Extended Named Entity Hierarchy

Satoshi Sekine[*] **Kiyoshi Sudo**[*] **Chikashi Nobata**[†]

[*]New York University
715 Broadway, 7th floor, New York, NY 10003, USA
{sekine,sudo}@cs.nyu.edu

[†]Communications Research Laboratory
2-2-2 Hikaridai Seika-cho, Soraku-gun Kyoto 619-0289 Japan
nova@crl.go.jp

Abstract

The tagging of Named Entities (NE), the names of particular things or classes and numeric expressions, is regarded as an important component technology for many NLP applications. These applications include Information Extraction, from which it was born, Question-Answering, Summarization and Information Retrieval. However, up to now, the number of NE types has been quite limited, 7 in MUC, 8 in IREX and 5 in the ACE program. Many more kinds of things have proper names or proper classes of expressions, and also finer distinctions are needed for some applications. We now propose a Named Entity hierarchy which contains about 150 NE types. The focus of this paper is the design of the hierarchy and we would like to provide this resource for any application. We report the design and development procedure of the hierarchy.

1. Introduction

The tagging of Named Entities, the names of particular things or classes, and numeric expressions[1] is regarded as an important component technology for many NLP applications. These applications include Information Extraction (IE), from which it was born, Question Answering (Q&A), Summarization and Information Retrieval (IR). The first Named Entity set had 7 types (Grishman and Sundheim, 1996), organization, location, person, date, time, money and percent expressions. The number of such entity types was limited because the target application of the evaluation was information extraction for business activities. So, when we encountered the application of "airplane crashes" in MUC 7, people recognized that a new NE type "airplane" was needed to make an IE system for the domain. In the IREX project (Sekine and Isahara, 2000) (IREX, Homepage), a Japanese IR and IE evaluation based project, another kind of expression, artifact, was added. This was done because, in the project, no application was presupposed and NE extraction was the final target. So, the generalization was taken into account. In the project, it was found that the new entity type was quite difficult to detect and there were many sub-types for that category. Also, in the ACE program (ACE, Homepage), two new entities, GPE (geographical and political entity, a location which has a government, like U.S.A. or New York) and facility, are added to pursue the generalization of the technology. From those experiences, we recognized that 7 or 8 kinds of NE are not broad enough to cover general issues. Many more kinds of things have proper names or proper classes of expressions, and also finer distinctions are needed for some applications.

We have to consider the domain dependency of this type of knowledge, because different kinds of named entities are needed in different domains. But, we are not aiming to cover special domains, like genomics; rather, the target is more general texts, like newspapers, which educated adults can understand without using special dictionaries. In other words, we designed our NE hierarchy so that it can cover most of the entities which appear in usual newspaper articles, with appropriate balance.

We now propose a Named Entity hierarchy which contains about 150 NE types. The focus of this paper is the design of the hierarchy and we would like to provide this resource for any research purpose.

There were three stages in the development. First, we designed three hierarchies based on three different methods; based on corpus (newspaper articles), based on previous systems and definitions, and based on thesaurus like WordNet and Roget. Then, we merged the three hierarchies. After that, we tagged a corpus using the definitions, and we refined the hierarchy.

2. Design Issues

In order to define the Named Entity hierarchy, we considered, at least, three issues.

- In order to adjust easily to different tasks, the types are organized as a hierarchy. For example, while we have 150 types of NEs, if you need only MUC's 7 NEs, it should be easily adaptable to the subset. In other words, we define the key NE types (like person), and the sub-types are organized under the type. The NEs defined in MUC, IREX and ACE NE are used as the key NEs. At the same time, flexibility is considered so that it can adapt to as many possible definitions as possible. However, unlike some of the

---

[1]In this paper, the term "Named Entity" includes names (which is narrow sense of Named Entity) and numeric expressions. The definition of this Named Entity is not simple, but, intuitively, this is a class that people are often willing to know in newspaper articles.

other hierarchies or thesauri, the only relationship between parent and child is is-a relation and no attribute relationship (like person and vocation) is included in the hierarchy.

