# Annotating Resources for Information Extraction

**Sean Boisen, Michael R. Crystal, Richard Schwartz, Rebecca Stone, Ralph Weischedel**

BBN Technologies
87 Fawcett Street, Cambridge MA 02138
Sean.Boisen@bbn.com

**Abstract**

Trained systems for NE extraction have shown significant promise because of their robustness to errorful input and rapid adaptability. However, these learning algorithms have transferred the cost of development from skilled computational linguistic expertise to data annotation, putting a new premium on effective ways to produce high-quality annotated resources at minimal cost. The paper reflects on BBN's four years of experience in the annotation of training data for Named Entity (NE) extraction systems discussing useful techniques for maximizing data quality and quantity.

## 1. Introduction

Named Entity (NE) extraction is now a well-established part of the field of information extraction and message understanding with significant practical potential, with performance rivaling that of humans at much higher throughput, and commercial systems emerging. Trained systems for NE extraction (Bikel et al., 1997; Palmer et al 1999; Cucerzan and Yarowsky, 1999) have shown significant promise because of their robustness to errorful input and rapid adaptability (Miller et al., 1999). However, these learning algorithms have transferred the cost of development from skilled computational linguistic expertise to data annotation, putting a new premium on effective ways to produce high-quality annotated resources at minimal cost.

BBN has now accumulated several years of experience in the annotation of training data for Named Entity extraction systems. These annotation projects include:

- Wall Street Journal (MUC-6 evaluation)
- New York Times (MUC-7 evaluation)
- Transcriptions of audio broadcast news (DARPA Hub-4 evaluation)
- Reuters new stories
- Xinhua (Chinese) news agency stories.

This work reports on lessons learned about the process of annotating training data, as measured by their impact on the performance of IdentiFinder™, BBN's automatically trained system for named entity extraction (Bikel et al., 1997, 1999).

## 2. Annotation in the Context of Trained Systems for Named Entity Extraction

### 2.1. The Named Entity Task

The named entity task is to identify and mark certain types of names and referring expressions in input text, typically via SGML tags. The definition of the types of names to be marked (henceforth *name classes*) is part of the application design. One widely used set of name classes originating with the MUC evaluations includes named locations, named persons, named organizations, dates, times, monetary amounts, and percentages.

### 2.2. The Training Paradigm

Trained information extraction systems like IdentiFinder are built on a train-by-example paradigm. Annotated data provides both positive and negative examples of named entities, from which IdentiFinder induces a model for processing new input. Since the system can output data in the same format as the training input, the same annotated data can serve both training and evaluation purposes.

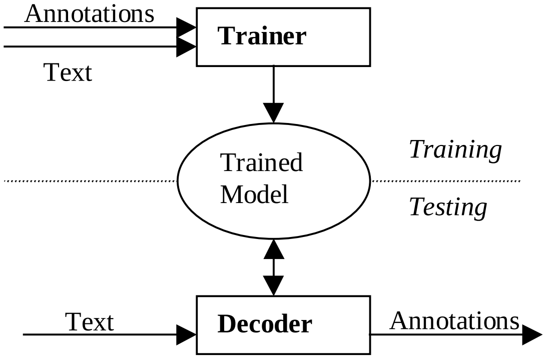

Figure 1: IdentiFinder's Train-by-Example Architecture

While a fair test always means processing input that was not part of the training, it is easy with trained systems to automatically vary the training and test sets. For example, given 10 training files, 10 different tests can be performed by training on 9 files and testing on the tenth, rotating which file is held out for testing. Since the system's knowledge of the data is derived from training examples, this is a reasonable approximation of a blind test. [1]

### 2.3. The Annotation Process

A typical annotation project includes phases for design, initial production of annotated data, and incremental refinement.

#### 2.3.1. Design

As with many other endeavors, careful planning prior to actual annotation is important to a successful outcome.

[1] If vocabulary lists are used to supplement the training, this testing regime is not fully blind and may therefore not be completely fair, though it is much less costly than a genuinely blind test set gathered and annotated under careful conditions.

IdentiFinder uses three sources of information in extracting named entities, all of them highly localized:

- the word itself
- word features like capitalization (when present), whether a word is punctuation or all digits, or whether a word begins a sentence.[2]
- the immediate word context (for example, a preceding "Mr." or a following "Inc.")

