Bootstrapping for Named Entity Tagging Using Concept-based Seeds

Cheng Niu, Wei Li, Jihong Ding, Rohini K. Srihari
Cymfony Inc.
600 Essjay Road, Williamsville, NY 14221. USA.
{cniu, wei, jding, rohini}@cymfony.com

Abstract

A novel bootstrapping approach to Named Entity （NE）tagging using concept-based seeds and successive learners is presented. This approach only requires a few common noun or pronoun seeds that correspond to the concept for the targeted NE, e.g. *he/she/man/woman* for PERSON NE. The bootstrapping procedure is implemented as training two successive learners. First, decision list is used to learn the parsing-based NE rules. Then, a Hidden Markov Model is trained on a corpus automatically tagged by the first learner. The resulting NE system approaches supervised NE performance for some NE types.

1 Overview

Recognizing and classifying proper names is a fundamental task for information extraction. Three types of proper names are defined in the Message Understanding Conference (MUC) Named Entity (NE) standards, namely, PERSON (PER), ORGANIZATION (ORG), and LOCATION (LOC). [MUC-7 1998]

There is considerable research on NE tagging using supervised machine learning [e.g. Bikel *et al.* 1997; Borthwick 1998]. To overcome the knowledge bottleneck of supervised learning, unsupervised machine learning has been applied to NE. [Cucchiarelli & Velardi 2001] discussed boosting the performance of an existing NE tagger by unsupervised learning based on parsing structures. [Cucerzan & Yarowsky 1999], [Collins & Singer 1999] and [Kim *et al.* 2002] presented various techniques using co-training schemes for NE extraction seeded by a small list of proper names or hand-crafted NE rules. NE tagging has two tasks: (i) NE chunking; (ii) NE classification. Parsing-supported unsupervised NE learning systems including ours only need to focus on NE classification, assuming the NE chunks have been constructed by the parser.

This paper presents a new bootstrapping approach using successive learning and concept-based seeds. The successive learning is as follows. First, parsing-based NE rules are learned with high precision but limited recall. Then, these rules are applied to a large raw corpus to automatically generate a tagged corpus. Finally, a high-performance HMM-based NE tagger is trained using this corpus.

Unlike co-training, our bootstrapping does not involve iterative learning between the two learners, hence it suffers little from error propagation which is commonly associated with iterative learning.

To derive the parsing-based learner, the system only requires a few common noun or pronoun seeds that correspond to the concept for the targeted NE, e.g. *he/she/man/woman* for PERSON NE. Such concept-based seeds share grammatical structures with the corresponding NEs, hence a parser is utilized to support bootstrapping. Since pronouns and common nouns occur more often than NE instances, the parsing-based NE rules can be learned in one iteration to avoid iterative learning.

The benchmarking shows that this system approaches the performance of supervised NE taggers for two of the three proper name NE types in MUC, namely, PER NE and LOC NE. This approach also supports tagging user-defined NE types.

2 Implementation

Figure 1 shows the overall system architecture. Before the bootstrapping is started, a large raw training corpus is parsed. The bootstrapping experiment reported in this paper is based on a corpus containing ~100,000 news articles and totally

~88,000,000 words. The parsed corpus is saved into a repository, which supports fast retrieval by keyword based indexing scheme.

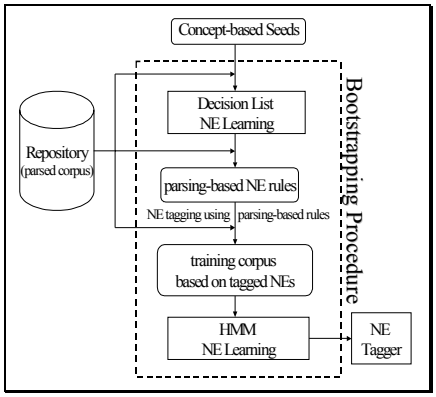

Figure 1. Bootstrapping System Architecture

The unsupervised bootstrapping is performed as follows:

1. User provides concept-based seeds;
2. Retrieve parsing structures involving concept-based seeds from the repository to train a decision list for NE classification;
3. Apply the learned rules to the NE candidates retrieved from the repository;
4. Construct an NE annotated corpus using the tagged proper names and their neighboring words;
5. Train an HMM based on the annotated corpus.

A parser is necessary for concept-based NE bootstrapping. This is due to the fact that concept-based seeds only share pattern similarity with the corresponding NEs at structural level, not at string sequence level. In fact, the anaphoric function of pronouns and common nouns to represent antecedent NEs indicates the substitutability of proper names by the noun phrases headed by the corresponding common nouns or pronouns. For example, *this man* can substitute the proper name *John Smith* in almost all structural patterns.

Five binary dependency relationships decoded by our parser are used for parsing-based NE rule learning: (i) a Has_Predicate(b): from logical subject *a* to verb *b*; (ii) a Object_Of(b): from logical object a to verb b; (iii) a Has_Amod(b): from noun *a* to its adjective modifier *b*; (iv) a Possess(b): from the possessive noun-modifier *a* to head noun *b*; (v) a IsA(b): equivalence relation (including appositions) from one NP *a* to another NP *b*.

The concept-based seeds used in the experiments are: (i) *he, she, his, her, him, man, woman* for PER; (ii) *city, province, town, village* for LOC; (iii) *company, firm, organization, bank, airline, army, committee, government, school, university* for ORG.

