

# Categorical Ambiguity and Information Content

## A Corpus-based Study of Chinese

Chu-Ren Huang, Ru-Yng Chang

Institute of Linguistics, Preparatory Office, Academia Sinica

128 Sec.2 Academy Rd., Nankang, Taipei, 115, Taiwan, R.O.C.

churen@gate.sinica.edu.tw, ruyng@hp.iis.sinica.edu.tw

### 1. Introduction

Assignment of grammatical categories is the fundamental step in natural language processing. And ambiguity resolution is one of the most challenging NLP tasks that is currently still beyond the power of machines. When two questions are combined together, the problem of resolution of categorical ambiguity is what a computational linguistic system can do reasonably good, but yet still unable to mimic the excellence of human beings. This task is even more challenging in Chinese language processing because of the poverty of morphological information to mark categories and the lack of convention to mark word boundaries. In this paper, we try to investigate the nature of categorical ambiguity in Chinese based on Sinica Corpus. The study differs crucially from previous studies in that it directly measure information content as the degree of ambiguity. This method not only offers an alternative interpretation of ambiguity, it also allows a different measure of success of categorical disambiguation. Instead of precision or recall, we can also measure by how much the information load has been reduced. This approach also allows us to identify which are the most ambiguous words in terms of information content. The somewhat surprising result actually reinforces the Saussurian view that underlying the systemic linguistic structure,

assignment of linguistic content for each linguistic symbol is arbitrary.

### 2. Previous Work

Assignment of grammatical categories or tagging is a well-tested NLP task that can be reliably preformed with stochastic methodologies (e.g. Manning and Shutz 1999). Depending on the measurement method, over 95% precision can be achieved regularly. But the question remains as to why the last few percentages are so hard for machines and not a problem for humans. In addition, even though over 95% seems to be good scores intuitively, we still need to find out if they are indeed better than the naïve baseline performance. Last but not the least, since natural languages are inherently and universally ambiguous, does this characteristic serve any communicative purpose and can a computational linguistic model take advantage of the same characteristics.

Since previous NLP work on categorical assignment and ambiguity resolution achieved very good results using only distributional information, it seems natural to try to capture the nature of categorical ambiguity in terms of distributional information. This is how the baseline model was set in Meng and Ip (1999), among others. Huang et al. (2002), the most extensive study on categorical ambiguity in

Chinese so far, also uses only distributional information.

Huang et al. (2002) confirmed some expected characteristics of ambiguity with convincing quantitative and qualitative data from the one million word Sinica Corpus 2.0. Their generalizations include that categorical ambiguity correlates with frequency; that verbs tend to be more ambiguous than nouns, and that certain categories (such as prepositions) are inherently more ambiguous.

What is not totally unexpected, and yet runs against certain long-held assumptions is the distribution of ambiguity. It is found that only a small fraction of all words (4.298%) are assigned more than one category. However, in terms of actual use, these words make up 54.59% of the whole corpus. These two facts are consistent with the frequency effect on ambiguity. An interesting fact is that of all the words that can have more than one category, 88.37% of the actual uses are in the default category.

A significant fact regarding Chinese language processing can be derived from the above data. Presupposing lexical knowledge of the lexicon and the default category of each word, a naïve baseline model for category assignment two simple steps: First, if a word has only one category in the lexicon, assign that category to the word. Second, if a word has more than one category in the lexicon, assign the default (i.e. most frequently used) category to that word. Since step 1) is always correct and the precision rate of step 2) depends on the percentage of use of the default category. Huang et al. (2002) estimated the expected precision of such a naïve model to be over 93.65%.

Huang et al.'s (2002) work, however, has its

limitation. It takes categorical ambiguity as a lexical attribute. In other words, an attribute is either + or -, and a certain word is either categorically ambiguous or not. For Huang et al. (2002), the degree of ambiguity is actually the distribution of the attribute of being ambiguous among a set of pre-defined (usually by frequency ranking) lexical items. Strictly speaking, this data only shows the tendency of being categorically ambiguous for the set members. In other words, what has been shown is actually:

Words with higher frequency are more likely to be categorically ambiguous.

The data has nothing to say about whether a lexical item or a set of lexical items are more ambiguous than others or not.

