Definition, dictionaries and tagger for Extended Named Entity Hierarchy

Satoshi Sekine

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

Chikashi Nobata

Communications Research Laboratory
3-5 Hikaridai, Seika-chou, Soraku-gun
Kyoto, 619-0289, Japan
nova@crl.go.jp

Abstract

The tagging of Named Entities, the names of particular things or classes, is regarded as an important component technology for many NLP applications. The first Named Entity set had 7 types, organization, location, person, date, time, money and percent expressions. Later, in the IREX project artifact was added and ACE added two, GPE and facility, to pursue the generalization of the technology. However, 7 or 8 kinds of NE are not broad enough to cover general applications. We proposed about 150 categories of NE (Sekine et al. 2002) and now we have extended it again to 200 categories. Also we have developed dictionaries and an automatic tagger for NEs in Japanese.

Introduction

The tagging of Named Entities, the names of particular things or classes, is regarded as an important component technology for many NLP applications. These applications include question answering (QA), summarization, information retrieval (IR) and information extraction (IE), from which it was born. The first Named Entity set had 7 types (Grishman and Sundheim 1996), organization, location, person, date, time, money and percent expressions. Later, in the IREX project (Sekine and Isahara 2000) artifact (for example, product name "Windows" or book title "Odyssey") was added and ACE (ACE homepage) added two, GPE (location with political function, such as Portugal or Lisbon) and facility (such as building name), to pursue the generalization of the technology. However, 7 or 8 kinds of NE are not broad enough to cover general applications. For example, when we encounter a new IE domain "airplane crashes" we need a new NE category "name of airplane". In QA applications, in order to answer the question "What is the name of an international prize the Korean President Kim Dae received in 2002?", we need to have a NE category "name of prize".

Aiming to cover such needs, we proposed about 150 categories of Extended NE (Sekine et al. 2002). However, in the process of developing system applications since then, we noticed the need for additional Extended NE categories and now we have extended it again to 200 categories. In this paper we will describe the Extended NE and report on our effort to create the dictionaries and an automatic tagger for ENEs in Japanese.

Initial Extended NE

We reported, at the previous LREC (Sekine et. al 2002), that we designed the Extended NE hierarchy based on three procedures. Note that our hierarchy is intended as a general hierarchy for newspaper domains, rather than a special hierarchy for a particular domain. This is because our target application is general IE or QA.

1) Based on a newspaper corpus

We extracted about 3500 candidate NE expressions from a corpus, using surface clues, namely capitalized words and the context of numerals. We assigned NE categories to each expression.

2) Based on existing systems and tasks

There are many systems and tasks related to NE, e.g. (ISI Webclopedia HP)

3) Based on thesaurus

A thesaurus provides data closely related to the NE hierarchy. We consulted two well-known thesauri, WordNet (WordNet HP) and Roget Thesaurus.

Our hierarchy was used for several applications, and some people referred to it to design their own hierarchy. We also developed an English tagger based on the hierarchy, and used it in several applications.

200 category Extended NE

We used the 150 categories for our applications, such as QA and IE systems in the newspaper domain. In the process of system development, we observed that 150 categories are not yet enough. We extended our hierarchy again, to about 200 categories. We are trying to make the Extended NE hierarchy cover major newspaper domains so that applications in such domains will work well with this hierarchy.

One of the major changes is that the non-terminal categories now have no instances. Instead, we created an "other" category for each non-terminal to put the instances which do not fit in any of its child categories. This is to make use of rule/feature inheritance. We used to put such instances to the non-terminal node, but then there is no way to make a rule specifically for such "other" instances. For example, in a QA application, we listed groups of noun phrases for each category. Then, for a question like "What is the name the area where the WTC used to exist?", the answer category (Ground Zero) is a location which shall be found by the word "area" in the question. But it doesn't fall into any of the subcategories of location, so it should be "location_other". In the old scheme, where such entities belonged to the non-terminal node "location", there was a conflict in that we would have to assign "area" as a noun phrase for all locations, which is not the case, as you don't usually use the noun phrase "area" to indicate the name of a star, which is one of the location categories. To avoid this problem, we introduced an "other" category for each non-terminal category.

Also, we extended several categories, mostly to make the existing categories finer. For example, we introduced

"natural disaster", as we found that such things are often reported in newspapers and useful for many applications, such as QA and IE. We extended numeric expressions, too. For example, “school age” (e.g. second grade) was introduced.

