An Unsupervised Model for Statistically Determining Coordinate Phrase Attachment

Miriam Goldberg
Central High School &
Dept. of Computer and Information Science
200 South 33rd Street
Philadelphia, PA 19104-6389
University of Pennsylvania
miriamg@unagi.cis.upenn.edu

Abstract

This paper examines the use of an unsupervised statistical model for determining the attachment of ambiguous coordinate phrases (CP) of the form $n1\ p\ n2\ cc\ n3$. The model presented here is based on [AR98], an unsupervised model for determining prepositional phrase attachment. After training on unannotated 1988 Wall Street Journal text, the model performs at 72% accuracy on a development set from sections 14 through 19 of the WSJ TreeBank [MSM93].

1 Introduction

The coordinate phrase (CP) is a source of structural ambiguity in natural language. For example, take the phrase:

box of chocolates and roses

'Roses' attaches either high to 'box' or low to 'chocolates'. In this case, attachment is high, yielding:

H-attach: ((box (of chocolates)) (and roses))

Consider, then, the phrase:

salad of lettuce and tomatoes

'Lettuce' attaches low to 'tomatoes', giving:

L-attach: (salad (of ((lettuce) and (tomatoes)))

Previous work has used corpus-based approaches to solve the similar problem of prepositional phrase attachment. These have included backed-off [CB 95], maximum entropy [RRR94], rule-based [HR94], and unsupervised [AR98] models. In addition to these, a corpus-based model for PP-attachment [SN97] has been reported that uses information from a semantic dictionary.

Sparse data can be a major concern in corpus-based disambiguation. Supervised models are limited by the amount of annotated data available for training. Such a model is useful only for languages in which annotated corpora are available. Because an unsupervised model does not rely on such corpora it may be modified for use in multiple languages as in [AR98].

The unsupervised model presented here trains from an unannotated version of the 1988 Wall Street Journal. After tagging and chunking the text, a rough heuristic is then employed to pick out training examples. This results in a training set that is less accurate, but much larger, than currently existing annotated corpora. It is the goal, then, of unsupervised training data to be abundant in order to offset its noisiness.

2 Background

The statistical model must determine the probability of a given CP attaching either high (H) or low (L), $p(attachment \mid phrase)$. Results shown come from a development corpus of 500 phrases of extracted head word tuples from the WSJ TreeBank [MSM93]. 64% of these phrases attach low and 36% attach high. After further development, final testing will be done on a separate corpus.

The phrase:

(busloads (of ((executives) and (their wives)))

gives the 6-tuple:

L busloads of executives and wives

where, a = L, $n1$ = *busloads*, p = *of*, $n2$ = *executives*, cc = *and*, $n3$ = *wives*. The CP attachment model must determine a for all ($n1$ p $n2$ cc $n3$) sets. The attachment decision is correct if it is the same as the corresponding decision in the TreeBank set.

The probability of a CP attaching high is conditional on the 5-tuple. The algorithm presented in this paper estimates the probability:

$$\hat{p} = (a \mid n1, p, n2, cc, n3)$$

The parts of the CP are analogous to those of the prepositional phrase (PP) such that $\{n1, n2\} \equiv \{n, v\}$ and $n3 \equiv p$. [AR98] determines the probability $p(v, n, p, a)$. To be consistent, here we determine the probability $p(n1, n2, n3, a)$.

3 Training Data Extraction

A statistical learning model must train from unambiguous data. In annotated corpora ambiguous data are made unambiguous through classifications made by human annotators. In unannotated corpora the data themselves must be unambiguous. Therefore, while this model disambiguates CPs of the form ($n1$ p $n2$ cc $n3$), it trains from implicitly *unambiguous* CPs of the form (n cc n). For example:

dog and cat

Because there are only two nouns in the unambiguous CP, we must redefine its components. The first noun will be referred to as $n1$. It is analogous to $n1$ and $n2$ in the ambiguous CP. The second, terminal noun will be referred to as $n3$. It is analogous to the third noun in the ambiguous CP. Hence $n1$ = *dog*, cc = *and*, $n3$ = *cat*. In addition to the unambiguous CPs, the model also uses any noun that follows a *cc*. Such nouns are classified, n_{cc}.

We extracted 119629 unambiguous CPs and 325261 n_{cc}s from the unannotated 1988 Wall Street Journal. First the raw text was fed into the part-of-speech tagger described in [AR96][1]. This was then passed to a simple chunker as used in [AR98], implemented with two small regular expressions that replace noun and quantifier phrases with their head words. These head words were then passed through a set of heuristics to extract the unambiguous phrases. The heuristics to find an unambiguous CP are:

- ω_n is a coordinating conjunction (cc) if it is tagged cc.
- ω_{n-x} is the leftmost noun ($n1$) if :
  - ω_{n-x} is the first noun to occur within 4 words to the left of *cc*.
  - no preposition occurs between this noun and *cc*.
  - no preposition occurs within 4 words to the left of this noun.
- ω_{n+x} is the rightmost noun ($n2$) if:
  - it is the first noun to occur within 4 words to the right of *cc*.
  - No preposition occurs between cc and this noun.

