Marker-based Chunking for Analogy-based Translation of Chunks

Kota Takeya
IPS
Waseda University
Kitakyushu, Japan
kota-takeya@toki.waseda.jp

Yves Lepage
IPS
Waseda University
Kitakyushu, Japan
yves.lepage@aoni.waseda.jp

Abstract

An example-based machine translation (EBMT) system based on analogies requires numerous analogies between linguistic units to work properly. Consequently, long sentences cannot be handled directly in such a framework. In this paper, we inspect the quality of translation of chunks obtained by marker-based chunking in English and French in both directions. Our results show 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.

1 Introduction

Translation memories (TMs) are nowadays an indispensable tool for translators. First-generation TMs retrieve similar sentences from a database of already translated examples and provide the translator with the translation of the most similar sentence for minimal edition to obtain a relevant translation. Second-generation TMs improve over traditional TMs by chunking sentences into pieces and providing the translator with sub-sentential proposals. In this paper, we propose to make a step forward towards what could be called third-generation TMs: second-generation TMs with machine translation (MT) of sub-sentential parts.

Lepage and Denoual (2005) have proposed an EBMT system based on analogy. The method requires numerous analogies between linguistic units to work properly, and consequently, long sentences cannot be handled directly in the analogy-based framework. However, there may be possibilities for sub-sentential units. It is thus possible to envisage some convergence between this example-based approach to MT and second-generation TMs to lead to third-generation TMs.

In this paper, we will inspect the quality of translation of chunks, obtained by the Marker Hypothesis, using the analogy-based framework of translation. We report a series of experiments on 11 European languages and test the hypothesis that the analogy-based framework can fit to translate chunks in terms of translation quality.

The rest of the paper is divided into two main parts. The first part introduces the basic notions used. Section 2 describes the basic notion of marker-based chunking and the method used in the reported experiments. Section 3 presents two ways to score multilingual alignments by using two well-known scoring techniques. Section 4 explains the notion of analogy and how to translate using the analogy-based framework. The second part of the paper describes the experiments. Section 5 describes the data that we have used in the experiments and experimental protocol. The conclusion is given in Section 6.

2 Marker-based Chunking

Our goal is to segment different languages into sub-sentential units in a fully automatic way.

2.1 The Marker Hypothesis

Chunking is the process by which a sentence is divided into chunks. We use the Marker Hypothesis for chunking. This hypothesis was first laid by Green (1979). We do chunking based on this notion and use the method of chunking called marker-based chunking (Gough and Way, 2004; Stroppa and Way, 2006; Van Den Bosch et al., 2007).

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

In this framework, 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. 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 not become a sequence of words with a meaning.

The following examples of 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 (Brown et al., 1993) 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 that, 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.

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 (Shannon's theorem), 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. 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, words with the smallest values for the following function can be said to be informative.

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

Conversely, markers, that is words that are not informative, should be the words with the largest values for the previous function. Our experiments showed that considering the absolute number of occurrences rather than 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:

$$-\log C(w) \ / \ l(w)$$

Table 1 shows markers obtained in accordance with the above considerations. For example, the tokens with the smallest values of information are "," and "." in English, French and German. This is because these two tokens occur very frequently and are very short compared with other words.

2.3 Left or Right Cutting

We use the branching entropy to find out whether to cut on the left or on the right of a marker. 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.

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 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 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. 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 1 shows examples of which side to cut for different markers. In English and German, "(" is separated on the left while ")" is separated on the right, which is a felicitous results. For French, however, as ")" is separated on the left cut, the result of automatic computation is less accurate. This should be cut by the right side. On the whole, except for few mismatches, the segmentation that we obtained seem roughly acceptable.

Table 1: The first 20 marker words, selected as the first 20 least informative words.

Rank	English		French		German	
	Word	Cut	Word	Cut	Word	Cut
1	,	right	,	right	,	right
2	.	right	.	right	.	right
3	a	left	a	right	-	left
4	i	left	-	left	)	right
5	-	left	)	left	(	left
6	s	right	(	left	!	right
7	)	right	y	left	:	right
8	(	left	m	left	;	right
9	:	right	:	right	%	right
10	’	left	;	right	?	left
11	;	right	de	right	”	left
12	?	right	%	right	1	left
13	of	right	?	right	in	right
14	to	right	la	left	zu	right
15	in	right	!	right	2	left
16	1	left	l’	left	5	left
17	!	right	et	right	3	left
18	is	right	le	left	es	left
19	2	left	à	right	d	left
20	%	right	d’	right	h	left
⋮	⋮		⋮		⋮	

3 Lexical Weights for Chunks

3.1 Word Alignment

We use the sampling-based subsentential alignment tool Anymalign[1] (2009) for word alignment. Its main advantage compared to other state-of-the-art tools such as GIZA++ is that it can align any number of languages simultaneously. Secondly, it has been shown to outperform GIZA++ in lexicon extraction tasks (Lardilleux et al., 2010).

