

Extracting Morphemes without

Toshikazu Ikuta
tikuta@indiana.edu

1 Introduction

In unsupervised morphology learning algorithms (Goldsmith, 2000) (Cavar et al, 2004), it is assumed that words are already segmented by white space. However, it is somewhat obvious that white space between words are not pronounced in spoken English. While word boundaries are indicated by white spaces in written English, speakers of English do not pronounce anything particular for space or they do not pose in between words.

If it is true that children are able to acquire morpheme/word segmentations, it means that they are able to learn segmentation from non-segmented input. Since children seem to be able to segment words even though the input to them are auditory and thus no segmentation involved, there must be some way in which word segmentation can be archived by subsequence occurrence information (as in the program in this paper) and/or suprasegmental or prosodic information. Although it may still be the case that information of word segmentation is suprasegmentally or prosodically conveyed in spoken English, the current paper tries to answer how much morphemes (and/or words) can be extracted from a text input without word segmentation (i.e. without space) only by analyzing frequencies of subsequences of unsegmented inputs. The ultimate goal of this project (although not yet achieved) is to provide word/morpheme segmentation from non-segmented input so that unsupervised learning algorithms can be executed on corpora without word segmentation.

2 Algorithm

The input to the program is a text corpus whose alphabets are all lowercased and whose all non-alphabetic characters including white space are eliminated.

The program first creates a set of n-grams in the range of $2 \leq n \leq 12$.¹ The occurrence of each

¹The maximum 12 is arbitrary selected.

element in n-grams is recorded (see Manning and Schutze (1999) for details of n-gram models).

After n-grams ($2 \leq n \leq 12$) are created, each element in n-grams is compared to every single element in other n-grams (call this m-grams) in the range of $n < m$. The score of an element is incremented if the entire sequence is a subsequence of a given element of m-gram. By denominating this by the occurrence as an n-gram (not as a subsequence), we will have the score for each element in n-grams. In short, a score of a sequence $seq_{n,i}$ which is the i-th element of an n-gram is;

$$Score_{n,i} = \sum_{m=n+1}^{12} \sum_{seq_{n,i} \in k} \frac{1}{occurrence_{n,i}}$$

By counting the score, we can estimate “how the subsequence independently occurs” or “how variable contexts for the subsequence are”. If the score is high, it means the subsequence is more likely to appear in various context. For example, compare “un” and “nh” both of which are subsequences of “unhappy”. On the one hand, the score of “un” is expected to be high, because the subsequence shows up in other words as in “unpack”, “unlikely”, “undesirable”, and so on because “un” is a morpheme. On the other hand, the score of “nh” is expected to be very low, since “nh” doesn’t seem to be a frequent subsequence in English. In general, it is expected that morphemes will have higher score than non-morphemes, since morphemes are what appear in various context.

3 Result

The program has been tested on randomly chosen 25KB part of the Brown Corpus. Subsequences are manually evaluated (whether or not morpheme). Among the top 40 high-score elements, 24 elements are manually evaluated to be morphemes (60% precision). Furthermore, 32.5% of the 40 high-score elements were functional category words which do not usually undergo morphological processes (in, he, an, the...).

4 Discussion and Near-Future Work

It seems that orthographically transcribed English corpus is not appropriate input for the purpose

and the program. Rather, phonetically transcribed corpora seem to be appropriate; The typical non-morpheme subsequences which the program rated to be high are long vowel representations such as “ou”, “ew” and “ea”. They would be transcribed as one segment in phonetically transcribed corpora. Also, “th” is scored high, although “th” usually represents one segment, that is, interdental fricatives. It is expected that the precision will be higher for phonetically transcribed corpora.

So far, 60% is the best answer to the question “How much can morphemes (and/or words) be extracted from a text input without word segmentation (i.e. without space) only by analyzing frequencies of subsequences?” Actually, the author originally expected much lower score and expected to have a conclusion that suprasegmental or prosodic cues are necessary for word/morpheme segmentation from no-spaced corpora. However, it turned out that morpheme segmentation seems to be much more possible. If much higher precision can be obtained from phonetically transcribed corpus input, it will answer to the question by saying that it is possible to extract morphemes from non segmented corpora. Otherwise, it is planned to implement another program that attests word segmentation only from suprasegmental or prosodic information.

Also, it is not surprising that functional category words obtained high scores with 32.5% precision. Since functional category words are frequent words (Manning and Schutze, 1999), this is what is expected. It should be noted that an important but difficult-to-answer question here is “What is the difference between frequent morphemes and functional category words?”. Their difference is very clear under English orthography, but their difference is not clear if we eliminate the spaces between words. In such a case, it may be possible to argue that “of” and “the” in “ofthe” are morphemes rather than words, although they are declared to be words under English orthography.

5 Appendix: Result from Old Japanese

As a supplementary test, the program was also tested on old Japanese corpus “The Tale of Genji” (available from

[<http://www.sainet.or.jp/~eshibuya/hp.html>])².

Although this is no more than a preliminary result, 87.5% precision out of top 40 tri-grams has been obtained³. It is planned to run the program on CHILDES corpus [<http://childes.psy.cmu.edu/>] especially in Chinese and Japanese, since the result can turn out drastically different from English given their different morphological structure (see Chen et al (1992) for Chinese). Note that they are both phonemically transcribed and thus transcription problems are not expected with these corpora.

References

- Cavar D, Herring J, Ikuta T, Rodrigues P and Schrementi G. 2004. *Alignment Based Induction of Morphology Grammar and its Role for Bootstrapping*. In proceedings of FGNancy 2004.
- Chen K J and Liu S H 1992. *Word identification for mandarin Chinese sentences*. Proceedings of the Fifteenth International Conference on Computational Linguistics, Nantes, 101-107.
- Goldsmith John. 2001 *Unsupervised Learning of the Morphology of a Natural Language*. *Association for Computational Linguistics*
- Manning C and Schutze H 1999. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Mass.

²The main purpose of running the program on Japanese is to see the result on language where space is not orthographically represented usually, where distinction between words and morphemes are not clearly represented in orthography; In English, words are segmented by spaces but morphemes are not.

³The precision rate may not be very accurate given the authors limited knowledge of old Japanese