The Number of Proportional Analogies between Marker-based Chunks in 11 European Languages

Kota Takeya, Jing Sun and Yves Lepage
Graduate School of Information, Production and Systems, Waseda University
{kota-takeya@toki,cecily.sun@akane,yves.lepage@aoni}.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 is very high. Whereas samples of thousand English sentences from the Europarl corpus do not lead to any analogy between sentences, we obtain several tens of thousands of analogies between the chunks extracted from these sentences using 10 markers. These results are very promising for the EBMT system that we intend to build.

1 Introduction

The example-based approach [8] contrasts with the statistical approach [1] to machine translation as well as with rule-based approach in that it uses a bilingual corpus of aligned sentences as its main knowledge at run time. We aim at building an EBMT system based on proportional analogies. The method has been proposed in [6]. Let D = ビールを二杯下さい。 be a source sentence to be translated into one or more target sentences $\widehat{D}$. Let the bilingual corpus consists of four sentences with their translations:

紅茶が飲みたい。	↔	can i have a tea?
ビールが飲みたい。	↔	i'd like a beer.
紅茶を二杯下さい。	↔	can we have two teas?
ビールを下さい。	↔	can i have a beer?

The method forms all possible analogical equations in x with all possible pairs of sentences from the parallel corpus. Among them:

紅茶が飲みたい。 : ビールが飲みたい。 :: x : ビールを二杯下さい。

The solution of this analogical equation is x = 紅茶を二杯下さい。. As the pair of sentences 紅茶を二杯下さい。 ↔ can we have two teas? is already part of the parallel aligned corpus, an analogical equation can be formed in the target language:

can i have a tea? : i'd like a beer. :: can we have two teas? : $\widehat{D}$

Its solution is a candidate translation of the source sentence: $\widehat{D}$ = we'd like two beers.

For such an EBMT system to work well, the more numerous the proportional analogies, the better the translation outputs are expected to be. In order to increase the number of proportional analogies, we propose to cut sentences into chunks using different markers and examine the number of proportional analogies between them in 11 European languages.

The rest of the paper is organized as follows. Section 2 describes the basic notions used in the reported experiments. Section 3 presents the data for the experiments which are sample sentences from the Europarl corpus in 11 European languages. Section 4 describes the results of the experiments and analyzes the results. The conclusion is given in Section 5.

2 Basic notions: marker-based chunking and analogy

2.1 Frequent words as markers

We use the Marker Hypothesis for chunking. The Marker Hypothesis was first defined by Thomas Green [2] in 1979. A definition is stated in [9].

> It is a psycholinguistic constraint which posits that all languages are marked for surface syntax by a specific closed set of lexemes or morphemes which signify context.

We shall define the set of specific lexemes, or markers, as the most frequent words. In the following experiments, we use different numbers of markers to determine where to cut.

2.2 Marker-based chunking

Chunking is the process by which a sentence is divided into chunks. We use the method of chunking called marker-based chunking.

A chunk is a sequence of words delimited by markers, such as determiners (the), conjunctions (and, but, or), prepositions (in, from, to), possessive and personal pronouns (mine, you). A chunk is created at each occurrence of a marker word. To decide whether to cut to the left or the right of a marker, we compare the entropy values on both of its sides. In addition, a further constraint requires that each chunk must contain at least one non-marker word. This restriction is very important to create chunks. Without non-marker words, a chunk would become somehow meaningless.

The following examples of English, French and German sentences were processed by marker-based chunking. The underlined words are markers.

- [We wish you courage] [and accomplishment] [in the coming months .]
- [À condition] [de parvenir] [à un tel consensus] [, j '] [espère que Malte nous rejoindra au sein] [de l ' Union .]
- [Ich hoffe] [, daß der] [Rat von Feira beschließt] [, dieses Thema auf die] [Tagesordnung der] [Regierungskonferenz] [zu setzen .]

