A Survey of Stochastic and Gazetteer Based Approaches for Named Entity Recognition

Benjamin Bengfort
bbengfort@umbc.edu

November 26, 2012

Abstract

The task of identifying proper names of people, organizations, locations, or other entities is a subtask of information extraction from natural language documents. This paper presents a survey of techniques and methodologies that are currently being explored to solve this difficult subtask. After a brief review of the difficulties and challenges of the task, as well as a look at previous conventional approaches, the focus will shift to a comparison of stochastic and gazetteer based approaches. Several machine-learning approaches are identified and explored, as well as a discussion of knowledge acquisition relevant to recognition. This paper will show that applications that require named entity recognition will be served best by some combination of knowledge-based and non-deterministic approaches.

Introduction

In school we were taught that a proper noun was "a specific person, place, or thing," thus extending our definition from a concrete noun. Unfortunately, this seemingly simple mnemonic masks an extremely complex computational linguistic task—the extraction of named entities, e.g. persons, organizations, or locations from corpora [1]. More formally, the task of Named Entity Recognition and Classification can be described as the identification of named entities in computer readable text via annotation with categorization tags for information extraction.

Named entity recognition is not only a subtask of information extraction, but also plays a vital role in reference resolution, other types of disambiguation, and meaning representation in other natural language processing applications. Semantic parsers, part of speech taggers, and thematic meaning representations could all be extended with this type of tagging to provide better results. Other, NER-specific, applications abound including question and answer systems, automatic forwarding, textual entailment, and document and news searching. Even at a surface level, an understanding of the named entities involved in a document provides much richer analytical frameworks and cross-referencing.

Named entities have three top-level categorizations according to DARPA's Message Understanding Conference: entity names, temporal expressions, and number expressions [2]. Because the entity names category describes the unique identifiers of people, locations, geopolitical bodies, events, and organizations, these are usually referred to as named entities and as such, much of the literature discussed in this paper focuses solely on this categorization, although it is easy to imagine extending the proposed systems to cover the full MUC-7 task. Further, the CoNLL-2003 Shared Task, upon which the standard of evaluation for such systems is

based, only evaluates the categorization of organizations, persons, locations, and miscellaneous named entities. For example:

[ORG S.E.C.] chief [PER Mary Shapiro] to leave [LOC Washington] in December.

This sentence contains three named entities that demonstrate many of the complications associated with named entity recognition. First, *S.E.C.* is an acronym for the *Securities and Exchange Commission*, which is an organization. The two words "*Mary Shapiro*" indicate a single person, and *Washington*, in this case, is a location.

Named Entity Recognition Challenges

Some key design decisions in an NER system are proposed in [3] that cover the requirements of NER in the example sentence above:

- Chunking and text representation
- Inference and ambiguity resolution algorithms
- Modeling of Non-Local dependencies
- Implementation of external knowledge resources and gazetteers

Named entities are often not simply singular words, but are chunks of text, e.g. *University of Maryland Baltimore County* or *The Central Bank of Australia*. Therefore, some chunking or parsing prediction model is required to predict whether a group of tokens belong in the same entity. Left to right decoding, Viterbi, and beam search algorithms has been employed as chunking algorithms in the literature. Further, some NER systems are comprised primarily of text parsers as in [4], [5], and [6].

Inference refers to the ability of a system to determine that a chunk is actually a named entity, or, sometimes more importantly, to determine the classification of a named entity, especially in places where there is ambiguity. For example "*Washington*" might refer to either a name or a location. "*Galaxy"* might refer to a generic noun or the professional major league soccer team. Maximum Entropy Models, Hidden Markov Models and other statistical methods are employed to perform this analysis, usually implemented as a machine-learning system, as for instance in [7], [8], and [9].

