Named Entity: History and Future

Satoshi Sekine

New York University
715 Broadway, 7th floor
NewYork, NY 10003 USA
sekine@cs.nyu.edu

This paper surveys Named Entity (NE) related research. It includes the history of NE, the problems it faces at the moment, and the possible directions for solving these problems in the future. We hope the paper will be useful for people who are interested in the research on NE and technology related to it.

1. Orthodox Named Entity

The term "Named Entity (NE)", widely used in Information Extraction (IE), Question Answering (QA) or other Natural Language Processing (NLP) applications, was born in the Message Understanding Conferences (MUC) which influenced IE research in the U.S. in the 1990's [Grishman and Sundheim 1996] (to be precise, it was introduced for MUC-6 in 1995). At that time, MUC focused on IE tasks where structured information of company activities and defense related activities is extracted from unstructured text, such as newspaper articles. In the course of system development, people noticed that it is important to recognize information units like names, including person, organization and location names, and numeric expressions including time, date, money and percent expressions. Extracting these entities was recognized as one of the important sub-tasks of IE. As this task is relatively independent, it has been evaluated separately in several different languages, e.g. Japanese, Chinese and Spanish in MET (Multilingual Entity Tracking) project. Outside the U.S., there have been several evaluation-based projects for NE, as one of the tasks of IREX (Information Retrieval and Extraction Exercise) in Japan [Sekine and Isahara 2000] [IREX HP], and as the shared task in CoNLL in 2002 and 2003 for four languages, English, German, Dutch and Spanish [CoNLL HP]. In the IREX project, a new category "artifact", such as "Odyssey" as a book title or "Windows" as a product name, was added to the original MUC categories. The NE task in MUC was inherited by the ACE project in the U.S., where 2 new categories are added, GPE (Geographical and Political Entities, such as "France" or "New York") and facility, such as "Empire State Building". Around this time, the number of categories is limited to 7 to 10, and the NE taggers, automatic annotation systems for NE entities in unstructured text, are based on 1) dictionaries and rules which were made by hand or 2) some supervised learning technique. More recent and currently dominating technology is the supervised learning techniques, which include Decision Tree [Sekine 1998], Hidden Markov Model (HMM) [Bikel et. al 1997], Maximum Entropy Model (ME) [Borthwick 1998], Support Vector Machine (SVM) [Asahara 2003], Boosting and voted perceptron [Collins 2000] and Conditional Random Fields (CRFs) [McCallum and Li 2003]. The NE extraction task has been the experimental sandbox for various forms of supervised learning.

2. Need for extended Named Entity

As already explained, the categories of NE have different variations, but the expansion of IE applications and the appearance of a new related task, Question Answering (QA) gave rise to the need for a drastic extension of NE categories. In the IE field, the category of IE depends on the scenario (the target information class of the

IE). When the scenario was “corporate merger”, the seven categories were enough; however, once the scenario changed to “rocket launch” or “disease outbreak”, we need “names of rockets” or “disease names”. The wider the IE tasks become, the more categories of NE are needed. QA research aims to make a system which can produce an answer like “Nobel Peace Prize” to a question like “What is the name of an international prize the Korean President Kim Dae received in 2002?” [TREC-QA HP]. NE plays an important role in creating such a system. Typical systems analyze the answer type from the question sentence, in our example, from “prize”, and the system searches for an answer of the analyzed type based on evidence such as a keyword in near context. As we can easily imagine, we need a large number of categories in order to create a system capable of answering a wide variety of questions.

Also, there are new fields where an NE-related task becomes an important component technology. For example, in bio-informatics, recognizing names of proteins or genes is crucial. There are on-going efforts to make such an extended NE [ISI HP] [Sekine 2002] [GENIA ontology].

3. Problems caused by the extension

In this section, we will discuss the two major problems caused by the extension of NE categories. The discussion is based on the author’s own experience of creating a “200 category extended named entity hierarchy”. The 200 category extended NE was created for the purpose of covering major general newspaper scenarios.

