Using Co-occurrence Statistics as an Information Source for Partial Parsing of Chinese

Elliott Franco DRÁBEK
The State Key Laboratory
for Intelligent Technology and Systems
Department of Computer Science
Tsinghua University, Beijing 100084
elliott_drabek@ACM.org

Qiang ZHOU
The State Key Laboratory
for Intelligent Technology and Systems
Department of Computer Science
Tsinghua University, Beijing 100084
zhouq@s1000e.cs.tsinghua.edu.cn

Abstract

Our partial parser for Chinese uses a learned classifier to guide a bottom-up parsing process. We describe improvements in performance obtained by expanding the information available to the classifier, from POS sequences only, to include measures of word association derived from co-occurrence statistics. We compare performance using different measures of association, and find that Yule's coefficient of colligation Y gives somewhat better results over other measures.

Introduction

In learning-based approaches to syntactic parsing, the earliest models developed generally ignored the individual identities of words, making decisions based only on their part-of-speech classes. On the other hand, many later models see each word as a monolithic entity, with parameters estimated separately for each word type. In between have been models which attempt to generalize by considering similarity between words, where knowledge about similarity is deduced from hand-written sources (e.g. thesauri), or induced from text. For example, The SPATTER parser (Magerman, 1995) makes use of the output of a clustering algorithm based on co-occurrence information. Because this co-occurrence information can be derived from inexpensive data with a minimum of pre-processing, it can be very inclusive and informative about even relatively rare words, thus increasing the generalization capability of the parser trained on a much smaller fully annotated corpus.

The current work is in this spirit, making complementary use of a relatively small treebank for syntactic information and a relatively large collection of flat text for co-occurrence information. However, we do not use any kind of clustering, instead using the co-occurrence data directly. Our parser is a bottom-up parser whose actions are guided by a machine-learning-based decision-making module (we use the SNoW learner developed at the University of Illinois, Urbana-Champaign (Roth, 1998) for its strength with potentially very large feature sets and for its ease of use). The learner is able to directly use statistics derived from the co-occurrence data to guide its decisions.

We collect a variety of statistical measures of association based on bigram co-occurrence data (specifically, mutual information, t-score, X^2, likelihood ratio and Yule's coefficient of colligation Y), and make the statistics available to the decision-making module. We use labelled constituent precision and recall to compare performance of different versions of our parser on unseen test data. We observe a marked improvement in some of the versions using the co-occurrence data, with strongest performance observed in the versions using Yule's coefficient of colligation Y and mutual information, and more modest improvements in those using the other measures.

1 Background

1.1 Our Task — Partial Parsing

The current work has developed in the context of developing a partial or "chunk" parser for Chinese, whose task is to identify certain kinds of local syntactic structure. The syntactic

analysis we use largely follows the outline of Steven Abney's work (Abney, 1994). We adopt the concept of a "c-head" and an "s-head" for each phrase, where the c-head corresponds roughly to the generally used concept of head (e.g., the main verb in a verb phrase, or the preposition in a prepositional phrase), and the s-head is the "main content word" of a phrase (e.g., the main verb in a verb phrase, but the object of the preposition in a prepositional phrase). The core of our chunk definition is also in line with Abney's: A chunk is essentially the contiguous range of words s-headed by a given major content word. Within this basic framework, we make some accommodations to the Chinese language and to practicality. For example, by our understanding of Abney's definition, a numeral-classifier phrase followed immediately by the noun it modifies should constitute two separate chunks. However such units seem likely to be useful in further processing, and easy to accurately identify, so we chose to include them in our definition of chunk.

For simplicity and consistency, we adopt a very restricted phrase-structured syntactic formalism, somewhat similar to a phrase-structured formulation of a dependency grammar. In our formalism, all constituents are binary branching, and the purpose of the non-terminal labels is restricted to indicating the direction of dependency between the two children. Figure 1 shows an example sentence with some indicative structures.