- One of the essential problems in defining NE types is whether to take the surface form or meaning in the particular context as the primary classification criterion. For example, "Japan" is normally used in a geographical sense, but sometime it refers to the government of Japan (organization), as in "Japan announced a tax cut." In order to minimize the ambiguity, we try to take the surface form as the primary clue as much as possible. This approach is used mainly because most of the current NE taggers do not use semantic information and disambiguation at the NE tagger level is almost impossible. So we introduced the NE types "GPE" (geographical and political entity, following the ACE definition) and "GOE" (geographical and organizational entity, which can refer to a location as well as an organization, like a museum or airport).

- The definition of what is a Named Entity is ambiguous. At first glance, Named Entity looks like proper names (names of things; typically capitalized) and numerical expressions (number of things). However, once we include, for example, artifact, we have a problem that there are things besides proper names which should be included as Named Entities. For example, is the name of a car, like "Integra" or "Odyssey" a proper noun? We certainly want to include this as a named entity, as the NLP applications, like MT, IE, IR or Q&A will benefit if these are identified as named entities. However, precisely speaking, these names are names of classes and not the names of specific cars (the particular car I have). This is different from the problem that more than one person can have the same name, as the name "Integra" denotes some concept, "class". If you accept this argument, then we have to decide how far we want to go into the class to be named entity. We basically decided that class names like color, type of airplane, material name, animal name are included, as we thought these are useful if we know its class. However, ordinary words like "cup", "feelings" or "tear" are not included as named entities. The border is unavoidably ambiguous, and some arbitrary decision was necessary.

- The degree of fineness can be arbitrary. For example, the "person" class can be divided by different criteria, like gender, nationality, occupation, age etc. However, we avoided dividing a class if the division depends primarily on its context rather than the name itself. For example, the gender distinction can often be achieved by just looking at the person's first name, but occupation can't.

2.1. Procedure

There were three stages in the development of the hierarchy:

1. Use three methods to design three initial hierarchies; corpus-based, based on previous systems and tasks, and based on thesauri.

2. Merge the three hierarchies while engaging in discussions including more than the three people who designed the initial hierarchies.

3. Refine the hierarchy by tagging additional corpus and developing automatic taggers.

We will describe each stage in turn.

2.1.1. Initial Hierarchies

First, we used three methods to design the initial hierarchies as follows;

- Based on a newspaper corpus
  This is basically a bottom-up method using actual examples. We extracted about 3500 candidate NE expressions from a corpus, using surface clues in English. Namely, capitalized words for proper name type entities (1500 examples total; 500 each from Wall Street Journal 1993, New York Times 1994 and Los Angeles Times 1995) and context for numerical entities (2000 examples from Wall Street Journal 1993). In practice, we found some peculiarities in the numerical expressions, which we did first, so we used three kinds of newspapers for the name type entities. We investigated the difference between the extracted entities from the three different newspapers, but we could not identify significant differences between them.

  We made KWIC for them and assigned NE categories to each entity based on the examiner's intuition. The person was allowed to assign the unknown tag (?), the tag indicating uncertainty (∗) and non-proper name tag (!), some of which are shown in the KWIC example in Figure 1. The tag names were made initially by intuition, but later the names were reviewed and re-assigned to keep consistency throughout the corpus. Also, the types were merged or divided based on number of examples and balance between tags.

- Based on existing systems and tasks
  There are many systems related to NE. The multiple task IE system (Aone Ramos-Santacruz, 2000) suggested several new NE classes. We analyzed the TREC-QA task (TREC-QA, Homepage), from which we induced some important NE classes; in practice, this was quite useful, so we did Japanese QA analysis in the later stage, as will be explained in the next section. Also, similar work was done by (Sasaki, 1999), who defined more NE types than used in IREX.

