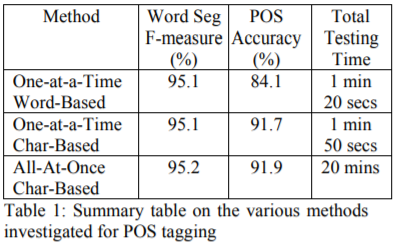
Chinese part-of-speech (POS) tagging assigns one POS tag to each word in a Chinese sentence. However, since words are not demarcated in a Chinese sentence, Chinese POS tagging requires word segmentation as a prerequisite.

Most corpus-based language processing research has focused on the English language. Theoretically, we should be able to just port corpus-based, machine learning techniques across different languages since the techniques are largely language independent. However, in practice, the special characteristics of different languages introduce complications. For Chinese in particular, words are not demarcated in a Chinese sentence. As such, we need to perform word segmentation before we can proceed with other tasks such as part-of-speech (POS) tagging and parsing, since one POS tag is assigned to each Chinese word (i.e., all characters in a Chinese word have the same POS tag), and the leaves of a parse tree for a Chinese sentence are words.



The findings that a character-based approach is better than a word-based approach for Chinese POS tagging is not too surprising. Unlike in English where each English letter by itself does not possess any meaning, many Chinese characters have well defined meanings. For example, the single Chinese character “知” means “know”. And when a character appears as part of a word, the word derives part of its meaning from the component characters. For example, “知识” means “knowledge”， “无知” means “ignorant”, “知名” means “well-known”, etc. In addition, since the out-of-vocabulary (OOV) rate for Chinese words is much higher than the OOV rate for Chinese characters, in the presence of an unknown word, using the component characters in the word to help predict the correct POS tag is a good heuristic.



Language differences between English and Chinese have made direct porting of an English POS tagging method to Chinese ineffective. In Chinese, individual characters encode information that aids in POS tagging. Using a character-based approach for Chinese POS tagging is more effective than a word-based approach. Our study has also revealed that the one-at-a-time, character-based approach gives relatively good POS tagging accuracy with a much improved training and testing time, compared with the all-at-once, character-based approach previously proposed.