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 Academia Rd., Nangkang, Taipei, 115, Taiwan, R.O.C.
churen@gate.sinica.edu.tw, ruyng@gate.sinica.edu.tw

Abstract

The degree of ambiguity in Chinese is investigated in this paper based on the tagged Sinica Corpus. We propose to use measurement of information content instead of frequency to model generalizations about categorical ambiguity Two important generalizations were found: First, the degree of ambiguity indeed correlates with the number of possible categories that a word has. Second,. Even though frequently used words are more likely to be categorical ambiguous (Huang et al. 2002), the degree of ambiguity of a word does not depend on its frequency.

Keywords

Categorical ambiguity, Information content, entropy, frequency, Sinica Corpus

1. Introduction

Assignment of grammatical categories is a fundamental step in natural language processing. Ambiguity resolution, moreover, is one of the most challenging NLP tasks that is currently still beyond the power of machines. 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 will investigate the nature of categorical ambiguity in Chinese based on the 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 the success of categorical disambiguation. Instead of precision or recall, we will measure 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 performed 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 a good score intuitively, we still need to find out if they are indeed better than 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 these 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, Chen and Shen (2002), in 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 findings demonstrated 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 of 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 (i.e. most frequently used) 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 involves 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 category to that word. Step 1) is always correct. 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 limitations. 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 shall 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.

$$Entropy\,(t_j) = -\sum_{t_j \in t\arg et} P(t_i \mid t_j) \log P(t_i \mid t_j)$$

where t_i : the frequency of the lexicon with POS
t_j : the frequency of a lexical form

Equation 1

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 of all the entropy of all its possible categories. We will make the further assumption that the degree of ambiguity of a word corresponds to the quantity of its information content.

The above definition 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's information value is actually 0).

3.2. Degree of Ambiguity and Number of Possible Categories Revisited

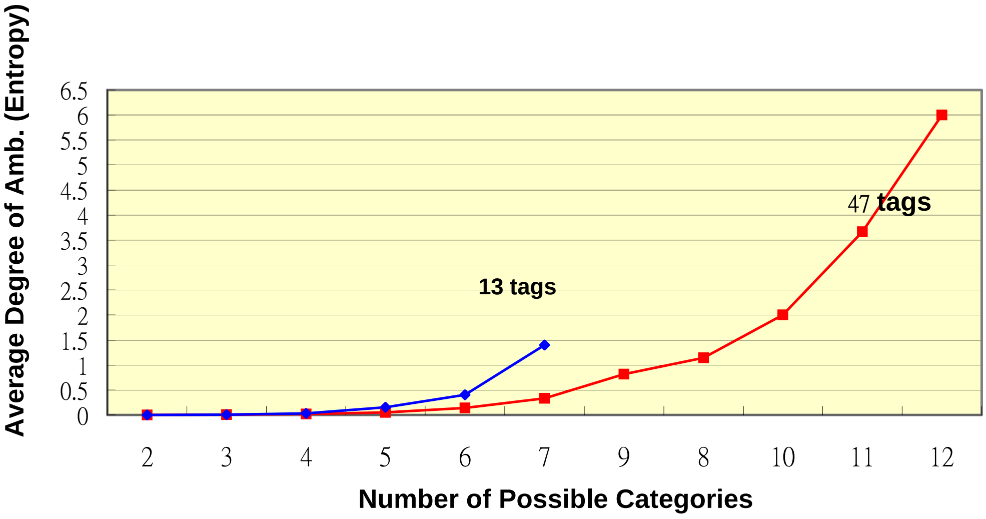

Figure 1 Degree of Ambiguity vs. Number of Categories

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 or 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 result is diagramed below:

In the above diagram, we can clearly see that whether a 47 tag system or 13 tag 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 important findings of Huang et al. (2002) was that the likelihood to be ambiguous correlates with frequency. That is, a more frequently used word is more likely to be categorically ambiguous. However, for all categorically ambiguous words, we do not know 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 a larger number of categories. This is also supported by our result that a larger number of possible categories correlates with degree of ambiguity. Since more frequent words tend to have more categories, this fact seems to favor the prediction that more frequent words are also more ambiguous (i.e. harder to predict their categories.) This prediction seems to be empirically verified when we took a more naïve approach by simply counting the number of words with multiple categories in a frequency range (Huang et al. 2002).