- Based on thesauri
  Obviously thesauri provide data closely related to the NE hierarchy. We consulted two well-known thesauri, WordNet and Roget Thesaurus. The Roget thesaurus categories are relatively shallow and one node

```
              at the American  Chamber    of Commerce in          : ORGANIZATION
     Century Corporation Will  Change     Our Work and Our Lives.''    : !
 Book, Calendar and To-Do List,  Checkbook, Household Manager, Financial.. : ?
   higher offer he wins,'' said  Chet      Needleman, chief executive.. : PERSON
            close to Seagram  Chief      Executive Edgar Bronfman Jr. : POSITION
   , whose favorite flavor is  Chocolate  Fudge Brownie.              : PRODUCT *
```

Figure 1: KWIC Examples

contains several meanings, so we mainly used Word-Net, with occasional reference to Roget. We analyzed the thesaurus both top-down and bottom-up. In the top-down method, we tried to improve our coverage of NE, i.e. not miss any important categories. In the bottom-up method, we aimed to find NE types related to those we already knew. For example, we searched the nodes from "New York" upward to find something other than location. The seeds in this search were some proper nouns and the common nouns related to some NE's. This method was particularly useful in finding measurement expressions. In general, Word-Net makes finer distinctions than we wanted, so we could not just copy a part of it for our purpose.

2.1.2. Merge the three hierarchies

We merged the hierarchies separately for numerical expressions and name type expressions.

- Numerical expressions
  As we found that the hierarchy created by the corpus based method had the largest coverage and best balance, we used that hierarchy as the basis for the final numeric types hierarchy. The fact that the entity types of frequency 1 are relatively rare in the KWIC list indicates that the coverage of this method might be fairly good. However, the missing but meaningful entities which were found in the other hierarchies are added. The previous system-based method found many domain dependent entities (like computer related terminology), but we thought some of them are too domain specific and we did not include them. In the thesauri, some parts of the hierarchy are very deep and some are shallow.

- Name type expressions
  In this category, we used the hierarchy based on the previous systems and tasks. In particular, types of QA samples were quite useful. The corpus based approach does not seem to get good coverage, as the number of types in this category appears to be much greater than for the numerical expressions. We found more number of single frequencies than that in the numeric type NEs. The thesaurus method found wide coverage, but it is hard to strike a good balance which is suitable in newspaper domain. As the thesaurus seem to put weight on completeness rather than on the balance of real world usage, we found many useless categories among useful categories. However, the knowledge from thesauri was used along those found

in corpus to refine the hierarchy made by the previous systems and tasks definition based method.

2.1.3. Refine hierarchy

At this point, we had made a hierarchy of about 100 types. However, we could not be sure that the hierarchy covers most of the entity types in newspaper text. We tried several methods to broaden the coverage, concurrent with the development of an NE tagger and Q&A system based on the hierarchy. Five methods were used:

- Making and classifying Q&A examples
  We made sample Q&A in Japanese using method similar to the TREC8-QA task. More than 10 people made more than 3000 Q&A samples by looking at the newspapers, and we assigned a type to each Q&A based on the hierarchy. When we thought we needed a new type which is important and general in the newspaper domain, we added it to the hierarchy.

- Find NE examples indirectly
  We wanted to tag sentences using the hierarchy in order to make training and test data for the NE tagger. However, some of the types are relatively rare. So we first extract sentences in which an example of a certain type is expected to be found. As we don't want to create a biased collection, i.e. "person" type collection in which only the person name "Clinton" is found, we took the following approach. We asked people to come up with some words which are likely to collocate with a certain type, and tag the sentences with the type. We still can't remove the possibility of bias, but judging from our results, it worked reasonably well. Again, if we encounter examples which do not fall into an existing type, we add a new type.

- Find NE examples from the Web
  We collect NE examples from Web. This aims to create NE dictionary, but at the same time, we found several new categories which we think important in the newspaper domain.

- Tagging Katakana in Japanese
  In Japanese, borrowed words are often written in a special character set, Katakana. At the moment, we could not find a good dictionary for Katakana, partially because these types of words are created almost every day. We tagged the most frequent Katakana words in some recent newspapers. In this process, we also added new NE types to the hierarchy, surprisingly not only new kinds of terms, like computer terms, but more general types, like "STAR".