Since only local, lexical information is used, only certain kinds of name classes are suitable choices for extraction. Once suitable name classes have been identified, the project must develop guidelines as to which phrases will be marked as examples of a name class (and which close cases will not be marked). This can be surprisingly complex. A priori, the MUC name classes of PERSON, LOCATION, and ORGANIZATION seem fairly well defined. Nevertheless, the printed MUC-6 guidelines for these name classes run more than a dozen pages.

#### 2.3.2. Production

Once the design of the annotation effort is settled, an production phase can be launched. This is typically an iterative process, and often includes more than one annotator. In preparation for annotation, the data is given standard document-level markup: minimally, this indicates document boundaries and a unique document identifier, as well as excluding header material that is not annotated. Duplicate or near-duplicate documents are identified and removed: not only do they waste annotation effort, but more importantly they can skew evaluation results if copies of same document happen to wind up in both the test and training corpus.

Several techniques are described in Section 3 below that help ensure the quality of the annotation. Once some initial amount of training data has been annotated, a model is trained and evaluated on a blind test set. If performance is deemed inadequate, additional material can be annotated, and this cycle continues as needed. Section 4 describes techniques for maximizing annotation quantity.

#### 2.3.3. Tradeoffs Between Quality and Quantity

As with many language phenomena, the distribution of names in text is typically highly skewed, and generally follows Zipf's Law (Zipf 1932, 1949; Palmer & Day 1997). The most frequent names account for a large percentage of the total, while more rare names may occur only once in training material or, even worse, only in the test (unseen in training). This has important consequences for named entity annotation and extraction. While even minimal annotation will capture the most frequent cases, no amount of annotation can ever be expected to cover the large number of highly infrequent names. Extraction performance is typically very good on names that have been seen in training, whereas names that are novel have a much lower likelihood of being extracted correctly.

One result of this skewed distribution of names is that different emphases may be needed at different points in the annotation process. In the initial phase, when little data has been annotated and performance is still low, most names will be new additions to the existing training corpus and will have significant impact on system coverage. In later stages, when more frequent names have numerous observations and performance is higher, quality issues become a larger component of the residual errors.

## 3. Maximizing Annotation Quality

### 3.1. Adjudication

Two human annotators of the same document may produce different results, either because the guidelines aren't clear enough, or because of simple human error. This is especially true early in an annotation project, when the guidelines may still be under-developed and the annotators less experienced. Having two annotators mark the same data can mitigate this problem. Typically a third "master annotator" then adjudicates any differences. This increases the annotation cost but also increases the quality.

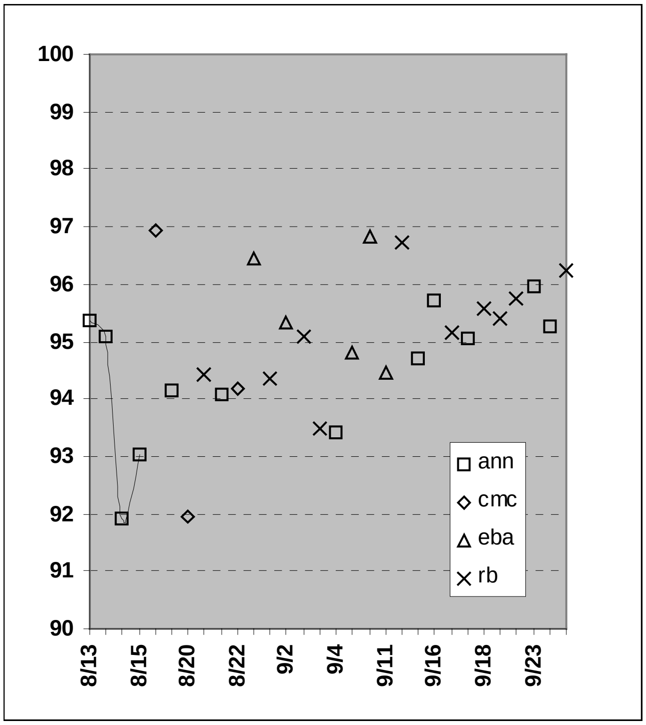

Figure 2: Adjudication Benefit Over Time

Figure 2 shows the performance of four different annotators (identified by their initials) over a six-week project annotating Reuters newswire data. None of them had any previous experience with either named entity annotation or the specific project. Each file was double annotated and adjudicated to produce an improved version, and then the initial annotator output was scored against the adjudicated version to produce an F-measure for inter-annotator agreement. For the earliest files, agreement ranged from 92 to 97 (using the standard F-measure). Over the course of the project, however, the performance range narrowed as the annotators became more skilled at their task and unclear cases were clarified.

