The contribution of the notion of hapax legomena to word alignment

Adrien Lardilleux, Yves Lepage

The contribution of the notion of *hapax legomena* to word alignement

Adrien Lardilleux and Yves Lepage

GREYC, University of Caen,
BP 5186, Caen Cedex, France
Adrien.Lardilleux@etu.info.unicaen.fr, Yves.Lepage@info.unicaen.fr

Abstract

Current techniques in word alignment disregard words with a low frequency because they would not be useful. Against this belief, this paper shows that, in particular, the notion of *hapax legomena* may contribute to word alignment to a large extent. In an experiment, we show that pairs of *corpus hapaxes* contribute the majority of the best word alignments. In addition, we show that the notion of *sentence hapax* justifies a practical and common simplification of a standard alignment method.

1. Introduction

Alignment is an important task in natural language processing for a variety of purposes like the constitution of lexical resources, machine translation or cross-lingual information retrieval.

The purpose of this paper is to contribute to the domain of alignment between two languages. Our work focuses on the impact of *hapax legomena*[1]. Contrary to common knowledge in the field, we show that hapaxes contribute to a considerable extent to word alignment. In addition, we show that the distribution of words in our corpus logically leads to a justification of a practically fast implementation of a standard alignment method.

The paper is organised as follows. Section 2 briefly recalls main results about hapaxes in corpora. Section 3 introduces the cosine method to compute word alignments and gives some theoretical insights. Section 4 describes the data used in our experiments and details experimental results: hapaxes are useful for alignment as they contribute to up to 91% of the best word alignments. Section 5 shows how the notion of hapax in a sentence leads to an efficient simplification of the cosine method.

2. Hapaxes

2.1. Common negative attitude towards hapaxes

A hapax is a word that occurs only once in a single text or corpus. A general belief in the field holds that:

> *Hapax legomenon* and other so-called rare events present an interesting problem for corpus-based applications: due to their low frequency, they fail to provide enough statistical data for applications like word alignment or statistical machine translation. (Schrader, 2006)

As a matter of fact, by definition, hapaxes are discarded from the data in those approaches which filter out any word with a low frequency. This is usually the case in statistical machine translation or word alignment. For example, (Cromieres, 2006) sets a lower bound on frequencies to consider a word for alignement. (Giguet and Luquet, 2006) define a threshold proportional to the inverse term length.

In addition to their infrequency, a presumed drawback of hapaxes is that they often include newly coined words (neologisms) and misspelled words (Schrader, 2006). Neologisms should be considered words on their own right. As for misspelled words, their number depends mostly on the quality of the corpus. According to (Nishimoto, 2004), who interprets the results of (Evert and Lüdeling, 2001), each error in a corpus occurs only once in average. Misspelled words are thus typically hapaxes, but their proportion remains relatively low.

2.2. Positive aspects of hapaxes

Various experiments have been conducted so far on hapaxes. Hapaxes generally represent about 40% of words of a corpus. This number may vary according to the following aspects.

- The *richness of the vocabulary*: the proportion of hapaxes reflects the quantity of different words used in the text. Counts on Shakespeare's most read plays yield up to 58% hapaxes[2].
- The *degree of synthesis* of the language: isolating, synthetic or polysynthetic. The more synthetic a language, the more inflected words, and consequently, the more different words. The proportion of hapaxes increases accordingly. On a corpus of Inuktikut, a highly synthetic language of Canada, (Langlais et al., 2005) report more than 80% of hapaxes. In such a case, rejecting hapaxes is tantamount to consider only 20% of the data, which may obviously hinder the relevance of any subsequent processing.

In addition to account for a large proportion of word tokens, the relation of hapaxes to unknown words has already been demonstrated (Baayen and Sproat, 1996), (Cartoni, 2006). This aspect makes them useful to estimate the behaviour of unseen words, in machine translation for example.

[1] From the Greek *hapax legomenon* "what has been uttered once". In the following we shall use the plural *hapaxes* for convenience.

