Improving Named Entity Recognition with Deep Parsing

Skatje Myers, Robert Leaman, Graciela Gonzalez

Introduction

As the number of publications in the biomedical domain expands, the need for an automated way of managing and analyzing them grows. By using machine learning and natural language processing techniques, the text can be automatically processed for patterns, such as relationships between genes and diseases, as mentioned in the biomedical literature.

In one such technique, named entities such as gene names or diseases are automatically identified. This named entity recognition is the foundation on which other text mining tools are built. These tools can only be as strong as their foundation, so NER is of the utmost importance.

The most successful NER systems use machine learning and take advantage of features such as word length, neighboring tokens, capitalization, affixes, [3][4] part of speech tagging, and dictionaries [7][1]. By training on pre-tagged corpora, the system can assign probabilities for each word of the input based on what it had learned from the training data.

Vlachos designed a NER system which used simple orthographic features (such as affixes and whether the word contained digits or punctuation), part-of-speech tagger, lemmatizer, and the RASP syntactic parser [8]. Using the parser, the system obtained lemmas of related subjects, verbs, objects, and modifiers. He found parsing to produce an increase in precision (percent of the tagged words that were actually named entities) of 0.46, which is small but significant, and a decrease in recall (percent of entities that were successfully identified) of 0.25.

Previously, Smith and Wilbur tested for the usefulness of parse features in NER [6]. They tested numerous parsers (Charniak, Bikel, Rasp, Minipar Enju, Stanford, Rasp Stanford, and Charniak-Lease) against each other and the base NER system without a parser. These parsers provided phrase structure trees, grammatical relations, predicate-argument structures, dependency relations, or annotated phrase structure trees.

Wilbur and Smith found that all nine of the parsers produced a small but statistically significant improvement, but there were no significant differences between the parsers. In this paper, they only evaluated the use of parsing for gene mention recognition, This paper is limited in that it only evaluates parsing for gene mention recognition, and not named entity recognition as a whole.

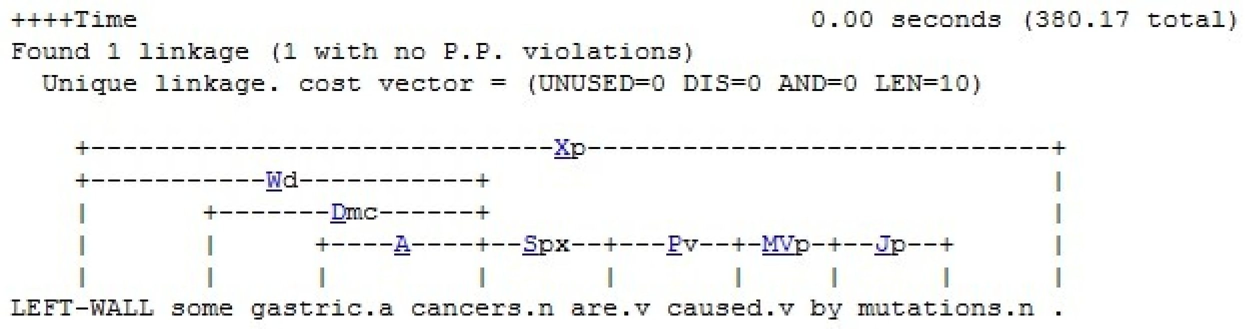
Description

BANNER is a named-entity recognition system for the biological domain, utilizing orthographic and semantic features[2]. Currently, BANNER is limited by its inability to process sentences in whole as a series of linkages. Some of its errors could have been corrected had it access to a full syntactic parse. For example, BANNER did not correctly label "M-MuLVneo delta Enh" in the following sentence:

> "However, a few sites in the genomes of EC cells permit M-MuLVneo delta Enh proviral expression. "

A syntactic parse would show that all the tokens "M-MuLVneo delta Enh proviral" are linked to "expression" as adjectives. With "expression" being a common noun used with genes, this would indicate that the series of adjectives linked to "expression" should be tagged. If BANNER could generate features from parse data, some of these flaws might be corrected.

