Fully-Automatic Marker-based Chunking in 11 European Languages and Counts of the Number of Analogies between Chunks ⋆

Kota Takeya and Yves Lepage

IPS, Waseda University
Kitakyushu, Fukuoka 808-0135, Japan
{kota-takeya@toki, yves.lepage@aoni}.waseda.jp

Abstract. Analogy has been proposed as a possible principle for example-based machine translation. For such a framework to work properly, the training data should contain a large number of analogies between sentences. Consequently, such a framework can only work properly with short and repetitive sentences. To handle longer and more varied sentences, cutting the sentences into chunks could be a solution if the number of analogies between chunks is confirmed to be large. This paper thus reports counts of number of analogies using different numbers of chunk markers in 11 European languages. These experiments confirm that the number of analogies between chunks is very large: several tens of thousands of analogies between chunks extracted from sentences among which only very few analogies, if not none, were found.

Keywords: Analogy, Marker Hypothesis, Marker-based Chunking, Branching Entropy

1 Introduction

The example-based approach (Nagao, 1984) contrasts with the statistical approach (Brown *et al.*, 1990; Brown *et al.*, 1993) to machine translation 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.

A translation method based on proportional analogies has been proposed by Lepage and Denoual (2005b). 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, German and French. Let x = "ein großes programm und" be a source chunk to be translated into one or more target chunks $\widehat{x}$. 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 method forms all possible analogies with all possible triples of chunks from the parallel corpus. Among them:

ernste programme : ein ernstes programm :: große programme und : ein großes programm und = x

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

Taking into account the knowledge in the bilingual corpus, analogical equation can be formed in the target language:

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

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

For such an EBMT system to work well, the more numerous the proportional analogies, the better the translation outputs are expected to be. The method can work on small sentences like the ones in the BTEC corpus (Lepage and Denoual, 2005a), but cannot handle long sentences like the ones in the Europarl corpus (Koehn, 2005). For long sentences, translating chunk by chunk could be a solution. We have inspected the quality of translation of chunks obtained by marker-based chunking in English and French in both directions in (Takeya and Lepage, 2011a; Takeya and Lepage, 2011b). Our results have shown that more than three quarters of the chunks can be translated by the one-step analogy-based translation method, and that a little bit less than half of the chunks has at least one translation that matches exactly with one of the references. As the number of analogies is the crucial point, this paper inspects ways of counting sentences into chunks using different markers and examines 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 notion of marker-based chunking used in the reported experiments. Section 3 explains the notion of proportional analogy. Section 4 presents the data for the experiments which are sample sentences from the Europarl corpus in 11 European languages and the experimental protocol. Section 5 describes the results of the experiments and analyzes them. A conclusion is given in Section 6.

2 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.

2.1 The Marker Hypothesis

We use the marker hypothesis for this. This hypothesis was first laid by Green (1979).

> 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 (Gough and Way, 2004; Stroppa and Way, 2006; Van Den Bosch *et al.*, 2007). 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 .]

2.2 Determining Markers by Informativity

Gough and Way (2004) 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:

$$-\log \frac{C(w)}{N}$$

In an ideal code, thus:

$$l(w) = -\log \frac{C(w)}{N}$$

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:

$$l(w) > -\log \frac{C(w)}{N}$$

Consequently again, words with the smallest values for the following function can be said to be informative.

$$-\log \frac{C(w)}{N} \ / \ l(w) \tag{1}$$

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:

$$-\log C(w) \ / \ l(w) \tag{2}$$

Table 1 shows markers obtained in accordance with the two proposed formulae. Those obtained with the latter formula (2) are true markers, on the contrary to those obtained with the former formula (1).

Figure 1(a) and 1(b) visualize the better efficiency of formula (2) over formula formula (1) to isolate words that correspond to the intuitive notion of a marker.

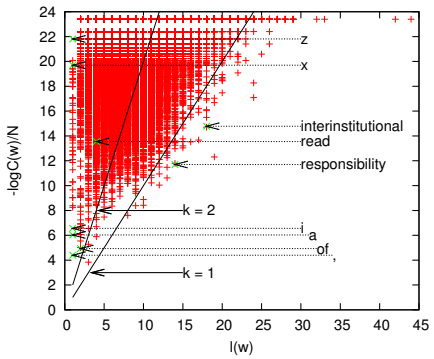
2.3 Left or Right Cutting

Following the famous intuition by Harris (1955) about branching entropy, Tanaka-Ishii (2005) and Jin and Tanaka-Ishii (2006) have shown how Japanese and Chinese can be segmented into words by formalizing the uncertainty using branching entropy at some point of a text.

