \delta(B,D) \end{cases}$$

where $|A|_a$ stands for the number of occurrences of character a in string A and $\delta(A,B)$ stands for the edit distance between strings A and B with only insertion and deletion as edit operations.

2.2 Translation by Proportional Analogy

A translation method based on proportional analogies has been proposed by Lepage and Denoual [1]. The following procedure gives the basic outline of the method to perform the translation of an input chunk. Let us suppose that we have a corpus of aligned chunks in two languages. Let $D =$ "ein großes programm und" be a source chunk to be translated into one or more target chunks $\widehat{D}$. Let the bilingual corpus consists of four chunks with their translations:

ernste programme	$\leftrightarrow$	programmes sérieux
ein ernstes programm	$\leftrightarrow$	un programme sérieux
große programme und	$\leftrightarrow$	gros programmes et
das ernste programm	$\leftrightarrow$	le programme sérieux

The proposed method forms all possible analogical equations in C with all possible pairs of chunks from the parallel corpus. Among them:

$$\text{ernste programme} : \text{ein ernstes programm} :: C : \text{ein großes programm und}$$

The solution of this analogical equation is $x =$ "große programme und". As the pair of chunks "große programme und" $\leftrightarrow$ "gros programmes et" is already part of the parallel aligned corpus, an analogical equation can be formed in the target language:

$$\text{programmes sérieux} : \text{un programme sérieux} :: \text{gros programmes et} : \widehat{D}$$

Its solution is a candidate translation of the source chunk: $\widehat{D} =$ "un gros programme et".

3 Marker-Based Chunking

In order to be able to apply the previous proposed method to various languages, we want to segment in a fully automatic and universal way sentences in different languages into sub-sentential units like chunks.

3.1 The Marker Hypothesis

We use the marker hypothesis for this. This hypothesis was first laid by Green [3].

> The marker hypothesis states that all natural languages contain a small number of elements that signal the presence of particular syntactic constructions.

We perform chunking based on this notion and use a method called marker-based chunking [4–6]. We define a chunk as a sequence of words delimited by markers. Markers should be words such as determiners (the), conjunctions (and, but, or), prepositions (in, from, to), possessive and personal pronouns (mine, you). A chunk can be created at each occurrence of a marker word. In addition, a further constraint requires that each chunk contains at least one non-marker word. Without non-marker words, a chunk would become meaningless as it would not contain any meaningful word.

As result examples, the following English, French and German sentences were processed by marker-based chunking using 50 markers. The underlined words are markers.

- [it is] [impossible to] [see why] [the resale right should] [be imposed on] [artists against their will] [as a form of] [copyright .]
- [on ne voit pas pourquoi] [le droit de] [suite doit être imposé comme une forme du] [droit d'] [auteur aux artistes , et] [ce contre leur volonté .]
- [es ist] [nicht einzusehen ,] [warum] [das folgerecht als ausformung des urheberrechts] [den künstlern gegen ihren willen aufgezwungen werden soll .]

3.2 Determining Markers by Informativity

Gough and Way [4] use marker-based chunking as a preprocessing step in SMT to improve the quality of translation tables and get improved results when combining their chunks with GIZA++/Moses translation table. They define a list of markers by hand and always cut left for European languages.

In contrast with their approach, we choose to automatically compute the list of markers. Frequency cannot do it: in the Europarl corpus "European" is a frequent word, but cannot be considered as a marker. We rely on some results from information theory and from our experimental results. In addition, to decide whether to cut to the left or the right of a marker, we compare the values of its branching entropy on both of its sides.

