Extended Named Entity Ontology with Attribute Information

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

Abstract

Named Entities (NE) are regarded as an important type of semantic knowledge in many natural language processing (NLP) applications. Originally, a limited number of NE categories were proposed. In MUC, it was 7 categories – people, organization, location, time, date, money and percentage expressions. However, it was noticed that such a limited number of NE categories is too small for many applications. The author has proposed Extended Named Entity (ENE), which has about 200 categories (Sekine and Nobata 04). During the development of ENE, we noticed that many ENE categories have specific attributes, and those provide very important information for the entities. For example, “rivers” have attributes like “source location”, “outflow”, and “length”. Some such information is essential to ‘knowing about’ the river, while the name is only a label which can be used to refer to the river. Also, such attributes are important information for many NLP applications. In this paper, we report on the design of a set of attributes for ENE categories. We used a bottom up approach to creating the knowledge using a Japanese encyclopedia, which contains abundant descriptions of ENE instances.

1. Introduction

Named Entities (NE) are regarded as an important type of semantic knowledge in many natural language processing (NLP) applications. Named entities were originally introduced as possible types of fillers in Information Extraction systems at the MUC-6 evaluation (Grishman and Sundheim 96). Also, many factoid question answering (QA) systems use the NE categories as the answer types. Information Retrieval (IR) and Summarization systems found that those entities are important elements of information for their processing. Originally, a limited number of NE categories were proposed. In MUC, it was 7 categories – people, organization, location, time, date, money and percentage expressions – and in IREX (Sekine and Isahara 00), it became 8 categories by adding artifact. However, it was noticed that such a limited number of NE categories is too small for many applications. For example, when we want to extract disease outbreak information from bio-medical reports, we need the names of diseases, medicines, and so on. Also, QA systems have to cover a variety of entity names. In this context, a large inventory of NE categories has been proposed (Li and Roth 02) (Harabagiu et al. 03). The author has proposed Extended Named Entity (ENE), which has about 200 categories (Sekine and Nobata 04).

During the development of ENE, we noticed that many ENE categories have specific attributes, and those provide very important information for the entities. For example, “rivers” have attributes like “source location”, “outflow”, and “length”, while “people” have “occupation”, “birth date”, “nationality” and so on. In theory, most ENE names are just meaningless labels, e.g. the name “Hudson River” doesn’t convey any information about the river which the name refers to. It would be the same river even if it were called “Muh-he-kun-ne-tuk” (this is the name given to the river by the local American Indian tribe). The meaning of an entity can only be discerned from those attribute values, e.g. “the river is in New York State”, “it is 507 km in length” and “it runs from Mount Marcy, Adirondack Mountains to Upper New York Bay” or “it is named after Henry Hudson, an Englishman sailing for the Dutch East India Company, who explored it in 1609”. Some such information (except perhaps for the history of the name) is essential to ‘knowing about’ the river, while the name is only a label which can be used to refer to the river. Also, such attributes are important information for many NLP applications. With this knowledge, a system can precisely answer questions such as “Where does the Hudson River run through?” or “What is the 112th longest river in the world?” The application of this knowledge is not limited to QA; IR, IE, summarization and MT also benefit from this knowledge.

In this paper, we report on the design of a set of attributes for ENE categories. We used a bottom up approach to creating the knowledge using a Japanese encyclopedia, which contains abundant descriptions of ENE instances.

2. Related Work

We can find attribute sets for particular types of entities on the Web, in books, or in an encyclopedia. Some types of Wikipedia entries (Wikipedia HP) have attributes. For example, country has 10 to 20 categories as of March, 2008, such as Anthem, Capital, Official Languages, and so on. However, we found that except for a small number of categories, such as country, the attribute sets are inconsistent and have a wide variation within the same category. As the categories and attributes are not well organized, it is not easy to use them in NLP applications. YAGO is a project to create an ontology using Wikipedia (Suchanek et al. 07). It extracts an Is-A hierarchy as well as relations between entities such as HasWonPrize, mostly based on the category information of Wikipedia. Although the methodologies are different, the final goals of the projects are similar. Many hand-made ontologies for a general domain do not focus on names, but rather mostly on nouns, verbs, adjectives and adverbs, such as WordNet (Fellbaum 98), Cyc or OpenCyc (Matuszek et al. 06), SUMO (SUMO HP) or Omega Ontology (Philpot et al. 08). Those ontologies have no or limited sets of attributes. We believe the knowledge of names will be very important for understanding documents, and attribute information would be crucial for NLP applications.