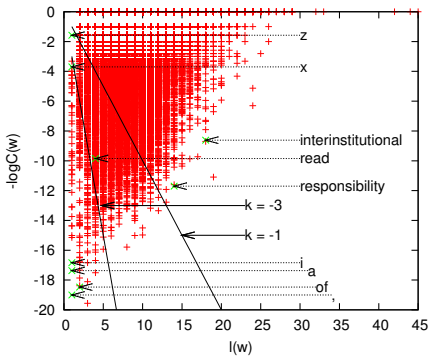
The entropy of a random variable X with m outcomes x_i is defined as its mathematical expectation and is a measure of its overall uncertainty:

Table 1: Words ranked according to two different formulae. Formula (1) on the left, (2) on the right.

Rank	$-\log \frac{C(w)}{N} \;/\; l(w)$		$-\log C(w) \;/\; l(w)$	
	Word	Value	Word	Value
1	z	21.81	,	-19.00
2	/	21.08	.	-18.57
3	$	20.08	a	-17.37
4	q	19.94	i	-16.84
5	x	19.70	-	-15.15
6	l	19.40	s	-15.01
7	u	19.40	)	-14.37
8	w	19.15	(	-14.36
9	r	19.15	:	-13.74
10	&	19.15	’	-13.10
11	o	18.94	;	-13.09
12	[	18.59	?	-12.12
13	h	18.59	1	-11.54
14	j	18.54	!	-11.30
15	n	18.35	2	-10.93
16	+	18.27	%	-10.89
17	f	17.97	5	-10.61
18	y	17.97	3	-10.61
19	d	17.09	4	-10.30
20	k	17.09	6	-9.97
⋮	⋮		⋮	



(a) $-\log \frac{C(w)}{N}$ against $l(w)$ for all words w (formula (1)). 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) $-\log C(w)$ against $l(w)$ for all words w (formula (2)). 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 $-\log C(w) \;/\; l(w)$.

Figure 1: Distribution of words using two different formulae.

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

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

The branching entropy at some position in a text is the entropy of the right context knowing the left context. Tanaka-Ishii (2005) 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.

We determine on which side of a 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 (2005), we compute branching entropies in words not in characters. 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.

Table 2 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.

Table 2: List of markers used for English.

Rank	$-\log C(w) \,/\, l(w)$		
	Word	Value	Cut
1	,	-19.00	right
2	.	-18.57	right
3	a	-17.37	left
4	i	-16.84	left
5	-	-15.15	left
6	s	-15.01	right
7	)	-14.37	right
8	(	-14.36	left
9	:	-13.74	right
10	’	-13.10	left
11	;	-13.09	right
12	?	-12.12	right
13	1	-11.54	left
14	!	-11.30	right
15	2	-10.93	left
16	%	-10.89	right
17	5	-10.61	left
18	3	-10.61	left
19	4	-10.30	left
20	6	-9.97	left
⋮		⋮	

3 Proportional Analogy

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 $A : B :: C : D$. The following are proportional analogies between words (3), chunks (4) and sentences (5):

$$\text{possible} : \text{impossible} :: \text{partiality} : \text{impartiality} \quad (3)$$

$$\text{the book} : \text{an expensive book} :: \text{the first trip} : \text{an expensive first trip} \quad (4)$$

$$\text{I like music.} : \text{Do you go to lives?} :: \text{I like jazz music.} : \text{Do you go to jazz lives?} \quad (5)$$

A formalization has been proposed in (Lepage, 2004). This formalization reduces to the counting of number of symbol occurrences and the computation of edit distances. Precisely:

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

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, $|A|_a - |B|_a = |C|_a - |D|_a \Leftrightarrow |A|_a - |C|_a = |B|_a - |D|_a$.