It is obvious that the more categories we have, the more difficult it is to make clear definitions, 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 7 or 8 category NE can be clearly classified as a GOE (Geographical Organizational Entities), which is a facility with an 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 minor casualties is a natural phenomena or a natural disaster, the ambiguity of a religious name between the religion and the group of people who believe it, or the definition of an 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 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.

Definition

We created a hyper-linked definition document in Japanese (available on the web http://nlp.cs.nyu.edu/ene), which is about 150 pages long. An English version is also available, but currently not as detailed as the Japanese version. The definition includes 3 major parts.

1) Global definition and explanation
   This includes the goal and background of the hierarchy, and the sentence to test whether an instance belongs to a category. Suppose we are testing if instance *I* belongs to Category *C*. We use a sentence “Tell me the name of that *C*.” and then if the question is a natural question with response *I*, we regard *I* as an instance of *C*. The phrase in the question has to be “the name of that *C*” rather than “the class of that *C*” or “the name of a *C*”. For example, “New York University” can be an answer to “What is the name of that school?”, so it is an instance of the school category, but “elementary school” can’t be. Such a test sentence is also used for numeric expressions; for example “What is its length?” is the test sentence to check if an instance can belong to the category “physical extent”. The noun phrases to be used in the test sentence for each category are listed in the manual.
2) Definition of each category
   Each category is defined, along with the link to its parent node, its child nodes, the typical noun phrases (used in the test sentence) for the category, examples, and links to the notes of difficult cases, described below. As mentioned before, we are not relying only on the top down description of the category. We listed many examples for each category. For example, "product_other" includes 40 positive and 12 negative examples.
3) Notes for the difficult cases
   This includes 130 notes addressing difficulties in definitions or tagging. For example, how we will tag imaginary entities, like fictional characters in movies, criteria for deciding if entities belong one category or another, how to treat special entities, e.g. “National central bank”, anaphoric expressions, abbreviations, coordination, initials, nicknames, compound nouns, overlapping of several entities, etc. In the manual, there are links from each note to related entities and other notes, so that the user can easily refer to related issues.

The 200 categories are listed in the appendix, in the form of a tree. The root node is TOP and its three children are NAME, TIME_TOP and NUMEX. The number of “>” signs at the beginning of each line indicates the depth of the node. The parent of a line can be found by searching upwards for the line with one fewer “>”. In order to save space, several categories are listed on a single line if there is more than one category at the same level and these do not have any children.

Dictionaries

We have accumulated instances of each category from the Web, newspapers and other sources. This was all done by hand and includes about 130,000 instances. Table 1 shows some categories and the number of instances in the dictionary. We also created a common noun phrase dictionary with about 50,000 instances. In the dictionary, common noun phrases which express a particular category are listed. For example, for the "people" category, words like "scientist" and "baseball player" are listed. One of the uses of the dictionary is for QA systems. For a question like "What is the name of the scientist who created the electric light?", you can determine the category of the answer using the dictionary.

Person	32,606	Theory	190
Landform	17,197	Conference	180
Company	7,261	Crime	152
Water form	5,244	…	
Disease	4,777	Train	67
City	3,750	Reptile	66
Mineral	2,561	GPE	47
…		…	

Table 1. Number of instances for some categories

Tagger

We developed an NE tagger using the dictionary and pattern-based rules. Rules are used to identify entities which can’t be tagged by dictionary. For example, although popular person names are listed in the dictionary, we can’t include all person names. So rules like "Mr. XXX" are used to tag person names. About 1,400 such rules were developed by hand from a study of newspapers and other sources. Some examples of rules are listed in Table 2; these examples are slightly modified to make them easier to understand. In the table, the left side shows patterns to be matched to the input sentence. Each element of a pattern can have up to four components: literal string, word class, POS and currently tagged NE label. For example, the first element of the first example

in Table 2 matches a token of *Katakana* which is currently tagged as "person". The right side indicates the NE tags to be changed for the token at each position of the pattern. These rules make use of word classes to group similar words, such as "Mr.", "Mrs." and "Miss". There are 140 classes with about 2,500 word instances; some examples are shown in Table 3.

