Using entropy to measure semantic information of Chinese and English words

Yi-Ling Chung
National Cheng Kung Univrsity

Chung-Ching Wang
National Cheng Kung Univrsity

Hsueh-Chih Chen
National Taiwan Normal University

Jon-Fan Hu
National Cheng Kung Univrsity

Abstract: One of the obstacles to fully ensure the semantic contents of words is how to grab the meanings of a word from various probabilities it associates with other words. According to Shannon's (1948) information theory, entropy can provide indications of amount of information and extent of uncertainty of a given variable by calculating the probability distributions of event occurrence. Therefore, entropy based on word-word co-occurrences of a document would disclose the semantic clues for word meanings. In the present study, the computed entropy values of eighty thousand Chinese words excepted from Academia Sinica Electronic Dictionary are calculated according to the word-pair occurrences. The findings show that the level of entropy correlated positively with the variety of semantics. Furthermore, the conditional entropy value for a given word can be used to differentiate the extent of how that word constrains the meaning of the subsequent words in the same text. It is also found that entropy values can reveal the differences of the amount of information carried for words having parallel translation definitions in Chinese and English.