4 Experimental Setting

4.1 Experimental Protocol

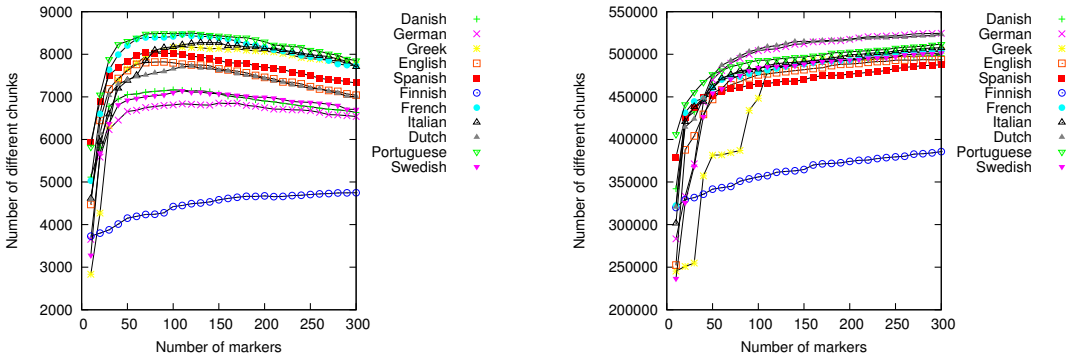
We present experiments similar to the ones reported for Japanese in (Lepage *et al.*, 2009), but on 11 European languages. Another similar experiment also has been done in (Takeya *et al.*, 2011). Here, we examine several sampling sizes and different numbers of markers. Our sampling sizes range from 10 to 100,000 sentences, and the number of markers ranges from 10 to 300 markers.

4.2 Experimental Data

We use the Europarl corpus (Koehn, 2005) in our experiments 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. Since the corpus is not exactly aligned, we aligned nearly 400,000 sentences across 11 languages properly. Our corpus comprises of about 10 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). Precise statistics are given in Table 3.

Table 3: Statistics of 11 European parallel aligned corpora.

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



(a) Number of different chunks against number of markers used for 1,000 sentences in 11 different languages. On the contrary to Figure 2(b), in most languages, 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.

Figure 2: Number of different chunks against number of markers used.

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 decrease in most languages.

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

Figure 3 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 numbers show clearly that an EBMT system using proportional analogies between sentences will not be able to translate any sentence.

5.3 Number of analogies between chunks

In comparison with Figure 3, Figure 4 plots the number of proportional analogies between chunks extracted from 10 to 2,500 sentences using different numbers of markers. Among Figures 4(a) to 4(d), the most productive one is 4(b). In Figure 4(b), chunks obtained from 100 sentences form very few analogies. After some 2,500 sentences, the number of proportional analogies found increases 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 of proportional analogies is 548,928 for Spanish. It is important to note that in contrast to Figure 3 not only the abscissae scale is different, but also the ordinates scale, which is different by three orders of magnitude. The curves on Figures 4 grow in fact more than thousand times faster than the ones on Figure 3.

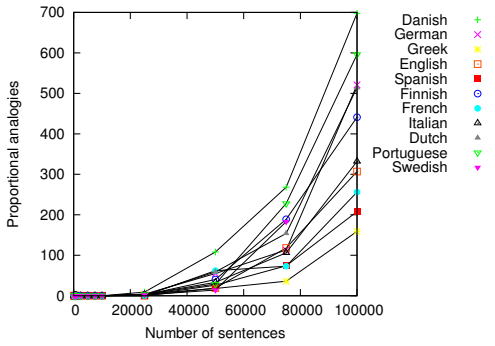
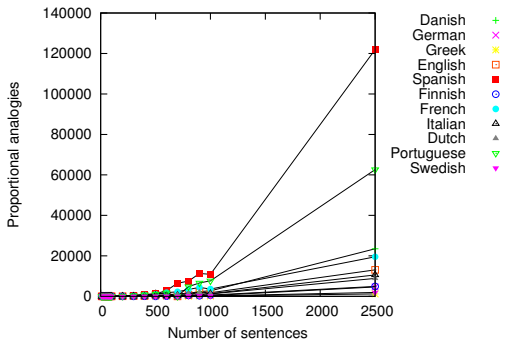
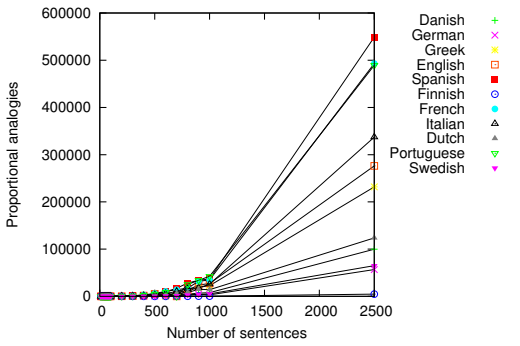


