[International Workshop on Fundamental Research for the Future Generation of Natural Language Processing (FGNLP), July 1992, Manchester]

Tagging an Unfamiliar Text With Minimal Human Supervision

Eric Brill and Mitch Marcus *
Department of Computer and Information Science
University of Pennsylvania
Philadelphia, Pennsylvania 19104
U.S.A.
brill@unagi.cis.upenn.edu mitch@unagi.cis.upenn.edu

Abstract

In this paper, we will discuss a method for tagging an unannotated text corpus whose structure is completely unknown, with a little bit of help from an informant. Starting from scratch, automated and semi-automated methods are employed to build a part of speech tagger for the text. There are three steps to building the tagger: uncovering a set of part of speech tags, discovering for each word its most likely tag, and learning rules to both correct mistakes in the dictionary and discover where contextual information can repair tagging mistakes. The long term goal of this work is to create a system which would enable somebody to take a text in a language he/she does not know, and with only minimal help from a speaker of the language (a couple of hours), accurately annotate the text with part of speech information.

Keywords: Automated Corpus Annotation

1 Introduction

Much of language learning involves the discovery and classification of linguistic entities, and learning the relationships that hold between objects of different classes. We are designing automata that are able to learn some of the classificatory aspects of human language with little or no human guidance. In particular, work is being carried out on a set of computer programs that take a large corpus of text as input and from observing regularities in the corpus they are able to learn information about morphology, word classes, and phrase structure[1].

The main tool used in this work to deduce linguistic information from large corpora is distributional analysis. We adopt many ideas originally proposed by Zellig Harris [Harris 51,

*This work was supported by DARPA and AFOSR jointly under grant No. AFOSR-90-0066, and by ARO grant No. DAAL 03-89-C0031 PRI.

[1] The system will be referred to as MTL for Mechanical Text Learner.

Harris 62, Harris 91]. Harris proposed a number of procedures that use the distributional behavior of relevant linguistic entities to uncover structural information about languages. The language learning procedures in MTL, although operating on different structural levels, all share one thing in common. They all learn structure using the tool of distributional analysis where the distributional behavior of an element is estimated from its behavior in a large corpus. Distributional analysis takes place over local and astructural environments. This means that environments such as *seven words to the right, subject of sentence* and *leftmost daughter of the phrasal head* are ruled out. Disallowing nonlocal environments constrains the set of possible environments that need to be considered when carrying out a distributional analysis. Disallowing structural environments also greatly constrains the set of possible environments, as well as ensuring that information about environments can readily and reliably be extracted from any large text corpus.

The figure below lays out the general framework under which this research is being carried out. The system begins in a language-naive start state. From the start state it is given a corpus of text as input and arrives at an end state. The end state is a level of language maturity which allows for morphological analysis, part of speech tagging and phrase structure analysis.