Whether the extra cost of redundant annotation and adjudication is justified depends on the value of any improvement in performance. For the Reuters project above, a system trained on adjudicated data had half as

[2] These word features are the only part of IdentiFinder that is language-specific.

many errors as one trained with a random selection of the unadjudicated training (F-measure of 94.5, versus 89). However, a careful study for a broadcast news annotation project (Crystal et al. 1999) shows that with higher error rates (F-measure below 90), the cost of the extra processing steps of double annotation, adjudication, and test-on-train are not repaid by significant improvements in performance. Instead, simply producing additional annotated data seems to provide the biggest impact. However, once performance begins to plateau, further improvements depend more heavily on the quality of annotation, and inconsistent data can degrade performance.

### 3.2. Test on Training

Another technique for improving the quality of annotated data comes from IdentiFinder's ability to reproduce its training input with high fidelity. Testing the system on the unannotated versions of its training, and then scoring this against the original training, can help identify inconsistencies in the annotation. This is not a fair test for evaluating system performance: scores are artificially high, typically F-measures in the range of 98-99. However, approximately one quarter of the discrepancies this technique identifies between the original documents and the IdentiFinder output are attributable to human error, even in the case of training developed by double-annotation and adjudication. Given that the adjudication process can only increase the quality of the training, never decrease it, even an automatic annotator (IdentiFinder) can provide benefit. This approach can be used equally well with single-annotation training, where the only extra cost is the time to adjudicate, typically much less than the effort of a second annotation.

## 4. Maximizing Annotation Output

BBN's experience over several named entity annotation projects suggests that an experienced annotator with good tools can process 5K words per hour of English news.[3] Given that as little as 100K words of training can provide interesting performance for some applications, this means initial annotation effort can be less than a week in the best case, though 500K words or more is a more representative figure.

The constant factor in annotation throughput is the time to read the document and understand it well enough to identify the names. A second factor is how long it takes to mark a name: this is a function of the tool used for annotation (see Section 5). Beyond these, the remaining component is the time required to make decisions about names, their categories, and their boundaries. We have found it important therefore to encourage annotators to minimize decision-making time by not thinking too hard about individual cases. In addition to clear guidelines, it is often effective to use a special tag like <UNSURE> to defer unclear cases and decide them in a second pass.

### 4.1. Selected Sentence Annotation

Another technique for getting the maximum benefit from annotation effort is *selected sentence annotation*. After an initial training corpus has been developed, the Zipfian distribution of names suggests that additional annotation of whole documents will produce many more instances of names that have already been incorporated into the training, but relatively few novel names. However, it is the novel names that have the greatest potential to improve performance once more frequent names have been covered well.

For this technique, a model is trained on the initial corpus, and then used to analyze other unannotated data to identify sentences with unknown vocabulary items. These selected sentences are then gathered into a new corpus, focusing annotator attention on only those data with the potential to add novel information. While there is no guarantee these unknown vocabulary items are actually names, selected sentences typically yield many more new names for a fraction of the effort of annotating whole documents, even after allowing for the additional time required to annotate isolated sentences. Though this approach misses the potential benefit of new context information or novel ways of using known names, these cases typically contribute fewer errors for better-trained applications than unknown items.

### 4.2. Blind Annotation

Given the need for maximum training at minimum cost, one simple approach is to streamline the annotation process. This can be done either by marking certain names throughout a whole document without reading each example, or by bootstrapping the annotation of additional data by first processing it with IdentiFinder and correcting the output, rather than annotating from scratch. While the speed-up of these approaches is appealing, our experience has been that these are problematic. For example, *Mexico* is a clear example of a location. However, blindly annotating all instances of this string in a document would incorrectly annotate *New Mexico* and *Bank of Mexico*. Of course, a careful annotator ought to catch such problems when reviewing the annotations afterwards. But our anecdotal evidence suggests annotators are more reluctant to change existing annotations than to introduce completely new ones. For these reasons, we normally prefer starting with unannotated text.

