In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 91–94, Lisbon, Portugal, 2000.

Inducing Syntactic Categories by Context Distribution Clustering

Alexander Clark
School of Cognitive and Computing Sciences
University of Sussex
alexc@cogs.susx.ac.uk

Abstract

This paper addresses the issue of the automatic induction of syntactic categories from unannotated corpora. Previous techniques give good results, but fail to cope well with ambiguity or rare words. An algorithm, context distribution clustering (CDC), is presented which can be naturally extended to handle these problems.

1 Introduction

In this paper I present a novel program that induces syntactic categories from comparatively small corpora of unlabelled text, using only distributional information. There are various motivations for this task, which affect the algorithms employed. Many NLP systems use a set of tags, largely syntactic in motivation, that have been selected according to various criteria. In many circumstances it would be desirable for engineering reasons to generate a larger set of tags, or a set of domain-specific tags for a particular corpus. Furthermore, the construction of cognitive models of language acquisition – that will almost certainly involve some notion of syntactic category – requires an explanation of the acquisition of that set of syntactic categories. The amount of data used in this study is 12 million words, which is consistent with a pessimistic lower bound on the linguistic experience of the infant language learner in the period from 2 to 5 years of age, and has had capitalisation removed as being information not available in that circumstance.

2 Previous Work

Previous work falls into two categories. A number of researchers have obtained good results using pattern recognition techniques. Finch and Chater (1992), (1995) and Schütze (1993), (1997) use a set of features derived from the co-occurrence statistics of common words together with standard clustering and information extraction techniques. For sufficiently frequent words this method produces satisfactory results. Brown et al. (1992) use a very large amount of data, and a well-founded information theoretic model to induce large numbers of plausible semantic and syntactic clusters. Both approaches have two flaws: they cannot deal well with ambiguity, though Schütze addresses this issue partially, and they do not cope well with rare words. Since rare and ambiguous words are very common in natural language, these limitations are serious.

3 Context Distributions

Whereas earlier methods all share the same basic intuition, i.e. that similar words occur in similar contexts, I formalise this in a slightly different way: each word defines a probability distribution over all contexts, namely the probability of the context given the word. If the context is restricted to the word on either side, I can define the context distribution to be a distribution over all ordered pairs of words: the word before and the word after. The context distribution of a word can be estimated from the observed contexts in a corpus. We can then measure the similarity of words by the similarity of their context distributions, using the Kullback-Leibler (KL) divergence as a distance function.

Unfortunately it is not possible to cluster based directly on the context distributions for two reasons: first the data is too sparse to estimate the context distributions adequately for any but the most frequent words, and secondly some words which intuitively are very similar

(Schütze's example is 'a' and 'an') have radically different context distributions. Both of these problems can be overcome in the normal way by using clusters: approximate the context distribution as being a probability distribution over ordered pairs of clusters multiplied by the conditional distributions of the words given the clusters :

$$p(< w_1, w_2 >) = p(< c_1, c_2 >)p(w_1|c_1)p(w_2|c_2)$$

I use an iterative algorithm, starting with a trivial clustering, with each of the K clusters filled with the kth most frequent word in the corpus. At each iteration, I calculate the context distribution of each cluster, which is the weighted average of the context distributions of each word in the cluster. The distribution is calculated with respect to the K current clusters and a further ground cluster of all unclassified words: each distribution therefore has $(K+1)^2$ parameters. For every word that occurs more than 50 times in the corpus, I calculate the context distribution, and then find the cluster with the lowest KL divergence from that distribution. I then sort the words by the divergence from the cluster that is closest to them, and select the best as being the members of the cluster for the next iteration. This is repeated, gradually increasing the number of words included at each iteration, until a high enough proportion has been clustered, for example 80%. After each iteration, if the distance between two clusters falls below a threshhold value, the clusters are merged, and a new cluster is formed from the most frequent unclustered word. Since there will be zeroes in the context distributions, they are smoothed using Good-Turing smoothing(Good, 1953) to avoid singularities in the KL divergence. At this point we have a preliminary clustering – no very rare words will be included, and some common words will also not be assigned, because they are ambiguous or have idiosyncratic distributional properties.

4 Ambiguity and Sparseness

Ambiguity can be handled naturally within this framework. The context distribution $p^{(w)}$ of a particular ambiguous word w can be modelled as a linear combination of the context distributions of the various clusters. We can find the mixing coefficients by minimising $D(p^{(w)}||\sum \alpha_i^{(w)} q_i)$ where the $\alpha_i^{(w)}$ are some coefficients that sum to unity and the q_i are the context distributions of the clusters. A minimum of this function can be found using the EM algorithm(Dempster et al., 1977). There are often several local minima – in practice this does not seem to be a major problem.