- Classify common nouns
  We also classify common nouns based on the hierarchy. This is mainly for the purpose of finding types of questions in Q&A system. We used a existing thesaurus for the start, but up to now, we did not add new types in this process.

It is hard to measure the coverage without tagging all the entities in a certain number of documents. However, after doing the above mentioned processes, we believe that the coverage of the hierarchy is quite satisfactory. Although this never means it is complete, we believe this hierarchy should be useful in some applications.

3. System

We developed NE taggers for English and Japanese. There have been many experiments involving the supervised training, using tagged corpora of NE taggers (Bickel et. al, 1997), (Sekine et. al, 1998), (Borthwick et. al, 1998). However, as we could not accumulate enough examples for supervised learning, we made a rule and dictionary based system, where the rules are created by hand while observing examples and dictionaries. The design and performance of this tagger will be reported separately.

4. Data and Guidelines

The complete hierarchy with examples is shown in the Appendix. An electronic version of the hierarchy, along with guidelines for the NE types, will be made available on the Web, and will be accessible through the first author's homepage (http://cs.nyu.edu/~ sekine).

5. Conclusion

We have reported on the design and the development of an extended Named Entity hierarchy. It has 150 types and is organized in a tree structure. The authors believe this will be useful not only in IE and Q&A in the newspaper domain, but also MT and other kinds of NLP applications and the further extension of the NE hierarchy in some specific domains.

6. Acknowledgments

7. References

Chinatsu Aone and Mila Ramos-Santacruz, "REES: A large-scale relation and evaluation system" In *Proceedings of ANLP/NAACL-2000*.

Daniel M. Bickel, Scott Miller, Richard Schwartz, Ralph Weischedel, "Nymble: a high-performance learning name-finder" In *Proceedings of the ANLP 97*.

Andrew Borthwick, John Sterling, Eugene Agichtein, Ralph Grishman, "Exploiting Diverse Knowledge Sources via Maximum Entropy in Named Entity Recognition In *Proceedings of the WVLC 98*.

Ralph Grishman and Beth Sundheim, "Message Understanding Conference - 6: A Brief History" In *Proceedings of COLING-96*.

Yutaka Sasaki "Japanese NE tagger using transducer (in Japanese)" In *Proceedings of Annual meeting of Japanese Association of NLP 1999 (in Japanese)*.

Satoshi Sekine, Hitoshi Isahara "IREX: IR and IE Evaluation project in Japanese" In *Proceedings of the LREC 2000*.

Satoshi Sekine, Ralph Grishman, Hiroyuki Shinnou "A Decision Tree Method for Finding and Classifying Names in Japanese Texts" In *Proceedings of the WVLC 98*.