The translation probabilities for a multilingual alignment are computed as follows. Assume that an input corpus has L languages. A translation probability is computed for each language i ($1 \leq i \leq L$). s_i is the probability of the sequence of words that it can be computed by the rest of the alignment. $C(s_i)$ is the total count of all alignments that s_i appears and $C(s_1, \ldots, s_L)$ is the count of rest of the alignment that $C(s_1, \ldots, s_L)$ appear.

$$w(s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_L | s_i) = \frac{C(s_1, \ldots, s_L)}{C(s_i)}$$

[1] Anymalign is available at http://users.info.unicaen.fr/~alardill/anymalign/

3.2 Lexical Weights

To validate the quality of a chunk translation pair, we use a lexical weight proposed in (Koehn et al., 2003; Koehn, 2010). Based on the word-to-word translation probability, we can check how much reliable chunk translation pairs are.

Following equation is the definition stated by Koehn et al. (2003). Given a chunk pair $\bar{f}, \bar{e}$ and a word alignment a between the foreign word positions $i = 1, \ldots, I$ and the English word positions $j = 0, 1, \ldots, J$, the lexical weight lex can be computed according to the following formula:

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

As we do not have an alignment at our disposal in our method, we have changed the equation above a little bit. We compute the arithmetic mean for each word of the foreign language over all the English words:

$$lex(f_1^I|e_1^J) = \prod_{i=1}^{I} \left(\frac{1}{J} \sum_{j=1}^{J} w(f_i|e_j) \right)$$

4 The Analogy-based Framework of Translation

4.1 Analogy

In this work, we use the notion of analogy formed in (Lepage, 2004). Between strings of characters, an analogy $A : B :: C : D$ means that "A is to B as C is to D". Saussure (1955) (Part III, Chap. 5) applied on words, solving analogical equations as a typically synchronic operation by which, given two forms of a given word, and only one form of a second word, the fourth missing form is coined.

relate : unrelated :: modulate : $x \Rightarrow x =$ ummodulated

Lepage (1998) gives an efficient algorithm for the resolution of analogical equations. The algorithm is based on the following formalisation of analogies in terms of edit distances, or equivalently, in terms of similarity. From the programming point of view, the formalization reduces to the counting of number of symbol occurrences and the computation of edit distances.

We denote $d(A, B)$ as the distance between strings A and B. We also denote $|A|_a$ as the number of occurrences of character a in string A and $|A|$ as the length of A.

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

The following are examples of analogies in English between words (1), chunks (2) and sentences (3):

relate : unrelated :: modulate : unmodulated (1)

a key : the key :: a first trip : the first trip (2)

I like music. : Do you go to lives? :: I like jazz music. : Do you go to jazz lives? (3)

4.2 Translation by Analogy

A translation method based on analogy has been proposed by Lepage and Denoual (2005). The following 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:

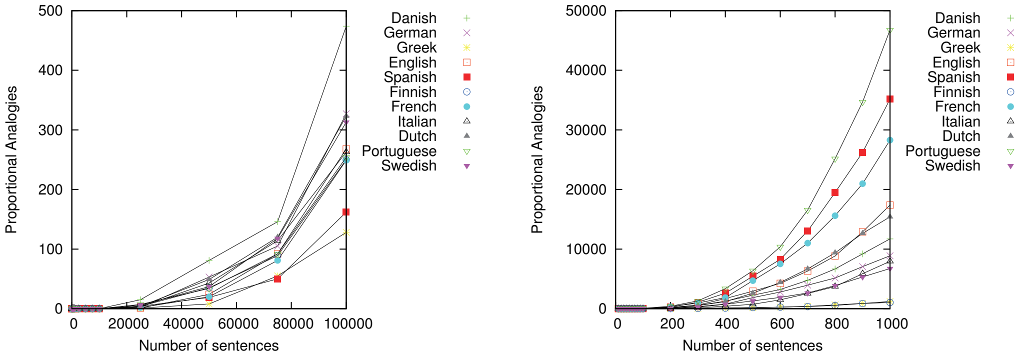
einfache programme	↔	programmes simples
ein einfaches programm	↔	un programme simples
große programme und	↔	gros programmes et
das ernste programm	↔	le programme sérieux

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

einfache programme : ein einfaches programm :: x : ein großes programm und

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

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



(a) Analogies between sentences.