2.3 Proportional analogy

Proportional analogy [5] 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 $A : B :: C : D$. The following are proportional analogies between words (1), chunks (2) and sentences (3):

$$\text{relate} : \text{unrelated} :: \text{modulate} : \text{unmodulated} \quad (1)$$

$$\text{a key} : \text{the key} :: \text{a first visit} : \text{the first visit} \quad (2)$$

$$\begin{array}{l}\text{Do you like music?}\end{array} : \begin{array}{l}\text{Do you go to concerts?}\end{array} :: \begin{array}{l}\text{Do you like jazz music?}\end{array} : \begin{array}{l}\text{Do you go to jazz concerts?}\end{array} \quad (3)$$

From the programming point of view, the formalization reduces to the counting of number of symbol occurrences and the computation of edit distances [5]. Precisely:

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

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. As B and C may be exchanged in an analogy, the constraint on edit distance has also to be verified for $A : C :: B : D$, i.e., $\delta(A,C) = \delta(B,D)$. There is no need to verify the first constraint as, trivially, $|A|_a - |B|_a = |C|_a - |D|_a \Leftrightarrow |A|_a - |C|_a = |B|_a - |D|_a$.

Table 1: Statistics for sentence data and chunk data obtained from 1,000 sentences (some sentences may be repeated; 10 markers used).

	Sentences		Chunks with 10 markers		
	Number (≠)	Length (in words) Avg. ± Std. dev.	Number (Total)	(≠)	Length (in words) Avg. ± Std. dev.
Danish	994	78 ± 48	18,093	15,601	4 ± 2
German	985	77 ± 47	16,821	14,899	4 ± 2
Greek	987	76 ± 49	16,736	14,263	4 ± 3
English	995	79 ± 49	19,929	16,267	4 ± 2
Spanish	987	85 ± 54	22,208	17,981	4 ± 2
Finnish	992	58 ± 37	11,972	11,084	5 ± 2
French	987	92 ± 60	20,899	17,044	4 ± 2
Italian	972	79 ± 52	17,052	14,495	5 ± 3
Dutch	986	82 ± 54	19,743	16,775	4 ± 2
Portuguese	993	84 ± 52	21,005	17,136	4 ± 2
Swedish	985	70 ± 44	15,336	13,835	4 ± 2

3 Experimental data

We use the Europarl corpus [3]. It is a collection of proceedings of the European Parliament, from 1996 to 2009. Altogether, the corpus comprises of about 30 million words for each of 11 official languages of the European Union: Danish (da), German (de), Greek (el), English (en), Spanish (es), Finnish (fi), French (fr), Italian (it), Dutch (nl), Portuguese (pt) and Swedish (sv).

To characterize our data, we give statistics obtained for samples of 1,000 sentences in each language in Table 1. The sentences are quite long: almost 80 words in average in English. Finnish has the least number of words per sentence: less than 60 words in average, which is a quarter less than English. Although such figures are not given in Table 1, Finnish has much longer words than English.

Statistics are also calculated for the total number of chunks obtained with 10 markers in each different language, and for the corresponding number of different chunks. On average, a chunk is repeated roughly 1.2 times.

4 Experimental Results

We present similar experiments as the ones reported for Japanese in [7], 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 70 markers. Using different numbers can help to obtain more reliable results.

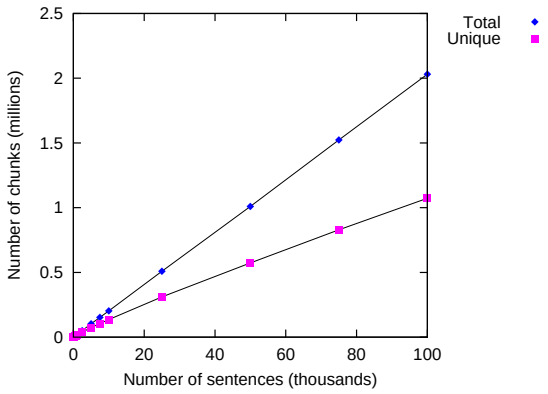

Figure 1: Total number of chunks (in ordinates) against number of different sentences (in abscissae) for twenty-seven different samplings in English using 10 markers. Naturally, the figures increase. Their relative increase looks linear.