Non-local dependency models refer to the ability to identify multiple tokens that should have the same label assignment or cross-reference. It is important to note that case becomes important here—e.g. *Bengfort*, *bengfort*, and *BENGFORT* should all be identified as the same entity, and these would break word-level rule based systems (e.g. find all words that are capitalized). But further, even different chunks should be identified similarly – *UMBC* vs. *University of Maryland Baltimore County* or the inclusion of titles in one location that are absent from another as in *President Barack Obama* vs. *Obama*. The non-locality of these models refers to the usage of these terms outside the scope of a sequence of tokens that is being analyzed together (usually a sentence), but rather in the entire document or corpus. The papers that address non-local dependencies include [10] and [11], but the focus of this paper will be on solutions that require external knowledge.

Names, because they uniquely identify entities, are a domain not easily captured by even the most expansive lexicons. For example, simply creating a list of the names of all companies formed in the United States would expand dramatically every single year. However, external knowledge and name lexicons are required for many of the approaches and solutions, not just non-local dependency models. Therefore the construction and use of gazetteers and other resources is necessary.

Approaches to Named Entity Recognition

Generally speaking, the most effective named entity recognition systems can be categorized as rule-based, gazetteer and machine learning approaches. Within each of these approaches are a myriad of sub-approaches that combine to varying degrees each of these top-level categorizations. However, because of the research challenge posed by each approach, typically one or the other is focused on in the literature.

Rule-based systems utilize pattern-matching techniques in text as well as heuristics derived either from the morphology or the semantics of the input sequence. They are generally used as classifiers in machine-learning approaches, or as candidate taggers in gazetteers. Some applications can also make effective use of stand-alone rule-based systems, but they are prone to both overreach and skipping over named entities. Rule-based approaches are discussed in [10], [12], [13], and [14].

Gazetteer approaches make use of some external knowledge source to match chunks of the text via some dynamically constructed lexicon or gazette to the names and entities. Gazetteers also further provide a non-local model for resolving multiple names to the same entity. This approach requires either the hand crafting of name lexicons or some dynamic approach to obtaining a gazette from the corpus or another external source. However, gazette based approaches achieve better results for specific domains. Most of the research on this topic focuses on the expansion of the gazetteer to more dynamic lexicons, e.g. the use of Wikipedia or Twitter to construct the gazette. Gazette based approaches are discussed in [15], [16], and [17].

Stochastic approaches fare better across domains, and can perform predictive analysis on entities that are unknown in a gazette. These systems use statistical models and some form of feature identification to make predictions about named entities in text. They can further be supplemented with smoothing for universal coverage. Unfortunately these approaches require large amounts of annotated training data in order to be effective, and they don't naturally provide a non-local model for entity resolution. Systems implemented with this approach are discussed in [7], [8], [4], [9], and [6].

Evaluation of NERC Systems

Throughout the literature on Named Entity Recognition and Classification systems, two seminal conferences and evaluations are mentioned when evaluating systems: the 1997 Message Understanding Conference (MUC-7) [18] and the CoNLL-2003 Shared Task [1]. Especially since 2003, many NERC systems use the CoNLL-2003 evaluation to demonstrate the performance of their work. This evaluation specifies a data set containing a training file, a development file, a test file, and a large amount of unannotated data in two languages: English and German taken

from news articles corpora. All learning methods were trained on the training file, and tested on the test data. The development file could be used for tuning the parameters of the system.

The data was preprocessed using a tokenizer, chunker and part of speech-tagger, with each token on a single line, and sentence boundaries represented by a single blank line. Named entities were tagged by hand at the University of Antwerp. Performance was measured using an F-Score with β=1 simplified to the harmonic mean:

$$F_{\beta=1} = \frac{2 \times precision \times recall}{precision + recall}$$

Where *precision* is the ratio of correct results to results returned and *recall* is the ratio of correct results to the number of results that should have been returned.

The baseline result was computed for a system that only identified entities with a unique class in the training data, and for CoNLL-2003 was 59.61 ± 1.2. The highest performing system was [7] with an F-Score of 88.76 ± 0.7, followed closely by [8] with an F-Score of 88.31 ± 0.7. However, the lowest performing systems still exceeded the baseline result.