The first problem is the definition of the categories. Even when we had a small number of categories, there were long discussions, arguments, and a bit artificial solution to define the categories. Now the problem becomes much prominent, although in some particular cases, extending the categories helps in finding the right category for an entity. For example, “The Supreme Court”, which was ambiguous between a location and an organization in the smaller NE can be clearly classified as a GOE, which is a facility with identity. However, there are more cases where it becomes difficult to find the right category, e.g. whether a civil strife is a war or an affair, whether a typhoon with a minor casualties is a natural phenomena or a natural disaster, ambiguity between a religion and the group of people who believe it, or the definition of ethnic group. This is the problem of categorizing the world into semantic categories, and finding the right category for each word (of each occurrence). We believe that there is no ultimate solution, so we seek rather empirical solution. In the development of our categories, we empirically surveyed a large volume of newspaper articles, application systems like Information Extraction and Question Answering and already created thesaurus. Also, we made definitions with a lot of examples in addition to the verbal definition of each category, as the verbal definition itself can’t be concrete enough to define most categories.

The other problem caused by the extension of the categories is the problem of NE tagging actual text. When we have a small number of categories, people use supervised learning methods based on a relatively large amount of annotated data. However it is not so easy now with a large number of categories. Actually, we annotated 30 days of Japanese newspaper articles with 200 categories [Sekine 2004], but it is not easy to maintain consistency and there are categories for which not so many examples are found in the newspaper articles. Another strategy is to use hand-crafted dictionaries and rules. Here we also tried to gather lists of names from the Web, but lists could not cover all and sometimes the expressions are incompatible with those in newspaper articles. As far as I know, there is no good solution to this problem and many people in this area are struggling creating dictionaries and rules by hand.

4. Directions to the solution

However, there are possible directions to solve the problems. The basic idea for the all directions mentioned

here is to reduce the burden of annotation/supervision in the training, as the fully supervised training methods seem unreasonable. Also, the availability of a huge corpus plays an important role. Currently huge corpora, like tens of years of newspaper or more than 10GB of Web texts, are available. We have an opportunity to utilize the data and extract useful information from them. An example of the utilization of a huge corpus, although not an NE application, is reported in [Banko and Brill 2001].

Weakly Supervised

The term "weakly supervised" (or semi-supervised) has been used relatively recently. The major technique in this category is called "bootstrapping". In this method, a small degree of supervision, such as a set of seeds, is used at the beginning. For example, in order to extract "disease names", five initial disease names are given to the system. Then the system searches the sentences that contain the names, and finds the strong indicators of context where the five disease names appear. Now, the system tries to find other instances that appear in the context. By repeating these processes, a large number of disease names can be gathered. Examples of this type of research include [Riloff 1999] [Yangarber 2000] [Collins 1999]. The same method can be applied to extract relations of entities, such as book title and author from seeds like "Shakespeare" and "Hamlet" [Brin 1998] [Lin and Pantel 2000].

Active Learning

In the supervised methods, in general, the larger the training data, the better accuracy the system can get. However, annotating a large corpus is not easy. So an attractive alternative idea is to annotate only the data which can help to improve the overall accuracy. The favorite practice is to annotate the data which is tagged with uncertainty by the current system. There are several studies in this area, although the target is not named entities [Ngai and Yarowsky 2000] [Sassano et .al 2002].

Unsupervised Learning

The typical approach of unsupervised learning is clustering. For example, we can try to gather named entities from the clustered groups based on the similarity of context.

There are unsupervised methods other than clustering. One of the promising examples is to use linguistic knowledge in order to extract information from a huge corpus. [Hearst 1992] proposed to extract examples of hyponym and hypernym relationships using phrases like "A such as B". We believe that similar technique can be applied in order to extract named entities.

These research efforts focused on extracting information using the knowledge underlying a corpus, rather than looking at the corpus as a sequence of characters. Anyway, it is only very recently that we can afford to deal with more than 1GB texts; we believe we are about breaking a new ground.

5. Relationships to the other research area

Terminology

There are close relationships between terminology research and NE research. A definition of "term" is controversial, but one of the definitions is "lexical units mostly/mainly used in a specific domain" [Kageura 2002]. The problem is what is a "domain". If we view the newspaper as a set of different domains, like politics, economics, sports or entertainment, the named entities for newspaper are genuine terms by this definition. Also in the bio-informatics field, the names of proteins or genes are surely terms, even though the current technique used in extracting such expressions inherits NE recognition technology. In terms of finding named entities, in particular those in small categories in the extended NE, we believe the technology developed in terminology will be very useful.

Sense disambiguation and thesaurus

As the number of categories becomes larger, NE tagging becomes more similar to the problem of sense disambiguation or finding the appropriate node in a thesaurus for the entity, as evaluated in SENSEVAL [Senseval HP]. Indeed there are Question Answering systems which use a thesaurus like WordNet [WordNet HP] for finding the type of an entity. However, the current WordNet includes mostly common nouns rather than proper nouns or names. Also it is not so conceivable that each node in the thesaurus will include corresponding proper names. However, in Japanese, a recently published thesaurus, Nihongo-goi-taikei, includes common nouns and proper nouns together, and it is a useful resource for the study of named entities.