[2] Word frequencies avalaible at Mt. Hararat High School web site: `http://www.mta75.org/ curriculum/-English/Shakes/index.html`.

In accordance to these facts, the purpose of this article is to show that hapaxes are useful in word alignment.

3. The cosine method

Ideally, word alignments consist of translation pairs (*source word*, *target word*). Practically, alignment methods deliver scores that reflect the probability of *target word* being an accurate translation of *source word*. Various methods, generally classified as heuristic or statistical, have been developped to compute word to word alignments, see a survey and assessments in (Och and Ney, 2003).

We propose to use the cosine method in order to calculate word alignments. It is a standard technique for the computation of similarities or distances between distributions. It has been widely used in various domains, from Named Entity discovery (Shinyama and Sekine, 2004) to conceptual vectors for semantic tasks (Lafourcade and Boitet, 2002), (Turney and Littman, 2005).

In the case of word alignment between two parallel corpora, the basic steps are the following:

- start with an aligned bicorpus of n lines, *i.e.*, n sentences in a source language with their corresponding translations in a target language. Each line becomes a dimension in a vectorial field;
- for a given word w, build its associated vector $\vec{w}$ in this vectorial field by taking the number of occurrences of w on the *i*th line as the value of the *i*th component of vector $\vec{w}$;
- do this for each word in the source language and each word in the target language;
- for a given pair of words w_s and w_t in the source and the target languages, compute the angle between their associated vectors $\vec{w_s}$ and $\vec{w_t}$:

$$\widehat{(\vec{w_s}, \vec{w_t})} = acos\left(\frac{\vec{w_s} \cdot \vec{w_t}}{\parallel \vec{w_s} \parallel \times \parallel \vec{w_t} \parallel}\right) \qquad (1)$$

  where $\vec{u} \cdot \vec{v}$ is the scalar product of vectors $\vec{u}$ and $\vec{v}$ and $\parallel \vec{v} \parallel$ is the norm of vector $\vec{v}$. This value is the score of the alignment (w_s, w_t)

Since all components of $\vec{w_s}$ and $\vec{w_t}$ are positive, the previous computation yields only positive values in the range of 0 to $\pi/2$.

Intuitively, we would like the cosine measure to correspond to the idea that the lesser the angle, the better the quality of the word alignment. In other words, we would like to interpret the cosine measure as a kind of translation distance. Thus, the word alignments in which we are interested are those with a measure close to zero.

Theoretically, an angle of 0 means that $\vec{w_s}$ and $\vec{w_t}$ are parallel, *i.e.*, the two words appear exclusively on the same lines and their number of occurences on these lines is proportional: $\vec{w_s} = \lambda \vec{w_t}$. In practice, on our data λ equals 1 most of the time. The desired interpretation of an angle of zero is that both words would be perfect translations of one another (lexical equivalence), but this may not be true all the time (see 4.2. for examples).

An angle of $\pi/2$ implies that the cosine equals 0. This happens when each component of the scalar product is zero, *i.e.*, when the intersection of the lines on which the source and target words appear together is empty. This happens almost all the time when we consider all possible pairs of words. Consequently, efficient implementations of the cosine method do not compute angles between all possible pairs of vectors. Instead, they restrict the computation to those pairs of vectors associated with words that appear at least once on the same line. This narrows vectors down to the dimensions that are relevant for the computation. Suffix arrays (Manber and Myers, 1993) are an efficient data structure to implement this, see also (Nagao and Mori, 1994), (Yamamoto and Church, 1996) and (Zhang and Vogel, 2005).

4. Hapaxes in word alignments

4.1. The data

We used the training corpus from the IWSLT 2005 machine translation evaluation campaign to conduct our experiments. It consists in 20,000 pairs of aligned short sentences in Japanese and English (average sentence length in English: 9.4 words). Japanese sentences are segmented into words using Chasen (Matsumoto, 2000). Sentences in the corpus are independent. A sample of aligned sentences is shown on Figure 2.