To obtain parse data, we used LinkGrammar, which is a syntactic parser that provides both a constituent tree and a linkage diagram of a sentence [5]. In this diagram, linkages such as adjective-noun, verb-object, and preposition-object are represented.



```
++++Time                                          0.00 seconds (380.17 total)
Found 1 linkage (1 with no P.P. violations)
  Unique linkage. cost vector = (UNUSED=0 DIS=0 AND=0 LEN=10)

    +-----------------------------Xp-----------------------------+
    +-----------Wd-----------+                                   |
    |       +-------Dmc------+                                   |
    |       |      +----A----+--Spx--+---Pv--+-MVp-+--Jp--+      |
    |       |      |         |       |       |     |      |      |
LEFT-WALL some gastric.a cancers.n are.v caused.v by mutations.n .
```

Figure 1: A linkage produced by LinkGrammar. “A” connects an adjective with the noun it modifies. “S” connects a subject with its verb.

With access to the top parse generated by LinkGrammar for each sentence in a corpus, BANNER can use this information to improve its probabilities when looking for named entities. Specifically, it is looking for what subjects, verbs, objects, and adjectives are linked to words. For the example sentence in Figure 1, BANNER would note for the word “cancers”, it would see that its verb is “is” and that it has an adjective modifying it—“gastric”. While knowing the verb won’t aid in identifying “cancers” as a disease since “is” is so common, the adjective “gastric”, which may be likely to be modifying some ailment, should weigh towards proper identification of “cancers”.

Results

To evaluate the effects of adding parsing to BANNER, we tested the system on the Arizona Disease Corpus, containing 3000 sentences. BANNER is based on conditional random fields, and in this testing we used a 1st-order model. The BANNER system before and after integrating LinkGrammar were evaluated using 10-fold cross validation—where the corpus was divided into ten pieces, trained on nine of them, then tested on the last. This is repeated ten times. The results are given by averaging the precision and recall of these tests. The results can be represented by a single number called the F-measure, which is the harmonic mean of the precision and recall. We found that by adding parse features, both BANNER's precision and recall increase by a small amount.

	Precision (%)	**Recall (%)**	**F-measure (%)**
BANNER without LinkGrammar features	77.75%	70.06%	73.71%
BANNER with LinkGrammar features	78.70%	70.64%	74.46%
Difference	0.95%	0.58%	0.75%

Table 1

Since we only added subject, object, verb, and adjective data, there is room for further expansion, such as including adverbs. This may enhance our results further, although also carries the risk of adding extra noise and decreasing recall. Finkel [1] suggests that deep syntactic features are not useful for corpora with so few types of entities to identify, and would be better suited to corpora with more, such as the GENIA corpus. Currently the greatest setback of our results is performance time, which is approximately one hour per thousand sentences, which makes using the system on larger corpora infeasible. These possible improvements in mind, future utility of parse features in NER shows promise.

References

[1] Finkel, J.; Dingare, S.; et al. Exploiting context for biomedical entity recognition: from syntax to the web. *Joint Workshop on Natural Language Processing in Biomedicine and Its Applications at Coling 2004*.

[2] Leaman, R.; and Gonzalez, G. (2008) BANNER: An executable survey of advances in biomedical named entity recognition. *Pacific Symposium on Biocomputing.*

Leaman, R; Miller, C.; and Gonzalez, G. Enabling Recognition of Diseases in Biomedical Text with Machine Learning: Corpus and Benchmark. *Proceedings of the 2009 Symposium on Languages in Biology and Medicine*, November 2009, Jeju island, Korea.

[3] Settles, B. (2005) ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics* **21**:3191–3192.

[4] Settles, B. (2004) Biomedical named entity recognition using conditional random fields and rich feature sets. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications.*

[5] Sleator, D.; and Temperley, D. (1993) Parsing English with a Link Grammar. *Third International Workshop on Parsing Technologies.*

[6] Smith, L.; and Wilbur, WJ. (2009) The value of parsing as feature generation for gene mention recognition*. J Biomed Inform* (2009), doi:10.1016/j.jbi.2009.03.011

[7] Tanabe, L.; and Wilbur, W.J. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics* 2002;18(8):1124–32.

[8] Vlachos, A. (2007) Tackling the BioCreative 2 gene mention task with conditional random fields and syntactic parsing. *Proceedings of the Second BioCreative Challenge Workshop* pp. 85-87.