Dependencies within individual chunks are shown with heavy arrows. A right-pointing dependency, such as the three dependencies within the noun phrase "瓜类蔬菜栽培方法", corresponds to a constituent labelled "right-headed". A left-pointing dependency, such as that between the verb "发生" and its aspect particle "了", corresponds to a constituent labelled "left-headed". These are cases where the s-head and the c-head of the phrase are identical. When they are not identical, we have a "two-headed" dependency, like those in the phrase "根本上的". Here, the relation between "根本" and "上" (and between "根本上" and "的") is that the left constituent provides the s-head of the phrase, while the right constituent provides the c-head.

These four non-terminal categories can describe high- or low- level syntactic structures. However, for chunking we wish to leave the higher-level structures of a sentence unspecified, leaving only a list of local structures. We treat this in a consistent way by adding a fifth non-terminal category "unspecified", and replacing all higher structures with a backbone of strictly left-branching "unspecified" nodes, anchored to a special "wall" token to the left of the sentence. This backbone structure is shown by the light lines in the figure.

1.2 Our Data Sources — One Large and One Small

During development, we made use of two corpora. The first is a relatively small-scale treebank of approximately 3500 sentences, 39,000 words, and 55,000 characters (Zhou, 1996). We transformed this corpus by annotating each phrase with c-heads and s-heads, using a large collection of hand-written rules, and then extracted chunks from this transformed version.

The second corpus, which we use only as a source of co-occurrence statistics, is much larger, with approximately 67,000 sentences, 1.5 million words, and 2.2 million characters, with sentences separated and words separated and marked with parts-of-speech, but with no further syntactic annotation (Zhou and Sun, 1999). In the current work we make no use of the part-of-speech annotation, taking co-occurrence counts of word-types alone.

1.3 Our Framework — Classifier-Guided Shift-Reduce Parsing

The parsing framework we use has been chosen for maximum simplicity, aided by the simplicity of the syntactic framework. In parsing, we model a left-to-right shift-reduce automaton which builds a parse-tree constituent-by-constituent in a deterministic left-to-right process. The parsing process is thus reduced to making a series of decisions of exactly what to build.

For training, we extract the series of actions the shift-reduce parser would have had to make to produce the trees from the surface structure of the sentences. This gives a long series of state-action pairs: "when the parser was in state X, it took action Y". The state description X is set of binary predicates describing the local surface structure of the sentence and the contents

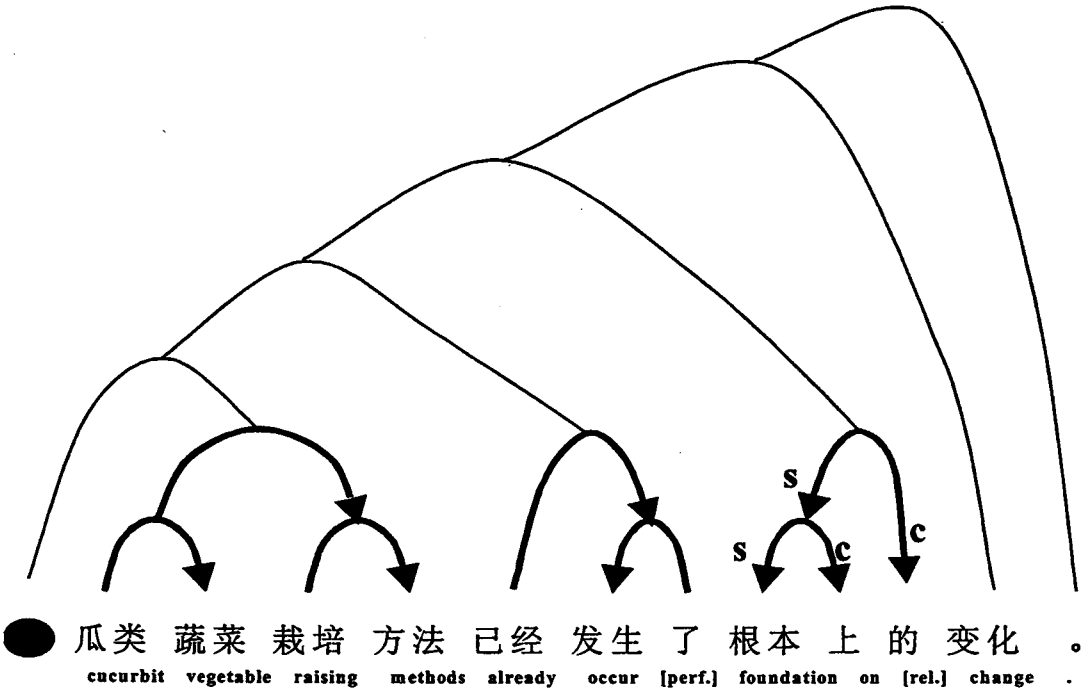

Figure 1. An example sentence annotated according to our system.