A good example of the inadequacy of Huang et al.'s (2002) approach is their measurement of the correlation between number of potential categories and the likelihood of default category to occur.

No. of Categories	Freq. (by type)	Freq. (by token)
2	77.65%	91.21%
3	77.71%	88.39%
4	74.21%	89.50%
5	73.83%	92.43%
6	73.46%	86.09%
7	68.51%	86.09%
Total	77.36%	88.37%

**Table 1. Frequency of Default Category**

In table one, the number seems to suggest that number of possible categories of a word form is not directly correlated with its degree of ambiguity, since its probability of being assigned the default category is not predictable and remains roughly the same in average. This is somewhat counter-intuitive in the sense that we expect the more complex the information structure (i.e. more possible categories), the less

likely that it will be assigned a simple default. Since the methodology is to take distributional information over a large corpus, it is most likely the number shown in table 1 is distorted by the dominance of the most frequent words.

Is there an alternative to pure distributional measurement? Recall that ambiguity is about information content. Hence if the quantity of information content is measured, there will be a more direct characterization of ambiguity.

### 3. Towards an Informational Description of Categorical Ambiguity

#### 3.1. Degree of Categorical Ambiguity

In this paper, we will adopt Shannon's Information Theory and measure categorical ambiguity by entropy. We define the information content of a sign as its entropy value.

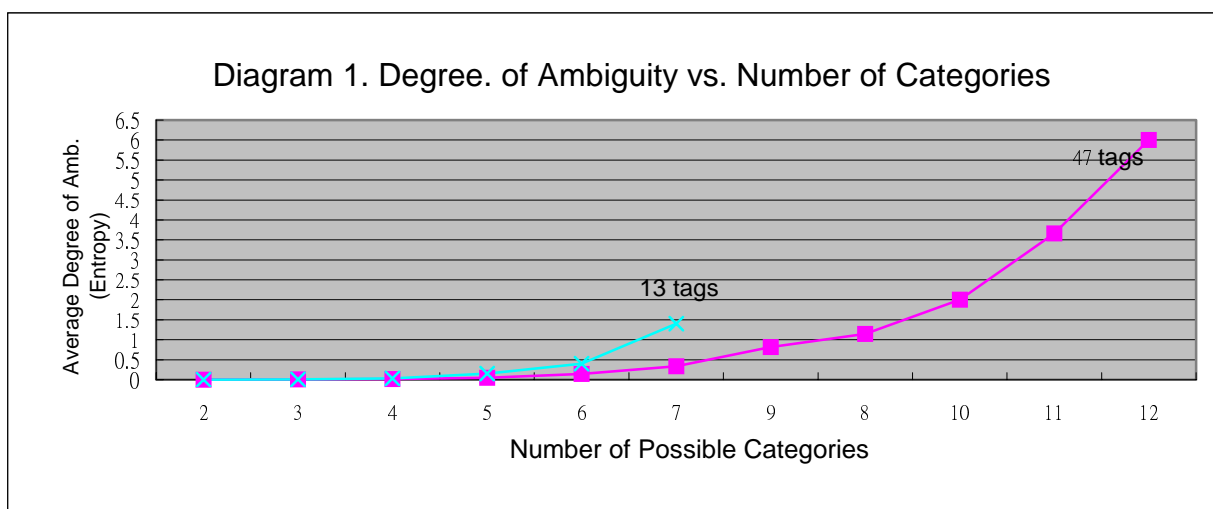
$$H = -(p_0 \log p_0 + p_1 \log p_1 + \dots + p_n \log p_n)$$

When measuring categorical ambiguity, for a word with  $n$  potential categories, the information content of that word in terms of grammatical categories is the sum all the entropy of all its possible categories. We will make the further assumption of that the degree of ambiguity of a word corresponds to the quantity of its information content.

The above definition nicely reflects the intuition that the more predictable the category is, the less ambiguous it is. That is, a word that can be 90% predicted by default is less ambiguous than a word that can only be predicted in 70% of the context. And of course the least ambiguous words are those with only one possible category and can be predicted all the time (it information value is actually 0).

