Unsupervised Morphological Segmentation Based on Segment Predictability and Word Segments Alignment

Delphine Bernhard
TIMC-IMAG
Institut de l'Ingénierie et de l'Information de Santé
Faculté de Médecine
F-38706 LA TRONCHE cedex
Delphine.Bernhard@imag.fr

Abstract

Word segments are relevant cues for the automatic acquisition of semantic relationships from morphologically related words. Indeed, morphemes are the smallest meaning-bearing units. We present an unsupervised method for the segmentation of words into sub-units devised for this objective. The system relies on segment predictability to discover a set of prefixes and suffixes and performs word segments alignment to detect morpheme boundaries.

1 Introduction

Morphemes are defined as the minimal meaning bearing units. Knowledge of morphologically related terms is thus worthy for many applications. This is especially true for morphologically complex languages like German or Finnish or scientific and technical vocabulary like biomedical language. Some research has for instance been devoted to the use of morphological decomposition for text indexing and retrieval in the biomedical domain (Schulz et al., 2002) or the acquisition of semantic relationships from morphologically related words (Zweigenbaum and Grabar, 2000; Namer and Zweigenbaum, 2004; Claveau and L'Homme, 2005). Work on the system presented in this paper has been undertaken with the objective of retrieving semantic relationships from morphologically related words (i.e. words sharing the same stem) in technical and scientific domains. Contrary to Schulz et al. (2002) or Namer and Zweigenbaum (2004) we have not built a morphological analyser relying on a dictionary of affixes and stems. Rather, morphological structure is discovered from a raw list of words and the method is not dependent on a given language, nor on a given domain. Related work on morphology induction is discussed in section 2. Our method is detailed in section 3. Finally in section 4 we present the results obtained.

2 Related work

2.1 Methods relying on segment predictability

Segment predictability is one possible cue for word segmentation. Harris (1955) proposes to use the number of different phonemes following a given phoneme sequence: morpheme boundaries are identified when this number reaches a peak. This method has been extended to written texts by Hafer and Weiss (1974) and Déjean (1998). Similarly, Saffran et al. (1996) suggest that learners use drops in the transitional probabilities between syllables to identify word boundaries. Like Déjean (1998) we use segment predictability to identify prefixes and suffixes. However, rather than determining segment boundaries by counting the number of letters following a given substring, as suggested in (Harris, 1955), we have developed a variant of this method based on transitional probabilities, following the proposition made by Saffran et al. (1996) (see Section 3.1).

2.2 Strategies based on word comparison

Other methods for the identification of morphologically related words are based on word comparison to identify similar and dissimilar parts in words. Neuvel and Fulop (2002) perform alignments starting either on the left or right edge of words to discover similarities and differences between the words compared. These similarities and differences correspond to word-formation strategies which can be used to generate new words without resorting to the notion of morpheme. Similarly, Schone and Jurafsky (2001) insert words in a trie either in good or reverse order to easily discover places where words differ from one another. Substrings which repeatedly differentiate words are considered as potential affixes. These methods based on the identification of initial or final common substrings are fine for prefix of suffix discovery but insufficient for words formed by compounding. In order to overcome these shortcomings our system performs word comparisons which are not anchored on word boundaries but rather on a shared stem which can be found at any position in the word (see Section 3.3).

2.3 Methods based on optimisation

Paradigmatic series of morphemes are extracted by Goldsmith (2001) in the form of “signatures” which are sets of suffixes which appear with the same stem. The method makes use of minimum description length (MDL) analysis to measure how effectively the morphology encodes the corpus. MDL is used by Creutz and Lagus (2002) as well to split words for highly-inflecting and morphologically complex languages. Our method is not directly related to MDL-based methods though it heavily relies on word segment length and frequency. Zipf (1968, page 173) had already noticed that, as well as words, “the length of a morpheme tends to bear an inverse ratio to its relative frequency of occurrence”. If we draw a parallel between words and morphemes, stems, which bear more meaning than affixes, are long and not so frequent while affixes are frequent and short[1]. Length is also used in the probabilistic framework proposed by Creutz and Lagus (2004) where the stem-likeness of a segment is function of its length. We use these general properties in a differential framework, drawing upon Saussure’s theory that syntagmatically related elements (like morphemes contained in a word) are defined by the differences amongst them. So rather than focusing on absolute values, we rely on differences in length and frequency (1) to distinguish between stems and affixes: a stem is identified as the longest and less frequent segment and (2) to impose constraints on the segments identified within a word: affixes have to be shorter and more frequent than stems.