The first noun to occur within 4 words to the right of *cc* is always extracted. This is n_{cc}. Such nouns are also used in the statistical model. For example, the we process the sentence below as follows:

> Several firms have also launched business subsidiaries and consulting arms specializing in trade, lobbying and other areas.

First it is annotated with parts of speech:

> Several_JJ firms_NNS have_VBP also_RB launched_VBN business_NN subsidiaries_NNS and_CC consulting_VBG arms_NNS specializing_VBG in_IN trade_NN ,_, lobbying_NN and_CC other_JJ areas_NNS ._.

From there, it is passed to the chunker yielding:

> firms_NNS have_VBP also_RB launched_VBN subsidiaries_NNS and_CC consulting_VBG arms_NNS specializing_VBG in_IN trade_NN ,_, lobbying_NN and_CC areas_NNS ._.

[1] Because this tagger trained on annotated data, one may argue that the model presented here is not *purely* unsupervised.

Noun phrase heads of ambiguous and unambiguous CPs are then extracted according to the heuristic, giving:

subsidiaries and arms
and areas

where the extracted unambiguous CP is $\{n1 = subsidiaries, cc = and, n3 = arms\}$ and *areas* is extracted as a n_{cc} because, although it is not part of an unambiguous CP, it occurs within four words after a conjunction.

4 The Statistical Model

First, we can factor $p(a, n1, n2, n3)$ as follows:

$$\begin{aligned} p(a, n1, n2, n3) &= p(n1)p(n2) \\ &\star\ p(a \mid n1, n2) \\ &\star\ p(n3 \mid a, n1, n2) \end{aligned}$$

The terms $p(n1)$ and $p(n2)$ are independent of the attachment and need not be computed. The other two terms are more problematic. Because the training phrases are unambiguous and of the form ($n1$ *cc* $n2$), $n1$ and $n2$ of the CP in question never appear together in the training data. To compensate we use the following heuristic as in [AR98]. Let the random variable ϕ range over $\{true, false\}$ and let it denote the presence or absence of any $n3$ that unambiguously attaches to the $n1$ or $n2$ in question. If $\phi = true$ when any $n3$ unambiguously attaches to $n1$, then $p(\phi = true \mid n1)$ is the conditional probability that a particular $n1$ occurs with an unambiguously attached $n3$. Now $p(a \mid n1, n2)$ can be approximated as:

$$p(a = \mathrm{H} \mid n1, n2) \approx \frac{p(true \mid n1)}{Z(n1, n2)}$$

$$p(a = \mathrm{L} \mid n1, n2) \approx \frac{p(true \mid n2)}{Z(n1, n2)}$$

where the normalization factor, $Z(n1, n2) = p(true \mid n1) + p(true \mid n2)$. The reasoning behind this approximation is that the tendency of a CP to attach high (low) is related to the tendency of the $n1$ ($n2$) in question to appear in an unambiguous CP in the training data.

We approximate $p(n3 \mid a, n1, n2)$ as follows:

$$p(n3 \mid a = \mathrm{H}, n1, n2) \approx p(n3 \mid true, n1)$$

$$p(n3 \mid a = \mathrm{L}, n1, n2) \approx p(n3 \mid true, n2)$$

The reasoning behind this approximation is that when generating $n3$ given high (low) attachment, the only counts from the training data that matter are those which unambiguously attach to $n1$ ($n2$), i.e., $\phi = true$. Word statistics from the extracted CPs are used to formulate these probabilities.

4.1 Generate ϕ

The conditional probabilities $p(true \mid n1)$ and $p(true \mid n2)$ denote the probability of whether a noun will appear attached unambiguously to some $n3$. These probabilities are estimated as:

$$p(true \mid n1) = \begin{cases} \frac{f(n1, true)}{f(n1)} & \text{if } f(n1, true) > 0 \\ .5 & \text{otherwise} \end{cases}$$

$$p(true \mid n2) = \begin{cases} \frac{f(n2, true)}{f(n2)} & \text{if } f(n2, true) > 0 \\ .5 & \text{otherwise} \end{cases}$$

where $f(n2, true)$ is the number of times $n2$ appears in an unambiguously attached CP in the training data and $f(n2)$ is the number of times this noun has appeared as either $n1$, $n3$, or n_{cc} in the training data.

