The Web is not a PERSON, Berners-Lee is not an ORGANIZATION, and African-Americans are not LOCATIONS: An Analysis of the Performance of Named-Entity Recognition

Robert Krovetz
Lexical Research
Hillsborough, NJ 08844
rkrovetz@lexicalresearch.com

Paul Deane **Nitin Madnani**
Educational Testing Service
Princeton, NJ 08541
{pdeane,nmadnani}@ets.org

Abstract

Most work on evaluation of named-entity recognition has been done in the context of competitions, as a part of Information Extraction. There has been little work on any form of extrinsic evaluation, and how one tagger compares with another on the major classes: PERSON, ORGANIZATION, and LOCATION. We report on a comparison of three state-of-the-art named entity taggers: Stanford, LBJ, and IdentiFinder. The taggers were compared with respect to: 1) Agreement rate on the classification of entities by class, and 2) Percentage of ambiguous entities (belonging to more than one class) co-occurring in a document. We found that the agreement between the taggers ranged from 34% to 58%, depending on the class and that more than 40% of the globally ambiguous entities co-occur within the same document. We also propose a unit test based on the problems we encountered.

1 Introduction

Named-Entity Recognition (NER) has been an important task in Computational Linguistics for more than 15 years. The aim is to recognize and classify different types of entities in text. These might be people's names, or organizations, or locations, as well as dates, times, and currencies. Performance assessment is usually made in the context of Information Extraction, of which NER is generally a component. Competitions have been held from the earliest days of MUC (Message Understanding Conference), to the more recent shared tasks in CoNLL. Recent research has focused on non-English languages such as Spanish, Dutch, and German (Meulder et al., 2002; Carreras et al., 2003; Rossler, 2004), and on improving the performance of unsupervised learning methods (Nadeau et al., 2006; Elsner et al., 2009).

There are no well-established standards for evaluation of NER. Since criteria for membership in the classes can change from one competition to another, it is often not possible to compare performance directly. Moreover, since some of the systems in the competition may use proprietary software, the results in a competition might not be replicable by others in the community; however, this applies to the state of the art for most NLP applications rather than just NER.

Our work is motivated by a vocabulary assessment project in which we needed to identify multi-word expressions and determine their association with other words and phrases. However, we found that state-of-the-art software for named-entity recognition was not reliable; false positives and tagging inconsistencies significantly hindered our work. These results led us to examine the state-of-the-art in more detail.

The field of Information Extraction (IE) has been heavily influenced by the Information Retrieval (IR) community when it comes to evaluation of system performance. The use of Recall and Precision metrics for evaluating IE comes from the IR community. However, while the IR community regularly conducts a set of competitions and shared tasks using standardized test collections, the IE community does not. Furthermore, NER is just one component

of an IE pipeline and any proposed improvements to this component must be evaluated by determining whether the performance of the overall IE pipeline has improved. However, most, if not all, NER evaluations and shared tasks only focus on intrinsic NER performance and ignore any form of extrinsic evaluation. One of the contributions of this paper is a freely available unit test based on the systematic problems we found with existing taggers.

2 Evaluation Methodology

We compared three state-of-the-art NER taggers: one from Stanford University (henceforth, Stanford tagger), one from the University of Illinois (henceforth, the LBJ tagger) and BBN IdentiFinder (henceforth, IdentiFinder).

The Stanford Tagger is based on Conditional Random Fields (Finkel et al., 2005). It was trained on 100 million words from the English Gigawords corpus. The LBJ Tagger is based on a regularized average perceptron (Ratinov and Roth, 2009). It was trained on a subset of the Reuters 1996 news corpus, a subset of the North American News Corpus, and a set of 20 web pages. The features for both these taggers are based on local context for a target word, orthographic features, label sequences, and distributional similarity. Both taggers include non-local features to ensure consistency in the tagging of identical tokens that are in close proximity. IdentiFinder is a state-of-the-art commercial NER tagger that uses Hidden Markov Models (HMMs) (Bikel et al., 1999).

Since we did not have gold standard annotations for any of the real-world data we evaluated on, we instead compared the three taggers along two dimensions:

- **Agreement on classification**. How well do the taggers work on the three most difficult classes: PERSON, ORGANIZATION, and LOCATION and, more importantly, to what extent does one tagger agree with another? What types of mistakes do they make systematically?[1]
- **Ambiguity in discourse**. Although entities can potentially have more than one entity classification, such as *Clinton* (PERSON or LOCATION), it would be surprising if they co-occurred in a single discourse unit such as a document. How frequently does each tagger produce multiple classifications for the same entity in a single document?