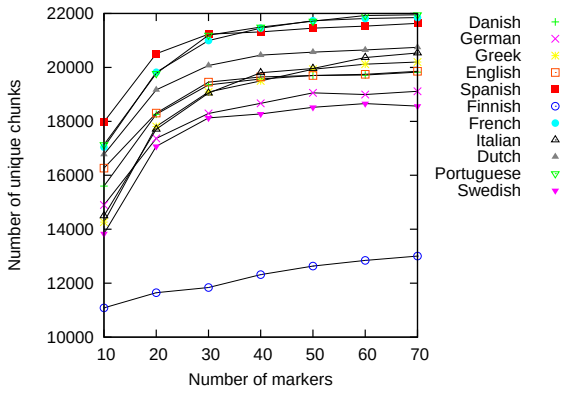

Figure 2: Number of unique chunks (in ordinates) against number of markers used (in abscissae) for 1,000 sentences in 11 different languages. As expected, the more the markers, the more the number of unique chunks obtained.

4.1 Total number of chunks and number of unique chunks

In English, on the whole 19,929 chunks were obtained for 1,000 sentences and 2,031,190 chunks were obtained for 100,000 sentences in total. The graph for English using 10 markers is given in Figure 1. The total number of chunks that we observed is of around 20 chunks in average for each sentence.

The most productive language is Spanish: 2,193,117 chunks were obtained for 100,000 sentences using 10 markers. Conversely, the least productive language is Finnish: 1,175,325 chunks were obtained in the same conditions.

4.2 Number of unique chunks obtained from different markers

By varying the number of markers, we measure how different markers affect the number of unique 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 unique chunks generated.

Figure 2 shows the number of unique chunks obtained using different numbers of markers on 1,000 sentences in each different language. After 20 markers, the increase slows down for every language except for Finnish. The low number of unique chunks for Finnish may be explained by the morphological richness of this language, and its relative lack in prepositions.

Figure 3 shows the number of chunks obtained using different numbers of markers on 100,000 sentences. This graph shows that when the number of markers increases, the number of chunks may decrease in some languages.

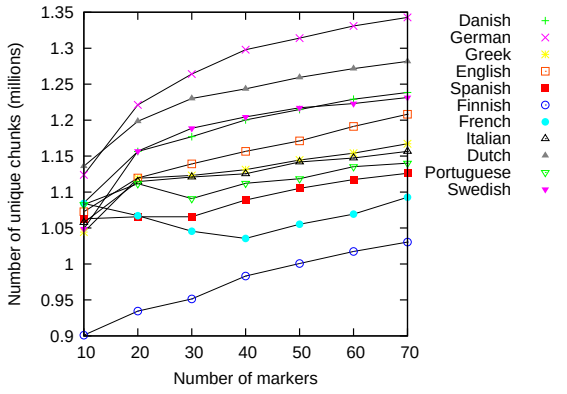

Figure 3: Number of unique chunks (in ordinates) against number of markers used (in abscissae) for 100,000 sentences in 11 different languages. On the contrary to Figure 2, in some languages, the number of unique chunks obtained does not always increase.

4.3 Number of proportional analogies between sentences and chunks

Figure 4 plots the number of proportional analogies between sentences for different numbers of sentences. Until 1,000 sentences, no analogies are found. After 25,000 sentences, the increase looks at least polynomial. The minimal number of proportional analogies is 3,895 for Spanish for 100,000 sentences and the maximal number of proportional analogies is 7,919 for German.