#### 3.2. Degree of Ambiguity and Number of Possible Categories Revisited

Armed with a new measurement of the degree of ambiguity for each lexical item, we can now take another look at the purported lack of correlation between number of possible categories and degree of ambiguity. Instead having to choose between type of token as units of frequency counting, we can now calculate the degree of categorical ambiguity for each lexical form in terms of entropy. The entropy of all lexical forms with the same numbers of possible categories can then be averaged. The results is diagrammed below:



In the above diagram, we can clearly see that whether a 47 tags system or 13 tags system is chosen, the number of potential categories correlates with the degree of ambiguity. The higher number of potential categories a word has, the more ambiguous it is. This correctly reflects previous observational and theoretical predictions.

### 3.3. Frequency and Degree of Ambiguity

One of the important findings of Huang et al. (2002) was that the likelihood to be ambiguous indeed correlates with frequency. That is, a more frequently used word is more likely to be categorically ambiguous. However, we do not know that, of all the categorically ambiguous words, whether their degree of ambiguity corresponds to frequency or not.

In terms of the number of possible categories, more frequent words are more likely to have larger number of categories. Since we have just showed in last session that larger

number of possible categories correlates with degree of ambiguity. This fact seems to favor the prediction that more frequent words are also more ambiguous (i.e. harder to predict their categories.)

Common sense of empirical models, however, suggests that it is easier to predict the behaviors of more familiar elements. Confidence of prediction corresponds to quantity of data. A different manifestation of this feature is that there is a data sparseness problem but never a data abundance problem. In addition, the high precision rate of categorical assignment requires that most frequent words, which take up the majority of the corpus, be assigned correct category at a reasonable precision rate. These two facts seem to suggest that the less frequent words may be harder to predict and hence more ambiguous.

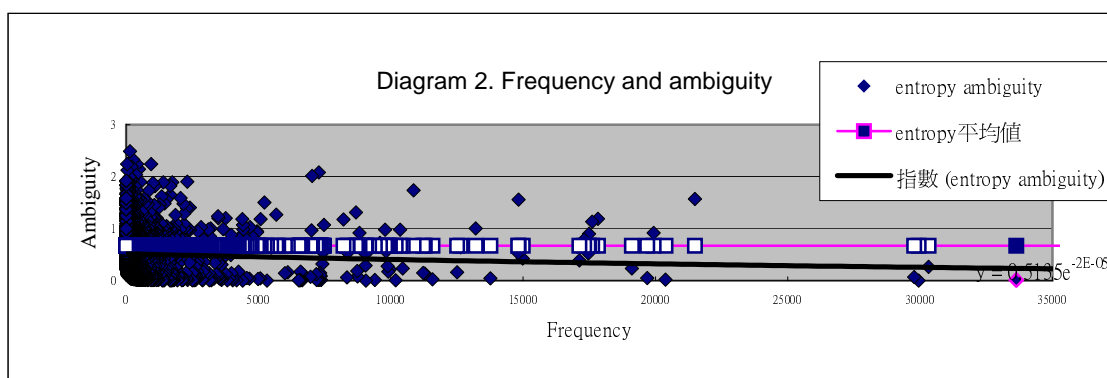
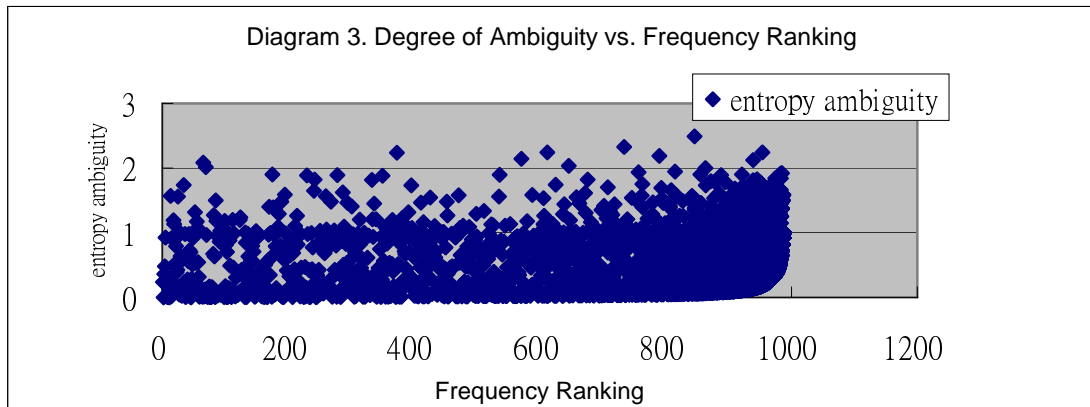