To determine which words are markers, we proceed as follows. If a language would be a perfect code, the length of each word would be a function of its number of occurrences, because, according to information theory, its emission length would be proportional to its self-information. The self-information of a word that appears $C(w)$ times in a corpus of N words is: $-\log(C(w)/N)$. In an ideal code, thus: $l(w) = -\log(C(w)/N)$ with $l(w)$ the length of the word, $C(w)$ its number of occurrences and N the total number of words in the text. Consequently, a word in a corpus of N words can be said to be informative if its length is much greater than its self-information in this text: $l(w) > -\log(C(w)/N)$. Consequently again, words with the smallest values for the following function can be said to be informative.

$$-\log\frac{C(w)}{N} \;\;/\;\; l(w) \tag{4}$$

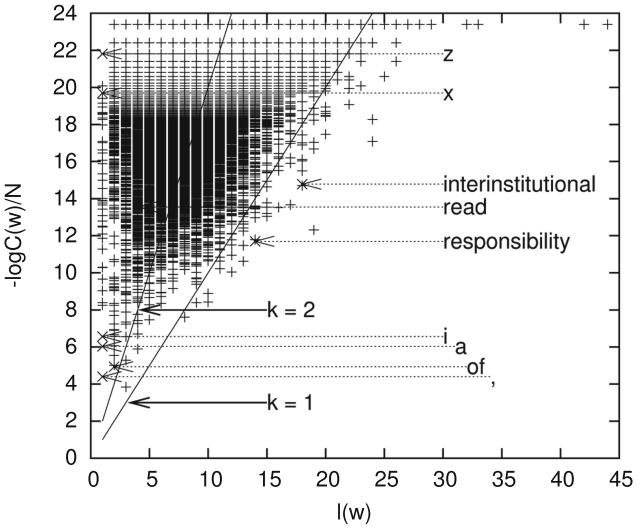
Conversely, markers, that is words that are not informative, should be the words with the largest values for the previous function. However, our experiments with this formula were deceptive. Rather, considering the absolute number of occurrences instead of the frequency delivers words that meet more the human intuition about linguistic markers. To summarize, the list of markers we use is the list of words with the smallest values for the following function:

$$-\log C(w) \;\;/\;\; l(w) \tag{5}$$

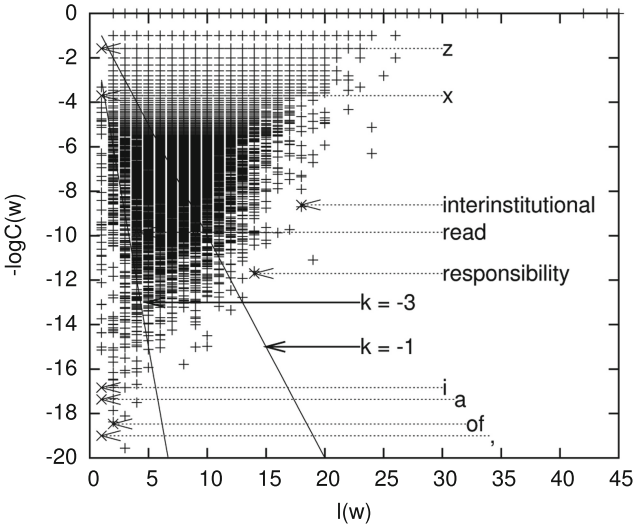
Table 1 shows markers obtained in accordance with the two proposed formulae. Those obtained with (5) are true markers, on the contrary to those obtained with (4). Figure 1(a) and (b) visualize the better efficiency of (5) over (4) to isolate words that correspond to the intuitive notion of a marker.

Table 1. Words ranked according to two different formulae. Formula (4) on the left, (5) on the right. For (5), the place where to cut is also given.

Rank	$-\log\frac{C(w)}{N} \;/\; l(w)$		$-\log C(w) \;/\; l(w)$		
	Word	Value	Word	Value	Cut
1	z	21.81	,	−19.00	right
2	/	21.08	.	−18.57	right
3	$	20.08	a	−17.37	left
4	q	19.94	i	−16.84	left
5	x	19.70	-	−15.15	left
6	l	19.40	s	−15.01	right
7	u	19.40	)	−14.37	right
8	w	19.15	(	−14.36	left
9	r	19.15	:	−13.74	right
10	&	19.15	’	−13.10	left
⋮	⋮	⋮	⋮	⋮	⋮



(a) Formula (4): $-\log \frac{C(w)}{N}$ against $l(w)$ for all words w. The lines stand for different values of $-\log \frac{C(w)}{N} \ / \ l(w)$. Words that correspond to the intuitive notion of a marker cannot be separated from other words using these lines.