(b) Analogies between chunks.

Figure 1: Number of analogies between sentences and chunks.

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

For such an EBMT system to work well, the more numerous the analogies, the better the translation outputs are expected to be. A similar experiment has been done in (Takeya et al., 2011). Figure 1(a) plots the number of analogies between sentences for different numbers of sentences. The maximum number of analogies is 474 for Danish for 100,000 sentences. In comparison with Figure 1(a), Figure 1(b) plots the number of analogies between chunks extracted from 10 to 1,000 sentences using 50 markers. After some 1,000 sentences, the number of analogies increases to more than 1,000 to 45,000 analogies ,however , with much variation.

5 Experiments

5.1 Experimental Data

We use the Europarl corpus (Koehn, 2005). It 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 (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 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.

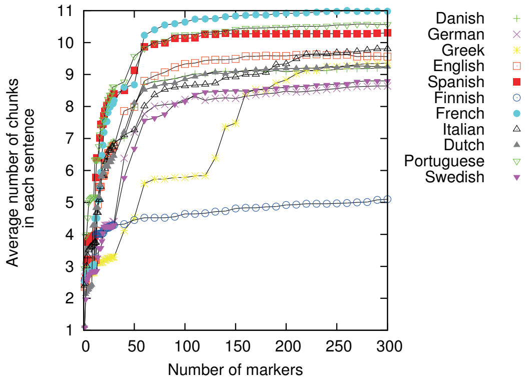


Figure 2: Average number of chunks per sentence for different numbers of markers in each of the 11 languages considered.

5.2 Experimental Protocol

5.2.1 Marker-based Chunking

For each language, we perform marker-based chunking with different numbers of markers and plot the graph in Figure 2. This graph allows us to determine the number of necessary markers in each language for a fixed average number of chunks per sentence equal in each language. Indeed, an equal average number of chunks per sentence in each language should a priori ensure a more stable correspondence between the chunks across languages. We determine these numbers of necessary markers for a range of average number of chunks per sentence from three to nine. These different numbers of markers in each different language allow us to chunk the texts using the corresponding number of markers. Chunking

Table 2: Statistics of 11 European parallel aligned corpora for training set and test set.

		da	de	el	en	es	fi	fr	it	nl	pt	sv
Train	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
Test	Sentences	500										
	Words	13.5k	13.6k	14.3k	14.2k	15.0k	10.1k	15.6k	14.3k	14.4k	14.8k	12.8k
	Voc.	3.3k	3.6k	4.1k	2.9k	3.4k	4.4k	3.3k	3.5k	3.2k	3.5k	3.4k

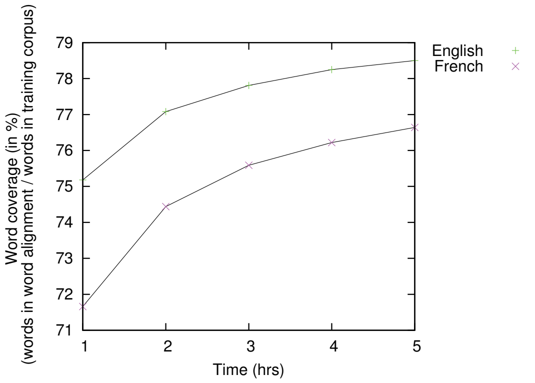


Figure 3: Word coverage from one to five hours. As expected the more time, the better word coverage.

has been performed on all the 11 languages of the Europarl corpus. However, in the sequel and the following experiments, we use only English and French as they are the pair of languages the most tested in machine translation.

5.2.2 Word Alignment and Chunk Alignment

To align the texts before translation of chunks, in a first step, we perform word-to-word alignment between English and French of the corpus using the sampling-based subsentential alignment tool Anymalign with the options `-n 1 -N 1` to limit the output to words only in the source and target languages. As Anymalign implements an any-time algorithm, it is possible to run it for different amounts of time. We run it for a range of one hour to five hours. The coverage of words in the vocabulary of the training data is shown in Figure 3. As should be expected, the more time, the better word coverage. Consequently, in our translation experiments, we use the word-to-word alignments obtained with the largest amount of time (five hours) which deliver the best coverage in words. This gives us a word-to-word translation table in French and English, with translation probabilities for each pair of words aligned.