The proportion of hapaxes in this corpus is 49% for the Japanese part and 45% for the English part, figures which are conform to other figures reported in the litterature (see above, Section 2.2.).

4.2. Results with the cosine method

Tables 3 and 4 give samples of word alignments obtained using the cosine method.

The method can lead to apparently surprising results. *E.g.*, the alignment between 刺さ (‘*to sting*’) and *bee* illustrates the fact that even a “perfect” alignment (angle of zero) does not mean that the words are translations of each other. It only depicts the reciprocal presence of the two words on any line: in this corpus, bees are mentioned only when a patient has been stung.

The example above shows that word alignments with an angle of zero are just a safe subset of all possible word pairs. The purpose of the next section is to show that hapaxes contribute to this safe subset for a good part.

4.3. Distribution of alignments

The distribution of pairs of words according to their angle on our data is shown by the graph on Figure 1. Alignments are divided according to their scores into three populations.

The first and largest one (not plotted) is a set of pairs with angle $\pi/2$ (81,263,082 pairs). It consists only of pairs of words which never appear on the same line. Such alignments are to be rejected.

The second population starts around $\pi/4$ and extends to $\pi/2$. It consists of pairs of words which are a priori no translations because of their bad scores (see the last 8 lines of Table 3). For clarity, pairs with an angle greater than 1.5 and less than $\pi/2$ are not plotted (402,793 pairs).

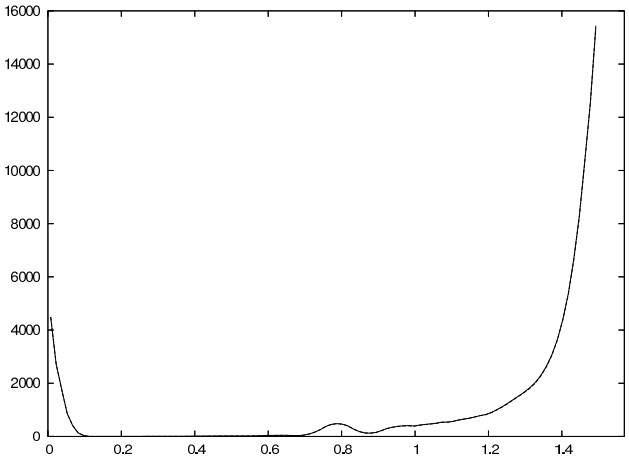

Figure 1: Smoothed distribution of pairs of words.

The third population is the pairs with angle around 0. There are 4,491 such pairs, 91% of which contain hapaxes in *both* languages with an angle of 0. Rejecting hapaxes would thus reduce this population to almost nothing: the efficiency of the cosine method in good candidates for translations pairs would be of less than 500 pairs for 20,000 sentences involving almost ten thousand words in each language!

Hapaxes are responsible for such a large proportion of the best alignments because, two *corpus hapaxes* appearing respectively in a source and a target sentence of a same line are, by definition, aligned with an angle of 0. For lack of place, we leave aside the discussion of the fact that alignments of *sequences of hapaxes* from a given line are also safe.

4.4. Distribution of hapaxes

As illustrated previously (Table 4), many alignments with an angle of 0, which are mainly obtained by hapaxes, are not valid word-to-word alignments (although they are part of valid translations of sequences of hapaxes). This happens because, when a sentence contains more than one *corpus hapax*, every source hapax gets aligned with every target hapax with an angle of 0. On the other hand, if there is only one source hapax and only one target hapax on a line, the resulting alignment is guaranteed to be correct. More generally, if there is only one source hapax and several target hapax on a line, (resp. one target hapax and several source hapax) it is reasonable to consider that the translation of the source hapax is the sequence of target hapaxes (resp. the translation of the target hapax is the sequence of source hapaxes). To study this phenomenon, we further inspect the distribution of hapaxes in our data.