(b) Formula (5): $-\log C(w)$ against $l(w)$ for all words w. Words that correspond to the intuitive notion of a marker are clustered at the bottom left part of the triangle of dots and can thus be easily isolated using lines that stand for different values of $-\log C(w) \ / \ l(w)$.

Fig. 1. Distribution of words using two different formulae.

3.3 Left or Right Cutting

Following the famous intuition by Harris [7] about branching entropy, Tanaka-Ishii [8], Jin and Tanaka-Ishii [9] or Magistry and Sagot [10] have shown how Japanese and Chinese can be segmented into words by formalizing the uncertainty at every position in a text using branching entropy.

The entropy of a random variable X with m outcomes x_i is defined as its mathematical expectation and is a measure of its overall uncertainty:

$$H(X) = -\sum_{i=1}^{m} p(x_i) \log p(x_i)$$

with $p(x_i)$ the probability of the outcome x_i.

The branching entropy at every position in a text is the entropy of the right context knowing the left context. Tanaka-Ishii [8] computes it as the entropy of the characters that may follow a given left context of n characters.

$$H(X|X_n = x_n) = -\sum_{x} p(x|x_n) \log p(x|x_n)$$

with x being all the different characters that follow the string x_n in a given text.

For each marker in a text, we determine on which side of the marker to cut, left or right, by comparing the branching entropy on its left and the branching entropy on its right. In opposition to Tanaka-Ishii [8], we compute branching entropies not in characters but in words. If the branching entropy on the left is greater than the one on the right, it means that there is more uncertainty on the left context of the marker, i.e., the connection of the marker to its left context is weaker. In other words, the marker is more tightly connected to its right context so that it should be grouped as a chunk with its right context, rather than its left context.

The rightmost column of Table 1 shows examples of which side to cut for different markers. In English, "(" is separated on the left while ")" is separated on the right, which is a felicitous results. On the whole, except for few mismatches, the segmentation that we obtained seems roughly acceptable.

4 Experimental Setting

We present similar experiments as the ones reported for Japanese in [11], but on 11 European languages. Here, we examine several sampling sizes and different numbers of markers. Our sampling sizes range from 10 to 100,000 sentences, and the number of markers ranges from 10 to 300 markers.

The data that we use in our experiments is the Europarl corpus [12] because our ultimate goal is to apply the analogy-based EBMT method to this kind of data. The Europarl corpus is a collection of proceedings of the European Parliament. The corpus comprises of about 10 million words for each of 11 official languages of the European Union: Danish (da), German (de), Greek (el), English

Table 2. Statistics on the Europarl corpus of 11 parallel European languages. The number of sentences is the same in all languages.

	da	de	el	en	es	fi	fr	it	nl	pt	sv
Sentences						384,237					
Words (million words)	10.4	10.5	10.0	10.9	11.5	7.9	12.1	10.9	11.0	11.3	9.9
Voc. (thousand words)	162.2	177.1	156.3	70.9	104.9	315.9	90.4	103.8	132.2	107.5	165.8

(en), Spanish (es), Finnish (fi), French (fr), Italian (it), Dutch (nl), Portuguese (pt) and Swedish (sv). Since the corpus is not exactly aligned, we aligned nearly 400,000 sentences across 11 languages properly. This gives about 13,000 words in each of the 11 languages for more than 380,000 utterances. Precise statistics are given in Table 2.

