The Non-Dictionary: Description and Evaluation of a Dictionaryless Semantic and Syntactic Tagger for Unrestricted English Text

Ezra W. Black , Stephen G. Eubank , 柏岡 秀紀
ATR 音声翻訳通信研究所

1 Introduction

This article presents the Non-Dictionary, a dictionaryless, decision-tree tagger; along with experimental results of runs of this tagger on the roughly-2800-tag, semantically- and syntactically-analyzed ATR/Lancaster Treebank [3]. In addition, we discuss evaluation of taggers, and show how we evaluated ours. In Section 1, we briefly reintroduce the reader to the problem of tagging, presenting new data on what we think are crucial aspects of the problem that have been overlooked or underappreciated. Section 2 describes the Non-Dictionary tagger. Section 3 discusses the evaluation of taggers, and the evaluation methodology we employ to measure our tagger's performance. In Section 4, we present and discuss experimental results for our tagger, using two different versions of the ATR English Tagset [3] , and, for old times' sake, using the UPenn WSJ Treebank and Tagset [8] as well.

2 The Task of Tagging

Tagging means automatically associating each word of previously-unseen text with a linguistically-descriptive label drawn from a set of such labels (a "tagset") reflecting some scheme of lexical analysis. Presently most tagging research utilizes 1960s-style tagsets[1], the most frequently-employed of these being the 45-tag UPenn Tagset[2]. The tagging research reported here uses two different versions of a 1990s-style tagset, the ATR English Tagset [3]: a 2800-tag version with full syntax and semantics ("ATR Full"), and a 440-tag version with full syntax and highly reduced semantics ("ATR Syntax").

This section concerns two aspects of the tagging task that we feel have been poorly understood so far, and which we take to motivate tagging approaches such as the dictionaryless, decision-tree Non-Dictionary tagger presented in the next section.

First is the problem of unknown words: words occurring in the test corpus but not in the training corpus. Table 1 shows mutual coverage statistics[3] for the ATR English Treebank [3] and the UPenn Wall Street Journal Treebank [8], each roughly a million words in length, and shows the extent to which ATR Treebank sentences are covered by the CUVOALD92 online dictionary[4], and by this dictionary together with the lexicon of the UPenn WSJ Treebank. Figure 1 shows, for the ATR and UPenn WSJ Treebanks, the percentage of test-corpus sentences containing one or more unknown words, given a training corpus from the same treebank. Together these data argue quantitatively for the severity of the unknown-word problem for anything like real-world tagging. For instance, even for the UPenn WSJ Treebank, one in three sentences of test data contains an unknown word.

Second is the problem of words with "new tags": words occurring in both the test and training corpora, but never, in the training corpus, with the tag it receives in the test corpus. Figure 1 shows, for the ATR and UPenn WSJ Treebanks, the percentage of test-corpus sentences containing one or more words with new tags, given a training corpus from the same treebank. This percentage is quite significant for ATR Full and ATR Syntax, and non-trivial even for UPenn. In a synactic-plus-semantic treebank in particular, the sheer variety

[1] tagsets based more or less closely on, and of roughly the same size as, the Brown Corpus tagset [6]

[2] [8]

[3] for meaningful, case-normalized, etc., words

[4] produced by Roger Mitton; available from: ftp://black.ox.ac.uk/ota/dicts/710

Covering Database	Covered Database	Category of Coverage	Coverage
UPenn WSJ Treebank	ATR Treebank	wordlist	75%
ATR Treebank	UPenn WSJ Treebank		75%
UPenn WSJ Treebank	ATR Treebank	running words	94%
ATR Treebank	UPenn WSJ Treebank		94%
UPenn WSJ Treebank	ATR Treebank	sentences	69%
CUVOALD92 Dictionary	ATR Treebank		60%
CUVOALD92 Dictionary + UPenn WSJ Treebank	ATR Treebank		80%