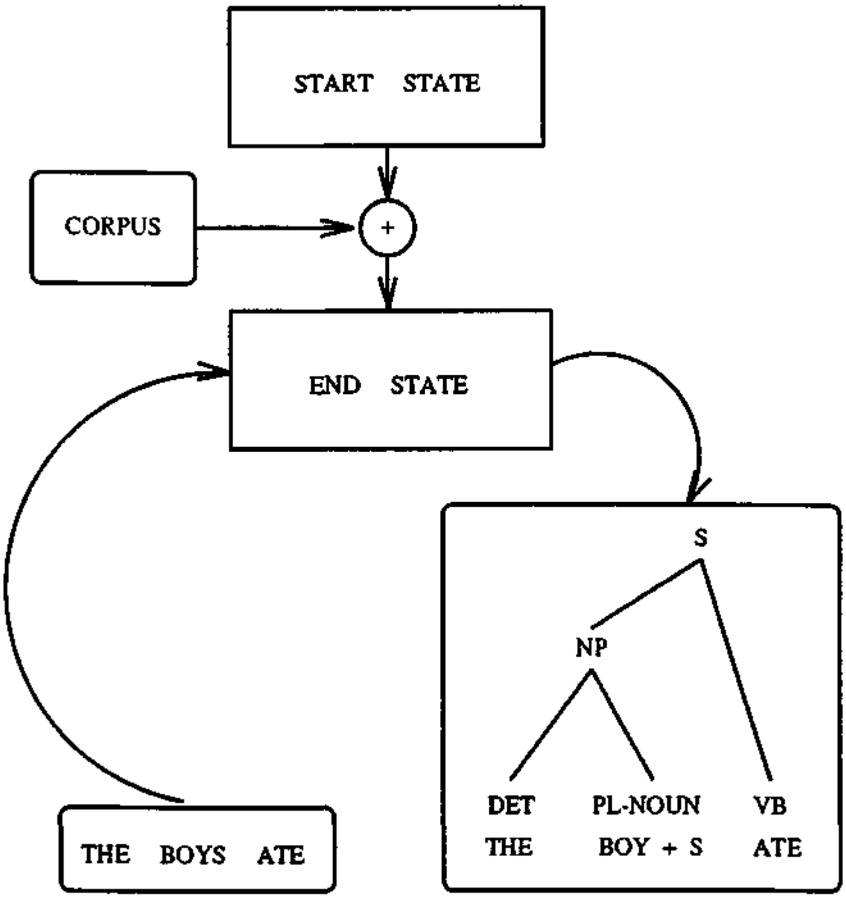


Figure 1: General Framework

For a description of the morphology module, see [Brill 2b]. A description of word class discovery can be found in [Brill et al 1990, Brill 1991]. Phrase structure learning is described in [Brill and Marcus 92].

In this paper, we will describe one module of MTL in detail. This module annotates text with part of speech labels. Prior to receiving the unannotated text as input, no information about the lexicon or about the syntax is known. Tagging information is then extracted from the corpus using distributional analysis. In its current incarnation the module allows us to take a text in a language we do not know, and with only a couple of hours of help from an informant familiar with the language of the text, we can build a tagger which is able to tag the text with fairly high accuracy. Note that we cannot even prespecify a tag set for the unfamiliar corpus. In a constrained corpus of computer manuals, a semantic tag such as *Computer Name* might be appropriate, whereas such a tag would not be appropriate in a less constrained corpus such as the Brown Corpus [Francis and Kučera 1982].

2 Tagging Unfamiliar Text

This work was motivated by a real problem. We have access to a large number of text corpora in different languages, and we wished to apply our language learning techniques to these corpora. However, some of the techniques need tagged text. None of the foreign language corpora are tagged. Not only were the corpora in foreign languages, but even the general content of many of the corpora was unknown to us. We wished to design a system to first find appropriate part of speech tags and then tag the text[2]. A completely automatic system would have been ideal, but this being out of reach we settled for a system that requires minimal supervision from a speaker of the language being processed.

There are a number of stages in tagging each unfamiliar corpus. First, a set of part of speech tags must be found by observing the distributional behavior of words in the text. Second, the most probable part of speech for all words which occur with sufficient probability is determined. For instance, although *can* can have a number of different parts of speech in English, we need only determine at this point that *modal* is the most likely part of speech for *can.* A dictionary is built containing each word, and only the most likely part of speech for that word. Low frequency words are not entered into the dictionary at this point. This is because the method employed for discovering the most likely tag for a word relies upon distributional analysis, and there will not be sufficient information about the distributional behavior of low frequency words. After the dictionary is built, a small amount of text is tagged by assigning each word its most likely tag. It is known that if every word in a corpus is tagged with its most probable tag, accuracy of about 90% can be obtained (see [Brill 2a]). Since the dictionary will not be completely accurate, our accuracy will be less than 90%. This small tagged text is then hand-corrected by our informant[3]. Rules are then learned automatically to correct errors in the dictionary and to correct cases where strong contextual cues indicate that a word is tagged incorrectly. The final step involves extracting

[2] Many taggers have been built that achieve high accuracy (95-97% of words tagged with their appropriate part of speech)[Brill 2a, Church 88, Cutting et al 92, Derose 88], but these taggers need to be trained on large amounts of tagged text and/or large dictionaries.