Recently, there are many activities to create ontologies of entities for a narrow domain, which is called a "domain ontology". For example, The Open biomedical ontologies Homepage (OBO HP) provides a list of many ontologies in the bio-medical domain. Also, the activities connected with the Semantic Web support the creation of domain ontologies (SemanticWeb HP). However, as far as the author knows, there is no large hand-crafted name ontology for a newspaper domain. As the newspaper domain generally includes a wide range of general names, it might be widely usable in many NLP applications.

Attributes have been emphasized as an important type of information (Guarino et al. 92; Pustejovsky 95) and we believe it is particularly important for ENE, because of the nature of ENE and the wide use of ENE attribute information in applications. There has been some work on attribute discovery from large texts (Almuhareb and Poesio 04; Yoshinaga and Torisawa 07). However, we believe a manual approach is feasible, because we have only a limited number of categories in the definition (around 200) and the greater accuracy of a manual approach is most desirable.

3. Attribute Design Procedure

In this section, we present the procedure for the attribute design. It relies heavily on an encyclopedia and manual labor. The procedure consists of the following four steps.

1) Extract sample entries
2) Extract attribute values and identify attributes
3) Redesign NE categories
4) Construct a set of attributes

In this project, we used one of the most famous Japanese encyclopedias, "Nippon Daihyakka Zensyo (Nipponika)" published by Shogakkan Inc.

1) Extract sample entries
We have the online encyclopedia, which has about 120K entries, where each entity has been manually categorized into one of the ENE categories. In general, the time and numeric expressions in the ENE don't have attributes, so only the name categories are used. For each of the 105 name categories which are expected to have some attributes, 50 samples are taken from the encyclopedia entries to annotate.

2) Extract attribute-values and identify attributes
The samples are shown to the human annotators. Each annotator reads the entry and extracts the expressions which are thought to be values of some attribute. We restrict the values to be noun phrases or noun phrase equivalents, but some sentential expressions are allowed, such as "white in general, but it has yellow dots" for the color of a fish. The attribute name such as "color", "length" or "habitat" is then coined by the annotator for the annotated values, because the attribute name is not explicitly mentioned in the encyclopedia entries.

3) Redesign NE categories
The coined attribute names are unified across different entities. Then the important attributes for each category are identified. Here "important" means that the attributes are essential and mandatory rather than optional for the entities in that category. For example, "date of birth" or "occupation" is important for the "people" category, but "hobby" is not.

The entities in a particular category should share the important attributes, and we redesigned the ENE towards this goal. For example, in the previous definition, there was a category called "body of water", which includes river, lake and so on. However, we found that a "river" has important attributes like "length", "the source of the river" and "the mouth of the river", which don't exist in lakes. We separate those two categories in the new definition.

4) Construct a set of attributes
For each category, we list a set of attributes which appear in more than 10% of the sample entries. Then, we organize the definition of the ENE hierarchy to incorporate the attribute information. We list typical ENE categories for each attribute if possible. Sample attributes are shown in Appendix B. The Japanese version of the definition in html can be found at our Extended Named Entity homepage (ENE HP), and we are expecting to have the English version soon.

4. Problems

We will describe some problems we encountered during our attribute construction. Some remain to be solved, but we list them here as a guide to future development.

Entity dependent attributes
Some attributes are limited to particular entities of a given ENE type. For example, "dam" exists on only a limited number of rivers, and only a few rivers have an associated famous song or poem (e.g. Loreley on Rhine River). The construction reported in this paper is our first attempt and only the widely applicable attributes are identified. However, some entity-dependent attributes are very important and should be considered in the future.

Fineness of attribute
Some of the attributes are hierarchical and there is no concrete guidance for selecting the appropriate level. For example, an entry of "birds" has attributes like "color of

chest” or “color of feather”, but another entry has “color of entire body”. Ideally, the attributes should be specified hierarchically, but it is too complicated to organize an attribute hierarchy as part of this initial effort.