3 Description of the method

The aim of the method described is to segment words into labelled segments. We only consider concatenative morphology and assign the following categories to morphological segments: stem, prefix, suffix and linking element. The latter category is not used by methods described in the previous section, but we think it is linguistically motivated in the sense that classical syntagmatic definitions of prefixes and suffixes fail to encompass linking elements. Indeed, prefixes are found before stems, at the beginning of words and suffixes are found after stems, at the end of words; linking elements can never be found at word boundaries and are always preceded and followed either by a stem or by another affix. For instance, “-**o**-” in “hormon**o**therapy” is a linking element.

Moreover, similarly to Creutz and Lagus (2004) we use the syntagmatic definition of morphological categories to impose constraints on possible sequences of word segments. In the next sections, we detail the procedure used to learn word segments.

3.1 Extraction of prefixes and suffixes

The input of the system is a plain wordlist *L*. The method does not make use of word frequency. The first step of the segmentation procedure is the extraction of a preliminary set of prefixes *P* and suffixes *S*. These are acquired using a method based on transitional probabilities between substrings. Moreover, only the longest words are segmented, following the intuition that these are the words most likely to be affixed. Words are sorted in reverse length order and are segmented using the variations of the transition probability between all the substrings coalescing at any given position *k* in the word.

Let *w* be a word whose boundaries are explicitly marked by the # symbol; n is the length of *w* (boundary markers included); $s_{i,j}$ is a substring of *w* starting at position *i* and ending at position *j*. For each position in the word *k* with k in [1, ..., n-1] we compute the following function, which corresponds to the mean of the maximum transition probabilities for all substrings ending and beginning at position *k*:

$$f(k) = \frac{\sum_{i=0}^{k-1} \sum_{j=k+1}^{n} \max[p(s_{i,k}|s_{k,j}), p(s_{k,j}|s_{i,k})]}{k \times (n-k)}$$

Where the transitional probabilities $p(s_{i,k}|s_{k,j})$ and $p(s_{k,j}|s_{i,k})$ are estimated by:

$$p(s_{i,k}|s_{k,j}) = \frac{f(s_{i,j})}{f(s_{k,j})} \quad \text{and} \quad p(s_{k,j}|s_{i,k}) = \frac{f(s_{i,j})}{f(s_{i,k})}$$

The frequency of a substring is equal to the number of times it occurs in *L*.

This yields a profile of the variations of the transition probabilities for *w*. Local minima indicate potential segment boundaries. A local minimum is validated if its difference both with the preceding and following maximum is at least equal to a standard deviation of the values. Figure 1 depicts this profile for the word “ultracentrifugation”. Valid boundaries are indicated by a bold vertical line, which corresponds to the following word segmentation: *ultra* + *centrifug* + *ation*.

Once a word has been segmented, the longest and less frequent segment is identified as stem if it also appears at least twice in the lexicon and once at the beginning of a word. The substrings which directly precede and follow this stem in the wordlist are added to the lists *P* and *S* if they are shorter and more frequent than the stem. Moreover, we discard prefixes of length 1 since we have noticed that these lead to erroneous segmentations in further stages of the process.

It is not necessary to apply this process of affix acquisition to all words. Indeed, the number of new affixes acquired decreases as the number of segmented words augments. This procedure ends when for *N* running words less than half of the affixes learned do not already belong to the lists *P* and *S*. Table 1 lists the

[1] See also (Vergne, 2005) for a method to distinguish function and content words based on differences in length and frequency.