The average hapax frequencies on the corpus is shown on Table 1. Although the number of hapaxes in the corpus is almost half of the words, they appear in only 20% of the English sentences (3,975) and 15% of the Japanese sentences (3,038). More importantly, the sentences containing a hapax generally do not contain more than one hapax: most of them (83% in Japanese and 84% in English) contain exactly one hapax, with an average of 1.23 hapaxes per sentence.

Among the alignments based only on hapaxes, 51% (2,094) are actually obtained by aligning a single hapax with a sequence of hapaxes (this produces as many alignments with an angle of 0 as there are hapaxes in the sequence), and 28% (1,154) are valid single-to-single hapax alignments.

Table 1: Japanese and English average hapax frequencies, computed on the sentences hapaxes appear in.

	Number of sentences containing a hapax	Sentences containing exactly 1 hapax	Avg. $\pm$ std. dev.
Japanese	3,975	3,288 (83%)	1.24 ± 0.67
English	3,038	2,567 (84%)	1.22 ± 0.63

Table 2: Japanese and English average word frequencies

	Number of words		Avg. $\pm$ std. dev.
	Total	words such that nbr. occ./sentences=1	
Japanese	9,982	9,540 (95.57%)	1.01 ± 0.07
English	8,191	7,766 (94.81%)	1.01 ± 0.08

Note that the 4,098 hapax-based alignments cover 2,552 Japanese words (26% of the total number of Japanese words) and 2,415 English words (29% of English words). Consequently, the 1,154 single-to-single alignments actually cover the vocabulary implied in hapax-based alignments up to 45% for Japanese and 48% for English. In other words, one can rely on these few alignments only (which are most certainly among the best alignments, see Table 5) to cover respectively 12% and 14% of the total vocabulary of our data.

5. A simplification of the cosine method

5.1. *Sentence hapaxes* vs. *corpus hapaxes*

We will now show that the distribution of words in sentences, and specifically, the notion of hapaxes in sentences, can lead to a simplification of the cosine method. To our knowledge, this common simplification is never justified.

In a first step, we determine how frequent a word is in the sentences it appears in, *i.e.*, we compute the total number of occurences of a word in the corpus divided by the number of sentences it appears in. Since a hapax in a corpus is necessarily a hapax on the sentence it appears in, the number of *sentence hapaxes* in a sentence is greater than or equal to the number of *corpus hapaxes*.

Table 2 summarizes the results. In average, 95% of the words have a frequency of 1 on the sentences they appear in. They include hapaxes in the corpus. On the whole, the average frequency is 1.01, very close to 1: almost all words are hapaxes in the sentences they appear in.

Consequently, instead of considering the real number of times a word appears in a sentence, counting its presence or absence should be enough. This reduces to see each word of the corpus as a *sentence hapax* (caution: not necessarily a *corpus hapax*). By doing so, the components in the vectors $\vec{w_s}$ and $\vec{w_t}$ now take their value in the set

$\{0, 1\}$, and equation (1) simplifies to:

$$(\widehat{\vec{w_s}, \vec{w_t}}) = acos\left(\frac{\mid S_s \cap S_t \mid}{\sqrt{\mid S_s \mid \times \mid S_t \mid}}\right) \quad (2)$$

where S_s is the set of lines in the source corpus on which $\vec{w_s}$ appears (same for S_s)[3].

The next section shows that this simplification does not alter the quality of the alignments obtained.

5.2. Comparison between the original and the simplified methods

5.2.1. Comparison in scores

In order to show that the previous simplification can be used as a substitute for the original method, we conducted a systematic comparison between the alignments obtained by the two methods, the original one serving as a baseline.

First, it is worth noticing that the simplification theoretically does not produce any new alignment within the two populations of alignments with an angle different from $\pi/2$ in comparison with the original method. This is verified in practice. Its effect reduces to a modification of the angle for pairs of words where one of the words is not a *sentence hapax* for any definite sentence (the angle of a pair (*corpus hapax, corpus hapax*) will not change).