of the stack. We describe these predicates in detail below. This series of state-action pairs is presented to the SNoW learner, which tries to learn to predict the parser actions from the parser states, attempting to find a linear discriminant over these binary predicates which best accounts for the corresponding actions in the training data.

These parse actions can be either "shift a word from the right on to the stack", or "reduce the top elements of the stack" into a single constituent. Because our syntactic framework is strictly binary branching, each reduce action operates on exactly the top two items on the stack, so the automaton need only choose a category for the new constituent. This decision turns out to be nearly trivial, and we were able to achieve 100% accuracy on our test set using only part-of-speech information, so in the remainder of this paper we discuss only issues relating to the more difficult decision of whether to shift or reduce.

Within the shift-reduce decisions, over half are pre-determined by the basic requirements of the framework. For example, if there are no words left to shift, we can only reduce. If there is only one item on the stack, we can only shift. These decisions are handled by simple deterministic rules within the parser and are not shown to the classifier either in training or in parsing.

In the first version of the parser, prior to the introduction of co-occurrence statistics, the information available to the classifier is limited to parts-of-speech of words in the surface structure of the sentence, nonterminal categories of constituents already built on the stack, and parts-of-speech of the s- and c-heads of constituents already built on the stack. These are collected into schemas representing sets of possible binary predicates. Table 1 shows a representative subset of this original set of 18 predicate schemas (space does not allow us to present all of them). The total of all the instantiations of all these templates presents a potentially huge feature set, so we rely on an important property of the SNoW architecture, that it can handle an indefinitely large set of

Predicate Schema	Range of Parameters
POS(Surface-word[k]) = t	$-1 \leq k \leq 2$
POS(Surface-word[k]) = t_1 ∧ POS(Surface-word[k + 1]) = t_2	$-2 \leq k \leq 1$
Category(Stack[k]) = c	$0 \leq k \leq 1$
Category(Stack[k]) = c_1 ∧ Category(Stack[k + 1]) = c_2	$0 \leq k \leq 1$
POS(S-head(Stack[k])) = t	$0 \leq k \leq 2$
POS(S-head(Stack[k])) = t_1 ∧ POS(S-head(Stack[k + 1])) = t_2	$0 \leq k \leq 1$
POS(S-head(Stack[k_1])) = t_1 ∧ POS(Surface-word[k_2]) = t_2	$0 \leq k_1 \leq 1$ $-1 \leq k_2 \leq 0$
Category(Stack[k_1]) = c ∧ POS(C-head(Stack[k_1])) = t_1 ∧ POS(Surface-word[k_2]) = t_2	$0 \leq k_1 \leq 1$ $-1 \leq k_2 \leq 0$

Table 1. A Subset of the Feature Schemas in the Original Version of the Parser. The variables t, t_1, and t_2 range over the set of part-of-speech categories, while the variables c, c_1, and c_2 range over the set of non-terminal categories. Surface words are indexed relative to the parsing position, such that Surface-word[0] is the next word to be shifted.

features, actually using only those features which are active. The set of these actually active features is reasonable for our set of schemas.

2 Enriching the Feature Set with Co-occurrence Statistics

2.1 Measures of Association

Table 2 shows the definitions of the five measures we have chosen to compare in the current work, taken from (Manning and Shütze, 1999), (Kageura, 1999).

These measures are based on empirical counts of word occurrences and co-occurrences. Because these events are very prone to zero-counts, both for unseen bigrams and for unseen words, we applied Simple Good-Turing smoothing (Gale and Sampson, 1995) to both bigram and word counts.