Rule Based Annotation

The first conceit in tackling named entity recognition is to look for clues within the structure and grammar of the text that indicate to readers (and therefore hopefully also systems) that the words and tokens refer to some named entity. Aside from the obvious issues with transcription, spelling, and unique formatting – it turns out that pattern matching is in fact fairly successful at detecting those entities, if only in formal corpora like newsprint. Application of a series of rules as a preprocessing step to reduce the complexity of other techniques is widely used and is exemplified in [10], a rule-based NER annotation system that was supplemented with a machine-learning component to facilitate discovery of domain specific names.

Nadeau et al provide a table of word-level features that may indicate named-entities in [19]. These features include case – if the word is capitalized or all capital letters (or indeed, includes a mixed case), punctuation, morphology, and the part of speech. Features like these, especially capitalization, form the basis for most heuristics. Word level features can then be combined with contextual features to further identify names.

So called “sure-fire” rules are described in [12], which rely on contextual indicators either prepended or appended to a series of capitalized tokens, whose part of speech is unknown or are generically referenced as nouns. These can take the form of titles – Mr., Mrs., or Dr. to name a few that are commonly prepended to names, or Jr., MD, and III, commonly appended. Corporate designators such as Ltd, LLC, or Bank can similarly identify organizations, and address designators like rue, rd., or strasse can identify particular locations.

These grammatical indicators can then be further extended to a series of rules that rely on part of speech tags or positional patterns. For instance, first names are generally easy to identify via a lexicon of proper names, and a capitalized word that follows a first name is most likely a last

name, which tend to be more unique. Other rules require inspection of the context surrounding the possible named entity. In the sentence "… purchased 100 shares of [Tokens+] …" where [Tokens+] is a series of capitalized words most likely the name of a publically traded company. Alternatively, prepositions and other parts of speech can also help us identify locations—for example, Frederick can be either a name or a location (a city in Western Maryland) but use of the phrase "in Frederick" would indicate that it was a location.

Learning and Stochastic Approaches

Out of sixteen participants at the CoNLL-2003 shared task, all employed some form of machine learning [1]. According to the paper, most used a maximum entropy model, others used a statistical learning method, and still others used Hidden Markov Models. Tjong Kim Sang and De Meulder even go so far as to say that because Maximum Entropy was used in both isolation and combined with other systems, it seems to be a good choice for the named entity recognition task.

As for the performance evaluations, a combined system approach faired the best, namely a Classifier Combination approach as in [7]. This approach employed a simple combination of four classifiers with equal voting, namely a Robust Risk Minimization Classifier, a Maximum Entropy Classifier, a Transformation-Based Learning Classifier and a Hidden Markov Model Classifier. However, the approach in [8] used only a Maximum Entropy learner and scored extremely closely to the combined approach, and many of the top scorers used Hidden Markov Models in combination with some other classifier, as in [20], therefore these approaches are the ones this paper will focus on.

Maximum Entropy Classification

Maximum Entropy estimates probability solely on the basis of the constraints derived from training data in an attempt to simplify the estimation by making as few assumptions as possible; in other words, attempting to select the best probability distribution with the information available to the learner. In [8], this is expressed as the following:

$$P(c_1, \dots, c_n | s, D) = \prod_{i=1}^{n} P(c_i | s, D) * P(c_i \, | c_{i-1})$$

Where $P(c_i | s, D)$ is the probability of a classifier given a sequence of tokens and the document that the sequence appears in. Note that the *D* element is used to perform NER with global information across the document, a variation on the Maximum Entropy approach that improved Chieu and Ng's performance over other statistical methods. However, as a general discussion of Maximum Entropy classification, it is not needed here. Dynamic programming algorithms are then used to select word sequences with the highest probabilities.

Classifiers are trained through a series of non-contextual features (previously called word-level features in the rule-based approach), lexical features, morphological features, and global features. For example, non-contextual features include case and zone (the location of the token in the sequence), string information (e.g. the string contains all capital letters and a period), and first word (whether or not the word is at the beginning of a sentence). Lexical features include token

frequencies, the existence of the token in the vocabulary, and common names. Global features can be taken from a Gazette or from information from the rest of the document.