[3] The text that needed to be corrected had fewer than 8,000 words in our experiment.

suffix information on the words in the dictionary to allow us to assign a part of speech to words that occur infrequently. Each of these four steps will now be described in detail. As a pilot experiment, the programs were run on the Brown Corpus, a corpus of about 1.1 million words containing text from various different genres of written English.

2.1 Finding a Set of Part of Speech Tags

A number of proposals have been suggested for automatically finding word classes in a corpus based upon the distributional similarity of words [Brill et al 1990, Brown et al. 90, Brill 1991]. These methods work by defining a measure of distributional similarity, and then clustering words into classes. We propose using these clustering methods to semi-automatically find word classes. In our current approach, only the 300 most frequently occurring words in the corpus are considered. The point of this clustering is not to find a class for every word, but only to try to elicit the salient word classes in the corpus. For each of these 300 words, we estimate from the corpus the probabilities of all other words in the corpus immediately preceding or following. If a word pair is seen fewer than three times, we consider its probability to be zero. Of all word bigrams in the Brown Corpus, only 14.5% occur with frequency greater than two. We have found that ignoring low frequency bigrams, while greatly reducing the computation time, does not affect the accuracy of word class formation. The vectors of adjacent word probabilities for each pair of the 300 words are then compared.

The divergence of the probability vectors is computed. Let P_1 and P_2 be two probability distributions over environments. The relative entropy between P_1 and P_2 is:

$$D(P_1||P_2) = \sum_{x \in Environments} P_1(x) * log\frac{P_1(x)}{P_2(x)}$$

Relative entropy $D(P_1||P_2)$ is a measure of the amount of extra information beyond P_2 needed to describe P_1. The *divergence* between P_1 and P_2 is defined as $D(P_1||P_2) + D(P_2||P_1)$, and is a measure of how difficult it is to distinguish between the two distributions. Below are the thirty word pairs from the 300 most frequently occurring words in the Brown Corpus deemed distributionally most similar according to the divergence of their distributional probability vectors.

HE SHE
WE THEY
GIVE MAKE
ME HIM
IF WHEN
GET TAKE
FIND MAKE
THEM HIM

COULD CAN
BUT ALTHOUGH
WHILE ALTHOUGH
KIND NUMBER
FIND TAKE
ALTHOUGH SINCE
GET MAKE
WHEN ALTHOUGH

IF THOUGH
MAKE TAKE
GIVE TAKE
MEN PEOPLE
FACE HEAD
GET FIND
SENSE KIND
COULD WOULD
MADE FOUND
MEN CHILDREN
MUST SHOULD
US THEM
CAME WENT
GIVE GET
TIME DAY
MIGHT MAY

We can use the distributional similarity scores to build a similarity tree for all 300 words. Begin with every word in its own class. Take the two classes that are most similar and merge them. Continue this until all words are merged into one class. Different areas of the tree will correspond to different word classes. This tree can be used to choose a tag set for the corpus. The informant is shown the tree and asked to find and name meaningful sections of the tree. Using the similarity tree can help uncover both syntactic and semantic word classes.

2.2 Finding the Most Likely Tag for Each Word

Let us now assume that we have settled upon a tag set[4]. The next step is to determine, for each word of sufficient frequency, the word's most likely tag. We carry this step out only on words that occur with sufficient frequency to allow for a reasonable approximation of the word's distributional behavior.

For each word class, we first ask the informant to indicate which classes he/she can easily and quickly enumerate all members of. From the Penn Treebank, some such classes are TO (only applies to the word *to)* and punctuation tags. For the classes whose members are not easily enumerated, the informant is asked to choose a small number of words that he/she considers to be good exemplars for that class of words[5]. In the experiment we ran, the number of exemplars chosen ranged from 4 to 9. The *open* classes in our experiment were: adjective, adverb, determiner, modal, noun, preposition and verb. From these exemplar lists, a *distributional fingerprint is* created for each word class: the probability distribution for words preceding and following any of the words in the word class exemplar list.