Span of value expression
We have restricted values to nouns or noun equivalents. However, some descriptions, including definitions, are actually good attributes. By allowing such values, attribute construction may become quite a bit harder and a more elaborate procedure will be needed.

Structure in value
Values could have structure. For example, “museum” has the attribute “exhibit”, but an exhibit, such as a painting, sculpture etc, has its own attributes like creator, year of creation and so on. The value can be a pointer to the entry if the entry exists in the knowledge base, but such instances are limited.

ENE category definition
In our past development, defining ENE categories was very subjective and it could be very difficult to judge the category of an entity. However, we found that the attributes provide very useful information for solving these problems, because the presence of important attributes is easier to judge, and people can find attributes for an entity relatively easily. However, some difficulties remain, as we have just described and further investigation is needed of the relationship between the ENE category design and attribute design.

Distinction of mandatory and optional
There are two kinds of attributes in many categories. One is mandatory (attributes) and the other is more optional (property). It is desirable to be able to distinguish the two types in the ontology. However, this may include subjective judgments and considerable human labor may be needed.

5. Inter-Annotator Agreement

In order to assess the difficulty of attribute construction, we measured the inter-annotator agreement for four categories; “Person”, “Landform”, “International Organization” and “Academy”. Both of the annotators have a Masters degree in linguistics and one of them has been working on this ontology for several years. We consider an attribute to “match” if more than 60% of the values of the attribute overlap, even if the attribute names are not identical. We classified the results based on the percentage of entities which have a value for a given attribute. Note that given an attribute for a category, not all of the entities have a value of the attribute. The more values are found for an attribute, the more common the attribute is for the category. Obviously, we expect greater agreement on the more common attributes. Table 1 shows the high agreement rate achieved, in particular for the most common attributes (those with a value ratio (percent of entries having a value) of 100-60%). If an attribute of one annotator matches two or more attributes of the other annotator, we count it as a “partial match” (e.g. “birth place” vs. “birth country” and “birth city”). The major disagreements were caused by the constraints on allowable values, i.e. if it is a noun equivalent or not.

Table 1. Inter-annotator agreement

Value ratio	100-60%	60-40%	40-10%
Match	13	12	26
Partial match	9	11	11
Disagree	4	22	28

6. ENE Definition and Attributes

We redesigned an ENE hierarchy to incorporate attributes. Appendix A shows the new ENE hierarchy. The categories with underline are the ones in which attributes are defined. There is no change in the numerical and time expression categories from the previous version. The depth of the hierarchy is now limited to 3. The second layer is shown by indentation, and the third layer is shown after “:”. Appendix B shows two examples of attributes, for “Person” and “International Organization”. The frequency shows the frequency of entities which have a value of the attribute. We can see that it roughly indicates the importance of the attributes. The typical attribute values are also described using ENE categories, if possible.

7. Acknowledgements

First, we would like to express sincere appreciation to Shogakkan Inc for letting us use the online version of their encyclopedia. In particular, the discussions and suggestions given by the people who have been working on the encyclopedia were very valuable for us. We also thank for the annotators. In particular, Mr. Kosuke Takeuchi has benn working on this Extended Named Entity project for a long time and now without his contribution, we could not make this project go forward. We appreciate the colleagues at New York University for their support and suggestions.

8. References

ENE HP: Extended Named Entity Homepage. http://nlp.cs.nyu.edu/ene

OBO HP: The Open Biomedical Ontologies HP: http://www.geneontology.org/

SemanticWeb HP: http://www.w3.org/2001/sw/

SUMO HP: Suggested Upper Merged Ontology. http://www.ontologyportal.org/

Wikipedia HP: http://wikipedia.org

C. Fellbaum, editor. WordNet: An Electronic Lexical-Database. MIT Press, 1998.

Ralph Grishman, Beth Sundheim (1996). Message Understanding Conference - 6: A Brief History In Proceedings of the 16th International Conference on Computational Linguistics, 1996

N. Guarino. (1992). Concepts, attributes and arbitrary relations: Some linguistic and ontological criteria for structuring knowledge base. Data and Knowledge Engineering, 8, 249–261.