Note that with rare words, the KL divergence reduces to the log likelihood of the word's context distribution plus a constant factor. However, the observed context distributions of rare words may be insufficient to make a definite determination of its cluster membership. In this case, under the assumption that the word is unambiguous, which is only valid for comparatively rare words, we can use Bayes's rule to calculate the posterior probability that it is in each class, using as a prior probability the distribution of rare words in each class. This incorporates the fact that rare words are much more likely to be adjectives or nouns than, for example, pronouns.

5 Results

I used 12 million words of the British National Corpus as training data, and ran this algorithm with various numbers of clusters (77, 100 and 150). All of the results in this paper are produced with 77 clusters corresponding to the number of tags in the CLAWS tagset used to tag the BNC, plus a distinguished sentence boundary token. In each case, the clusters induced contained accurate classes corresponding to the major syntactic categories, and various subgroups of them such as prepositional verbs, first names, last names and so on. Appendix A shows the five most frequent words in a clustering with 77 clusters. In general, as can be seen, the clusters correspond to traditional syntactic classes. There are a few errors – notably, the right bracket is classified with adverbial particles like "UP".

For each word w, I then calculated the optimal coefficents $\alpha_i^{(w)}$. Table 1 shows some sample ambiguous words, together with the clusters with largest values of α^i. Each cluster is represented by the most frequent member of the cluster. Note that "US" is a proper noun cluster. As there is more than one common noun cluster, for many unambiguous nouns the optimum is a mixture of the various classes.

Word	Clusters		
ROSE	CAME	CHARLES	GROUP
VAN	JOHN	TIME	GROUP
MAY	WILL	US	JOHN
US	YOU	US	NEW
HER	THE	YOU	
THIS	THE	IT	LAST

Table 1: Ambiguous words. For each word, the clusters that have the highest α are shown, if $\alpha > 0.01$.

Model	CDC	Brown	CDC	Brown
Freq	NN1	NN1	AJ0	AJ0
1	0.66	0.21	0.77	0.41
2	0.64	0.27	0.77	0.58
3	0.68	0.36	0.82	0.73
5	0.69	0.40	0.83	0.81
10	0.72	0.50	0.92	0.94
20	0.73	0.61	0.91	0.94

Table 2: Accuracy of classification of rare words with tags NN1 (common noun) and AJ0 (adjective).

Table 2 shows the accuracy of cluster assignment for rare words. For two CLAWS tags, AJ0 (adjective) and NN1(singular common noun) that occur frequently among rare words in the corpus, I selected all of the words that occurred n times in the corpus, and at least half the time had that CLAWS tag. I then tested the accuracy of my assignment algorithm by marking it as correct if it assigned the word to a 'plausible' cluster – for AJ0, either of the clusters "NEW" or "IMPORTANT", and for NN1, one of the clusters "TIME", "PEOPLE", "WORLD", "GROUP" or "FACT". I did this for n in $\{1, 2, 3, 5, 10, 20\}$. I proceeded similarly for the Brown clustering algorithm, selecting two clusters for NN1 and four for AJ0. This can only be approximate, since the choice of acceptable clusters is rather arbitrary, and the BNC tags are not perfectly accurate, but the results are quite clear; for words that occur 5 times or less the CDC algorithm is clearly more accurate.

Evaluation is in general difficult with unsupervised learning algorithms. Previous authors have relied on both informal evaluations of the plausibility of the classes produced, and more formal statistical methods. Comparison against existing tag-sets is not meaningful – one set of tags chosen by linguists would score very badly against another without this implying any fault as there is no 'gold standard'. I therefore chose to use an objective statistical measure, the perplexity of a very simple finite state model, to compare the tags generated with this clustering technique against the BNC tags, which uses the CLAWS-4 tag set (Leech et al., 1994) which had 76 tags. I tagged 12 million words of BNC text with the 77 tags, assigning each word to the cluster with the highest *a posteriori* probability given its prior cluster distribution and its context.

Test set	1	2	3	4	Mean
CLAWS	411	301	478	413	395
Brown et al.	380	252	444	369	354
CDC	372	255	427	354	346

Table 3: Perplexities of class tri-gram models on 4 test sets of 100,000 words, together with geometric mean.

I then trained 2nd-order Markov models (equivalently class trigram models) on the original BNC tags, on the outputs from my algorithm (CDC), and for comparision on the output from the Brown algorithm. The perplexities on held-out data are shown in table 3. As can be seen, the perplexity is lower with the model trained on data tagged with the new algorithm. This does not imply that the new tagset is better; it merely shows that it is capturing statistical significant generalisations. In absolute terms the perplexities are rather high; I deliberately chose a rather crude model without backing off and only the minimum amount of smoothing, which I felt might sharpen the contrast.