5 Experimental Results

5.1 Number of Different Chunks Obtained from Different Markers

By varying the number of markers, we measure how different markers affect the number of different chunks obtained. By doing so, it is possible to determine which markers are the most productive ones. Increasing the number of markers should increase the number of different chunks generated.

Figure 2(a) shows the number of different chunks obtained using different numbers of markers on 1,000 sentences in each different language. This graph shows that when the number of markers increases, the number of chunks may first increase and then decreases after some value.

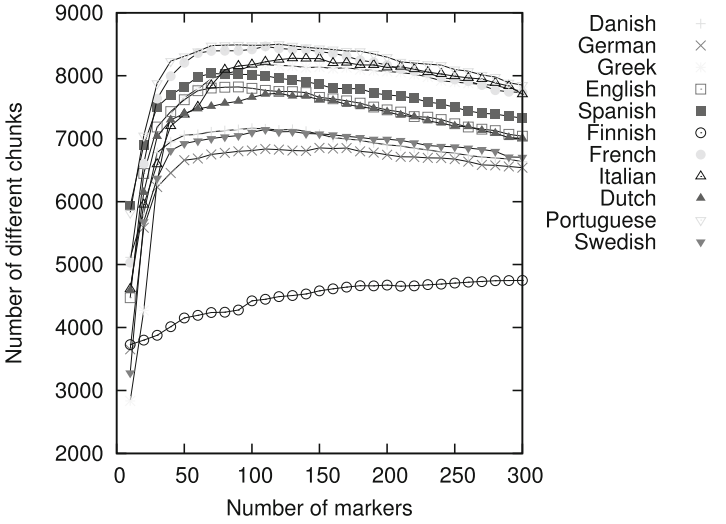
Figure 2(b) shows the number of different chunks obtained using different numbers of markers on 100,000 sentences. After 20 markers, the increase slows down for every language except for Finnish. The low number of different chunks for Finnish may be explained by the morphological richness of this language, and its relative lack in prepositions.

5.2 Number of Analogies Between Sentences and Chunks

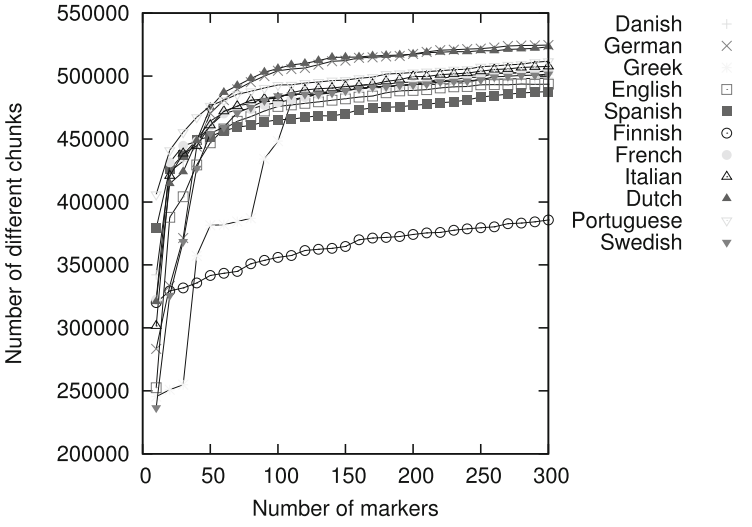
Figure 3(a) plots the number of proportional analogies between sentences for different numbers of sentences. Until 25,000 sentences, no analogies are found. After 50,000 sentences, the increase looks at least polynomial. The minimal number of proportional analogies is 159 for Greek for 100,000 sentences and the maximal number of proportional analogies is 698 for Danish. These absolute numbers show clearly that an EBMT system using proportional analogies between sentences will not be able to translate any sentence.

In comparison with Fig. 3(a) and (b) plots the number of proportional analogies between chunks extracted from 10 to 2,500 sentences using 100 markers.