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

IREX homepage http://www.cs.nyu.edu/cs/project/proteus/irex

TREC-QA homepage http://trec.nist.gov/

Appendix: Extended NE hierarchy

```
TOP
 NAME
  PERSON                     # Bill Clinton, George W. Bush, Satoshi Sekine,
   LASTNAME                  # Clinton, Bush, Sekine,
   MALE_FIRSTNAME            # Bill, George, Satoshi,
   FEMALE_FIRSTNAME          # Mary, Catherine, Ilene, Yoko

  ORGANIZATION               # United Nations, NATO
   COMPANY                   # IBM, Microsoft
   COMPANY_GROUP             # Star Alliance, Tokyo-Mitsubishi Group
   MILITARY                  # The U.S Navy
   INSTITUTE                 # the National Football League, ACL
   MARKET                    # New York Stock Exchange, NASDAQ
   POLITICAL_ORGANIZATION    #
    GOVERNMENT               # Department of Education, Ministry of Finance
    POLITICAL_PARTY          # Republican Party, Democratic Party, GOP
    PUBLIC_INSTITUTION       # New York Post Office,
   GROUP                     # The Beatles, Boston Symphony Orchestra
    SPORTS_TEAM              # the Chicago Bulls, New York Mets
   ETHNIC_GROUP              # Han race, Hispanic
   NATIONALITY               # American, Japanese, Spanish

  LOCATION                   # Times Square, Ground Zero
   GPE                       # Asia, Middle East, Palestine
    CITY                     # New York City, Los Angeles
    COUNTY                   # Westchester
    PROVINCE                 # State (US), Province (Canada), Prefecture (Japan)
    COUNTRY                  # the United States of America, Japan, England
   REGION                    # Scandinavia, North America, Asia, East coast
   GEOLOGICAL_REGION         # Altamira
    LANDFORM                 # Rocky Mountains, Manzano Peak, Matterhorn
    WATER_FORM               # Hudson River, Fletcher Pond
    SEA                      # Pacific Ocean, Gulf of Mexico, Florida Bay
   ASTRAL_BODY               # Halley's comet, the Moon
    STAR                     # Sirius, Sun, Cassiopeia, Centaurus
    PLANET                   # the Earth, Mars, Venus
   ADDRESS                   #
    POSTAL_ADDRESS           # 715 Broadway, New York, NY 10003
    PHONE_NUMBER             # 212-123-4567
    EMAIL                    # sekine@cs.nyu.edu
    URL                      # http://www.cs.nyu/cs/projects/proteus
  FACILITY                   # Empire State Building, Hunter Montain Ski Resort
   GOE                       # Pentagon, White House, NYU Hospital
    SCHOOL                   # New York University, Edgewood Elementary School
    MUSEUM                   # MOMA, the Metropolitan Musium of Art
    AMUSEMENT_PARK           # Walt Disney World, Oakland Zoo
    WORSHIP_PLACE            # Canterbury Cathedral, Westminster Abbey
    STATION_TOP              #
     AIRPORT                 # JFK Airport, Narita Airport, Changi Airport
     STATION                 # Grand Central Station, London Victoria Station
     PORT                    # Port of New York, Sydney Harbour
     CAR_STOP                # Port Authority Bus Terminal, Sydney Bus Depot
```

```
  LINE                    # Westchester Bicycle Road
    RAILROAD              # Metro-North Harlem Line, New Jersey Transit
    ROAD                  # Lexington Avenue, 42nd Street
    WATERWAY              # Suez Canal, Bering Strait
    TUNNEL                # Euro Tunnel
    BRIDGE                # Golden Gate Bridge, Manhattan Bridge
  PARK                    # Central Park, Hyde Park
  MONUMENT                # Statue of Liberty, Brandenburg Gate
PRODUCT                   # Windows 2000, Rosetta Stone
  VEHICLE                 # Vespa ET2, Honda Elite 50s
    CAR                   # Ford Escort, Audi 90, Saab 900, Civic, BMW 318i
    TRAIN                 # Acela, TGV, Bullet Train
    AIRCRAFT              # F-14 Tomcat, DC-10, B-747
    SPACESHIP             # Sputnik, Apollo 11, Space Shuttle Challenger, Mir
    SHIP                  # Titanic, Queen Elizabeth II, U.S.S. Enterprise
  DRUG                    # Pedialyte, Tylenol, Bufferin
  WEAPON                  # Patriot Missile, Pulser P-138
  STOCK                   # NABISCO stock
  CURRENCY                # Euro, yen, dollar, peso,
  AWARD                   # Nobel Peace Prize, Pulitzer Prize
