Knowledge Extraction from Text for Intelligent Responses

Kugatsu Sadamitsu, Ryuichiro Higashinaka, Toru Hirano, and Tomoko Izumi

Abstract

In this article, we describe a linguistic analysis technique for extracting useful knowledge from the vast amounts of text on the World Wide Web. First, we introduce a technique for extracting named entities as a key to knowledge to be handled by a computer, and then we show how this can be applied to extract relations between named entities.

1. Introduction—Text processing aimed at providing an intelligent response

Computers are now able to correctly answer questions such as *How high is Mount Everest?* or to provide information that the user may be interested in but unaware of, such as *It looks like NTT is launching a new service this weekend* via a spoken language interface.

In this article, we introduce the latest techniques to acquire information with a spoken language interface.

For the most part, it seems that the knowledge users require is generally centered around specific entities, such as Mount Everest and NTT in the above examples. Textual expressions relating to such entities are called *named entities*. This means that the expression *Mount Everest* can be uniquely associated with a single entity.

These named entities are crucial to generate intelligent responses. This is because, for example, the named entity *Mount Everest* can be imparted with additional information such as *height* so that an answer can be found for the question *How high is Mount Everest?*, while for the named entity *NTT*, information about the launch of a new service can be gleaned from the World Wide Web (hereafter, the Web) to enable a response in the form, *It looks like NTT is launching a new service this weekend*. If a computer-accessible knowledge database can be constructed in this way, then it will become possible to provide intelligent responses (**Fig. 1**).

In this article, we introduce basic techniques for collecting named entities and a technique for extracting the relationships between them.

2. Automatic collection of named entities

If named entities have to be collected to construct a knowledge database, roughly how many named entities are there in the first place? The Wikipedia online encyclopedia contains many named entities; there are over 800,000 entries in the Japanese Wikipedia, and over 4 million in the English Wikipedia. However, these articles are limited to things that are famous or fairly well known, so if we include other less well known people, products, and places, then the total number of entities is quite staggering. Previously, sources of text material were limited to printed media such as newspapers, which imposed severe limitations on the availability of named entities. However, the recent growth of Internet services such as Twitter and blogs has made it possible for users everywhere to publish information by themselves. The Web is now flooded with a wide variety of named entities, and the extraction of these named entities is becoming a very important topic in the construction of knowledge databases.

Since the number of named entities is almost limitless, there is no point trying to manually add each one

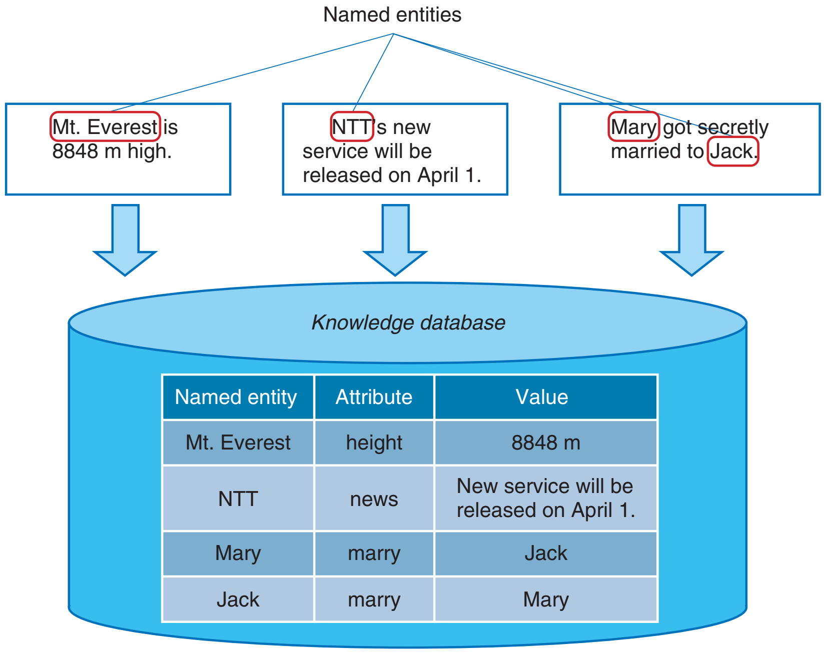

Fig. 1. Knowledge database based on named entities.