In Fig. 3(b), chunks obtained from 100 sentences form very few analogies. After some 2,500 sentences, the number of proportional analogies found increases



(a) Number of different chunks against number of markers used for 1,000 sentences in 11 different languages. On the contrary to Fig. 2(b), except for Finnish, the number of different chunks obtained does not always increase.



(b) Number of different chunks against number of markers used for 100,000 sentences in 11 different languages. As expected, the more the markers, the more the number of different chunks obtained.

Fig. 2. Number of different chunks against number of markers used.

to more than 5,000 to 550,000 analogies with much variation. The minimal number of proportional analogies is 4,777 for Finnish. The maximum number is 548,928 for Spanish. It is important to note that in contrast to Fig. 3(a), not only the abscissae scale is different, but also the ordinates scale, different by two orders of magnitude in both graphs. The curve on Fig. 3(b) grows in fact ten thousand times faster than the one on Fig. 3(a).

5.3 Translation of Chunks by Analogy

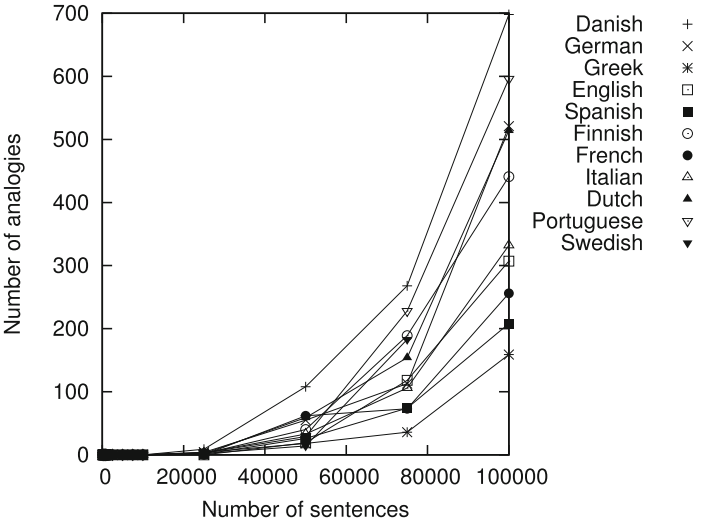
In this section, we perform an experiment in translation using an analogy-based machine translation system similar to the one reported in [1], and in conformity to the description given in Sect. 2. The experiment reported here is between French and English. For each language, we perform marker-based chunking with different numbers of markers. We determine the number of necessary markers in each language so as to obtain various average numbers of chunks per sentence, from three to nine.

In a first step, we perform word-to-word alignment between English and French on a training part of the Europarl corpus using the sampling-based subsentential alignment tool Anymalign [13]. In a second step, we compute a chunk-to-chunk translation table for each possible value of number of chunks per sentence, with lexical weights [14] and translation probabilities in both directions, by maximizing the lexical weights between chunks computed on the basis of the word-for-word alignments. As a final translation table, we use the merge of the two translation tables: word-to-word and chunk-to-chunk.

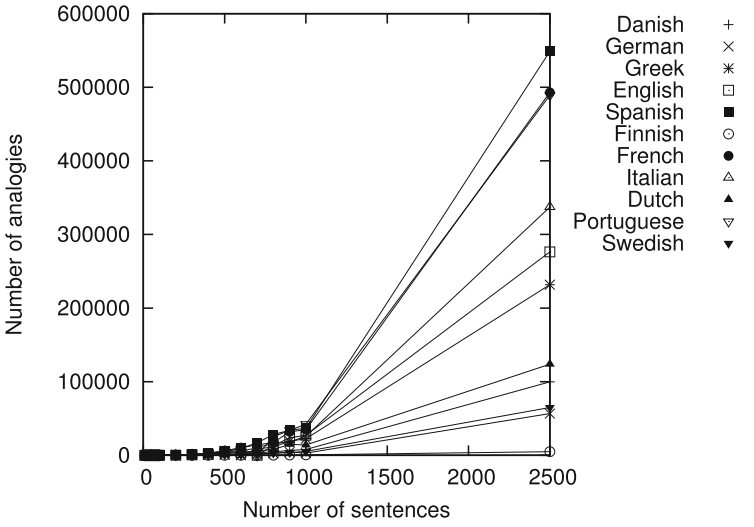
We translate each chunk of a test set using the analogy-based translation system fed with the previous translation tables. As we are dealing with chunks, which are shallow syntactic units, in this experiment, we do not use the recursion normally allowed in the analogy-based framework. For each chunk in the test set, there are three cases:

- the chunk cannot be translated;
- the chunk can be translated, but none of the translation hypotheses obtained correspond to a translation in the references;
- the chunk can be translated, and at least one of the translation hypotheses matches exactly one of the references.

Tables 3(a) and (b) give the percentages corresponding to the two last cases in different configurations that correspond to each different average number of chunks per sentence. As this number increases, the number of chunks that can be translated increases strongly, from 60 % to 80 % in English to French and from 56 % to 76 % in French to English. As for the number of chunks that could be translated and that have at least one perfect match in the references, this number is just below half of the chunks, varying from 40 % to 45 % in English to French and from 41 % to 46 % in French to English.



(a) Number of proportional analogies between sentences obtained with an increasing number of sentences.



(b) Number of proportional analogies between chunks extracted from an increasing number of sentences using 100 markers. The horizontal axis extends from 0 to 2,500, i.e., forty times less than in Fig. 3(a). The vertical axis reaches 600,000, i.e., almost thousand times more than in Fig. 3(a).

Fig. 3. Number of proportional analogies between sentences and chunks.

Table 3. Translation of chunks by analogy.

(a) English to French.

Average number of chunks per sentence	Number of chunks in test set	Number of translated chunks	Number of translated chunks with an exact match in the references
3	919	541 (58.86%)	369 (40.15%)
4	1,633	1,076 (65.89%)	660 (40.41%)
5	2,054	1,447 (70.44%)	836 (40.70%)
6	2,739	2,065 (75.39%)	1,131 (41.29%)
7	3,922	3,035 (77.38%)	1,663 (42.40%)
8	5,624	4,464 (79.37%)	2,511 (44.64%)
9	7,387	5,932 (80.30%)	3,291 (44.55%)

(b) French to English.

Average number of chunks per sentence	Number of chunks in test set	Number of translated chunks	Number of translated chunks with an exact match in the references
3	924	513 (55.51%)	380 (41.12%)
4	1,678	1,095 (65.25%)	717 (42.72%)
5	2,017	1,376 (68.22%)	856 (42.43%)
6	2,659	1,929 (72.54%)	1,162 (43.70%)
7	4,015	2,994 (74.57%)	1,771 (44.10%)
8	5,699	4,309 (75.60%)	2,675 (46.93%)
9	7,192	5,451 (75.79%)	3,337 (46.39%)

6 Conclusion

It was understood from previous work [1] that analogies between reasonably long sentences are too scarce to help in translating longer sentences using an analogy-based machine translation system. In a series of previous experiments [11,15,16], it has been shown, on a small number of languages, that the number of analogies between chunks is quite high, so that segmentation into chunks could contribute to translation of longer sentences by analogy.

The contribution of this paper was to show that it is possible to segment into chunks in a completely automatic way while keeping a high number of analogies between chunks. We tested our proposal on a reasonable number of languages: 11 European languages. We obtained more than several tens of thousands of analogies between chunks extracted from only 2,500 sentences in each language in average, a number by far much higher than between sentences.

The completely automatic segmentation into chunks that we proposed in this paper relied on insights from information theory. First, we relied on the marker hypothesis and gave an automatic way of determining markers, that differs from the usual approach with stop words. Our method consists in extracting the words with a length closest to an ideal length in an ideal code. Second, we rely

on the change of entropy at the boundary between markers to determine whether markers start or end a chunk.