We first compared the two freely available, academic taggers (Stanford and LBJ) on a corpus of 425 million words that is used internally at the Educational Testing Service. Note that we could not compare these two taggers to IdentiFinder on this corpus since IdentiFinder is not available for public use without a license.

Next, we compared all three taggers on the American National Corpus. The American National Corpus (ANC) has recently released a copy which is tagged by IdentiFinder.[2] Since the ANC is a publicly available corpus, we tagged it using both the Stanford and LBJ taggers and could then compare all three taggers along the two intended dimensions. We found that the public corpus had many of the same problems as the ones we found with our internally used corpus. Some of these problems have been discussed before (Marrero et al., 2009) but not in sufficient detail.

The following section describes our evaluation of the Stanford and LBJ taggers on the internal ETS corpus. Section 4 describes a comparison of all three taggers on the American National Corpus. Section 5 describes the unit test we propose. In Section 6, we propose and discuss the viability of the "one named-entity tag per discourse" hypothesis. In Section 7, we highlight the problems we find during our comparisons and propose a methodology for improved intrinsic evaluation for NER. Finally, we conclude in Section 8.

3 Comparing Stanford and LBJ

In this section, we compare the two academic taggers in terms of classification agreement by class and discourse ambiguity on the ETS SourceFinder corpus, a heterogeneous corpus containing approximately 425 million words, and more than $270,000$

[1] Although one could draw a distinction between named entity identification and classification, we focus on the final output of the taggers, i.e., classified named entities.

[2] `http://www.anc.org/annotations.html`

Person		Organization		Location	
Stanford	**LBJ**	**Stanford**	**LBJ**	**Stanford**	**LBJ**
Shiloh	A.sub.1	RNA	Santa Barbara	Hebrew	The New Republic
Yale	What	Arnold	FIGURE	ASCII	DNA
Motown	Jurassic Park	NaCl	Number:	Tina	Mom
Le Monde	Auschwitz	AARGH	OMITTED	Jr.	Ph.D
Drosophila	T. Rex	Drosophila	Middle Ages	Drosophila	Drosophila

Table 1: A sampling of false positives for each class as tagged by the Stanford and LBJ taggers

	Common Entities	Percentage
Person	548,864	58%
Organization	249,888	34%
Location	102,332	37%

Table 2: Agreement rate by class between the Stanford and LBJ taggers

articles. The articles were extracted from a set of 60 different journals, newspapers and magazines focused on both literary and scientific topics.

Although Named Entity Recognition is reported in the literature to have an accuracy rate of 85-95% (Finkel et al., 2005; Ratinov and Roth, 2009), it was clear by inspection that both the Stanford and the LBJ tagger made a number of mistakes. The ETS corpus begins with an article about Tim Berners-Lee, the man who created the World Wide Web. At the beginning of the article, "Tim" as well as "Berners-Lee" are correctly tagged by the Stanford tagger as belonging to the PERSON class. But later in the same article, "Berners-Lee" is incorrectly tagged as ORGANIZATION. The LBJ tagger makes many mistakes as well, but they are not necessarily the same mistakes as the mistakes made by the Stanford tagger. For example, the LBJ tagger sometimes classifies "The Web" as a PERSON, and the Stanford tagger classifies "Italian" as a LOCATION.[3] Table 1 provides an anecdotal list of the "entities" that were misclassified by the two taggers.[4]

Both taggers produced about the same number of entities overall: 1.95 million for Stanford, and 1.8 million for LBJ. The agreement rate between the taggers is shown in Table 2. We find that the highest rate of agreement is for PERSONS, with an agreement rate of 58%. The agreement rate on LOCATIONS is 37%, and the agreement rate on ORGANIZATIONS is 34%. Even on cases where the taggers agree, the classification can be incorrect. Both taggers classify "African Americans" as LOCATIONS.[5] Both treat "Jr." as being part of a person's name, as well as being a LOCATION (in fact, the tagging of "Jr." as a LOCATION is more frequent in both).