2.2 Making Measures of Association Available to the Parser

To make the measures of association available to the parser, we started by discretizing each measure, that, is substituting for each continuous measurement a set of binary predicates coarsely describing its approximate value. We used a very simple form of discretization, counting occurrences of each value, and then dividing the values into bins of approximately equal counts. Informal exploration showed consistently better performance when bin membership was made cumulative; that is, using non-mutually-exclusive predicates of the form:

$$\text{statistic}(w_1, w_2) \leqslant X_1$$
$$\text{statistic}(w_1, w_2) \leqslant X_2$$
$$\ldots$$

rather than mutually-exclusive predicates of the form:

$$X_0 < \text{statistic}(w_1, w_2) \leqslant X_1$$
$$X_0 < \text{statistic}(w_1, w_2) \leqslant X_2$$
$$\ldots$$

Using these cumulative predicates, parsing accuracy consistently improved with increases in the number of bins, though the rate of improvement slowed at the same time. The cost of increasing the number of bins came primarily in the algorithm's training time. We chose thirty-two to be a good number of bins.

The predicates resulting from discretization are predicates over values of a statistic. To apply these predicates in parsing, we created features relating to particular slots within parse-state descriptions. Specifically, we made three new feature schemas available to the Winnow learner, as shown in Table 3. Each of these feature schemas is an extension to one available to the original parser. In each case, the original schema was of the form:

$$\text{POS}(w_1) = t_1 \ \wedge \ \text{POS}(w_2) = t_2$$

And the extended schema was of the form:

$$\text{POS}(w_1) = t_1 \ \wedge \ \text{POS}(w_2) = t_2 \ \wedge \ \text{statistic}(w_1, w_2) \leqslant X$$

In this way, the learner is able to condition separately depending on the parts-of-speech of the two words in question. This is based on the intuitions that different cases for part-of-speech combinations would behave very differently, and that the training data was sufficient that

Measure	Definition
MI	$\log\left(\frac{c(w_1,w_2)}{c(w_1,\bullet)c(\bullet,w_2)}\right)$
T-score	$\sqrt{c(w_1,w_2)}\left(\frac{c(w_1,\bullet)c(\bullet,w_2)}{c(w_1,w_2)}-1\right)$
X^2	$\frac{c(\bullet,\bullet)\left(c(w_1,w_2)c(\neg w_1,\neg w_2)-c(w_1,\neg w_2)c(\neg w_1,w_2)\right)^2}{c(w_1,\bullet)c(\neg w_1,\bullet)c(\bullet,w_2)c(\bullet,\neg w_2)}$
Likeli-hood Ratio	$2\left[\begin{array}{l} LogL\left(\frac{c(w_1,w_2)}{c(\bullet,w_2)};c(w_1,w_2);c(\bullet,w_2)\right)+LogL\left(\frac{c(w_1,\neg w_2)}{c(\bullet,\neg w_2)};c(w_1,\neg w_2);c(\bullet,\neg w_2)\right) \\ -LogL\left(\frac{c(w_1,\bullet)}{c(\bullet,\bullet)};c(w_1,w_2);c(\bullet,w_2)\right)-LogL\left(\frac{c(w_1,\bullet)}{c(\bullet,\bullet)};c(w_1,\neg w_2);c(\bullet,\neg w_2)\right) \end{array}\right]$
Note:	$LogL(p;n;k)=k\log(p)+(n-k)\log(1-p)$
Yule's Y	$\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1}$
Note:	$\alpha=\frac{c(\bullet,\bullet)c(w_1,w_2)}{c(w_1,\overline{w_2})c(\overline{w_1},w_2)}$

Table 2. Definitions of the Five Measures of Association. $c(x_1,w_2)$ represents the count of the event that x and y occur adjacent and in this order in the training corpus. $\overline{w}$ represents summation over all words other than w, and $\bullet$ represents summation over all words.

performance would not be hurt by the resulting sub-division; however we have no specific empirical support for this.

2.3 Experimental Results

We trained a series of SNoW networks using features sets extended with each of the five measures, and tested five versions of our parser, one using each of the resulting networks. This was done on a held-out test set comprising approximately ten percent of our treebank. The resulting measurements for labeled constituent precision and recall are shown in Table 4, arranged according to the geometric mean of the two measurements.