We also briefly reported a preliminary experiment in translation of chunks by analogy on the French-English language pair. Alignment of chunks was performed based on word-for-word alignment and maximization of lexical weights. Almost three quarters of the chunks in our test set could be translated. 40 % of the chunks had an exact match in the target part of the translation tables, showing that a direct application of analogy can predict between a third and a half of unknown chunks.

References

1. Lepage, Y., Denoual, E.: Purest ever example-based machine translation: detailed presentation and assessment. Mach. Transl. **19**(3), 251–282 (2005)
2. Lepage, Y.: Analogy and formal languages. Electron. Notes Theoret. Comput. Sci. **53**, 180–191 (2004)
3. Green, T.: The necessity of syntax markers: two experiments with artificial languages. J. Verbal Learn. Verbal Behav. **18**(4), 481–496 (1979)
4. Gough, N., Way, A.: Robust large-scale EBMT with marker-based segmentation. In: Proceedings of TMI-04, pp. 95–104 (2004)
5. Stroppa, N., Way, A.: MaTrEx: the DCU machine translation system for IWSLT 2006. In: Proceedings of the International Workshop on Spoken Language Translation, pp. 31–36 (2006)
6. Van Den Bosch, A., Stroppa, N., Way, A.: A memory-based classification approach to marker-based EBMT. In: Proceedings of the METIS-II Workshop on New Approaches to Machine Translation, Leuven, Belgium, pp. 63–72 (2007)
7. Harris, Z.: From phoneme to morpheme. Language **31**(2), 190–222 (1955)
8. Tanaka-Ishii, K.: Entropy as an indicator of context boundaries: an experiment using a web search engine. In: Dale, R., Wong, K.-F., Su, J., Kwong, O. (eds.) IJCNLP 2005. LNCS (LNAI), vol. 3651, pp. 93–105. Springer, Heidelberg (2005)
9. Jin, Z., Tanaka-Ishii, K.: Unsupervised segmentation of Chinese text by use of branching entropy. In: Proceedings of the COLING/ACL on Main Conference Poster Sessions, pp. 428–435. Association for Computational Linguistics (2006)
10. Magistry, P., Sagot, B.: Unsupervized word segmentation: the case for Mandarin chinese. In: Annual Meeting of the Association for Computational Linguistics (ACL 2012), Jeju, Korea, ACL, July 2012 (2012)
11. Lepage, Y., Migeot, J., Guillerm, E.: A measure of the number of true analogies between Chunks in Japanese. In: Vetulani, Z., Uszkoreit, H. (eds.) LTC 2007. LNCS, vol. 5603, pp. 154–164. Springer, Heidelberg (2009)
12. Koehn, P.: Europarl: A parallel corpus for statistical machine translation. In: Proceedings of MT Summit X, Phuket, Thailand, pp. 79–86 (2005)
13. Lardilleux, A., Lepage, Y.: A truly multilingual, high coverage, accurate, yet simple, subsentential alignment method. In: Proceedings of the Xth conference of the Association for Machine Translation in the Americas, Waikiki, Hawai'i, October 2008, pp. 125–132 (2008)
14. Koehn, P., Och, F., Marcu, D.: Statistical phrase-based translation. In: Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL), Edmonton, Alberta, pp. 127–133 (2003)

15. Lepage, Y., Migeot, J., Guillerm, E.: Analogies of form between chunks in Japanese are massive and far from being misleading. In: Proceedings of the 3rd Language and Technology Conference (LTC 2007), Poznań, Poland, October 2007, pp. 503–507 (2007)
16. Lepage, Y., Migeot, J., Guillerm, E.: A corpus study on the number of true proportional analogies between chunks in two typologically different languages. In: Proceedings of the seventh international Symposium on Natural Language Processing (SNLP 2007), Kasetsart University, Pattaya, Thailand, December 2007, pp. 117–122 (2007). ISBN:978-974-623-062-9