Chunked Tagging with Hidden Markov Models

Hidden Markov Models seem to be naturally applied to named entity recognition annotation because the act of tagging named entities seems closely related to the act of part of speech tagging—and these systems have been successful using HHMs, exemplified in [21]. NER, like part of speech tagging, is a classification problem that involves an inherent state sequence representation. However, unlike in part of speech tagging, the state itself is a sequence of states that needs to be coordinated by some sort of chunking. Zhou and Su tackled this problem in [4] through the following, familiar equation:

$$\text{argmax}\left(\sum_{i=1}^{n} \log\ P(\ g_i \mid t_i) + \log\ P(T_1^n)\right)$$

In order to find the token sequence ($G = g_1, \ldots, g_n$) that maximizes the probability of a tag sequence ($T = t_1, \ldots, t_n$). This allows the chunking program to generate original named entity tags from the original text, and as with other approaches that use Bayes rule, we can calculate $P(g_i \mid t_i)$ using n-gram frequencies in a training set, and $P(T_1^n)$ by similarly counting an annotated training set. Log probabilities are used here to prevent underflow.

Hidden Markov Models rely heavily on training data to both acquire sequence boundary features (although many systems in the literature use a combined approach with classifiers similar to those described in the maximum entropy approach). They also must use smoothing to include untrained data.

Gazetteer Based Approaches

Stochastic methodologies for named entity recognition provide candidates for annotation, and to a certain extent, can list the likelihood of a candidate belonging to a category or subcategory of a named entity. However, this is not a complete solution as machine-learning approaches require knowledge in order to propose candidacy for untrained tokens, and further, after candidates have been proposed, knowledge is required to classify them. Additionally, other issues may be resolved by gazettes, for instance non-local dependencies can be listed in gazettes to indicate cross-references between various names indicating the same entity.

Gazettes, therefore, are utilized to supply external knowledge to learners, or to supply unannotated data with a training source. In fact, most of the teams in the CoNLL-2003 Shared Task made use of a gazette [1], and gazettes have been described as the bottleneck in many learning systems [12] due to the elasticity and rapid evolution and expansion of names as unique descriptions for entities. Therefore, research has been directed toward the development of gazettes as lexicons of named entities rather than towards the specific application of gazettes in a named entity system.

This does not mean, however, that gazette-only systems do not exist. However, because of the overhead of gazette-based look-up (many require a query to the Internet), filtering of corpus tokens is required to generate named entity candidates. Systems described in [15], [16], and [19] use a combination of rule-based mechanisms, part of speech tags, and word frequency analysis to propose these candidates, with no machine learning approach.

Wiki Based Gazetteer Approaches

Looking to the Internet as a source for knowledge that both humans and computers can understand poses its own design challenges. Kazama and Torisawa propose that the advent of open collaborative encyclopedias on the web, like Wikipedia, organize external knowledge for gazette construction in [15]. They make the point that since Wikipedia's articles are intended to be encyclopedic, most articles are about named entities and are more structured than raw text. Therefore, named entity candidates can be resolved by a simple web search to Wikipedia.

The Wikipedia article can then be mined for classification information. Kazama and Torisawa use the first sentence to locate the object of the verb "to be." Richman and Schone utilize the Wikipedia categories attached to the article for classification in their attempt to mine resources for multilingual named entity recognition [16] in order to avoid the use of semantic parsing. Generally speaking, the beginning of the article on an encyclopedic entry can be used as a contextual signature to categorize the named entity. Further, it can be used to provide non-local dependency analysis.

For example the Wikipedia result for "Mekong Delta" contains the contextual clues "region", "river", "sea", "distributaries", "Vietnam", and "water" that could lead to the classification for water based location. Salman Rushdie's article contains the signature words "novelist" and "essayist" in order to classify this entity as a writer (person). This technique of classification is similar to the word disambiguation techniques used in the Lesk algorithms.

Wikipedia also provides for non-local dependency analysis through two unique features. First, each article concerns only one particular named entity. Therefore any referential names found in the article can be said to identify that particular entity. Further, Wikipedia provides as a tool a page called "disambiguation" that will present the most frequent searches first, along with categorization. For instance, a search for "Cambridge (disambiguation)" will reveal several geographic place names in England, the United States, Canada, and Australia. It will also expose the University of Cambridge as another common use of that name. Mining contextual clues in these Wikipedia articles for a signature, and comparing them to the signature of the input provides not only the correct categorization, but also a source for variations on the name for use in cross-referencing.