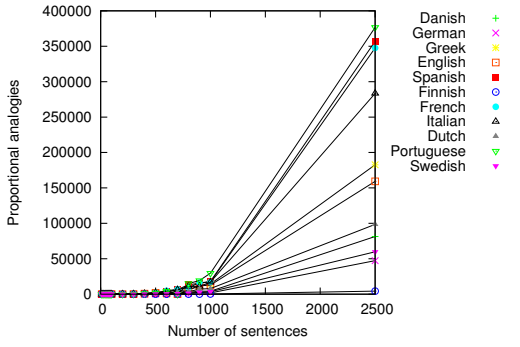
Figure 3: Number of proportional analogies between sentences obtained with an increasing number of sentences.



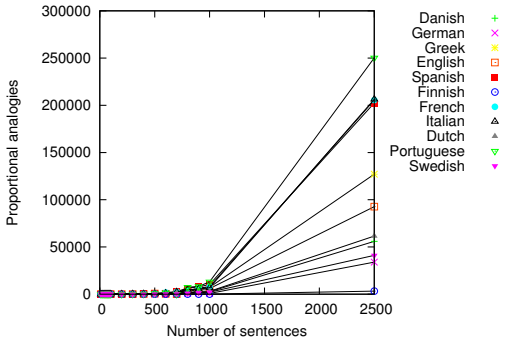
(a) Number of proportional analogies between chunks extracted from an increasing number of sentences using 10 markers.



(b) Number of proportional analogies between chunks extracted from an increasing number of sentences using 100 markers.



(c) Number of proportional analogies between chunks extracted from an increasing number of sentences using 200 markers.



(d) Number of proportional analogies between chunks extracted from an increasing number of sentences using 300 markers.

Figure 4: Number of proportional analogies between chunks extracted from 10 to 2,500 sentences using different numbers of markers. Caution: the ordinate axis scales vary

6 Conclusion

The experiments reported in this paper are conclusive for our goal of building an EBMT system based on proportional analogies: as expected, the number of proportional analogies between chunks is by far much higher than between sentences. Beyond expectation, this number is ways much higher. We obtained more than several tens of thousands of analogies between chunks extracted from only 2,500 sentences in each language in average.

In our goal of building an EBMT sytem, future research should address the following questions.

- Propose a method to align chunks. A natural way to do so is to use lexical weights as proposed by (Koehn *et al.*, 2003; Koehn, 2010).
- Design an algorithm to reorder the chunks after translation. This is tantamount to design a reordering model of chunks.

References

Brown, P.F., J. Cocke, S.A.D. Pietra, V.J.D. Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, and P.S. Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2), 79–85.

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

Gough, N. and A. Way. 2004. Robust large-scale EBMT with marker-based segmentation. In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-04)*, pp. 95–104.

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

Harris, Z.S. 1955. From phoneme to morpheme. *Language*, 31(2), 190–222.

Jin, Z. and K. Tanaka-Ishii. 2006. Unsupervised segmentation of Chinese text by use of branching entropy. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pp. 428–435.

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

Koehn, P. 2010. *Statistical Machine Translation*. Cambridge University Press.

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

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

Lepage, Y. and E. Denoual. 2005a. ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2005)*, pp. 47–54.

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

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

Nagao, M. 1984. A framework of a mechanical translation between Japanese and English by analogy principle. *Artifiical and Human Intelligence*, pp. 173–180.

Stroppa, N. and A. Way. 2006. MaTrEx: the DCU machine translation system for IWSLT 2006. In *Proceedings of the International Workshop on Spoken Language Translation*, pp. 31–36.

Takeya, K. and Y. Lepage. 2011a. Evaluation of Analogy-based Translation of Chunks obtained by Marker-based Chunking. In *Information Processing Society of Japan (IPSJ) SIG Technical Report 2011-NL-202*, pp. 1–7.

Takeya, K. and Y. Lepage. 2011b. Marker-based Chunking for Analogy-based Translation of Chunks. In *Proceedings of MT Summit XIII*, pp. 338–345.

Takeya, K., J. Sun, and Y. Lepage. 2011. The Number of Proportional Analogies between Marker-based Chunks in 11 European Languages. In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pp. 677–680, Toyohashi, Japan.

Tanaka-Ishii, K. 2005. Entropy as an indicator of context boundaries: An experiment using a web search engine. In *Proceedings of the Second International Joint Conference on Natural Language Processing*, pp. 93–105.

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