Using the result of word-to-word alignment and the word translation probabilities delivered by Anymalign, for each aligned pair of sentences in French and English, we compute the lexical weights for all pairs of chunks. In addition, by inspection of the entire training corpus, we compute the translation probabilities for each pair of chunks appearing on the same line in the training corpus. As a result, we obtain a chunk-to-chunk translation table in French and English, with lexical weights and translation probabilities in both directions (from French to English and English to French).

The translation table that we use in the translation table is the merge of the two above-mentioned translation tables: word-to-word and chunk-to-chunk translation tables.

5.2.3 Chunk Translation and Evaluation

We translate each chunk of the testset using the analogy-based framework and the translation table obtained in the previous step. In this experiment, it is important to mention that we do not use the recursivity normally allowed in the analogy-based framework. This can be called a one-step analogy-based translation. For each chunk in the testset, 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.

Table 3 gives the percentage of sentences corresponding to the two last cases in different configura-

Table 3: Statistics of number of chunks in testset, number of translated chunks and number of translated chunks with at least one exact match in the references.

	en ⇒ fr			fr ⇒ en		
	Number of chunks in testset	Number of translated chunks	Number of translated chunks with an exact match in the references	Number of chunks in testset	Number of translated chunks	Number of translated chunks with an exact match in the references
3	919	541 (58.86%)	369 (40.15%)	924	513 (55.51%)	380 (41.12%)
4	1,633	1,076 (65.89%)	660 (40.41%)	1,678	1,095 (65.25%)	717 (42.72%)
5	2,054	1,447 (70.44%)	836 (40.70%)	2,017	1,376 (68.22%)	856 (42.43%)
6	2,739	2,065 (75.39%)	1,131 (41.29%)	2,659	1,929 (72.54%)	1,162 (43.70%)
7	3,922	3,035 (77.38%)	1,663 (42.40%)	4,015	2,994 (74.57%)	1,771 (44.10%)
8	5,624	4,464 (79.37%)	2,511 (44.64%)	5,699	4,309 (75.60%)	2,675 (46.93%)
9	7,387	5,932 (80.30%)	3,291 (44.55%)	7,192	5,451 (75.79%)	3,337 (46.39%)

tions that correspond to each different average number of chunks in each sentence. As the number of chunks 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.

6 Conclusion

The analogy-based framework of automatic translation has been shown not to be able to handle long sentences. A possible remedy is to split sentences into sub-sentential units like chunks. In the experiments reported in this paper, we use marker-based chunking to split sentences of the Europarl corpus in 11 languages and performed translation between French and English in both directions. We examined the quality of the translation of chunks obtained by marker-based chunking by checking the proportion of chunk translations with an exact match in the references. We inspected several values for average numbers of chunks in sentences using a range from 3 to 9 chunks in each sentence.

As a result, when using an average number of 9 chunks per sentence, the proportion of chunks that can be translated by the one-step analogy-based translation method reaches more than three quarters of the chunks (and even 80% in English to French). The number of chunks exactly translated is a little bit less than half, with an amount of around 45% in both directions. These results are promising to apply chunking as a first step in the framework of analogy-based translation, and should be tested for all the pairs with the 11 European languages available with the Europarl corpus. The possibility of concatenating translation of chunks to translate longer sentences and the need for some rewriting should also be explored.

Acknowledgments

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

References

P.F. Brown, 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.

F. de Saussure. 1955. *Cours de linguistique générale [A course in general linguistics]*. Payot, Lausanne, Switzerland.

N. Gough 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)*, pages 95–104.

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

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

Z. Jin 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*, pages 428–435.

P. Koehn, 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)*, pages 127–133, Edmonton, Alberta.

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

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

A. Lardilleux and Y. Lepage. 2009. Sampling-based multilingual alignment. In *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, pages 214–218, Borovets, Bulgaria.

A. Lardilleux, J. Gosme, and Y. Lepage. 2010. Bilingual lexicon induction: Effortless evaluation of word alignment tools and production of resources for improbable language pairs. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 252–256, Valletta, Malta.

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

Y. Lepage. 1998. Solving analogies on words: an algorithm. In *Proceedings of the 17th international conference on Computational linguistics*, pages 728–734.

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

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

K. Takeya, 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*, pages 677–680, Toyohashi, Japan.

K. Tanaka-Ishii. 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*, pages 93–105.

A. Van Den Bosch, 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*, pages 63–72, Leuven, Belgium.