From the parsed corpus in the repository, all instances (821,267) of the concept-based seeds involved in the five dependency relations are retrieved. Each seed instance was assigned a concept tag corresponding to NE. For example, each instance of *he* is marked as PER. The instances with concept tagging plus their associated parsing relationships are equivalent to an annotated NE corpus. Based on this training corpus, the Decision List Learning algorithm [Segal & Etzioni 1994] is used. The accuracy of each rule was evaluated using Laplace smoothing as follows,

$$accuracy = \frac{\text{positive} + 1}{\text{positive} + \text{negative} + \text{NE category No.}}$$

As the PER tag dominates the corpus due to the high occurrence frequency of *he* and *she,* learning is biased towards PER as the answer. To correct this bias, we employ the following modification scheme for instance count. Suppose there are a total of N_{PER} PER instances, N_{LOC} LOC instances, N_{ORG} ORG instances, then in the process of rule accuracy evaluation, the involved instance count for any NE type will be adjusted by the coefficient $\frac{\min(N_{\text{PER}}, N_{\text{LOC}}, N_{\text{ORG}})}{N_{\text{NE}}}$.

A total of 1,290 parsing-based NE rules, shown in samples below, are learned, with accuracy higher than 0.9.

Possess(*wife*) → PER
Has_Predicate(*divorce*) → PER
Object_Of(*deport*) → PER
Possess(*mayor*) → LOC
Has_AMod(*coastal*) → LOC
Possess(*ceo*) → ORG
Has_AMod(*non-profit*) → ORG
Has_AMod(*non-governmental*) → ORG
……

Due to the unique equivalence nature of the IsA relation, we add the following *IsA*-based rules to the top of the decision list: IsA(seed)→ tag of the seed, e.g. IsA(*man*) → PER

The parsing-based first learner is used to tag a raw corpus. First, we retrieve all the named entity candidates associated with at least one of the five parsing relationships from the repository. After applying the decision list to the retrieved 1,607,709 NE candidates, 33,104 PER names, 16,426 LOC names, and 11,908 ORG names are tagged. In order to improve the bootstrapping performance, we use the heuristic *one tag per domain for multi-word NE* in addition to the *one sense per discourse* principle [Gale *et al* 1992]. These heuristics are found to be very helpful in both increasing positive instances (i.e. tag propagation) and decreasing the spurious instances (i.e. tag elimination). The tag propagation/elimination scheme is adopted from [Yarowsky 1995]. After this step, a total of 367,441 proper names are classified, including 134,722 PER names, 186,488 LOC names, and 46,231 ORG names.

The classified proper name instances lead to the construction of an automatically tagged training corpus, consisting of the NE instances and their two (left and right) neighboring words within the same sentence.

In the final stage, a bi-gram HMM is trained based on the above training corpus. The HMM training process follows [Bikel 1997].

3 Benchmarking

We used the same blind testing corpus of 300,000 words containing 20,000 PER, LOC and ORG instances to measure performance degradation of unsupervised learning from the existing supervised NE tagger (Table 1, P for Precision, R for Recall, F for F-measure and F/D for F-measure degradation).

Table 1: Supervised-to-Unsupervised NE Degradation

	Supervised NE			Unsupervised NE			
TYPE	P	R	F	P	R	F	F/D
PER	92.3%	93.1%	**92.7%**	86.6%	88.9%	**87.7%**	5.0%
LOC	89.0%	87.7%	**88.3%**	82.9%	81.7%	**82.3%**	6.0%
ORG	85.7%	87.8%	**86.7%**	57.1%	48.9%	**52.7%**	34.0%

The performance for PER and LOC are above 80%, and approaching the performance of supervised learning. The reason of the unsatisfactory performance of ORG (52.7%) is not difficult to understand. There are numerous sub-types of ORG that cannot be represented by the less than a dozen concept-based seeds used for this experiment.

In addition to the key NE types in MUC, we also tested this method for recognizing user-defined NE types. We use the following concept-based seeds for PRODUCT (PRO) NE: *car, truck, vehicle, product, plane, aircraft, computer, software, operating system, database, book, platform, network*. Table 2 shows the benchmarks for PRODUCT tagging.

Table 2: Performance for PRODUCT NE

TYPE	PRECISION	RECALL	F-MEASURE
PRODUCT	67.27%	72.52%	69.80%

References

Bikel, D. M. 1997. Nymble: a high-performance learning name-finder. *Proceedings of ANLP'*97, 194-201, Morgan Kaufmann Publishers.

Borthwick, A. *et al*. 1998. Description of the MENE named Entity System. *Proceedings of MUC-7*.

Collins, M. and Y. Singer. 1999. Unsupervised Models for Named Entity Classification. *Proceedings of the Joint SIGAT Conference on EMNLP and VLC*. ???Association for Computational Linguistics, 1999.

Cucchiarelli, A. and P. Velardi. 2001. Unsupervised Named Entity Recognition Using Syntactic and Semantic Contextual Evidence. *Computational Linguistics*, Volume 27, Number 1, 123-131.

Cucerzan, S. and D. Yarowsky. 1999. Language Independent Named Entity Recognition Combining Morphological and Contextual Evidence. *Proceedings of the Joint SIGDAT Conference on EMNLP and VLC*, 90-99.

Gale, W., K. Church, and D. Yarowsky. 1992. One Sense Per Discourse. *Proceedings of the 4th DARPA Speech and Natural Language Workshop*. 233-237.

Kim, J., I. Kang, and K. Choi. 2002. Unsupervised Named Entity Classification Models and their Ensembles. *Proceedings of COLING 2002*.

MUC-7, 1998. Proceedings of the Seventh Message Understanding Conference (MUC-7), published on the website http://www.muc.saic.com/

Segal, R. and O. Etzioni. 1994. Learning decision lists using homogeneous rules. *Proceedings of the 12th National Conference on Artificial Intelligence*.

Yarowsky, David. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Method. *Proceedings of ACL* 1995.