Table 1: Mutual Coverage Statistics For ATR and UPenn Treebanks

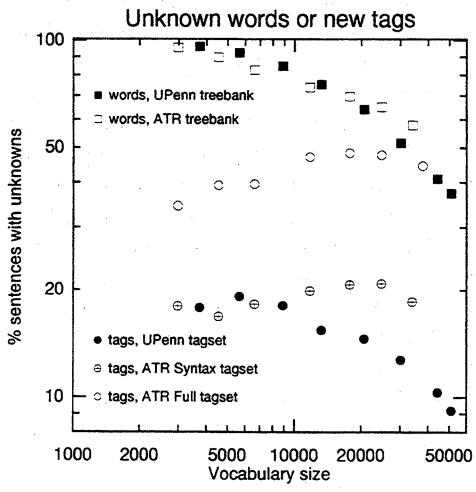

Figure 1: Percentage of sentences in ATR and UPenn test corpora with one or more unknown words or with one or more words having tags not used in the training set, as a function of training-set vocabulary size. Words consisting entirely of digits or punctuation are ignored. ATR training set, thus purged, contains 331,770 running words and a vocabulary of 33,946; UPenn, 885,010 and 51,064, respectively.

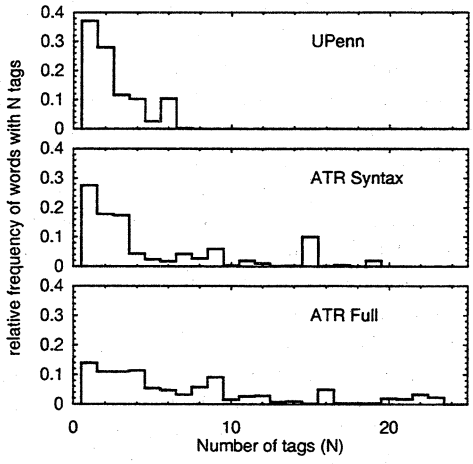

Figure 2: Histograms showing relative frequency of occurrence of words as function of number of distinct tags with which word is associated. Not shown: >25 (ATR–Full).

of meanings a word can take on argue the folly of banking on pre–set wordlists. Figure 2 shows the relative frequency of words with N tags, for the UPenn and ATR Treebanks.

3 The Non-Dictionary Tagger

With the Non–Dictionary tagger, we explore the possibility that, given the ubiquity of unknown words and new tags in real–world tagging, especially with semantic tagsets, it is not a winning move to rely on a dictionary as one's source of information about a given word to be tagged in context. The Non-Dictionary tagger uses proba-

bilistic decision trees to predict tags, building on the work in [1], but differs crucially from [1] in several respects.

The most crucial of these differences, especially in the present context, is that no use is made of a dictionary or dictionary lookup of any sort. Rather, we ask a large number of questions as a basis for predicting each tag: (a) questions about the word to be tagged: its substrings; its length; the semantic and/or syntactic categories it or words resembling it are associated with; (b) questions about other specific words in the sentence, such as the words immediately surrounding the word to be tagged; the first word of the sentence; the last word of the sentence; and (c) questions about the sentence as a whole, such as the number of commas or quotes it contains; its length; and the number of instances of the word being tagged in the sentence. These questions come from two sources: Some are derived automatically, from word–class-cluster information obtained by running the algorithm of [4] on tens of millions of words of Wall Street Journal text [9, 10]. Others are created by our team grammarian, utilizing a "question language" designed to permit one to navigate through a parse tree, and so to pose questions in terms of already–established structure; and then to allow the asking of any question about the node arrived at, or in fact about about any word or wordstring of the input sentence.

4 Evaluation of Tagging Output

In our view, any effective evaluation methodology for automatic tagging must confront head–on the problem of multiple correct answers in tagging. That is, it is often the case that there is more than one "correct tag" for a word in context, where that word could be considered to be functioning as: a proper or a common noun; an adjective or a noun; a participle or an adjective; a gerundial noun or a noun;[5] an adverbial particle or a locative adverb; and even an adjective or an adverb. This is true even where there are highly detailed and well–understood guidelines for the application of each tag to text.

Barring the recording of the set of correct tags for each word in a treebank, the next–best solution to the problem of multiple correct tags is to at least provide such a recording in one's test set, i.e. to provide a "gold standard" test set with all correct tags for each word in context. This is the solution we adopted in creating the ATR/Lancaster English Treebank. In the case of a treebank using a 2800–tag tagset, the first–named solution, that of providing multiple correct answers throughout the treebank as a whole, was not practical. So we chose the next–best solution, which has proved practical.

The way we evaluate our tagger is to compare its performance to the set of correct tags for each word in context within our "gold standard" test data.[6] Thus, in all cases we are able to take into account the full set of "correct" answers.[7] Since 32% of running words in our test data have 2 or more correct tags, potential differences in performance evaluation are large vis–a–vis traditional metrics.