6 Conclusion

The work of Chater and Finch can be seen as similar to the work presented here given an independence assumption. We can model the context distribution as being the product of independent distributions for each relative position; in this case the KL divergence is the sum of the divergences for each independent distribution. This independence assumption is most clearly false when the word is ambiguous; this perhaps explains the poor performance of these algorithms with ambiguous words. The new algorithm currently does not use information about the orthography of the word, an important source of information. In future work, I will integrate this with a morphology-learning program. I am currently applying this approach to the induction of phrase structure rules, and preliminary experiments have shown encouraging results.

In summary, the new method avoids the limitations of other approaches, and is better suited to integration into a complete unsupervised language acquisition system.

References

Peter F. Brown, Vincent J. Della Pietra, Peter V. de Souza, Jenifer C. Lai, and Robert Mercer. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18:467–479.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, 39:1–38.

S. Finch and N. Chater. 1992. Bootstrapping syntactic categories. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society*, pages 820–825.

S. Finch, N. Chater, and Redington M. 1995. Acquiring syntactic information from distributional statistics. In Joseph P. Levy, Dimitrios Bairaktaris, John A. Bullinaria, and Paul Cairns, editors, *Connectionist Models of Memory and Language.* UCL Press.

I. J. Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:237–264.

G. Leech, R. Garside, and M Bryant. 1994. CLAWS4: the tagging of the British National Corpus. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 622–628.

Hinrich Schütze. 1993. Part of speech induction from scratch. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics*, pages 251–258.

Hinrich Schütze. 1997. *Ambiguity Resolution in Language Learning.* CSLI Publications.

A Clusters

Here are the five most frequent words in each of the 77 clusters, one cluster per line except where indicated with a double slash \\

```
THE A HIS THIS AN
PEOPLE WORK LIFE RIGHT END
OF IN FOR ON WITH \\ , &MDASH ( : ;
NEW OTHER FIRST OWN GOOD
&SENTENCE \\ . ? !
AND AS OR UNTIL SUCH␣AS
NOT BEEN N'T SO ONLY
IS WAS HAD HAS DID
MADE USED FOUND LEFT PUT
ONE ALL MORE SOME TWO
TIME WAY YEAR DAY MAN \\ TO
WORLD GOVERNMENT PARTY FAMILY WEST
BE HAVE DO MAKE GET
HE I THEY SHE WE
US BRITAIN LONDON GOD LABOUR
BUT WHEN IF WHERE BECAUSE
) UP OUT BACK DOWN
WILL WOULD CAN COULD MAY
USE HELP FORM CHANGE SUPPORT
THAT BEFORE ABOVE OUTSIDE BELOW
IT EVERYBODY GINA
GROUP NUMBER SYSTEM OFFICE CENTRE
YOU THEM HIM ME THEMSELVES
&BQUO \\ &EQUO \\ ARE WERE \\ 'S '
CHARLES MARK PHILIP HENRY MARY
WHAT HOW WHY HAVING MAKING
IMPORTANT POSSIBLE CLEAR HARD CLOSE
WHICH WHO
CAME WENT LOOKED SEEMED BEGAN
JOHN SIR DAVID ST DE
YEARS PER␣CENT DAYS TIMES MONTHS
GOING ABLE LOOKING TRYING COMING
THOUGHT FELT KNEW DECIDED HOPE
SEE SAY FEEL MEAN REMEMBER
SAID SAYS WROTE EXPLAINED REPLIED
GO COME TRY CONTINUE APPEAR \\ THERE
LOOK RUN LIVE MOVE TALK
SUCH USING PROVIDING DEVELOPING WINNING
TOOK TOLD SAW GAVE MAKES
HOWEVER OF␣COURSE FOR␣EXAMPLE INDEED
PART SORT THINKING LACK NONE
SOMETHING ANYTHING SOMEONE EVERYTHING
MR MRS DR HONG MR.
NEED NEEDS SEEM ATTEMPT OPPORTUNITY
WANT WANTED TRIED WISH WANTS
BASED RESPONSIBLE COMPARED INTERESTED
THAN \\ LAST NEXT GOLDEN FT-SE \\ THOSE
THINK BELIEVE SUPPOSE INSIST RECKON
KNOW UNDERSTAND REALISE
LATER AGO EARLIER THEREAFTER
BETTER WORSE LONGER BIGGER STRONGER
&HELLIP ..
ASKED LIKED WATCHED SMILED INVITED
'M AM \\ 'D
FACT IMPRESSION ASSUMPTION IMPLICATION
NOTHING NOWHERE RISEN
BECOME \\ ENOUGH \\ FAR INFINITELY
'LL \\ 'RE \\ 'VE \\ CA WO AI
COPE DEPEND CONCENTRATE SUCCEED COMPETE
RO HVK AMEN
KLERK CLOWES HOWE COLI GAULLE
NEZ KHMER
```