to a database. What we need is some way of extracting named entities from text automatically. In recent years, machine learning has become the tool of choice for the extraction of named entities. For example, let's consider the phrase *Today I met Lisa* (**Fig. 2**). The word *Lisa* in this phrase would normally be considered a named entity that is probably someone's name. This is because we expect *Lisa* to be a person's name based on the surrounding context. In this way, when we identify named entities, we also simultaneously classify them into a category such as the names of people or places. Similarly, machine learning can extract information (specifically, feature quantities) as cues from the surrounding context. On the basis of a statistical model obtained beforehand, we can simultaneously decide whether or not it is a named entity and, if so, which named entity category it belongs to.

Although broad named entity categories such as the names of people or places can be determined from the context, it can be difficult to infer categories with a finer level of detail. Consider the phrase *arrived at K2*. It is understood from the context that *K2* is a place name, but since the phrase provides no additional information, we need other cues in order to achieve a detailed categorization. (*K2* is, in fact, the name of the world's second highest mountain.)

Unlike existing dictionaries where each dictionary tends to have its own category definitions, CGM (consumer-generated media) dictionaries such as Wikipedia offer a high degree of freedom in the assignment of categories. This can cause problems because a systematic categorization tends not to be maintained. Furthermore, the necessary categories are themselves often application-dependent and cannot be used *as-is*.

In the following, we introduce two ways of resolving these issues.

3. Extraction of named entities from text

First, in cases where the characteristics of a named entity cannot be grasped from a single sentence, we considered that it should be possible to grasp the characteristics of the named entity by looking at the document as a whole. For example, if we know that the topic of a document concerns a mountain or a river, then it is possible to characterize the named entity.

So how do we go about grasping the topic of a document? A statistical model called a *topic model* was recently proposed and has been used in many applications [1]. The use of a topic model makes it possible to infer that a document containing words

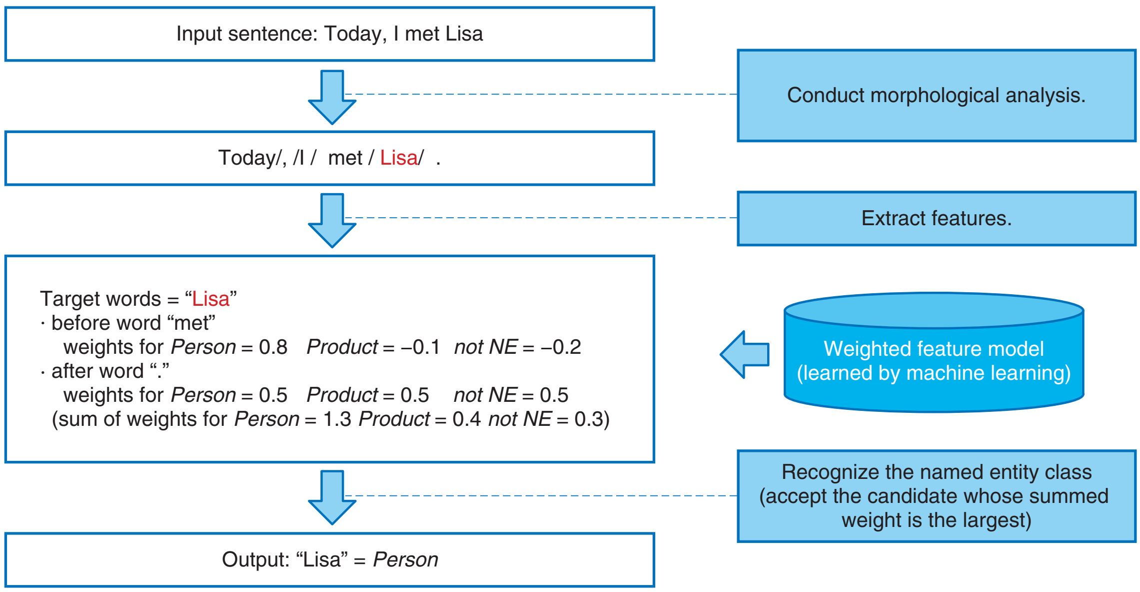

Fig. 2. Automatic extraction of named entities.