For our second evaluation criterion, i.e., within-discourse ambiguity, we determined the percentage of globally ambiguous entities (entities that had more than one classification across the entire corpus) that occurred with multiple taggings within a single document. This analysis showed that the problems described above are not anecdotal. Table 3 shows that at least 40% of the entities that have more than one classification co-occur within a document. This is true for both taggers and all of the named entity classes.[6]

[3] "Italian" is classified primarily as MISC by the LBJ tagger. These terms are sometimes called Gentilics or Demonyms.

[4] Both taggers can use a fourth class MISC in addition to the standard entity classes PERSON, ORGANIZATION, and LOCATION. We ran Stanford without the MISC class and LBJ with MISC. However, the problems highlighted in this paper remain equally prevalent even without this discrepancy.

[5] The LBJ tagger classifies the majority of instances of "African American" as MISC.

[6] The LBJ tagger also includes the class MISC. We looked at the co-occurrence rate between the different classes and MISC, and we found that the majority of each group co-occurred within a document there as well.

	Stanford		LBJ	
	Overlap	**Co-occurrence**	**Overlap**	**Co-occurrence**
Person-Organization	98,776	40%	58,574	68%
Person-Location	72,296	62%	55,376	69%
Organization-Location	80,337	45%	64,399	63%

Table 3: Co-occurrence rates between entities with more than one tag for Stanford and LBJ taggers

	Stanford-BBN		LBJ-BBN	
	Common Entities	**Percentage**	**Common Entities**	**Percentage**
Person	8034	28%	27,687	53%
Organization	12533	50%	21,777	51%
Location(GPE)	3289	28%	5475	47%

Table 4: Agreement rate by class between the Stanford (and LBJ) and BBN IdentiFinder taggers on the ANC Corpus

4 Comparing All 3 Taggers

A copy of the American National Corpus was recently released with a tagging by IdentiFinder. We tagged the corpus with the Stanford and LBJ tagger to see how the results compared.

We found many of the same problems with the American National Corpus as we found with the SourceFinder corpus used in the previous section. The taggers performed very well for entities that were common in each class, but we found misclassifications even for terms at the head of the Zipfian curve. Terms such as "Drosophila" and "RNA" were classified as a LOCATION. "Affymetrix" was classified as a PERSON, LOCATION, and ORGANIZATION.

Table 4 shows the agreement rate between the Stanford and IdentiFinder taggers as well as that between the LBJ and IdentiFinder taggers. A sample of terms that were classified as belonging to more than one class, across all 3 taggers, is given in Table 5.

All taggers differ in how the entities are tokenized. The Stanford tagger tags each component word of the multi-word expressions separately. For example, "John Smith" is tagged as John/PERSON and Smith/PERSON. But it would be tagged as [PER John Smith] by the LBJ tagger, and similarly by IdentiFinder. This results in a higher overlap between classes in general, and there is a greater agreement rate between LBJ and IdentiFinder than between Stanford and either one.

The taggers also differ in the number of entities that are recognized overall, and the percentage that are classified in each category. IdentiFinder recognizes significantly more ORGANIZATION entities than Stanford and LBJ. IdentiFinder also uses a GPE (Geo-Political Entity) category that is not found in the other two. This splits the LOCATION class. We found that many of the entities that were classified as LOCATION by the other two taggers were classified as GPE by IdentiFinder.

Although the taggers differ in tokenization as well as categories, the results on ambiguity in a discourse support our findings on the larger corpus. The results are shown in Table 6. For both the Stanford and LBJ tagger, between 42% and 58% of the entities with more than one classification co-occur within a document. For IdentiFinder, the co-occurrence rate was high for two of the groupings, but significantly less for PERSON and GPE.

5 Unit Test for NER

We created a unit test based on our experiences in comparing the different taggers. We were particular about choosing examples that test the following:

1. Capitalized, upper case, and lower case versions of entities that are true positives for PERSON, ORGANIZATION, and LOCATION (for a variety of frequency ranges).

2. Terms that are entirely in upper case that are not named entities (such as RNA and AAARGH).

Person/Organization	Person/Location	Organization/Location
Bacillus	Bacillus	Affymetrix
Michelob	Aristotle	Arp2/3
Phenylsepharose	ArrayOligoSelector	ANOVA
Synagogue	Auschwitz	Godzilla
Transactionalism	Btk:ER	Macbeth

Table 5: A sampling of terms that were tagged as belonging to more than one class in the American National Corpus

	Stanford		LBJ		IdentiFinder	
	Overlap	**Co-occurrence**	**Overlap**	**Co-occurrence**	**Overlap**	**Co-occurrence**
Person-Org	5738	53%	2311	58%	8379	57%
Person-Loc(GPE)	4126	58%	3283	43%	2412	22%
Org-Loc(GPE)	5109	57%	4592	50%	4093	60%

Table 6: Co-occurrence rates between entities with more than one tag for the American National Corpus

3. Terms that contain punctuation marks such as hyphens, and expressions (such as "A.sub.1") that are clearly not named entities.