4.2 Generate n3

The terms $p(n3 \mid n1, true)$ and $p(n3 \mid n2, true)$ denote the probabilies that the noun $n3$ appears attached unambiguously to $n1$ and $n2$ respectively. Bigram counts are used to compute these as follows:

$$p(n3 \mid true, n1) = \begin{cases} \frac{f(n1, n3, true)}{f(n1, true)} & \text{if } f(n1,n3,true)>0 \\ \frac{1}{N} & \text{otherwise} \end{cases}$$

$$p(n3 \mid true, n2) = \begin{cases} \frac{f(n2, n3, true)}{f(n2, true)} & \text{if } f(n2,n3,true)>0 \\ \frac{1}{N} & \text{otherwise} \end{cases}$$

where N is the set of all $n3$s and n_{cc}s that occur in the training data.

5 Results

Decisions were deemed correct if they agreed with the decision in the corresponding Tree-Bank data. The correct attachment was chosen

72% of the time on the 500-phrase development corpus from the WSJ TreeBank. Because it is a forced binary decision, there are no measurements for recall or precision. If low attachment is always chosen, the accuracy is 64%. After further development the model will be tested on a testing corpus.

When evaluating the effectiveness of an unsupervised model, it is helpful to compare its performance to that of an analogous supervised model. The smaller the error reduction when going from unsupervised to supervised models, the more comparable the unsupervised model is to its supervised counterpart. To our knowledge there has been very little if any work in the area of ambiguous CPs. In addition to developing an unsupervised CP disambiguation model, In [MG, in prep] we have developed two supervised models (one backed-off and one maximum entropy) for determining CP attachment. The backed-off model, closely based on [CB95] performs at 75.6% accuracy. The reduction error from the unsupervised model presented here to the backed-off model is 13%. This is comparable to the 14.3% error reduction found when going from [AR98] to [CB95].

It is interesting to note that after reducing the volume of training data by half there was no drop in accuracy. In fact, accuracy remained exactly the same as the volume of data was increased from half to full. The backed-off model in [MG, in prep] trained on only 1380 training phrases. The training corpus used in the study presented here consisted of 119629 training phrases. Reducing this figure by half is not overly significant.

6 Discussion

In an effort to make the heuristic concise and portable, we may have oversimplified it thereby negatively affecting the performance of the model. For example, when the heuristic came upon a noun phrase consisting of more than one consecutive noun the noun closest to the *cc* was extracted. In a phrase like *coffee and rhubarb apple pie* the heuristic would chose *rhubarb* as the $n3$ when clearly *pie* should have been chosen. Also, the heuristic did not check if a preposition occurred between either $n1$ and *cc* or *cc* and $n3$. Such cases make the CP ambiguous thereby invalidating it as an unambiguous training example. By including annotated training data from the TreeBank set, this model could be modified to become a partially-unsupervised classifier.

Because the model presented here is basically a straight reimplementation of [AR98] it fails to take into account attributes that are specific to the CP. For example, whereas $(n1\ cc\ n3) \equiv (n3\ cc\ n1)$, $(v\ p\ n) \not\equiv (n\ p\ v)$. In other words, there is no reason to make the distinction between "dog and cat" and "cat and dog." Modifying the model accordingly may greatly increase the usefulness of the training data.

7 Acknowledgements

We thank Mitch Marcus and Dennis Erlick for making this research possible, Mike Collins for his guidance, and Adwait Ratnaparkhi and Jason Eisner for their helpful insights.

References

[CB95] M. Collins, J. Brooks. 1995. Prepositional Phrase Attachment through a Backed-Off Model, *ACL 3rd Workshop on Very Large Corpora*, Pages 27-38, Cambridge, Massachusetts, June.

[MG, in prep] M. Goldberg. in preparation. Three Models for Statistically Determining Coordinate Phrase Attachment.

[HR93] D. Hindle, M. Rooth. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19(1):103-120.

[MSM93] M. Marcus, B. Santorini and M. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: the Penn Treebank, *Computational Linguistics*, 19(2):313-330.

[RRR94] A. Ratnaparkhi, J. Reynar and S. Roukos. 1994. A Maximum Entropy Model for Prepositional Phrase Attachment, *In Proceedings of the ARPA Workshop on Human Language Technology*, 1994.

[AR96] A. Ratnaparkhi. 1996. A Maximum Entropy Part-Of-Speech Tagger. *In Proceedings of the Empirical Methods in Natural Language Processing Conference*, May 17-18.

[AR98] A. Ratnaparkhi. 1998. Unsupervised Statistical Models for Prepositional Phrase Attachment, *In Proceedings of the Seventeenth International Conference on Computational Linguistics*, Aug. 10-14, Montreal, Canada.

[SN97] J. Stetina, M. Nagao. 1997. Corpus Based PP Attachment Ambiguity Resolution with a Semantic Dictionary. In Jou Shou and Kenneth Church, editors, *Proceedings of the Fifth Workshop on Very Large Corpora*, pages 66-80, Beijing and Hong Kong, Aug. 18-20.