Grounding Named Entities for Knowledge Extraction

Yoshihiro Matsuo†, Nozomi Kobayashi, Toru Hirano, and Izumi Takahashi

Abstract

This article presents a technique for grounding named entities that makes possible services that link text content to various kinds of databases by using semantic disambiguation of named entities. Such collation of information that is expressed in various different ways is essential for accurate analysis and utilization of user-generated content such as weblogs (blogs).

1. Grounding

Grounding is a term used in the field of artificial intelligence research to describe the creation of associations between the names mentioned in dialogues between humans and machines and the real-world objects to which these names refer. This is necessary because for a meaningful dialogue to take place it is essential that when somebody refers to a person called Mr. X, the machine identifies this as the same Mr. X to whom the human is referring.

In Web pages, named entities referring to the same object can be expressed in a multitude of different ways. This tendency is particularly evident in consumer generated media (CGM), which can contain a very wide diversity of references to the same entity, including abbreviated names, pet names, and even obfuscations.

We will consider situations where Web pages of this sort are being searched. For example, let's suppose that we are searching for Web pages that mention a prime minister called 電電一郎 (*Denden Ichiro*). Is it sufficient simply to type 電電一郎 into a search engine? In a search engine such as *goo* [1], we would probably get more results by typing in 電電 AND 首相 (*Denden AND prime minister*) instead of 電電一郎. Since it can be inferred that most of the results obtained by the latter method will be documents relating to Prime Minister Denden Ichiro, it can be seen that if you simply uses the person's full name as a search string, you are likely to end up missing an appreciable number of results.

The grounding technique we have been working on aims to solve this sort of problem, and we have devised a function for associating a unique ID (ground) with various expressions that appear in text documents. In the previous example, the same ID would be associated with expressions such as 首相の電電さん (*Denden-san the prime minister*), 電電内閣総理大臣 (*Denden, the head of the cabinet*), and the pet name 電ちゃん (*Den-chan*) (**Fig. 1**). If IDs can be associated with names mentioned in Web text, then it will become possible to gather these documents (**Fig. 2**). Furthermore, by linking these IDs to a database, it will be possible to provide advanced Web services such as a unified display of related information.

2. Difficulty of implementing grounding techniques

There are basically two techniques needed for grounding. One is a technique for resolving ambiguities, e.g., by determining the individual to whom a name refers. This requires the word sense disambiguation technique in natural language processing. When a document contains a common name like Mr. Fukuda, there are many possibilities for the referent (the

† NTT Cyber Space Laboratories
Yokosuka-shi, 239-0847 Japan
Email: matsuo.yoshihiro@lab.ntt.co.jp

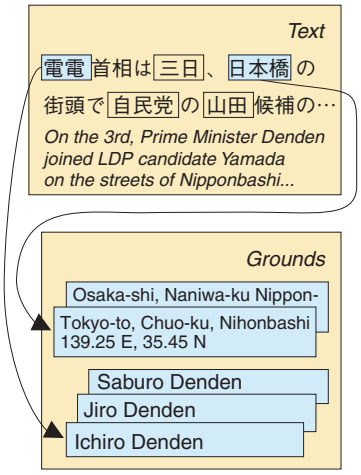

Fig. 1. Expressions and grounds.

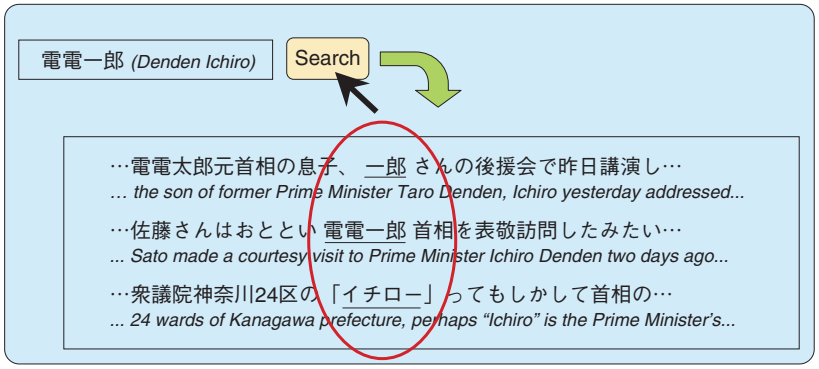

Fig. 2. Gathering Web pages with various expressions.

object being referring to). For example, the Japanese Wikipedia has articles on at least 70 people called Fukuda. Since the people mentioned in Wikipedia have achieved some degree of notability, it is likely that expanding the scope to include other less notable individuals such as the friends of bloggers would result in innumerable possibilities. The telephone directory lists thousands of people called Fukuda. The first type of technique is used to resolve this ambiguity based on information such as the context.