Once these fingerprints are formed, words are assigned to the class whose fingerprint most closely matches their own. The measure used to judge fingerprint similarity is divergence, which is described above.

The accuracy of this method was tested as follows. The entire Brown Corpus was used to estimate distributional information. A word list was made from a 15,607 word sample of random sentences from the corpus which consisted of all words not covered by the very small class tags and which did not occur on any exemplar list. Words that occurred fewer

[4] For this pilot experiment, we based our tag set upon the Penn Treebank tag set to make it easier to quantify our results.

[5] This information can also be taken from the similarity tree.

than three times in the random sample were removed from the word list[6]. Each word on the word list was then compared to each word class fingerprint, and was assigned to the class it most closely resembled. To allow us to compare our results to a gold standard, we compared our answers to those deemed correct according to the Penn Treebank version of the Brown Corpus, with their tag set reduced to ours. Fine-grained class distinctions were removed. For instance, the classes noun, proper noun, plural noun, plural proper noun and pronoun were all mapped down to the tag *noun.* The precise mapping is shown below.

Tag	*Tags Mapped Down To This Tag*
Modal	MD
Determiner	DT WDT
Noun	NN NP NNS NPS PP
Verb	VBN VB VBD VBZ VBP VBG
Adjective	JJ JJR JJS
Adverb	RB RBR WRB
Preposition	IN RP

When we attempt to assign each word to its most probable class, the accuracy by type is 66%. In other words, 66% of the words on the word list were assigned their proper class. Of the errors, 56% are from classifying words which should be nouns as adjectives. If instead we consider tokens and not types, things look better. Of the total number of tokens that occur in the 15,607 word sample and appear on the word list, 76% are tagged with their most probable tag. Considering all types that occur with frequency greater than two in the 15,607 word sample, about 84% of these, by token, are tagged with their most probable tag.

Now let us consider tagging words with the correct tag for their context. While *modal* is the most probable tag for can, in the sentence *kick the can, can* should be tagged as a noun. Once again considering only words that occur with frequency greater than two, if each of these words is tagged everywhere in the small test corpus with its most probable tag, an accuracy of 84% is obtained. We would now like to consider improving upon this performance.

2.3 Learning Rules to Improve Accuracy

In [Brill 2a], we describe a rule-based part of speech tagger. This tagger works by first tagging every word with its most probable part of speech and then automatically learning a small set of contextual rules to improve tagging performance. The tagger has a small set of rule templates. Templates are of the form:

- If a word is tagged a and it is in context C, then change that tag to b, or

[6] About 20% of the total tokens.

- If a word is tagged a and it has lexical property P, then change that tag to b, or
- If a word is tagged a and a word in region R has lexical property P, then change that tag to b.

At each stage of learning, a small sample text is tagged. This tagged text is then compared to the correct tagging, and the rule template instantiation whose application would result in the greatest error reduction in the sample text is added to the tagger's list of rules. The rule is applied to improve the tagging accuracy of the sample text, and rule learning continues on the corrected sample text. Rule discovery is completely automated. Some example rules learned by the tagger using the original Brown Corpus tag set were:

(1) Change a tag from infinitival-to to preposition if the next word is tagged as a determiner.
(2) Change a tag from verb to noun if one of the two previous words is tagged as a determiner.
(3) Change a tag from noun to verb if the previous word is tagged as infinitival-to.
(4) Change a tag from subject pronoun to object pronoun if the next word is tagged as a period.

With fewer than 100 such rules, performance comparable to stochastic taggers was obtained.

We used the rule-based tagger to improve upon the system's accuracy at tagging an unfamiliar text. In our experiment, we used about 10,000 words of text to train the tagger, and a separate 5,600 to test. We tagged the 10,000 words by mapping the proper tag indicated in the Penn Treebank to its reduced tag. In real use, the informant would have to tag the 10,000 word sample. However, we are only attempting to tag the higher frequency words, and so at this point about 20% of the 10,000 words are tagged as *unknown,* and need not be tagged by the informant. To tag the rest of the text, the informant need not start from scratch. This text can first be tagged by assigning every word its most probable tag from the dictionary built in the previous phase.