S. Harabagiu, D. Moldovan, C. Clark, M. Bowden, J. Williams, and J. Bensley. (2003). Answer Mining byCombining Extraction Techniques with Abductive Reasoning. In Proceedings of TREC 2003.

Xin Li and Dan Roth (2002). Learning Question Classifiers. In Proceedings of the 19th International Conference on Computational Linguistics

C. Matuszek, J. Cabral, M. Witbrock, and J. DeOliveira. An introduction to the syntax and content of Cyc. In AAAI Spring Symposium, 2006.

A. Philpot, E. H. Hovy, and P. Pantel. 2008. The Omega Ontology. In Huang, C. R., A. Gangemi, A. Lenci, and N. Calzolari (eds), Ontologies and Lexical Resources for Natural Language Processing. Cambridge University Press.

Almuhareb, A., Poesio, M. (2004) Attribute-Based and Value-Based Clustering: An Evaluation, In the Proceedings of Empirical Methods in Natural Language Processing. 2004.

James Pustejovsky. (1995). The Generative Lexicon. The MIT Press.

Satoshi Sekine and Hitoshi Isahara. (2000) IREX: IR and IE evaluation-based project in Japanese In Proceedings of the Second International Conference on Language Resources and Evaluation ; 2000

Satoshi Sekine and Chikashi Nobata. (2004) Definition, Dictionary and Tagger for Extended Named Entities Forth International Conference on Language Resources and Evaluation, Canaly Island, 2004.

Fabian M. Suchanek, Gjergji Kasneci, Gerhard Weikum (2007) YAGO: A Core of Semantic Knowledge Unifying WordNet and Wikipedia, In the Proceedings of 16th International WWW Conference.

Naoki Yoshinaga and Kentaro Torisawa. (2007) Open-Domain Attribute-Value Acquisition from Semi-Structured Texts, Proceedings of the Workshop on Ontolex 2007 -- The Lexicon/Ontology Interface held at the fifth International Semantic Web Conference，pp. 55-66，Nov., 2007

Appendix A: Definition of ENE hierarchy

NAME

- Name_Other, Person, God
- Organization
  - Organization_Other, Internationa_Organization, Show_Organization, Family,
  - Ethnic_Group : Ethnic_Group_Other, Nationality
  - Sport_Organization : Sport_Organization_Other, Pro._Sport_Organization, Sport_League
  - Corporation : Corporation_Other, Company, Company_Group
  - Political_Organization : Political_Organization_Other, Government, Political_Party, Cabinet, Military
- Location
  - Location_Other, Spa
  - GPE : GPE_Other, City, County, Province, Country
  - Region : Region_Other, Continental_Region, Domestic_Region
  - Geological_Region : Geological_Region_Other, Landform, River, Lake, Sea, Bay
  - Astral_Body : Astral_Body_Other, Star, Planet, Constellation
  - Address : Address_ Other, Postal_Address, Phone_Number, Email, URL
- Facility
  - Facility_Other, Facility_Part
  - Relics : Relics_Other, Tumulus
  - GOE : GOE_Other, Public_Institution, School, Research_Institute, Market, Park, Sports_Facility, Museum, Zoo, Amusemnt_Park, Theater, Workship_Place, Car_Stop, Station, Airport, Port
  - Line : Line_Other, Railroad, Road, Canal, Water_Route, Tunnel, Bridge
- Product
  - Product_Other, Material, Clothing, Money, Drug, Weapon, Stock, Award, Decoration, Offense, Service, Class, Character, ID_Number
  - Vehicle : Vehicle_Other, Car, Train, Aircraft, Spaceship, Ship
  - Food : Food_Other, Dish
  - Art : Art_Other, Picture, Broadcast_Program, Movie, Show, Music, Book
  - Printing : Printing_Other, Newspaper, Magazine
  - Doctrine_Method : Doctrine_Method_Other, Culture, Religion, Academic, Style, Movement, Theory, Plan
  - Rule : Rule_Othere, Treaty, Law
  - Title : Title_Other, Position_Vocation
  - Language : Language_Other, National_Language
  - Unit : Unit_Other, Currency
- Event
  - Event_Other
  - Occasion : Occasion_Other, Festival, Game, Conference
  - Incident : Incident_Other, War
  - Natural_Disaster : Natural_Disaster_Other, Earthquake
- Natural_Object
  - Natural_Object_Other, Element, Compound, Mineral
  - Living_Thing : Living_Thing_Other, Fungus, Mollusc_Crustacean, Insect, Fish, Amphibia, Reptile, Bird, Mammal, Flora
  - Living_Thing_Part : Living_Thing_Part_Other, Animal_Part, Flora_Part
- Disease
  - Disease_Other, Animal Part
- Color
  - Color_Other, Nature_Color