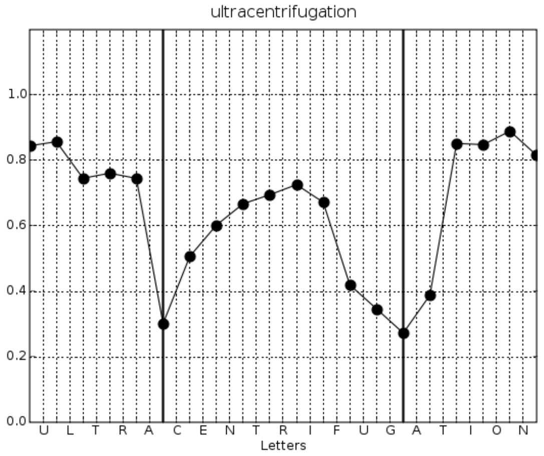

Figure 1: Profile for the variation of transitional probabilities for the word "ultracentrifugation"

most frequent prefixes and suffixes extracted from the MorphoChallenge English wordlist for N=5.

Prefixes		Suffixes	
in-	pre-	-s	-ly
un-	natur-	-e	-ble
inter-	counter-	-ed	-tion
dis-	over-	-al	-es
mis-	psycho-	-ally	-ately
re-	ultra-	-ing	-ity
ex-	hyper-	-ation	-l
pseudo-	con-	-ness	-ism

Table 1: Most frequent prefixes and suffixes extracted from the MorphoChallenge English wordlist for N=5.

3.2 Acquisition of stems

Stems are obtained by stripping off from each word in the list L all the possible combinations of the affixes previously acquired and the empty string. Of course this list is rather noisy. The following constraints are therefore applied on each extracted stem s:

1. it must have a minimum length of 3.
2. it can be followed by at least 2 different letters (including the word boundary marker); otherwise, this would mean that the stem is included in another stem.
3. it cannot contain a hyphen, since hyphens are boundary markers.
4. at least one word must begin with s.

3.3 Segmentation of words

Word segmentation is performed by comparing words containing the same stem b in order to find limits between shared and different segments (see Figure 2).

Segments thus obtained are assigned one of the three affix types (prefix, suffix, linking element) according to their position within the word, relatively to the stem. For instance, segments 'ist', 's' and 'y' in Figure 2 are labelled as suffixes. In order to deal with compounding, we make use of a temporary category for segments which contain another stem. These segments are labelled as 'potential stems'. This is the case for the segments 'organo' and 'physio' in Figure 2.

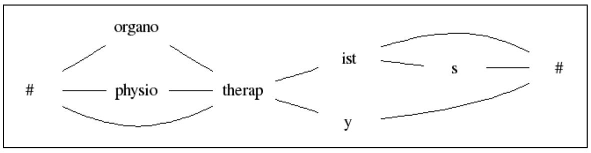

Figure 2: Example word segments alignment for the stem "therap".

As a result of the alignment new affixes which do not belong to the lists P and S may be discovered and these have to be validated. The validation procedure is somewhat similar to the validation of new morphemes in (Déjean, 1998) and is performed as follows: amongst aligned words which share the same stem we form subgroups of words beginning with the same segment. Table 2 lists word-final segments for the sub-group of the words containing the stem "hous" and starting with the empty string prefix.

Words	Suffixes from list S	Potential stems	New suffixes
housekeeping		-ekeeping	
housing	-ing		
household		-ehold	
house's			-e's
house	-e		
housed	-ed		

Table 2: Word final segments for words containing the stem "hous" and starting with the empty string prefix.

Let $|A_1|$ be the number of suffixes from list S, $|A_2|$ the number of potential stem segments and $|A_3|$ the number of new suffixes. For the examples in Table 2 $|A_1|$=3, $|A_2|$= 2 and $|A_3|$=1. New suffixes and potential stem segments are validated only if the following conditions are met:

$$\frac{|A_1| + |A_2|}{|A_1| + |A_2| + |A_3|} \geq a \text{ and } \frac{|A_1|}{|A_1| + |A_2|} \geq b$$

The same procedure is applied for the validation of word-initial segments.

Valid segmentations for each word are stored: we thus keep trace of all the segments proposed for a word, since a word may contain more than one stem and may therefore be aligned and segmented more than once. When all stems have been analysed for segmentation, we examine the segments stored for each word and remove potential stem segments. Potential stem segments are either replaced by other segments (as a whole or only partially) or assigned a final category (prefix, suffix or linking element) if no replacement is possible. Finally, we compute a frequency of occurrence for each segment. Frequency of occurrence is equal to the number of different words whose analysis includes the segment considered.