4. Terms that contain an initial, such as "T. Rex", "M.I.T", and "L.B.J."

5. Acronym forms such as ETS and MIT, some with an expanded form and some without.

6. Last names that appear in close proximity to the full name (first and last). This is to check on the impact of discourse and consistency of tagging.

7. Terms that contain a preposition, such as "Massachusetts Institute of Technology". This is intended to test for correct extent in identifying the entity.

8. Terms that are a part of a location as well as an organization. For example, "Amherst, MA" vs. "Amherst College".

An excerpt from this unit test is shown in Table 7. We provide more information about the full unit test at the end of the paper.

6 One Named-Entity Tag per Discourse

Previous papers have noted that it would be unusual for multiple occurrences of a token in a document to be classified as a different type of entity (Mikheev et al., 1999; Curran and Clark, 2003). The Stanford and LBJ taggers have features for non-local dependencies for this reason. The observation is similar to a hypothesis proposed by Gale, Church, and Yarowsky with respect to word-sense disambiguation and discourse (Gale et al., 1992). They hypothesized that when an ambiguous word appears in a document, all subsequent instances of that word in the document will have the same sense. This hypothesis is incorrect for word senses that we find in a dictionary (Krovetz, 1998) but is likely to be correct for the subset of the senses that are homonymous (unrelated in meaning). Ambiguity between named entities is similar to homonymy, and for most entities it is unlikely that they would co-occur in a document.[7] However, there are cases that are exceptions. For example, Finkel et al. (2005) note that in the CoNLL dataset, the same term can be used for a location and for the name of a sports team. Ratinov and Roth (2009) note that "Australia" (LOCATION) can occur in the same document as "Bank of Australia" (ORGANIZATION).

Existing taggers treat the non-local dependencies as a way of dealing with the sparse data problem, and as a way to resolve tagging differences by looking at how often one token is classified as one type

[7] Krovetz (1998) provides some examples where different named entities co-occur in a discourse, such as "New York" (city) and "New York" (state). However, these are both in the same class (LOCATION) and are related to each other.

```
                    This is not a Unit Test
                (a tribute to Rene Magritte and RMS)

Although we created this test with humor, we intend it as a serious
test of the phenomena we encountered.  These problems include
ambiguity between entities (such as Bill Clinton and Clinton,
Michigan), uneven treatment of variant forms (MIT, M.I.T., and
Massachusetts Institute of Technology - these should all be
labeled the same in this text - are they?), and frequent false
positives such as RNA and T. Rex.

...
```

Table 7: Excerpt from a Unit test for Named-Entity Recognition

versus another. We propose that these dependencies can be used in two other aspects: (a) as a source of error in evaluation and, (b) as a way to identify semantically related entities that are systematic exceptions. There is a grammar to named entity types. "Bank of Australia" is a special case of *Bank of* [LOCATION]. The same thing is true for "China Daily" as a name for a newspaper. We propose that co-occurrences of different labels for particular instances can be used to create such a grammar; at the very least, particular types of co-occurrences should be treated as an exception to what is otherwise an indication of a tagging mistake.

7 Discussion

The Message Understanding Conference (MUC) has guidelines for named-entity recognition. But the guidelines are just that. We believe that there should be standards. Without such standards it is difficult to determine which tagger is correct, and how the accuracy varies between the classes.

We propose that the community focus on four classes: PERSON, ORGANIZATION, LOCATION, and MISC. This does not mean that the other classes are not important. Rather it is recognition of the following facts:

- These classes are more difficult than dates, times, and currencies.
- There is widespread disagreement between taggers on these classes, and evidence that they are misclassifying unique entities a significant percentage of the time.
- We need at least one class for handling terms that do not fit into the first three classes.
- The first three classes have important value in other areas of NLP.