Pattern	NE tag
(* ~*Katakana* * PERSON) (.) (* ~Alphabet * OTHER)	1:B-PERSON 2:I-PERSON 3:I-PERSON
(*TSUMA*|*ANI*) (* * NNP) (*SHI*|*SAN*)	2:B-PERSON 3:B-TITLE
(* ~*Katakana*) (* COM_SUFFIX)	1:B-COMP 2:I-COMP
(`) (*) (') (*TOIU*) (*EIGA*)	2: B-MOVIE

Translation: "*Katakana*": A Japanese character type. It usually represents borrowed words, like names of foreign persons. "*TSUMA*": "wife". "*ANI*": "big brother". "*SHI*|*SAN*": Prefix indicating person, like "Mr.", but located after a person name. "*X TOIU EIGA*": "Movie called X"

Table 2. Example of tagging rules

Class	Examples
COM_SUFFIX	*KABUSHIKIGAISYA (Cooperation),* Co., Ltd.
PERSON_SUFFIX	*SAN (Mr.), SAMA (Mr. with respect)* , Reporter, Driver
POSITION_PREFIX	*FUKU (Vice), SHIN (New), ZEN (Previous)*
FACILITY_SUFFIX	*KYOUGIJOU (Stadium), GEKIJOU (Theater)*

Table 3. Example of Classes

ENE category	Frequency	Precision	Recall
TOTAL	6708	72%	80%
Person	803	65%	77%
Date	709	88%	86%
Country	574	92%	90%
Position title	433	72%	69%
…			
Money	118	95%	98%
Event	90	36%	41%
…			
Product	40	0%	0%
Landform	36	86%	86%
…			
Weight	5	60%	100%
Picture	5	0%	0%

Table 4. Evaluation Results

The accuracy of the tagger is 72% recall and 80% precision. Table 4 shows the frequency in the test data and accuracy for the overall categories and several categories. Although there is a room for improvement, we can't expect it to be on a par with the state-of-the-art 8-category NE taggers. As the number of categories is large, it is quite difficult to prepare tagged sentences for supervised training, which is a major technique for creating taggers for a small number of categories. We adopt the strategy of using a dictionary and rules at the moment.

Discussion and Future Directions

Domain dependent category vs. General NE

It is controversial whether a general purpose Extended NE hierarchy really exists or not. It might be better to create a domain dependent NE hierarchy if the application targets only one domain. However, there are applications like open-domain Question Answering or open-domain Information Extraction where the target domain is very broad, like the newspaper domain. Then we can't afford to create NE hierarchies for each sub-domain, but must create a general Extended NE hierarchy.

Extended NE hierarchy and thesaurus

As the number of categories grows, 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 as we observed when we designed the hierarchy in the first place, WordNet is not really suitable for use as a hierarchy for names, as it is designed for common nouns.

Accuracy of the tagger

Obviously the accuracy of the tagger is not satisfactory. Some large categories, including person and organizations, make the overall accuracy worse. It may be a good idea to combine machine learning methods with our system. Also, for some categories, we found that collecting more instances could easily help. We may have to work hard to make the list longer. Finally, we are considering other kinds of machine learning; including weakly supervised learning, bootstrapping, active learning and unsupervised methods using linguistic clues, in order to improve the accuracy of the tagger.

Acknowledgements

A part of this research is supported by the Defense Advanced Research Projects Agency as part of the Translingual Information Detection, Extraction and Summarization (TIDES) program, under Grant N66001-001-1-8917 from the Space and Naval Warfare Systems Center, San Diego, and by the National Science Foundation under Grant ITS-00325657. This paper does not necessarily reflect the position of the U.S. Government. We would like to thank our colleagues at New York University and Communications Research Laboratory, who provided useful suggestions and discussions, including, Prof. Ralph Grishman, Mr. Kiyoshi Sudo, Mr. Yusuke Shinyama at NYU and Mr. Kiyotaka Uchimoto and Hitoshi Isahara at CRL.

Bibliography

(Grishman and Sundheim 1996) Ralph Grishman and Beth Sundheim, "Message Understanding Conference 6: A Brief History", COLING-1996.

(Sekine and Isahara 2000) Satoshi Sekine and Hitoshi Isahara, "IREX: IR and IE Evaluation project in Japanese", LREC-2000.