3.4 Selection of the best segments

For each word, we have stored the labelled segments resulting from its successive segmentations (one segmentation per stem). In order to choose the best possible segments, we perform a best-first search privileging the most frequent segment, given a choice. The final segmentation must also obey constraints related to word structure (at least one stem amongst the segments, a prefix cannot be directly followed by a suffix, only one running linking element between two prefixes or suffixes) and to the frequency of the segments relatively to one another (stems must be less frequent than the other types of segments). At the end of this stage, each word in the list L is segmented. Another output of this stage is the list of selected segments associated with their category (prefix, stem, suffix, linking element) and the number of times they have been selected (this corresponds to segment frequency). This list of segments can be used to segment any word in the same language, as explained in the next section.

3.5 Using the list of learned segments

Given the list of the best segments selected in the previous stage of the method, it is possible to segment any list of words. This stage is therefore optional and is proposed as a solution for the segmentation of words which do not belong to the list of words from which segments have been learned. The A* algorithm is used to find the best segmentation for each word. The global cost for a segmentation is the sum of the costs associated with each segment s_i. We have used two different segment cost functions for MorphoChallenge resulting in two different submissions:

$$cost_1(s_i) = -log \frac{f(s_i)}{\sum_i f(s_i)}$$

$$cost_2(s_i) = -log \frac{f(s_i)}{\max_i[f(s_i)]}$$

Moreover, the same constraints on possible successions of word segments as those described in section 3.4 are used.

4 Results

There are two main ways of directly assessing the quality of the results, either by evaluating the conflation sets built out of morphologically related words sharing an identical stem or by evaluating the position of the boundaries within a word. The latter is used by MorphoChallenge 2005.

4.1 Conflation-based evaluation

We have performed an evaluation of the results of the method on a list of words extracted from an English corpus on breast cancer. This corpus has been automatically built from the Internet and contains about 86,000 different word forms. We have manually built morphological word families for the top 5,000 keywords in the corpus. Keywords have been identified by comparison with a corpus on volcanology, using the method described in (Rayson and Garside, 2000). For instance, one of the manually built morphological families contains the words “brachytherapy”, “chemoradiotherapy”, “chemotherapeutic”, “therapies”, “therapist”, etc. We have used conflation-based evaluation, since we wish to assess the ability of the method to retrieve words linked both by form and by meaning, which is closer to our objective of retrieving semantic relationships between words thanks to morphology. Evaluation consists in counting the number of correct, incorrect and missing pairs of morphologically related words. Words are considered as morphologically related if they contain the same stem according to the method. For instance, the words “chemoradiotherapy” and “therapist” form a correct pair of words. Precision is defined as the number of correct word pairs divided by the number of suggested word pairs. Recall is defined as the number of correct word pairs divided by the number of word pairs in the list of manually built morphological families. For this evaluation, we used the segmentations provided directly after selection of the best segments (see Section 3.4) with N=5, a=0.8 and b=0.1. Results are given in Table 3. Precision suffers from the fact that most words ending with -logy, -logic or -logical share the same stem “log” according to the system. Results also evidence that recall should be improved. For instance, “artery” is segmented as arter + y while “arterial” is segmented as arteri + al. Both words are therefore not conflated in the same set.

	Number	Example
Correct word pairs	3,936	lymphedematous lymphoedema
Incorrect word pairs	2,359	additive addresses
Missing word pairs	5,210	therapeutics therapy

Precision	Recall	F-measure
62.5	43.0	51.0

Table 3: Results of conflation-based evaluation.

4.2 MorphoChallenge 2005 results

The MorphoChallenge 2005 datasets were considerably bigger than the dataset used for the previous assessment. Word segments learning has been performed on the whole English dataset. However, for Finnish and

				F-measure			
				Sample evaluation		Final evaluation	
Language	N	a	b	method 1	method 2	method 1	method 2
English	5	0.85	0.1	64.29	61.05	66.6	62.4
Finnish	5	0.8	0.1	63.18	64.44	63.3	64.7
Turkish	5	0.7	0.1	55.93	66.06	55.3	65.3

Table 4: Parameter values used and results obtained for the submissions to MorphoChallenge 2005.