Although we recognize that an extrinsic evaluation of named entity recognition would be ideal, we also realize that intrinsic evaluations are valuable in their own right. We propose that the existing methodology for intrinsically evaluating named entity taggers can be improved in the following manner:

1. Create test sets that are organized across a variety of domains. It is not enough to work with newswire and biomedical text.
2. Use standardized sets that are designed to test different types of linguistic phenomena, and make it a de facto norm to use more than one set as part of an evaluation.
3. Report accuracy rates separately for the three major classes. Accuracy rates should be further broken down according to the items in the unit test that are designed to assess mistakes: orthography, acronym processing, frequent false positives, and knowledge-based classification.
4. Establish a way for a tagging system to express uncertainty about a classification.

The approach taken by the American National Corpus is a good step in the right direction. Like the original Brown Corpus and the British National Corpus, it breaks text down according to informational/literary text types, and spoken versus written text. The corpus also includes text that is drawn from the literature of science and medicine. However, the relatively small number of files in the corpus makes it difficult to assess accuracy rates on the basis of repeated occurrences within a document, but with different tags. Because there are hundreds of thousands of files in the internal ETS corpus, there are many opportunities for observations. The tagged version of the American National Corpus has about 8800 files. This is one of the biggest differences between the evaluation on the corpus we used internally at ETS and the American National Corpus.

The use of a MISC class is needed for reasons that are independent of certainty. This is why we propose a goal of allowing systems to express this aspect of the classification. We suggest a meta-tag of a question-mark. The meta-tag can be applied to any class. Entities for which the system is uncertain can then be routed for active learning. This also allows a basic separation of entities into those for which the system is confident of its classification, and those for which it is not.

8 Conclusion

Although Named Entity Recognition has a reported accuracy rate of more than 90%, the results show they make a significant number of mistakes. The high accuracy rates are based on inadequate methods for testing performance. By considering only the entities where both taggers agree on the classification, it is likely that we can obtain improved accuracy. But even so, there are cases where both taggers agree yet the agreement is on an incorrect tagging.

The unit test for assessing NER performance is freely available to download.[8]

As with Information Retrieval test collections, we hope that this becomes one of many, and that they be adopted as a standard for evaluating performance.

[8] `http://bit.ly/nertest`

Acknowledgments

This work has been supported by the Institute for Education Sciences under grant IES PR/Award Number R305A080647. We are grateful to Michael Flor, Jill Burstein, and anonymous reviewers for their comments.

References

Daniel M. Bikel, Richard M. Schwartz, and Ralph M. Weischedel. 1999. An Algorithm that Learns What's in a Name. *Machine Learning*, 34:211–231.

Xavier Carreras, Llus Mrquez, and Llus Padr. 2003. Named entity recognition for Catalan using Spanish resources. In *Proceedings of EACL*.

James R. Curran and Stephen Clark. 2003. Language Independent NER using a Maximum Entropy Tagger. In *Proceeding of the 7th Conference on Computational Natural Language Learning (CoNLL)*, pages 164–167.

Micha Elsner, Eugene Charniak, and Mark Johnson. 2009. Structured Generative Models for Unsupervised Named-Entity Clustering. In *Proceedings of NAACL*, pages 164–172.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of ACL*, pages 363–370.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense per Discourse. In *Proceedings of the ARPA Workshop on Speech and Natural Language Processing*, pages 233–237.

Robert Krovetz. 1998. More than One Sense Per Discourse. In *Proceedings of the ACL-SIGLEX Workshop: SENSEVAL-1*.

Monica Marrero, Sonia Sanchez-Cuadrado, Jorge Morato Lara, and George Andreadakis. 2009. Evaluation of Named Entity Extraction Systems. *Advances in Computational Linguistics, Research in Computing Science*, 41:47–58.

Fien De Meulder, V Eronique Hoste, and Walter Daelemans. 2002. A Named Entity Recognition System for Dutch. In *Computational Linguistics in the Netherlands*, pages 77–88.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999. Named Entity Recognition Without Gazetteers. In *Proceedings of EACL*, pages 1–8.

David Nadeau, Peter D. Turney, and Stan Matwin. 2006. Unsupervised Named-Entity Recognition: Generating Gazetteers and Resolving Ambiguity. In *Proceedings of the Canadian Conference on Artificial Intelligence*, pages 266–277.

L. Ratinov and D. Roth. 2009. Design Challenges and Misconceptions in Named Entity Recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pages 147–155.

Marc Rossler. 2004. Adapting an NER-System for German to the Biomedical Domain. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 92–95.