The rule-based tagger learned 117 rules[7]. The tagger was then tested on a separate 5,600 word sample. Initially, tagging accuracy was 84%, ignoring all words labelled as *unknown.* After the rules were applied to the testing corpus, the accuracy was increased to 94%. Below are a few of the rules that were learned. Rules are of two types: correcting *most likely tag* mistakes for words, and using contextual information to tag more accurately.

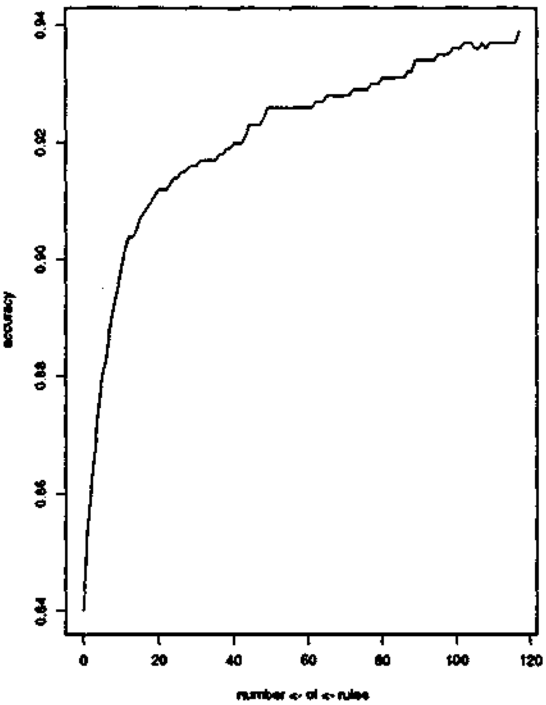
(1) Change a tag from adjective to noun if one of the previous two words is tagged as a determiner.

[7] Rules are not learned for tagging words labelled as *unknown*.

(2) Change a tag from modal to verb if the word is *had*.
(3) Change a tag from adverb to preposition if the word is *as*.
(4) Change a tag from noanswer to verb if one of the previous two words is tagged as a noun.

Rule (1) helps correct the most prevalent mistake of the system: tagging nouns as adjectives. Rules (2-3) correct for mistakes made in automatically learning the most probable tag for a word. During the discovery of most probable tags by comparing each word to distributional fingerprints, some words are assigned the tag *noanswer*. Distributional fingerprints only make use of bigrams that occur at least three times in the corpus. It is possible for a word to occur three or more times in the corpus (our criterion for classifying a word), but no bigrams containing that word have sufficiently high frequency. If this is the case, words have no distributional fingerprint, and are incorrectly assigned to the class *noanswer*. Rule (4) is an attempt to properly tag some such words, by guessing that a word is likely a verb if it follows a noun.

Below is a graph showing the improvement in tagging performance as rules were applied to the test corpus.



2.4 Tagging Low Frequency Words

The final phase involves tagging low frequency words. Low frequency words present a difficulty, since there is not enough distributional information to get reliable distributional cues about their parts of speech. Part of speech taggers have been built that are fairly accurate at tagging words not seen in the training corpus. They do so using morphological information. Since we do not want any language-specific information hard-coded in the learner, we cannot prespecify a set of affixes that indicate a particular part of speech. Instead, we use the 10,000 word accurately tagged text to compute for all three letter strings the most likely tag for a word ending in those three letters. Testing on the words marked *unknown* in the test corpus, tagging every word according to its suffix results in a 79.5% accuracy. Words for which no suffix information is available are tagged as nouns by default. The default tag can be determined by asking the informant the part of speech of a small number of low frequency words, and then determining the most frequent part of speech among these words.