  THEORY                  # Newton's law, GB theory, Blum's Theory
  RULE                    # Kyoto Global Warming Pact, The U.S. Constitution
  SERVICE                 # Pan Am Flight 103, Acela Express 2190
  CHARACTER               # Pikachu, Mickey Mouse, Snoopy
  METHOD_SYSTEM           # New Deal program, Federal Tax
  ACTION_MOVEMENT         # The U.N. Peace-keeping Operation
  PLAN                    # Manhattan Project, Star Wars Plan
  ACADEMIC                # Sociology, Physics, Philosophy
  CATEGORY                # Bantam Weight, 48kg class
  SPORTS                  # Men's 100 meter, Giant Slalom, ski, tennis
  OFFENCE                 # first-degree murder
  ART                     # Venus of Melos
    PICTURE               # Night Watch, Monariza, Guernica
    BROADCAST_PROGRAM     # Larry King Live, The Simpsons, ER, Friends
    MOVIE                 # E.T., Batman Forever, Jurassic Park, Star Wars
    SHOW                  # Les Miserables, Madam Butterfly
    MUSIC                 # The Star Spangled Banner, My Life, Your Song
  PRINTING                # 2001 Consumer Survey
    BOOK                  # Master of the Game, 1001 Ways to Reward Employees
    NEWSPAPER             # The New York Times, Wall Street Journal
    MAGAZINE              # Newsweek, Time, National Business Employment Weekly

DISEASE                   # AIDS, cancer, leukemia

EVENT                     # Hanover Expo,  Edinburgh Festival
  GAMES                   # Olympic, World Cup, PGA Championships
  CONFERENCE              # APEC, Naples Summit
  PHENOMENA               # El Nino
  WAR                     # World War II, Vietnam War, the Gulf War
  NATURAL_DISASTER        # Kobe Earthquake, the PuuOo-Kupaianaha Eruption

  CRIME                   # Murder of Black Dahlia, the Oklahoma City bombing

TITLE                     # Mr., Ms., Miss., Mrs,
  POSITION_TITLE          # President, CEO, King, Prince, Prof., Dr.

LANGUAGE                  # English, Spanish, Chinese, Greek

RELIGION                  # Christianity, Islam, Buddhism
```

```
NATURAL_OBJECT           # mitochondria, shiitake mushroom
 ANIMAL                  # elephant, whale, pig, horse
 VEGETABLE               # spinach, rice, daffodil
 MINERAL                 # Hydrogen, carbon monoxide,

COLOR                    # black, white, red, blue

TIME_TOP
 TIMEX
  TIME                   # 10 p.m.,  afternoon
  DATE                   # August 10, 2001,  10 Aug. 2001,
  ERA                    # Glacial period, Victorian age

 PERIODX                 # 2 semesters, summer vacation period
  TIME_PERIOD            # 10 minutes, 15 hours, 50 hours
  DATE_PERIOD            # 10 days, 50 days
  WEEK_PERIOD            # 10 weeks, 50 weeks
  MONTH_PERIOD           # 10 months, 50 months
  YEAR_PERIOD            # 10 years, 50 years

NUMEX                    # 100 pikel, 10 bits
 MONEY                   # $10, 100 yen, 20 marks
 STOCK_INDEX             # 26 5/8,
 POINT                   # 10 points
 PERCENT                 # 10%, 10 1/2%
 MULTIPLICATION          # 10 times
 FREQUENCY               # 10 times a day
 RANK                    # 1st prize, booby prize
 AGE                     # 36, 77 years old

 MEASUREMENT             # 10 bytes, 10 Pa, 10 millibar
  PHYSICAL_EXTENT        # 10 meters, 10 inches, 10 yards, 10 miles
  SPACE                  # 10 acres, 10 square feet,
  VOLUME                 # 10 cubic feet, 10 cubic yards
  WEIGHT                 # 10 milligrams, 10 ounces, 10 tons
  SPEED                  # 10 miles per hour, Mach 10
  INTENSITY              # 10 lumina, 10 decibel
  TEMPERATURE            # 60 degrees
  CALORIE                # 10 calories
  SEISMIC_INTENSITY      # 6.8 (on Richter scale)

 COUNTX
  N_PERSON               # 10 biologists, 10 workers, 10 terrorists
  N_ORGANIZATION         # 10 industry groups, 10 credit unions
  N_LOCATION             # 10 cities, 10 areas, 10 regions, 10 states
   N_COUNTRY             # 10 countries
  N_FACILITY             # 10 buildings, 10 schools, 10 airports
  N_PRODUCT              # 10 systems, 20 paintings, 10 supercomputers
  N_EVENT                # 5 accidents, 5 interviews, 5 bankruptcies
  N_ANIMAL               # 10 animals, 10 horses, 10 pigs
  N_VEGETABLE            # 10 flowers, 10 daffodils
  N_MINERAL              # 10 diamonds
```