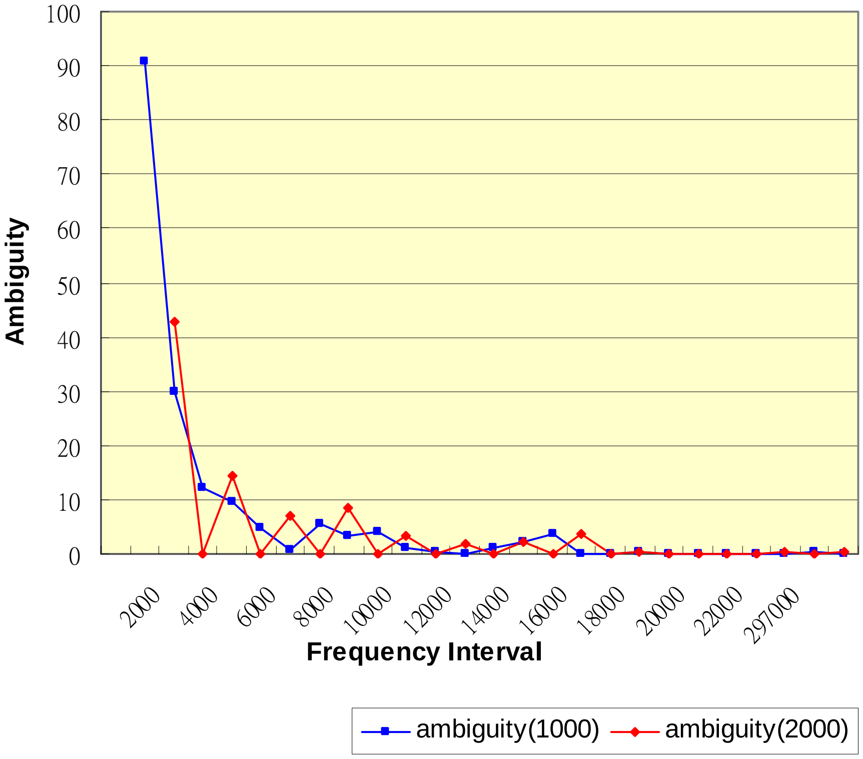

Figure 2 Frequency Interval and Categorical Ambiguity

On the other hand, common sense empirical models suggest 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 to the 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.

In order to better understand the effect of these two competing motivations, we took the more direct measure of degree of ambiguity based on entropy and recalculated.

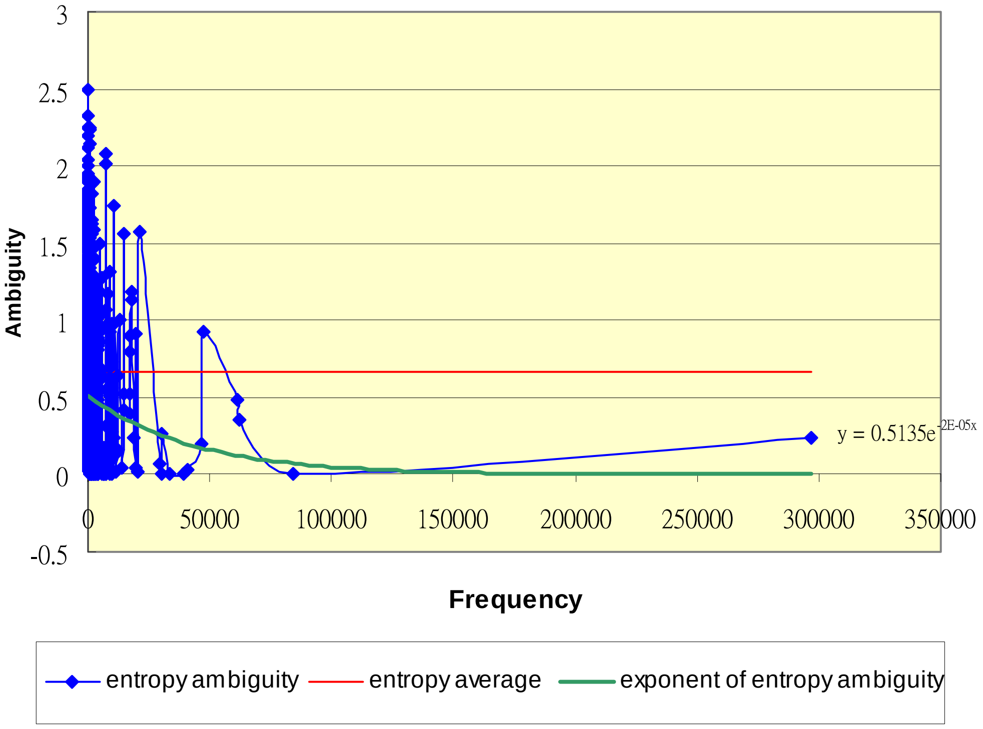

Figure 3 Frequency and Ambiguity

Figure 3 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 Figure 4.