Among the 499,480 alignment angles from the above mentioned populations, 26% (131,392) appear to be strictly identical, up to 10 decimals[4]. For the remaining 74%, the relative variation between the original angle and the one in the simplified method amounts to 0.27% (average) $\pm$ 1.52% (standard deviation)[5]. Clearly, the difference in angles between the original cosine method and the simplified version should not affect the quality of any subsequent processing task.

The distribution of pairs of words according to their angle, obtained with the simplified cosine method, is shown on Figure 1, along with the original graph. No difference is visible.

5.3. Comparison in runtime

The main advantage of the simplified method lies in its speed. It is much faster than the original one, because it is possible to use binary operations[6]. On several machines with different architectures, the observed speed-up was around 10 to 15 times (from the minute to few seconds on our data).

6. Conclusion

This paper addressed the impact of hapaxes on word alignment using the cosine method.

We first showed that *corpus hapaxes* contribute to up to 91% of the best alignments obtained by the cosine method. Such pairs of *corpus hapaxes* align more than 25% of the entire vocabulary on our data. These best alignments correspond to one of the three main populations of alignments we obtained by the cosine method, the two remaining and largest ones covering alignments that turn out to be no translation at all. In addition, word alignments with a mitigated score are almost inexistent. These results clearly demonstrate that the common attitude of rejecting hapaxes may lead to an important loss in efficiency.

We also showed that the notion of *sentence hapax* justifies in a logical manner a practical and common simplification of the cosine method. This simplification appears to be very reliable since it yields an average difference of only 0.20% in scores when compared with the original method, 26% of the scores remaining unchanged. An improvement of 10 to 15 times in speed is observed.

[3]Note that the argument of *acos* is different from the Jaccard coefficient: $\mid S_s \cap S_t \mid / \mid S_s \cup S_t \mid$.

[4]The smallest possible cosine is 1/400,000,000 in our experiments, thus having 10 decimals.

[5]0.20% $\pm$ 1.31% on the whole.

[6]An intersection is computed using a logical AND and the number of set bits in a machine-word can be computed in O(log$n$).

7. References

Baayen, Harald and Richard Sproat, 1996. Estimating lexical priors for low-frequency morphologically ambiguous forms. In *Computational Linguistics*, volume 22.

Cartoni, Bruno, 2006. Constance et variabilité de l'incomplétude lexicale. In *RECITAL 2006*. Leuven, Belgium: TALN 2006.

Cromieres, Fabien, 2006. Sub-sentential alignment using substring co-occurrence counts. In *Proceedings of the COLING/ACL 2006 Student Research Workshop*. Sydney, Australia: Association for Computational Linguistics.

Evert, Stefan and Anke Lüdeling, 2001. Measuring morphological productivity: Is automatic preprocessing sufficient? In *Proceedings of the Corpus Linguistics 2001 Conference*. Lancaster, UK.

Giguet, Emmanuel and Pierre-Sylvain Luquet, 2006. Multilingual lexical database generation from parallel texts in 20 european languages with endogenous resources. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*. Sydney, Australia: Association for Computational Linguistics.

Lafourcade, Mathieu and Christian Boitet, 2002. UNL lexical selection with conceptual vectors. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002)*. Las Palmas de Gran Canaria, Spain: ELRA.

Langlais, Philippe, Fabrizio Gotti, and Guihong Cao, 2005. NUKTI: English-Inuktitut word alignment system description. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*. Ann Arbor, Michigan: Association for Computational Linguistics.

Manber, Udi and Gene Myers, 1993. Suffix array: A new method for on-line string searches. In *SIAM Journal on Computing*, volume 22.

Matsumoto, Y., 2000. Japanese morphological analyser chasen. *Journal of Information Processing Society of Japan*, 41(11):1208–1214.

Nagao, Makoto and Shinsuke Mori, 1994. A new method of n-gram statistics for large number of n and automatic extraction of words and phrases from large text data of japanese. In *Coling 1994*, volume 1.

Nishimoto, Eiji, 2004. Defining new words in corpus data: Productivity of english suffixes in the british national corpus. Chicago: CogSci 2004.