Unknown word problem

Named entities are often unknown words which are not in the dictionary. Especially this fact causes a serious problem in languages without word delimiters like Japanese or Chinese. Most morphological analyzers are confused when they encounter unknown words and segment them in an inappropriate manner. There are several studies to deal with the unknown word problem in Japanese [Mori and Nagao 1996] [Uchimoto et. al 2001].

5. Summary

We briefly presented the history of named entity research, the problems we encounter and directions for a possible solution. Currently, the named entity task is changing from tagging only proper names to tagging a broader range of words and expressions that are of interest to people with particular information needs. In any case, the named entity or extended named entity task is definitely one of the important component technologies for the applications of natural language technologies.

Bibliography

M. Asahara and Y. Matsumoto: “Japanese Named Entity Extraction with Redundant Morphological Analysis”, HLT-NAACL 2003.

M. Banko and E. Brill: “Scaling to very very large corpora for natural language disambiguation. ACL/EACL 2001.

D. Bikel, S. Miller, Richard Schwartz and Ralph Weischedel: “Nymble: a High-Performance Learning Name Finder” ANLP 1997.

A. Borthwick, J. Sterling, E. Agichtein, R. Grishman, “Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition”, WVLC 98.

B. S. Brin: “Extracting Patterns and Relations from the World Wide Web” EDBT 98.

M. Collins, Y. Singer “Unsupervised Models for Named Entity Classification”, EMNLP 1999.

M. Collins: “Ranking Algorithms for Named–Entity Extraction: Boosting and the Voted Perceptron”, ACL 2000

E. Riloff, R. Jones: “Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping”, AAAI 1999.

R. Grishman, B. Sundheim, “Message Understanding Conference - 6: A Brief History”, COLING-96.

M. Hearst: “Automatic Acquisition of Hyponyms from Large Text Corpora” COLING 1992.

K. Kageura and B. Umino: “Methods of Automatic Term Recognition”, *Terminology.* vol. 3, no. 2, 1996

K. Kageura: “The Dynamics of Terminology: A Descriptive Theory of Term Formation and Terminological Growth.” Amsterdam: John Benjamins, 2002.

D. Lin, P. Pantel: “Concept Discovery from Text”, COLING 2000.

A. McCallum and W. Li, “Early Results for Named Entity Recognition with Conditional Random Fields, Fetures Induction and Web-Enhanced Lexicons”, CoNLL 2003.

S. Mori and M. Nagao: “Word Extraction from Corpora and Its Part-of-Speech Estimation Using Distributional Analysis”, COLING 1996.
M. Sassano: “An Empirical Study of Active Learning with Support Vector Machines for Japanese Word Segmentation”, ACL 2002.
S. Sekine, R. Grishman, H. Shinnou “A Decision Tree Method for Finding and Classifying Names in Japanese Texts”, WVLC 98.
S. Sekine, H. Isahara “IREX: IR and IE Evaluation project in Japanese”, LREC 2000.
S. Sekine, K. Sudo, C. Nobata: “Extended Named Entity Hierarchy”, LREC 2002.
S. Sekine: “Definition, dictionaries and tagger for Extended Named Entity Hierarchy”, LREC 2004.
G. Ngai and D. Yarowsky: “Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking”, ACL 2000.
K. Uchimoto, S. Sekine and H. Isahara: “The Unknown Word Problem: a Morphological Analysis of Japanese Using Maximum Entropy Aided by a Dictionary”, EMNLP 2001.
R. Yangarber, W. Lin, R. Grishman: “Unsupervised learning of generalized names” COLING 2002.
ACE HP: http://www.itl.nist.gov/iaui/894.01/tests/ace.
IREX homepage http://nlp.nyu.edu/irex
TREC-QA homepage http://trec.nist.gov/
CoNLL homepage: http://cnts.uia.ac.be/conll2003/
Senseval homepage: http://www.senseval.org/
WordNet HP: http://www.cogsci.princeton.edu/~wn/
The ISI Question Answer Typology：Webclopedia HP: http://www.isi.edu/natural-language/projects/webclopedia/
GENIA ontology HP: http://www-tsujii.is.s.u-tokyo. ac.jp/ ~genia/topics/Corpus/genia-ontology.html
Sekine’s NE homepage: http://nlp.cs.nyu.edu/jneh/