Automatic Disambiguation of Chinese Modal Expressions
- A Supervised Machine Learning Experiment

Ting Chi
Georgetown University

The ambiguity of modal expressions is a major obstacle to natural language processing (NLP). For instance, the meaning of the word *must* can vary greatly depending on the contexts. In sentences like *he must be lost*, the event *lost* modified by the modal word *must* is not necessarily a fact, but is highly probable; while in sentences like *he must leave*, it tells us nothing about whether the event *leave* will happen or not. This obviously presents a serious challenge to computers that need to analyze natural language with modal expressions.

To facilitate the understanding of modal expressions, researchers have attempted to classify modality into different types. These "modality types", despite the lack of unanimous agreement on a single method of classification, are useful tools for computers to learn modality. Returning to the example of *must* - *he must be lost* is an "epistemic" expression, pertaining predictions on possibility; *he must leave* is a "deontic" expression, pertaining obligation or permission. Therefore to overcome the challenge that modal ambiguity presented to NLP, a system that can automatically discern modality types is required.

It is the intention of this thesis to develop such a system for Chinese model expressions. The development of such a system is best conducted through supervised machine learning, teaching computers the syntactic structures and semantical interpretations of modality types in real speech. This thesis gathered Chinese speech materials from CHTB 4.0. Then these materials, sentences, are subjugated to annotation, which select the ones with modal expressions, mark the modality types and identify the determining attributes, such as prejacent, source and background. Signaling terms in these attributes - such as think or require - that are most likely associated with a particularly modality type are selected to form a feature set, while irrelevant terms are sifted out. These feature sets are used as the training data of supervised machine learning, telling computers which modality type is the most likely choice based on the presence of features in a given sentence. The computer analyzes these features through certain algorithms. This thesis chose to use three different algorithms - Naive Bayes, logistic regression and decision tree - to search for best result. Eventually, the "trained computer" will be able to automatically detect relevant features in a Chinese sentence, and understand its modal expression based on the modality type it identifies.

The systems this thesis produced show more than 92% accuracy in discerning modality types in Chinese.