Both [15] and [16] scored well on the CoNLL-2003 evaluation scale. Kazama et al. published their results with an F-Score of 88.02, which compares very well to the best performing stochastic methods. They attribute the increase in score from the baseline as a result of improved precision of results rather than improved recall. However, neither of these methods supplied details of the performance cost associated with a Wikipedia gazette based approach.

Mechanical Turk Approaches

One interesting approach to named-entity knowledge acquisition deserves a small mention here. Lawson et al. made use of Amazon's Mechanical Turk service to annotate a data set of emails with named entities in [22]. The Mechanical Turk is a low cost way of providing human labor to a repetitive task by paying some small amount of money for a small work unit. This paper demonstrates the difficulties in having human annotated training data available for statistical or gazette based analysis.

For even the seemingly simple task of identifying the categories for PERSON, LOCATION, or ORGANIZATION in emails of no more than 60 to 900 characters, Lawson et al. found that they could not pay a flat rate per email, but instead had to set up a reward mechanism in which each discovered entity paid off (otherwise annotators simply checked no-entries to game the system). Although the authors argue that they were able to use inter-annotator agreement to assign a particular number of workers to the same task to reduce costs, it seems that only a portion of the annotators provided significant contribution as 12.5% of the workers completed 74.9% of the annotation! Although the final annotated data set was of a high quality, and a low economic cost, the dataset took more than four months to build using this methodology.

Conclusion

This paper explored the topic of named entity recognition, particularly as described in the CoNLL-2003 shared task: language-independent named entity recognition of organizations, locations, and people. Several of the pitfalls associated with named entity recognition systems were discussed, including chunking and text representation for multi-word names, inference for resolving name ambiguity, cross-referencing with non-local dependencies, and the use of external knowledge in NER systems even though unique names provide a lexical challenge.

Several approaches to named entity recognition were discussed including rule-based systems, machine-learning systems and gazetteer approaches. Rule based systems provide an advantage of not requiring significant computation, but have low accuracy and recall. These systems are often used in coordination with gazette based and machine learning systems. Stochastic methods focus on probabilistic modeling and can cover new tokens and decide if sequences of tokens belong together, however they require a large training set to operate, as well as a smoothing implementation to capture universal knowledge. Finally, gazetteer, or dynamic lexical generation, was discussed in accordance with named entity recognizers. This approach performs well at discovering non-local dependencies as well as disambiguating multiple name senses. However, candidates must be proposed for gazette look-up, and some overhead is incurred in accessing the gazette.

For now, combinations of machine learning and gazetteer systems, which use rule-based classifiers for features and candidate proposal, are the best performing systems. Further research needs to be performed on multi-stage approaches to better integrate the two methodologies.

References

[1] Erik F. Tjong Kim Sang and De Meulder Fien, "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," in *CONLL '03 Proceedings of the 7th Conference on Natural Language Learning*, vol. 4, Stroudsburg, PA, 2003, pp. 142-147.

[2] Nancy Chinchor, Erica Brown, Lisa Ferro, and Patty Robinson, "1999 Named Entity Recognition Task Definition," MITRE and SAIC, 1999.

[3] Lev Ratinov and Dan Roth, "Design Challenges and Misconceptions in Named Entity Recognition," in *CoNLL '09 Proceedings of the 13th Conference on Computational Natural Language Learning*, Stroudsburg, PA, 2009, pp. 147-155.

[4] GuoDong Zhou and Jian Su, "Named Entity Recognition Using an HMM-Based Chunk Tagger," in *ACL '02 Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, 2002, pp. 473-480.

[5] Jenny Rose Finkel and Christopher D. Manning, "Joint Parsing and Named Entity Recognition," in *NAACL '09 Proceedings of Human Language Technologies*, Stroudsburg, Pa, 2009, pp. 326-334.