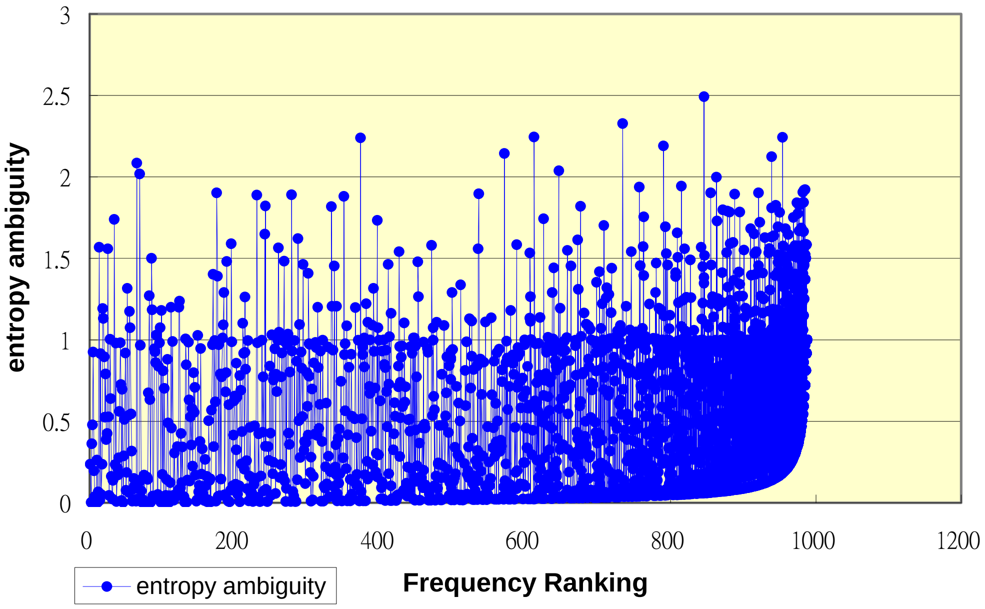

Figure 4 Degree of Ambiguity vs. Frequency Ranking

In Figure 4, the 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 to multiple elements sharing the same ranking. Due to the sharing of rankings, the highest rank only goes to 1,000. Figure 4 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.

Lastly, in order to have a full picture of the correspondence between degree of ambiguity and frequency ranking, we plotted the degree of ambiguity by ranking every 50 words. The result is shown in Figure 5.

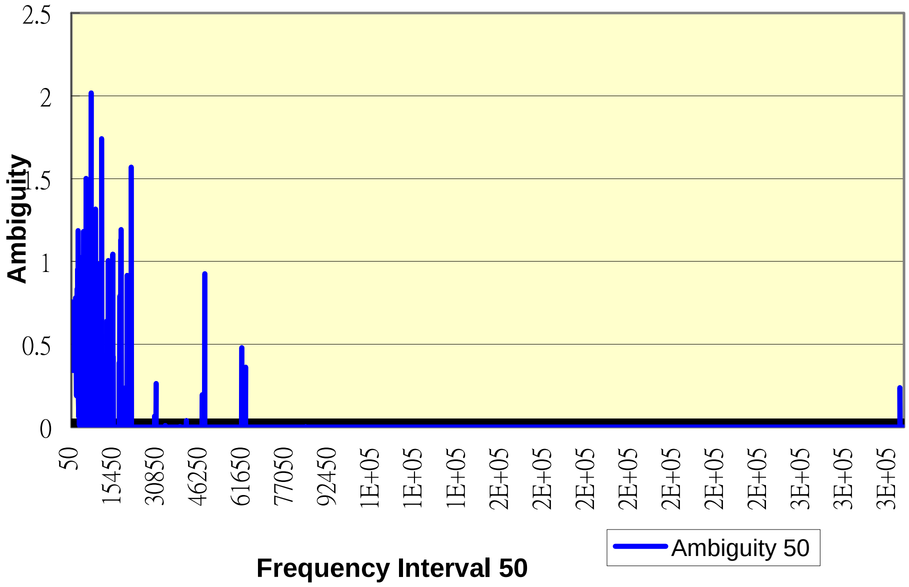

Figure 5 Frequency Interval and Categorical Ambiguity

The most striking fact shown in the above diagram is that 1) there is no clear correlation between frequency interval and degree of ambiguity, and that 2) there is a frequency threshold for degree of ambiguity. The diagrams shows that words ranked after 60,000 are basically categorically non-ambiguous. We can conclude that the relation between frequency and ambiguity is a threshold relation. Words that are not used frequent enough simply do not have enough population (i.e. occurrences) to support variations.

4. Conclusion

In this paper, we propose an information-based measure for ambiguity in Chinese. The measurement complements the more familiar distributional data and allows us to investigate directly the categorical information content of each lexical word. We show in this paper that the degree of ambiguity correlates with the number of possible categories of that word. However, the 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 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, as discussed by Saussure. In terms of processing models and empirical predictions, 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 work without knowing exactly which ones apply and in what order. This view is compatible with the Lexical Diffusion (Wang 1969) view on the application of linguistic rules.

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

Acknowledgements

An earlier version of this paper was presented at the First SigHAN workshop on Chinese Language Processing held in conjunction with COLING2002. We would like to thank the participants of that workshop as well as colleagues at CKIP, Academia Sinica for their helpful comments. We would like to thank Kathleen Ahrens, Hua-rui Zhang and Hongyin Tao for their comments on later versions. Lastly, we would like to thank Tien-ning Hsu for helping with the manuscript.

References:

Chen, Keh-jiann, 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 11th Pacific Asia Conference on Language, Information and Computation* pp. 167-176.

Chinese Knowledge Information Processing (CKIP) Group**,** 1995, An Introduction to Academia Sinica Balanced Corpus for Modern Mandarin Chinese, *CKIP Technical Report*, 95-01.

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*, 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, 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.