Turkish learning has been performed only on a subset of the datasets (the 300,000 most frequent words), due primarily to heavy memory consumption. Three different parameter values have to be set: N (see Section 3.1), a and b (see Section 3.3). Parameter values used for each language were roughly the same, only we took those values which yielded the best results on the evaluation datasets. Yet keeping default values N=5, a=0.8 and b=0.1 does not bring a change of more than about 1 or 2% in F-measures. Table 4 details parameter values used and the results obtained. Method 1 corresponds to results obtained by using $cost_1$ and method 2 to results obtained by using $cost_2$ (see Section 3.5).

Results for method 2, using $cost_2$, indicate better recall but lower precision on all datasets. This is especially noticeable on the Turkish dataset. Recall for the Turkish dataset was indeed an issue which led to the use of $cost_2$. This might be due to the fact that segments in the Turkish gold standard sample are shorter on the average than Finnish and English gold standard sample segments.

5 Conclusions

Thanks to MorphoChallenge 2005 the method has been tested on new languages (Finnish and Turkish), bigger wordlists and for different objectives (speech recognition). Results show that the method performs well even on Finnish and Turkish. Planned improvements include better implementation to deal with large datasets and incorporation of equivalence matching between stems to capture orthographic variants like "canc**e**r" and "canc**é**r". In work in progress, we are investigating the usefulness of morphological segmentation for the automatic acquisition of semantic relationships.

References

Vincent Claveau and Marie-Claude L'Homme. 2005. Structuring Terminology by Analogy-Based Machine Learning. In *Proceedings of the 7th International Conference on Terminology and Knowledge Engineering, TKE'05*.

Mathias Creutz and Krista Lagus. 2002. Unsupervised Discovery of Morphemes. In *Proceedings of the Workshop on Morphological and Phonological Learning of ACL-02*, pages 21–30.

Mathias Creutz and Krista Lagus. 2004. Induction of a Simple Morphology for Highly-Inflecting Languages. In *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*, pages 43–51, Barcelona.

Hervé Déjean. 1998. Morphemes as Necessary Concept for Structures Discovery from Untagged Corpora. In D. Powers, editor, *Proceedings of the CoNLL98 Workshop on Paradigms and Grounding in Language Learning*, pages 295–298.

John Goldsmith. 2001. Unsupervised Learning of the Morphology of a Natural Language. *Computational Linguistics*, 27(2):153–198.

Margaret A. Hafer and Stephen F. Weiss. 1974. Word segmentation by letter successor varieties. *Information Storage and Retrieval*, 10:371–385.

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

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Proceedings of Medinfo. 2004*, volume 11, pages 535–539, San Francisco CA.

Sylvain Neuvel and Sean A. Fulop. 2002. Unsupervised Learning of Morphology Without Morphemes. In *Proceedings of the ACL Workshop on Morphological and Phonological Learning 2002*, pages 31–40.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing Corpora, held in conjunction with the 38th annual meeting of the Association for Computational Linguistics (ACL 2000)*, pages 1–6, Hong Kong, 1-8 October 2000.

Jenny R. Saffran, Elissa L. Newport, and Richard N. Aslin. 1996. Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*, 35(4):606–621.

Patrick Schone and Daniel Jurafsky. 2001. Knowledge-Free Induction of Inflectional Morphologies. In *Proceedings of the Second meeting of the North American Chapter of the Association for Computational Linguistics*, pages 1–9.

Stefan Schulz, Martin Honeck, and Udo Hahn. 2002. Biomedical Text Retrieval in Languages with a Complex Morphology. In *Proceedings of the ACL Workshop on Natural Language Processing in the Biomedical Domain*, pages 61–68, Philadelphia, July.

Jacques Vergne. 2005. Une méthode indépendante des langues pour indexer les documents de l'internet par extraction de termes de structure contrôlée. In *Actes de la Conférence Internationale sur le Document Électronique (CIDE 8)*, Beyrouth, Liban.

George Kingsley Zipf. 1968. *The Psycho-biology of Language. An Introduction to Dynamic Philology.* The M.I.T. Press, Cambridge, second paperback printing (first edition: 1935) edition.

Pierre Zweigenbaum and Natalia Grabar. 2000. Liens morphologiques et structuration de terminologie. In *Actes de IC 2000 : Ingénierie des Connaissances*, pages 325–334.