It is clear from the table that co-occurrence information can be made useful, and that the measure used to represent this information has a large influence on its usefulness. There is also a large disparity between the improvement in precision, 1.7%, and the improvement in recall, 4.1%. We conjecture that this is because the parser originally tended to err in the direction of splitting words into separate chunks, the commoner case, while with the co-occurrence information, it is able to pick out some cases where a strong association suggests that words be joined in the same chunk.

3 Related Work

Statistical measures of association applied to bigram co-occurrence counts have been used most extensively in terminology and collocation extraction. (Manning and Shütze, 1999) contains a good introduction to this topic. (Kageura, 1999) is an especially good empirical comparison of the performance of several measures of association on a set of tasks in both terminology extraction and in morpheme splitting of Chinese character sequences. This latter task, which can be seen as a very restricted form of parsing, has been treated in a body of interesting work, including (Sun, Shen and Tsou, 1998), (Lee, 1999) . This work has generally used very simple heuristic control policies, such as repeatedly splitting at the point of lowest mutual information. The use of similar

Predicate Schema	Range of Parameters
POS(Surface-word[k]) = t_1 ∧ POS(Surface-word[k + 1]) = t_2 ∧ Statistic(Surface-word[k],Surface-word[k + 1]) ⩽ X	-2 ≤ k ≤ 1
POS(S-head(Stack[k])) = t_1 ∧ POS(S-head(Stack[k + 1])) = t_2 ∧ Statistic(S-head(Stack[k]),S-head(Stack[k + 1])) ⩽ X	0 ≤ k ≤ 1
POS(S-head(Stack[k_1])) = t_1 ∧ POS(Surface-word[k_2]) = t_2 ∧ Statistic(S-head(Stack[k_1]),Surface-word[k_2]) ⩽ X	0 ≤ k_1 ≤ 1 -1 ≤ k_2 ≤ 0

Table 3. Augmented Feature Schemas.

approaches for general parsing received some early exploration (Brill, Magerman, Marcus and Santorini, 1990), (Magerman and Marcus, 1990), but this approach seems to have lost popularity. This may be because using co-occurrence statistics as a sole source of guidance may become insufficient as the object of parsing moves from the very local structure of word splitting to the longer-distance dependencies of general parsing. The current work attempts to remedy this by using a general learning device to balance co-occurrence statistics with other information to be integrated into a larger control policy.

Conclusions and Future Work

Our experiments show that simple statistical information gathered from the unprocessed surface structure of large-scale text has value in guiding parsing decisions. However, we feel that there is still a great deal of further advantage to be gained from this approach. Our next step will be to include co-occurrence information from a much larger corpus, containing on the order of 10^8 characters.

We would also like to experiment with other definitions of co-occurrence. (Yuret, 1998) describes some very interesting work, in a different framework from ours, in which a parser using only co-occurrence mutual information was able to achieve a high precision but low recall when co-occurrence was defined as adjacent co-occurrence, and low precision but high recall when co-occurrence was defined as occurrence within the same sentence. We would like to experiment with ways of balancing these two measures.

We also suspect that significant gains are possible through a more sophisticated inclusion of the statistics in the decision making process. The current discretization scheme is very simple, but there is ample empirical evidence that discretization which takes into account target categories can significantly improve classification accuracy (Dougherty, Kohavi, and Sahami, 1995).

The several articles we have cited which use exclusively co-occurrence information to predict constituent boundaries are very interesting for the simplicity of their control structures, but in one important way they are more complex than the current work: they make decisions by explicitly comparing the measures of association between different pairs of words. We predict that augmenting the feature set to allow our parser to be sensitive to this kind of information would be a very valuable extension.

A related issue is the choice of learning methodology. The Winnow learner has served us well with its ability to handle very large feature sets, but it is weak in its ability to take advantage of the interaction between features. We would like to experiment with learning methods which do not suffer from this weakness, and with methods for automatic feature extraction which could supplement Winnow.