5 Experimental Results Using the Non–Dictionary Tagger on the ATR/Lancaster English Treebank

Before citing our performance results on ATR/Lancaster Treebank "gold standard" test data, we wish to present the obligatory calling card, and show our results on the task of tagging UPenn WSJ Treebank test data using the UPenn Tagset. Table 2 shows that our method of predicting tags performs as well as the others "on the market" on this canonical task.

This ritual having been completed, we proceed to present our tagging results on ATR Syntax and ATR Full, in Table 3.

Table 2 shows percentages of all running words correctly tagged. The overall result is broken down into the results for (all) unknown words (UW); (all) known words (KW); known words

[5] terminology of [7], for e.g. a *sleeping* pill vs. to make a good *living*

[6] We now have about 9,500 words of this test data, and expect to have a total of 55,000 words by Summer 1997.

[7] We limit the set of correct tags to five tags; however, for only 2% of running words of test data were as many as 5 tags provided by our human experts; so in general, we are accounting for "all correct tags" for the given word in context.

Tag set	overall	UW	KW	KWKT	KWUT	trivial
UPenn	96.0	91.9	96.7	99.6	61.0	89.6
ATR Syntax	90.8	79.6	93.8	94.6	41.2	83.6

Table 2: Tagging results: UPenn models were tested on a 50,000 word test set; ATR models were tested on 60,000 words of randomly chosen documents. See text for description of columns.

Tag set	single tag	multi–tag
ATR Syntax	91.8	92.8
ATR Full	63.3	67.5

Table 3: Tagging results for the evaluation criteria discussed in this paper on our "gold standard" test set, which currently contains 9500 running words.

with a previously-seen tag (KWKT); and known words with a new tag (KWUT). Also shown is the overall score for a trivial model, which assigns to each word the tag with which it appears most often in the training data. The trivial model indicates, for comparison, what a dictionary-based model ignoring context might produce.

Table 3 shows the percentage of running words whose predicted tag is correct. The column labelled "single tag" compares the predicted tag to the single tag which appears in the treebank for a given word in context. The column labelled "multi–tag" shows the percentage of running words whose predicted tag exactly matches any one of the set of correct tags assigned that word in context by the treebankers. For models on both tag sets, roughly 10% of the tagging "errors" are in fact assignments on which trained (human) experts could reasonably disagree.

In terms of future research, we continue to develop our repertoire of grammarian–created questions, and plan to run our word–clustering program on larger datasets to increase accuracy and coverage of predictions derived from these classes. We are pursuing experiments aimed at quantifying the value added by our tagger to speech synthesis, information retrieval and other language–related systems.

References

[1] E. Black, F. Jelinek, J. Lafferty, R. Mercer, S. Roukos. 1992. Decision tree models applied to the labelling of text with parts–of–speech. In *Proceedings, DARPA Speech and Natural Language Workshop*, Arden House, Morgan Kaufman Publishers.

[2] E. Black, R. Garside, and G. Leech, Editors. 1993. *Statistically–Driven Computer Grammars Of English: The IBM/Lancaster Approach.* Rodopi Editions. Amsterdam.

[3] E. Black, S. Eubank, R. Garside, H. Kashioka, G. Leech, D. Magerman. 1996. Beyond Skeleton Parsing: Producing A Comprehensive Large–Scale General–English Treebank With Full Grammatical Analysis. In *Proceedings, COLING 96, EACL, Copenhagen.*

[4] P. Brown, V. Della Pietra, P. de Souza, J. Lai, R. Mercer. 1992. Class–Based n–Gram Models of Natural Language. *Computational Linguistics*, 18.4:467–479.

[5] R. Garside, G. Leech, G. Sampson, Editor 1987. *The Computational Analysis of English.* London, Longman.

[6] H. Kucera and W. N. Francis. 1967. *Computational Analysis of Present–Day American English.* Brown University Press. Providence, RI.

[7] R. Long. 1961. *The Sentence and Its Parts.* University of Chicago Press. Chicago.

[8] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. 1993. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, 19.2:313-330.

[9] A. Ushioda. 1996. Hierarchical clustering of words. *Proceedings, COLING 96, Copenhagen.*

[10] A. Ushioda. 1996. Hierarchical clustering of words and application to NLP tasks. *Proceedings, Fourth Workshop on Very Large Corpora, Copenhagen.*