Och, Franz Josef and Hermann Ney, 2003. A systematic comparison of various statistical alignment models. In *Computational Linguistics*, volume 29.

Schrader, Bettina, 2006. How does morphological complexity translate? A cross-linguistic case study for word alignment. Tübingen, Germany: International Conference on Linguistic Evidence.

Shinyama, Yusuke and Satoshi Sekine, 2004. Named entity discovery using comparable news articles. In *Proceedings of Coling 2004*. Geneva, Switzerland: COLING.

Turney, Peter D. and Michael L. Littman, 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60:1.

Yamamoto, Mikio and Kenneth Church, 1996. Using suffix arrays to compute term frequency and document frequency for all substrings in a corpus. In *Proceedings of the 6th Workshop on Very Large Corpora*. Computational Linguistics.

Zhang, Ying and Stephan Vogel, 2005. An efficient phrase-to-phrase alignment model for arbitrarily long phrase and large corpora. In *Proceedings of the Tenth Conference of the European Association for Machine Translation (EAMT-05)*. Budapest, Hungary: The European Association for Machine Translation.

Appendix: Sentence and word alignments

住所 と 名前 を 書い て ください 。	Please write down your name and address .
彼 は 来週 戻り ます 。	He will be back next week .
網棚 に お 乗せ し ましょ う 。	I 'll put it up on the rack for you .
今晩 の 席 は あり ます か 。	Are there any seats for this evening ?
もっと 安い 部屋 は あり ませ ん か 。	Do you have any cheaper rooms ?
ドーバー まで の 定期 券 を ください 。	May I have a commuter ticket for Dover ?
高級 な ところ が いい ん です が 。	I would like someplace fancy .
塩 コショウ を お 入れ し ます か 。	With salt and pepper ?
ナイフ や フォーク は どこ に あり ます か 。	Where can I get the knives and forks ?
東京 で 予約 し ました 。	I made a reservation from Tokyo .

Figure 2: Excerpt of the data used to conduct the experiment. Each line is a pair of aligned sentences.

Table 3: Examples of word pairs ordered by angles in increasing order, obtained by sampling. The sample meets intuition: the pair with a smaller angle is a better translation. The sample also reflects the distribution of the angles between 0 and $\pi/2 \approx 1.57$ (see Figure 1).

Japanese	meaning	freq.	English	freq.	angle $0 \sim \pi/2$
温泉	'hot springs'	5	springs	5	0.00
帽子	'hat'	12	hat	10	0.42
専門家	'specialist'	1	willing	7	1.18
ていき	'regular'	10	Susan	7	1.45
取っ	'take'	32	reserve	90	1.53
よい	'good'	145	life	13	1.55
良く	'well'	7	Yes	396	1.55
です	'is'	7,797	missing	19	1.55
今晩	'this evening'	62	fifty	90	1.56
れる	*verbal ending*	74	New	93	1.56

Table 4: Sample of pairs of words with angle 0.

Japanese	meaning	freq.	English	freq.
刺さ	'to sting'	2	bee	2
ブラジル	'Brazil'	2	Brazil	2
マナー	'manners'	2	manners	2
気候	'climate'	1	temperate	1
辛子	'mustard'	1	vinaigrette	1
旅行社	'travel agency'	1	bureau	1
不可	'impossible'	1	incomplete	1
壷	'pot', 'vase'	1	pitchers	1
衛星	'satellite'	1	satelliting	1
乗馬	'riding horse'	1	riding	1

Table 5: Sample of pairs of aligned "single hapaxes" (hapaxes from sentences with one hapax only). In this sample, only one alignment is wrong (*crocodile*).

Japanese	meaning	English
桃	'peach'	peach
風景画	'landscape painting'	landscape
ミサ	'mass', 'church service'	mass
宝くじ	'lottery'	lottery
アルバム	'album'	album
スイート	'suite'	suite
ユーフォー	'UFO'	UFO
ベーグル	'bagel'	bagels
ワッ	'ulp'	crocodile
エイズ	'AIDS'	AIDS