The other technique acquires knowledge regarding what sort of expressions are likely to be used for each entity. For example, people called Ichiro Yamada and Hanako Sato might be referred to by the pet names ヤマさん (*Yama-san*) and さとっち (*Satotchi*). Without this knowledge, it would be impossible to identify さとっち as a reference to someone called Hanako Sato. Since new pet names and abbreviated names come into use every day, the second technique is used to acquire these synonyms automatically.

The technique for resolving ambiguities related to people and places and the technique for acquiring synonymous named entities are described in below.

3. Person name disambiguation

Disambiguating person names, which are often mentioned and searched for in CGM, is a key issue. For example, suppose that there is a political journalist called 電電花子 (*Denden Hanako*) and a soccer player called 電電太郎 (*Denden Taro*). When people see the words 電電さん (*Denden-san*), how do they know whether it refers to Hanako or Taro? Without any context, it is hard to identify which of these individuals is the referent. This decision can only be made based on the context. For example, a reference made in a political context is more likely to refer to Hanako the journalist, whereas a reference made in a

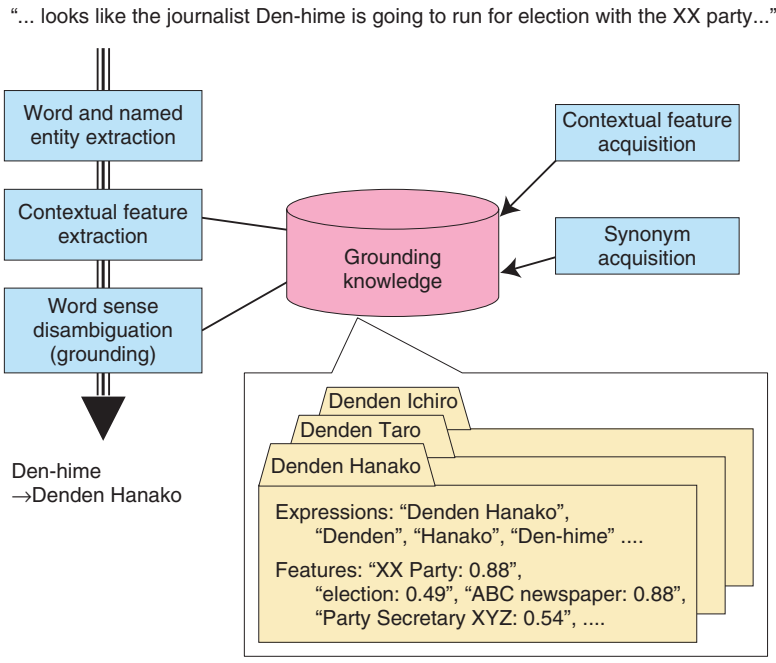

Fig. 3. Process of named entity grounding.

sports context is more likely to refer to Taro the soccer player. Also, Hanako is more likely to be mentioned at the same time as Secretary-General X of some political party, while Taro is more likely to be mentioned at the same time as soccer player Y.

The purpose of grounding is to perform such judgments automatically (**Fig. 3**). In this technique, the distribution of terms appearing in the vicinity of each person's name is calculated beforehand from a large volume of text. In Hanako's case, we can expect that terms such as "Liberal Democratic Party" and "parliament" will be collected with high frequency.

When judging who is actually being referred to by the name *Denden-san* appearing in a passage of text, the technique compares the distribution of the terms surrounding the words *Denden-san* with previously gathered distributions, and the actual referent is judged by the similarity of the distributions.

A technical issue involved in this estimation is how to apply suitable weighting to the terms appearing in the vicinity of these entities so that terms with high discrimination ability receive higher weighting than terms that are less useful. For example, terms such as "television" might appear in connection with both of these people. In this technique, an efficient weight is estimated statistically on the basis of which terms appear more often for a particular individual and not so often for other people.

4. Place name disambiguation

Nowadays, with the development of the geographic information system (GIS), it is becoming increasingly important to identify the location of geographical affairs in the real world. In particular, the appearance of online maps such as goo's map application programming interface (API) has led to a rapid spread of services where information from various databases is organized on maps. If the real-world locations of places mentioned in text content could be inferred, then it would become possible to use text content as a source of data for GIS positioning in addition to databases (**Fig. 4**).