[6] Hirotaka Funayama, Tomohide Shibata, and Sadao Kurohashi, "Bottom-Up Named Entity Recognition Using a Two-Stage Machine Learning Method," in *MWE '09 Proceedings of the Workship on Multiword Expressions: Identification, Interpretation, Disambiguation, and Applications*, Stroudsburg, PA, 2009, pp. 55-62.

[7] Radu Florian, Abe Ittycheriah, Hongyan Jing, and Tong Zhang, "Named Entity Recognition Through Classifier Combination," in *CONLL '03 Proceedings of the 7th Conference on Natural Language Learning*, vol. 4, Stroudsburg, PA, 2003, pp. 168-171.

[8] Hai Leong Chieu and Hwee Tou Ng, "Named Entity Recognition: A Maximum Entropy Approach Using Global Information," in *COLING '02 Proceedings of the 19th International Conference on Computational Linguistics*, vol. 1, Stroudsburg, PA, 2002, pp. 1-7.

[9] Andrew McCallum and Wei Li, "Early Results for Named Entity Recognition with Conditional Random Fields, Feature Induction, and Web-Enhanced Lexicons," in *CONLL '03 Proceedings of the 7th Conference on Natural Language Learning*, vol. 4, Stroudsburg, PA, 2003, pp. 188-191.

[10] Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan, "Domain Adaption of Rule-Based Annotators for Named Entity Recognition Tasks," in *EMNLP '10 Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, 2010, pp. 1002-1012.

[11] Vijay Krishnan and Christopher D. Manning, "An Effective Two-Stage Model for Exploiting Non-Local Dependencies in Named Entity Recognition," in *ACL-44 Proceedings of the 21st International Conference on Computational Linguistics*, Stroudsburg, PA, 2006, pp. 1121-

1128.

[12] Andrei Mikheev, Marc Moens, and Claire Grover, "Named Entity Recognition Without Gazetteers," in *EACL '99 Proceedings of the 9th Conference on the European Chapter of the Association for Computational Linguistics*, Stroudsburg, PA, 1999, pp. 1-8.

[13] Silviu Cucerzan and David Yarowsky, "Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence," in *Joint SIGDAT Conference on EMNLP and VLC*, 1999.

[14] Dimitra Farmakioutou et al., "Rule-Based Named Entity Recognition for Greek Financial Texts," in *Proceedings of the Workshop on Computational Lexicography and Multimedia Dictionaries*, 2000, pp. 75-78.

[15] Jun'ichi Kazama and Kentaro Torisawa, "Exploiting Wikipedia as External Knowledge for Named Entity Recognition," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 698-707.

[16] Alexander E. Richman and Patrick Schone, "Mining Wiki Resources for Multilingual Named Entity Recognition," in *Proceedings of the 46th Annual Meeting of the Association of Computational Linguistics: Human Language Technologies*, Stroudsburg, PA, 2008, pp. 1-9.

[17] Alan Ritter, Sam Clark, Mausam, and Oren Etzioni, "Named Entity Recognition in Tweets: An Experimental Study," in *EMNLP '11 Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, 2011, pp. 1524-1534.

[18] Nancy Chinchor, "MUC-7 Named Entity Task Definition," in *Message Understanding Conference*.

[19] David Nadeau and Satoshi Sekine. (2008, Jan.) A Survey of Named Entity Recognition and Classification. [Online]. http://nlp.cs.nyu.edu/sekine/papers/li07.pdf

[20] Dan Klein, Joseph Smarr, Huy Nguyen, and Christopher D. Manning, "Named Entity Recognition with Character-Level Models," in *Proceedings of CoNLL-2003*, 2003, pp. 180-183.

[21] Scott M Thede and Mary P Harper, "A Second-Order Hidden Markov Model for Part-of-Speech Tagging," in *Proceedings of the 37th annual meeting of ACL*, Stradsbourg, PA, 1999, pp. 175-182.

[22] Nolan Lawson, Kevin Eustice, Mike Perkowitz, and Meliha Yetisgen-Yildiz, "Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk," in *CSLDAMT '10 Proceedings of the NAACL HLT 2010 Workshop on Creating Speach and Language Data with Amazon's Mechanical Turk*, Stroudsburg, PA, 2010, pp. 71-79.