TIME EXPRESSION

- Time_Top_Other
- Timex
  - Timex_Other, Time, Date, Day_Of_Week, Era
- Periodx
  - Periodx_Other, Period_Time, Period_Day, Period_Week, Period_Month, Period_Year

NUMERICAL EXPRESSION

- Numex_Other, Money, Stock_Index, Point, Percent, Multiplication, Frequency, Age, School_Age, Ordinal_Number, Rank, Latitude_Longtitude
- Measurement
  - Measurement_Other, Physical_Extent, Space, Volume, Weight, Speed, Intensity, Temperature, Calorie, Seismic_Intensity, Seismic_Magnitude
- Countx
  - Countx_Other, N_Person, N_Organization, N_Facility, N_Product, N_Event
  - N_Location : N_Location_Other, N_Country
  - N_Natural_Object : N_Natural_Object_Other, N_Animal,N_Flora

Appendix B: Examples of Attributes

Table 2 Attributes for “Person”

Attribute(20)	Example of value	Freq. (%)	ENE
Vocation	professional baseball player, economist, poet	46(100)	Vocation
Nationality	American, Chinese, Japanese	29(63)	Country
Career	A professor at Yale University, The Princess of Wales	26(57)	Vocation
Masterpiece	Guernica, Mona Lisa	25(54)	Product, Facility
Graduate	M.A. in German at Cambridge, MK High School	20(44)	School
Hometown	Paris, Manchester, Shanghai	19(41)	City
Native Province	State of Illinois, Sichuan	18(39)	Province
Previous stay	England, New York	12(26)	Location
Mentor	Andrea del Verrocchio, Michelangelo di Lodovic Buonarroti Simoni	10(22)	Person
Death date	04,23,1704, 04/23/1704, unknown	10(22)	date
Era	Edo period, the 11th century	8(17)	Era
Award	Academy Award, MVP, Nobel Prize	8(17)	Award
Real Name	Saint Nicholas	8(17)	Person
Another name	Santa, father Christmas	8(17)	Person
Title	Knight, an honorary degree at Yale	6(13)	Title
Competition	World Series, 1955 piano competition in Paris	6(13)	Game
Place of birth	New York, Birmingham	5(11)	Location
Father	John B. Kelly, Sr.	5(11)	Person
Cause of death	Car accident, Guillotine	5(11)	

Table 3 Attributes for “International Organization”

Attribute(17)	Example of Value	Freq.	ENE
Another name	CARICOM, EMU, CCDN	30(75)	Inter._Org
Year founded	1/10/1920, 01,10,1920, 2004,	26(65	Date
Purpose of foundation	Encouragement of the African economy	23(58)	
Number of signatories	170 countries, 190	20(50)	N_Country
Type	League of Nations, International Labor Organization	16(40)	
Headquarters	New York, Prague	13(33)	City
Agreement, Proposal	Covenant of the League of Nations	12(30)	Rule
Top organization	EU (the European Union)	11(28)	Inter._Org
Member	China, Senegal, Norway	10(25)	Country
Predecessor	African Union (OAU), Caribbean Free Trade Association	9(23)	Inter._Org
Subsidiary organization	International Amateur Athletics Federation	8(20)	Org.
Rank	Board of directors, Special organization.	7(18)	
Headquarters (country)	Japan, Czech, Ethiopia	7(18)	Country
Year of dissolution	1974, 06/20/1977, Dec,01	6(15)	Date
Proposer Country	USA, England, Luxemburg	5(13)	Country
Successor organization	United Nations Economic and Social Commission for Asia and the Pacific	5(13)	Inter._Org
Proposer (Person)	Eisenhower, Colonel Qadhafi, Pierre Wellner	4(10)	Person