To infer the locations of place names mentioned in text content, it is necessary to identify the locations that are actually indicated by place names referred to in vague expressions. For example, consider the following expression: "日本橋 (*Nihonbashi/Nipponbashi*) can easily be reached from 難波 (*Namba*)". According to the gazetteer [2] published by the Ministry of Land, Infrastructure and Transport, there are two place names in Japan that include the characters 日本橋 and eighteen place names containing the characters 難波. Place name disambiguation could thus be described as the problem of selecting likely candidates from among these possibilities.

This technique mainly uses two viewpoints to resolve ambiguities. One is the viewpoint that place

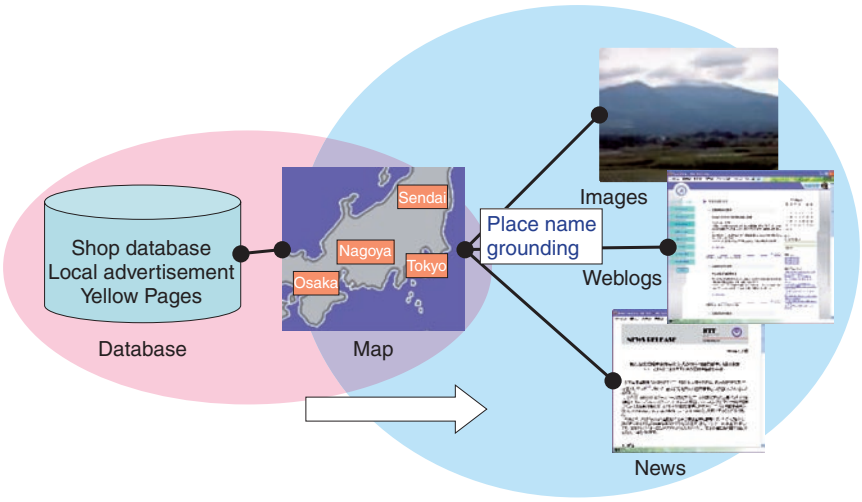

Fig. 4. Linking maps with text contents.

names appearing in the same text are often situated close together (i.e., geographically close together, not close together on the page). The place names mentioned in the text and the latitude & longitude candidates for each place name are all listed out, and if the positional candidates that are closest together out of all possible combinations are judged to correspond to the locations of these place names in the real world, then the place names 日本橋 and 難波 in the above example can be judged to correspond to Nipponbashi in Naniwa-ku, Osaka (34.39 N, 135.50 E) and Namba in Chuo-ku, Osaka (34.40 N, 135.51 E).

The other viewpoint is that place names that appear in text without any preface must be well-known. There are many locations in Japan called Ginza; however, when it is mentioned without any preface, it almost always refers to the district in Chuo-ku, Tokyo. When referring to other locations called Ginza, people will generally use an expression of the form "Ginza town in city X, prefecture Y".

So how can we go about estimating how well-known a place name is? Several methods have been proposed: one of them uses the number of shops in each location as an indicator of how well-known it is [3]. The reasoning behind this method is that the presence of a large number of commercial establishments in an area means that it must be visited by a large number of people, so there is likely to be a fixed relationship between the number of shops in an area and the proportion of people that know of this location when they read about it. By combining these disambiguation techniques, we have confirmed that it is possible to correctly infer the location of place names in the real world with 92.7% accuracy in tests where blog articles were evaluated.

An example of a service where this technique is applied is the goo image search [4]. In this service, place names contained in text associated with crawled images (captions, etc.) are used to provide an image search function with a map interface.

5. Automatic acquisition of synonym knowledge needed for grounding

To achieve named entity disambiguation, we first need a list of candidates to define which meanings are possible for each expression. If the candidate list includes the relationship *Den-chan* → *Denden Ichiro*, *Yamada Denko*, then disambiguation is the problem of deciding which of these people is the referent. This list of candidates can be generated from a list of synonyms indicating what kinds of expressions are used to refer to a particular entity.

There are many different types of synonyms for named entities. They range from simple spelling variations such as the elision of a vowel-extending character (ー) to pet names, which can defy explanation and can take many forms depending on how they are derived. We have classified these derivation processes, and we have proposed a method for automatically acquiring derivative named entity synonyms [5].