The *unknown* category accounted for 22% of the words in the test corpus. Therefore, the accuracy in tagging the entire test corpus is (probability of low frequency word * accuracy in tagging low frequency word) + (probability of high frequency word * accuracy in tagging high frequency word):

$$.22 * 79.5 + .78 * 93.9 = 90.7\%$$

2.5 Improvements

Our pilot experiment shows that the approach outlined above holds promise as a way to accurately tag an unfamiliar corpus of text with only minimal help from an informant. We are currently pursuing a number of approaches toward improving the system. For some uses, the coarse tag set used in our experiment may not be sufficient. One possible approach is to first tag all words with a tag from the coarse tag set and then decide how to break the coarse tag into a number of more restrictive tags. Once this is done, a distributional fingerprint can be built for each of the restrictive tags, and words can be assigned to their proper class. Also, the rule-based tagger may be able to learn rules to improve the accuracy of tagging low frequency words that are tagged according to their last three letters.

3 Rules For Correction

The system outlined above uses both rule-based and statistical techniques. We have combined the two methods by first using statistical techniques to extract information from a corpus, and then using a program to automatically learn rules that can patch up mistakes made by the statistical techniques. Learning correcting rules can be an effective approach when the distribution of errors somewhat follows Zipf's Law [Zipf 49]. If Zipf's Law is obeyed, then a small number of high probability error types will account for a large percentage of total error tokens. Such a distribution is amenable to the rule-based correction approach: the fact that there is a small number of high probability error types ensures that such errors can easily be uncovered in a small sample, and the fact that these errors account for a high percentage of total error tokens will ensure that remedying these errors will result in significant system performance improvement.

References

[Brill et al 1990] Brill, Eric; Magerman, David; Marcus, Mitchell P.; and Santorini, Beatrice. Deducing Linguistic Structure from the Statistics of Large Corpora. In *Proceedings of the DARPA Speech and Natural Language Workshop, June 1990,* pages 275-282.

[Brill 1991] Brill, Eric. Discovering the Lexical Features of a Language. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, CA. (Student Session).* 1991.

[Brill and Marcus 92] Brill, E. and Marcus, M. Automatically Acquiring Phrase Structure Using Distributional Analysis. In *Darpa Workshop on Speech and Natural Language,* Harriman, N.Y. 1992.

[Brill 2a] Brill, E. A Simple Rule-Based Part of Speech Tagger. *Proceedings of the Third Conference on Applied Natural Language Processing, ACL,* Trento, Italy. A longer version of this paper appears in the *Proceedings from the 1992 Darpa Workshop on Speech and Natural Language.* 1992.

[Brill 2b] Brill, E. A Distributional Analysis Approach to Language Learning. Dissertation Proposal, University of Pennsylvania. Philadelphia, Pa. 1992.

[Brown et al. 90] Brown, P., Della Pietra, V., Della Pietra, S. and Mercer, R. Class-Based n-gram Models of Natural Language. In *Proceedings of the IBM Natural Language ITL, pp. 283-298,* Paris, France. 1990.

[Church 88] Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. In *Proceedings of the Second Conference on Applied Natural Language Processing, ACL,* 136-143, 1988.

[Cutting et al 92] Cutting, D., Kupiec, J., Pederson, J. and Sibun, P. A Practical Part-of-Speech Tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing, ACL,* 1992.

[Derose 88] DeRose, S.J. Grammatical Category Disambiguation by Statistical Optimization. *Computational Linguistics* 14: 31-39, 1988.

[Francis and Kučera 1982] Francis, W. Nelson and Kučera, Henry. *Frequency analysis of English usage. Lexicon and grammar* (Boston: Houghton Mifflin). 1982.

[Harris 51] Harris, Zellig. *Structural Linguistics.* Chicago: University of Chicago Press. 1951.

[Harris 62] *String Analysis of Language Structure.* The Hague: Mouton & Co. 1962.

[Harris 91] Harris, Zellig. *A Theory of Language and Information.* Oxford: Clarendon Press. 1991.

[Zipf 49] Zipf, G. *Human Behavior and the Principle of Least Effort.* New York: Hafner Pub. Co. 1949.