Word Segmenter of Chinese, Japanese and Korean

<u>Research Paper</u>

Link: https://www.aclweb.org/anthology/W09-3426.pdf

Abstract

Word segmentation is a process to divide a

sentence into meaningful units called “word

unit” [ISO/DIS 24614-1]. What is a word

unit is judged by principles for its internal integrity and external use constraints. A word

unit’s internal structure is bound by principles of lexical integrity, unpredictability

and so on in order to represent one syntactically meaningful unit. Principles for external

use include language economy and frequency

such that word units could be registered in a

lexicon or any other storage for practical reduction of processing complexity for the further syntactic processing after word segmentation. Such principles for word segmentation

are applied for Chinese, Japanese and Korean,

and impacts of the standard are discussed.

Word Segmentation: Framework and Principles

Word unit is a layered pre-syntactical unit. That

means that a word unit consists of the smaller

word units. But the maximal word unit is frequently occurred in corpora under the constraints

that the syntactic processing will not refer the

internal structure of the word unit

Basic atoms of word unit are word form, morpheme including bound morpheme, and nonlexical items like punctuation mark, numeric

string, foreign character string and others as shown in Figure 1. Usually we say that “word” is lemma or word form. Word form is a form that a lexeme takes when used in a sentence. For example, strings “have”, “has”, and “having” are word forms of the lexeme HAVE, generally distinguished by the use of capital letters. [ISO/CD 24614-1] Lemma is a conventional form used to represent a lexeme, and lexeme is an abstract unit generally associated with a set of word forms sharing a common meaning.

Conclusion

Conclusion

Word segmentation standard is to recommend what type of word sequences should be identified as one word unit in order to process the syntactic parsing. Principles of word segmentation want to provide the measure of such word units. But principles of frequency and language economy are based on a statistical measure, which will be decided by some practical purpose.

Word segmentation in each language is somewhat different according to already made word segmentation regulation, even violating one or more principles of word segmentation. In the future, we have to discuss the more synchronized word unit concept because we now live in a multi-lingual environment.

Jieba Code Snippet

```
import jieba
seg_list = jieba.cut("我来到北京清华大学", cut_all=True)
print("Full Mode: " + "/ ".join(seg_list))    # 全模式
seg_list = jieba.cut("我来到北京清华大学", cut_all=False)
print("Default Mode: " + "/ ".join(seg_list))    # 精确模式
seg_list = jieba.cut("他来到了网易杭研大厦")    # 默认是精确模式
print(", ".join(seg_list))
seg_list = jieba.cut_for_search("小明硕士毕业于中国科学院计算所，后在日本京都大学深造")    # 搜索引擎模式
print(", ".join(seg_list))
```

Typography: Typographical Devices. Typographical art is the employment of typography or text in ways that form patterns, ambigrams, shapes and images. It is sometimes referred to as shaped verse (when poems are employed) or pattern poems.

Types: Grotesque like Venus, Monotype Grotesque, and News Gothic.

Neo-Grotesque like Helvetica, San Francisco, and Roboto.

Humanist like Tahoma, Verdana, Calibri, and Trebuchet.

Geometric like Gotham, Avenir, and ITC Avant Garde.

Orthography: An orthography is a set of conventions for writing a language. It includes norms of spelling, hyphenation, capitalization, word breaks, emphasis, and punctuation.

Suprasegmental: Also called Prosodic **Feature**, in phonetics, a speech **feature** such as stress, tone, or word juncture that accompanies or is added over consonants and vowels;
these **features** are not limited to single sounds but often extend over syllables, words, or phrases.