## 5. Annotation Tools

To support the production of abundant, high quality annotated training data, BBN has developed IdentiTagger™, a graphical user interface for annotating named entity resources for information extraction. IdentiTagger is based on Java and UNICODE to provide language and encoding independence, and is configurable for different NE applications. The design reflects BBN's experience in producing annotated resources.

BBN annotation staff includes both experienced annotators, as well as new annotators with only minimum computer experience. It was important therefore to produce a specialized interface for data annotation that could be easily learned, yet highly productive.

An important part of the design was the decision to provide for keyboard-only (mouse-free) operation of the interface. This supports cursor movement, text selection, and name class labeling, as well as correction, search, etc.

[3] However, individual annotator throughput can vary widely. Crystal et al. (1999) compared six annotators and found more than a factor of 10 difference in speed between the most and least experienced annotators.

Our earlier mouse-based interfaces contributed to repetitive stress injuries for several staff members, many of whom annotate for 20 or more hours each week, making this an essential health concern.

To enhance readability, the markup itself is hidden, and annotated text is displayed instead with colored backgrounds. To minimize the problem of misplacing label boundaries inside words, a tokenizer limits the default text selection to whole words, while providing the means to override the tokenizer for special cases. IdentiTagger includes an adjudication mode, which can be used to compare two annotated versions of the same file to produce an adjudicated version. Adjudication mode steps through differences one at a time, providing the ability to change between alternative annotations or move to other differences with a single keystroke.

## 6. References

Bikel, D., Miller, S., Schwartz, R., & Weischedel, R. (1997). Nymble: a High-Performance Learning Name-finder. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 194-201). San Francisco, CA: Morgan Kaufman Publishers.

D. Bikel, Schwartz, R., & Weischedel, R., “An Algorithm that Learns What’s in a Name,” *Machine Learning* 34, pp 211-231, (1999).

Choi, M., Ravin, Y., & Wacholder, N. (1997). Disambiguation of Proper Names in Text. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 202-208). San Francisco, CA: Morgan Kaufman Publishers.

Crystal, M., Kubala, F., & MacIntyre, R. (1999). Studies in Data Annotation Effectiveness. In *Proceedings of 1999 DARPA Broadcast News Workshop.* San Francisco, CA: Morgan Kaufman Publishers.

Cucerzan, S., & Yarowsky, D. (1999). Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora* (pp. 90-99). San Francisco, CA: Morgan Kaufman Publishers.

Day, D., & Palmer, D.(1997). A Statistical Profile of the Named Entity Task. In *Proceedings of the Fifth Conference on Applied Natural Language Processing* (pp. 190-193). San Francisco, CA: Morgan Kaufman Publishers.

Miller, D., Schwartz, R., Weischedel, R., & Stone, R. (1999). Named Entity Extraction From Broadcast News. In *Proceedings of 1999 DARPA Broadcast News Workshop.* San Francisco, CA: Morgan Kaufman Publishers.

Palmer, D., Burger, J., & Ostendorf, M. (1999). Information Extraction From Broadcast News Speech Data. In *Proceedings of 1999 DARPA Broadcast News Workshop*. San Francisco, CA: Morgan Kaufman Publishers.

*Proceedings of the Sixth Message Understanding Conference (MUC-6),* (1995). San Francisco, CA: Morgan Kaufman Publishers.

Weischedel, R., Miller, D., Boisen, S., Schwartz, R., & Stone, R. (2000). Named Entity Extraction from Noisy Input: Speech and OCR. To appear in *Proceedings of the Language Technology Joint Conference, ANLP-NAACL 2000*. San Francisco, CA: Morgan Kaufman Publishers.

Zipf, G (1932). *Selected Studies of the Principle of Relative Frequency in Language.* Cambridge, MA: Harvard University Press

Zipf, G (1949). *Human Behavior and the Principle of Least Effort.* New York: Hafner.