We experimented with a nondeterministic control policy for the parser, using cost-front search to find the most probable series of parsing decisions, but we found this not to be very useful. Over a series of comparative experiments, the non-deterministic control policy consistently raised precision by a small margin, lowered recall by a small margin, increased run times by an order of magnitude or more, and for about 10% of the test-set sentences exhausted system resources before finding any parse at all. We posit that these problems may in part be due to the fact that while the Winnow learner is otherwise quite well adapted for our purposes, its output is not intended to be interpreted probabilistically. In the future we intend to run parallel experiments with more probabilistically oriented learners; we

Measure of Association	Precision	Recall	Geometric Mean
Yule's Y	0.882	0.875	0.879
Mutual Information	0.885	0.857	0.871
Likelihood Ratio	0.879	0.845	0.862
X^2	0.870	0.848	0.859
T-score	0.870	0.836	0.853
None (Original Feature Set)	0.868	0.834	0.851

Table 4. Accuracy Measurements of Parsing with Different Measures of Association

are especially interested in experimenting with a Maximum Entropy model.

In the larger context, we plan to experiment with more sophisticated, model-based unsupervised learning methods, including clustering and beyond, and ways of providing their gathered knowledge to the parser, to make the fullest possible use of the vast wealth of un-annotated text available.

Acknowledgements

The research was supported by National Natural Science Foundation of China (NSFC) (Grant No. 69903007) and National 973 Foundation (Grant No. G1998030507-2).

References

Steven Abney (1994) *Parsing by Chunks.* http://www.sfs.phil.uni-tuebingen.de/~abney/

Eric Brill, David Magerman, Mitch Marcus and B. Santorini (1990) *Deducing Linguistic Structure from the Statistics of Large Corpora.* Proceedings of the DARPA Speech and Natural Language Workshop, pp. 275-281.

Kenneth W. Church and P. Hanks (1990) *Word Association Norms, Mutual Information, and Lexicography.* Computational Linguistics, Computational Linguistics, 16/1, pp. 22-29.

Ted Dunning (1993) *Accurate Methods for the Statistics of Surprise and Coincidence.* Computational Linguistics, Computational Linguistics, 19/1, pp. 61-74.

Kyo Kageura (1999) *Bigram Statistics Revisited: A Comparative Examination of Some Statistical Measures in Morphological Analysis of Japanese Kanji Sequences.* Journal of Quantitative Linguistics, 6/22, pp. 149-166.

James Dougherty, Ron Kohavi and Mehran Sahami (1995) *Supervised and Unsupervised Discretization of Continuous Features.* In "Machine Learning: Proceedings of the Twelfth International Conference", Morgan Kaufmann Publishers.

William A. Gale and Geoffrey Sampson (1995) *Good-Turing Frequency Estimation without Tears.* Journal of Quantitative Linguistics, 2, pp. 217-237.

Christopher D. Manning and Hinrich Shütze (1999) *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, Massachusetts.

David M. Magerman (1995) *Natural Language Parsing as Statistical Pattern Recognition.* Ph.D. Dissertation, Stanford University.

David Magerman and Mitch Marcus (1990) *Parsing a Natural Language Using Mutual Information Statistics.* In "Proceedings, Eighth National Conference on Artificial Intelligence (AAAI 90)".

Adwait Ratnaparkhi (1997) *A Linear Observed-Time Statistical Parser Based on Maximum Entropy Models.* In "Proceedings of the Second Conference on Empirical Methods in Natural Language Processing".

Dan Roth (1998) *Learning to Resolve Natural Language Ambiguities, a Unified Approach.* In "AAAI'98".

K-Y. Su, M-W. Wu, and J-S. Chang (1994) *A Corpus-Based Approach to Automatic Compound Extraction.* In "Proceedings of the 32nd Annual Meeting of the ACL", pp. 27-30.

Maosong Sun, Dayang Shen, and Benjamin K. Tsou (1998) *Chinese Word Segmentation without Using Lexicon and Hand-Crafted Training Data.* In "Proceedings of the 36th Annual Meeting of the ACL", pp. 1265-1271.

Aboy Wong, Dekai Wu (1999) *Are Phrase Structured Grammars Useful in Statistical Parsing?.* In "Proceedings of the Fifth Natural Language Processing Pacific Rim Symposium", pp. 120-125.

Deniz Yuret (1998) *Discovery of Lexical Relations Using Lexical Attraction.* Ph.D. Thesis, Massachusetts Institute of Technology.

Qiang Zhou (1996) *Phrase Bracketing and Annotating on Chinese Language Corpus.* (in Chinese), Ph.D. Thesis, Beijing University.