In comparison with Figure 4, Figure 5 plots the number of proportional analogies between chunks extracted from 10 to 1,000 sentences using only 10 markers. Chunks obtained from 10 sentences form very few analogies. After some 1,000 sentences, the number of analogies found increase to more than

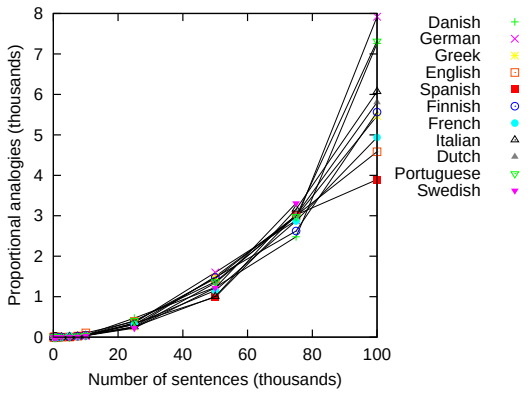

Figure 4: Number of analogies (in ordinates) between sentences obtained with an increasing number of sentences (in abscissae).

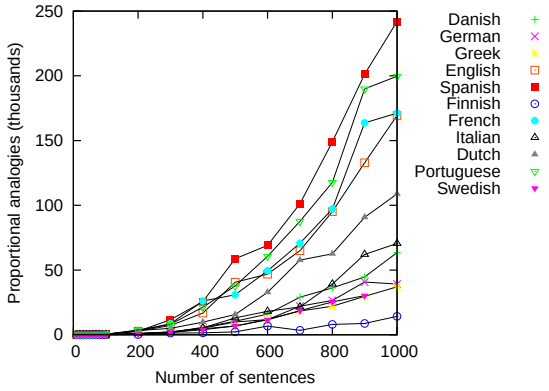

Figure 5: Number of analogies (in ordinates) between chunks extracted from an increasing number of sentences (in abscissae). Caution: the ordinates scale is two orders of magnitude that of Figure 4. The abscissae scale is also different.

15,000 to 250,000 analogies with much variation. The minimal number of proportional analogies is 14,215 for Finnish. The maximum number of proportional analogies is 241,892 for Spanish. It is important to note that in contrast to Figure 4 not only the abscissae scale is different, but also the ordinates scale, different by two orders of magnitude in both graphs. The curve on Figure 5 grows in fact ten thousand times faster than the one on Figure 4.

5 Conclusion

The experiments reported in this paper are conclusive for our goal of building an EBMT system based on analogy. As expected, the number of proportional analogies between chunks is higher than between sentences. Beyond expectation, this number is much higher. We obtained more than several tens of thousands of analogies for only 1,000 sentences in each language in average, however with much variation.

Future research should address the following problems.

- Propose a method to align chunks. A natural way to do so is to use lexical weights as proposed by Koehn et al. [4].
- Design an algorithm to reorder the chunks after translation. This is tantamount to design a reordering model of chunks.

References

[1] P.F. Brown, V.J.D. Pietra, S.A.D. Pietra, and R.L. Mercer. The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311, 1993.

[2] TRG Green. The necessity of syntax markers: Two experiments with artificial languages. *Journal of Verbal Learning and Verbal Behavior*, 18(4):481–496, 1979.

[3] P. Koehn. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of MT Summit X*, pages 79–86, Phuket, Thailand, 2005.

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

[5] Y. Lepage. Analogy and formal languages. *Electronic notes in theoretical computer science*, 53:180–191, 2004.

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

[7] Y. Lepage, J. Migeot, and E. Guillerm. A Measure of the Number of True Analogies between Chunks in Japanese. *Human Language Technology. Challenges of the Information Society*, pages 154–164, 2009.

[8] M. Nagao. A framework of a mechanical translation between Japanese and English by analogy principle. *Artifical and Human Intelligence*, pages 173–180, 1984.

[9] A. Van Den Bosch, N. Stroppa, and A. Way. A memory-based classification approach to marker-based EBMT. In *Proceedings of the METIS-II Workshop on New Approaches to Machine Translation*, pages 63–72, Leuven, Belgium, 2007.