Diagram 2 plots the degree of ambiguity of each ambiguous word in terms of its frequency in the Sinica Corpus (Chen et al. 1996). Not only does the distribution of the degree of ambiguity vary widely, the medium tendency line (thick black line in the diagram) varies barely perceptibly

across frequency ranges. As suggested by the two competing tendencies discussed above, our exhaustive study actually shows that there is no correlation between degree of ambiguity and frequency. This generalization can be shown with even more clarity in Diagram 3.



In Diagram 3, entropy value of each word form is plotted against its frequency ranking. When word forms share the same frequency, they are given the same ranking, and no ranking jumps were given after multiple elements sharing the same ranking. Due to the sharing of rankings, the highest rank only goes to 1,000. Diagram 3 shows unequivocally that the range of degree of ambiguity remains the same across different frequency ranges. That is, degree of ambiguity does not correlate to word frequency.

#### 4. Conclusion

In this paper, we propose an information-based measure for ambiguity in Chinese. The measurement compliments the more familiar distributional data and allows us to investigate directly the categorical information content of each lexical word. We showed in this paper that degree of ambiguity indeed correlates with the number of possible categories of that word. However, degree of ambiguity of a word does not correlate with its frequency, although its tendency to be categorically ambiguous is dependent on frequency.

The above findings have very important implications for theories and applications in language processing. In terms of representation of linguistic knowledge, it underlines the arbitrariness of the encoding of lexical information, following Saussure. In terms of processing model and empirical prediction, it suggests a model not unlike the theory of unpredictability in physics. Each word is like an electron. While the behavior of a group of words can be accurately predicted by stochastic model, the behavior of any single word is not predictable. In terms of linguistic theory, this is because there are too many rules that may apply to each lexical item at different time and on different levels, hence we cannot predict exactly

how these rules the results without knows exactly which ones applied and in what order. This view is compatible with the Lexical Diffusion (Wang 1969) view on application of linguistic rules.

In NLP, this clearly predicts the performance ceiling of stochastic approaches. As well as that the ceiling can be surpassed by hybridizing with specific lexical heuristic rules covering the 'hard' cases for stochastic approaches, as suggested in Huang et al. (2002).

#### References:

- Chen, Keh-jian, Chu-Ren Huang, Li-ping Chang, and Hui-Li Hsu. 1996. Sinica Corpus: Design Methodology for Balanced Corpora. In B.-S. Park and J.B. Kim. Eds. Proceeding of the 11<sup>th</sup> Pacific Asia Conference on Language, Information and Computation. 167-176. <http://www.sinica.edu.tw/SinicaCorpus>
- Chinese Knowledge Information Processing (CKIP) Group. 1995. An Introduction to Academia Sinica Balanced Corpus for Modern Mandarin Chinese. CKIP Technical Report. 95-01. Nankang: Academia Sinica
- Huang, Chu-Ren, Chao-Ran Chen and Claude C.C. Shen. 2002. Quantitative Criteria for Computational Chinese The Nature of Categorical Ambiguity and Its Implications for Language Processing: A Corpus-based Study of Mandarin Chinese. Mineharu Nakayama (Ed.) Sentence Processing in East Asian Languages. 53-83. Stanford: CSLI Publications
- Manning Christopher D. and Hinrich Shutze. 1999. Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.
- Meng, Helen and Chun Wah Ip. 1999. An Analytical Study of Transformational Tagging for Chinese Text. In Proceedings of ROCLING XII. 101-122. Taipei: Association of Computational Linguistics and Chinese Language Processing.
- Schutz, Hinrich. 1997. Ambiguity Resolution in Language Learning: Computational and Cognitive Models. Stanford: CSLI Publications.
- Wang, S.-Y. W. 1969. Competing Changes as a Cause of Residue. Language. 45.9-25.