A derivative named entity synonym is a synonym produced by means such as forming similar sounds or similar expressions, regular abbreviations, and the addition of typical pet name forms (e.g., *Kawamura*

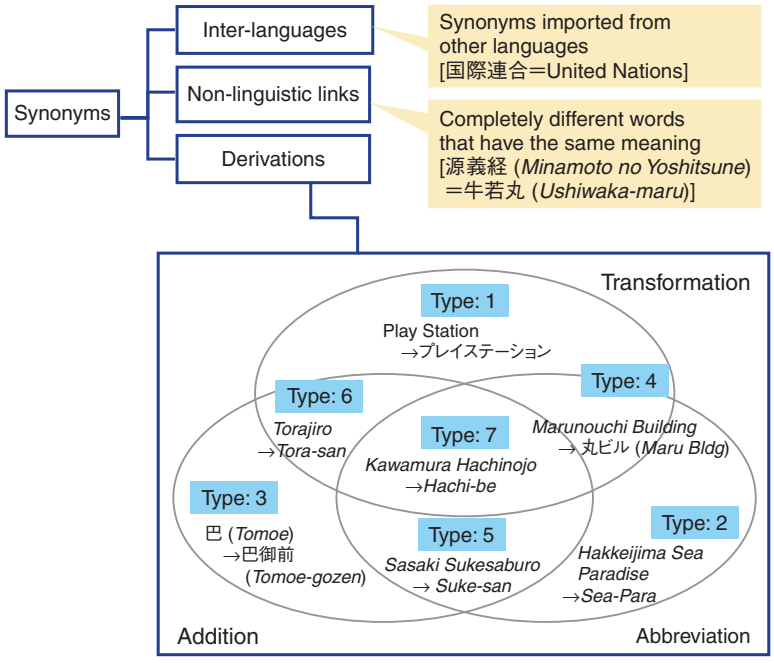

Fig. 5. Model for generating named entity synonyms.

Hachinojo → *Hachi-be*). A classification chart assembled from these derivation processes is shown in **Fig. 5**. By analyzing a text corpus, we have found that the derivation of about 90% of synonyms can be explained by combining these classes.

To decide whether or not a pair of words has a synonymous relationship, the abovementioned derivation processes are reproduced on a computer. To judge the similarity of sounds, the similarity of pronunciations is determined based on parameters such as edit distance, while typical pet name forms are judged using a dictionary. Furthermore, regular abbreviations are judged by machine learning to determine whether or not one term is an abbreviation of the other.

In tests where synonymous pairs were extracted from a corpus by a procedure involving a combination of these methods, we found that we were able to extract synonyms with approximately 70% accuracy.

6. Future work

In this article, we described a technique for linking items of Web content by inferring the meanings of named entities, and we discussed a named entity grounding technique that makes it possible to analyze and utilize this sort of content accurately. In the future, we plan to work at expanding the range of expressions that can be processed and to continue applying this technique to new Web mining services.

References

[1] http://www.goo.ne.jp/ (in Japanese).
[2] Ministry of Land, Infrastructure and Transport, "Street-level positional reference data," (in Japanese).
[3] T. Hirano, Y. Matsuo, and G. Kikui, "Location Disambiguation Using Geographic Distance and Popularity," The 70th National Convention of IPSJ, Vol. 2, pp. 85–86, 2008 (in Japanese).
[4] http://bsearch.goo.ne.jp/maptop/ (in Japanese).
[5] I. Takahashi, H. Asano, Y. Matsuo, and G. Kikui, "Judging Synonymous Named Entities based on Word Normalization," The 14th Annual Meeting of the Association for Natural Language Processing, pp. 821–824, 2008 (in Japanese).

Yoshihiro Matsuo
Senior Research Engineer, Supervisor, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.S. and M.S. degrees in physics from Osaka University, Osaka, in 1988 and 1990, respectively. He joined NTT Communications and Information Processing Laboratories in 1990. His research interests include multimedia indexing, information extraction, and opinion analysis. He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Nozomi Kobayashi
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

She received the M.E. and Dr.Eng. degrees in information science from Nara Institute of Science and Technology, Nara, in 2004 and 2007, respectively. She joined NTT Cyber Space Laboratories in 2007. She is a member of IPSJ.

Toru Hirano
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

He received the B.E. degree in systems engineering from Wakayama University, Wakayama, and the M.E. degree in information science from Nara Institute of Science and Technology, Nara, in 2003 and 2005, respectively. He joined NTT Cyber Space Laboratories in 2005. He is a member of NLP.

Izumi Takahashi
Research Engineer, Speech, Acoustics and Language Laboratory, NTT Cyber Space Laboratories.

She received the B.E. and M.E. degrees in engineering from Fukui University, Fukui, in 2004 and 2006, respectively. She joined NTT Cyber Space Laboratories in 2006. She is a member of the Japanese Cognitive Science Society.