(Sekine et al. 2002) Satoshi Sekine, Kiyoshi Sudo and Chikashi Nobata. “Extended Named Entity Hierarchy”, LREC-2002

(ACE homepage) http://www.itl.nist.gov/iaui/894.01/tests/ace

(Senseval HP) http://www.senseval.org/

(WordNet HP) http://www.cogsci.princeton.edu/~wn/

(ISI Webclopedia HP) http://www.isi.edu/natural-language/projects/webclopedia/

APPENDIX: 200 ENE categories

TOP
>NAME
>>NAME_OTHER
>>PERSON
>>ORGANIZATION
>>>ORGANIZATION_OTHER COMPANY COMPANY_GROUP MILITARY INSTITUTE GOVERNMENT POLITICAL_PARTY GAME_GROUP INTERNATIONAL_ORGANIZATION ETHNIC_GROUP NATIONALITY
>>LOCATION
>>>LOCATION_OTHER
>>>GPE
>>>>GPE_OTHER CITY COUNTY PROVINCE COUNTRY
>>>CONTINENTAL_REGION DOMESTIC_REGION
>>>GEOLOGICAL_REGION
>>>>GEOLOGICAL_REGION_OTHER LANDFORM WATER_FORM SEA
>>> ASTRAL_BODY
>>>>ASTRAL_BODY_OTHER STAR PLANET
>>>ADDRESS
>>>>ADDRESS_OTHER POSTAL_ADDRESS PHONE_NUMBER EMAIL URL
>>FACILITY
>>>FACILITY_OTHER
>>>GOE
>>>>GOE_OTHER SCHOOL PUBLIC_INSTITUTION MARKET MUSEUM AMUSEMENT_PARK WORSHIP_PLACE
>>>>STATION_TOP
>>>>>STATION_TOP_OTHER AIRPORT STATION PORT CAR_STOP
>>>LINE
>>>>LINE_OTHER RAILROAD ROAD WATERWAY TUNNEL BRIDGE
>>>PARK SPORTS_FACILITY MONUMENT FACILITY_PART
>>PRODUCT
>>>PRODUCT_OTHER
>>>VEHICLE
>>>>VEHICLE_OTHER CAR TRAIN AIRCRAFT SPACESHIP SHIP
>>>FOOD CLOTHES DRUG WEAPON STOCK AWARD THEORY RULE SERVICE CHARACTER METHOD_SYSTEM DOCTRINE CULTURE RELIGION LANGUAGE PLAN ACADEMIC CLASS SPORTS OFFENCE
>>>ART
>>>>ART_OTHER PICTURE BROADCAST_PROGRAM MOVIE SHOW MUSIC BOOK
>>>PRINTING
>>>>PRINTING_OTHER NEWSPAPER MAGAZINE
>>EVENT
>>>EVENT_OTHER
>>>OCCASION
>>>>OCCASION_OTHER GAMES CONFERENCE
>>>NATURAL_PHENOMENA NATURAL_DISASTER WAR INCIDENT
>>NATURAL_OBJECT
>>>NATURAL_OBJECT_OTHER
>>>LIVING_THING
>>>>LIVING_THING_OTHER
>>>>ANIMAL
>>>>>ANIMAL_OTHER
>>>>>INVERTEBRATE
>>>>>>INVERTEBRATE_OTHER INSECT
>>>>>VERTEBRATE
>>>>>>VERTEBRATE_OTHER FISH REPTILE AMPHIBIAN BIRD MAMMAL
>>>>FLORA BODY_PARTS FLORA_PARTS
>>>MINERAL
>>TITLE
>>>TITLE_OTHER POSITION_TITLE
>>UNIT UNIT_OTHER CURRENCY
>>VOCATION
>>DISEASE
>>GOD
>>ID_NUMBER
>>COLOR

>TIME_TOP
>>TIME_TOP_OTHER
>>>TIMEX
>>>>TIMEX_OTHER TIME DATE DAY_OF_WEEK ERA
>>>PERIODX
>>>>PERIODX_OTHER TIME_PERIOD DATE_PERIOD WEEK_PERIOD MONTH_PERIOD YEAR_PERIOD

>NUMEX
>>NUMEX_OTHER MONEY STOCK_INDEX POINT PERCENT MULTIPLICATION FREQUENCY RANK AGE SCHOOL_AGE LATITUDE_LONGITUDE
>>MEASUREMENT
>>>MEASUREMENT_OTHER PHYSICAL_EXTENT SPACE VOLUME WEIGHT SPEED INTENSITY TEMPERATURE CALORIE SEISMIC_INTENSITY SEISMIC_MAGNITUDE
>>COUNTX
>>>COUNTX_OTHER N_PERSON N_ORGANIZATION
>>>N_LOCATION
>>>>N_LOCATION_OTHER N_COUNTRY
>>>N_FACILITY N_PRODUCT N_EVENT N_ANIMAL N_FLORA N_MINERAL
>>ORDINAL_NUMBER