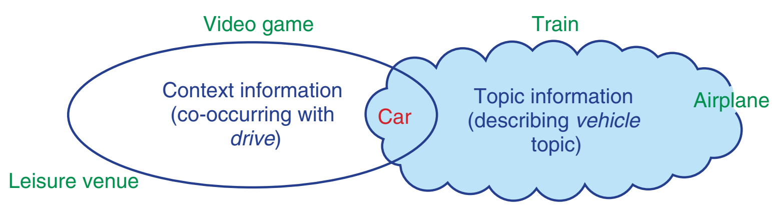

Fig. 3. Illustration of our named entity extraction method, which uses both context information and topic information.

such as *baseball* and *soccer* is probably related to the topic of *sports*.

By combining the global topic information obtained by this topic model with the local context information (and using them as features), we can gather named entities with greater precision [2]. This idea is shown in **Fig. 3**.

For example, suppose we want to gather named entities corresponding to models of vehicles. To extract the names of cars, it is important to look for words such as *drive* in the surrounding context. However, it is difficult to rule out other categories such as holiday destinations or motor racing video games without additional cues, since these categories can also co-occur with the word *drive*. Even if we assume it has been established that the topic is *vehicles*, there is still some ambiguity with the topic information alone, which could refer to *trains or airplanes*. Only when it is used in conjunction with context information can the ambiguity be greatly reduced, making it possible to accurately infer the correct category of car names.

4. Creating a named entity dictionary from existing dictionaries

The second method involves making use of existing dictionaries. As discussed above, existing dictionaries each have their own category definitions but do not share a unified categorization system. If these categories could be mapped to our desired named entity categories, then it would be possible to use existing dictionaries for the automatic extraction of named entities.

It hardly needs to be said that existing dictionaries are treasure troves of information. For example, if the Wikipedia entry for *K2* is assigned to the category *Himalayas*, then this suggests that it maps to a named entity in the *names of mountains* category. The phrase *the mountain in* ~ at the head of the description also provides a useful clue. We have developed a technique that can perform accurate category mapping by combining machine learning with clues obtained from multiple viewpoints in this way [3].

Although this technique has made it possible to extract named entities with high precision, it still has a weak point in that its applicable range is limited to well-known named entities. Therefore, a challenge for the future is to develop a better technique that can be used in conjunction with the automatic extraction of named entities from text as described above.

5. Extraction of relationships between named entities

So far, we have introduced a method that automatically constructs a named entity dictionary labeled with category information. However, a named entity dictionary on its own is not able to answer complex questions such as *Who did Maria marry?* To be able to do this sort of advanced processing, we need new information associated with the named entity. As an example of a practical technique that deals with named entities, we considered a method that extracts relationships between named entities from text. For example, let's consider how relationships between named entities can be extracted from the sentence *Nancy was shocked that Maria got secretly married to Jack*. We'll assume that the named entities *Maria*, *Nancy* and *Jack* are extracted. This sentence says that *Maria* and *Jack* have the relationship *married*, and that *Nancy* and *Jack* have no relationship. If we simply consider the surrounding named entities to be related, then we would mistakenly extract a relationship between *Maria* and *Nancy*, even though nothing is said about the relationship between these named entities.

Therefore, as shown in **Fig. 4**, we have developed a technique based on the results of dependency analysis that uses cues derived from the relative positioning of named entities to figure out if these entities are related, and if so, how they are related [4]. For example, we can see that the clauses about *Maria* and *Jack* are both connected to the *married* clause. We can therefore conclude that *Maria* and *Jack* have the relationship *married*. More accurate extraction of relationships between named entities can be achieved by combining the results of dependency analysis with a method that automatically identifies whether or not the surrounding words indicate a lexical relationship between the named entities based on a large-scale text corpus.

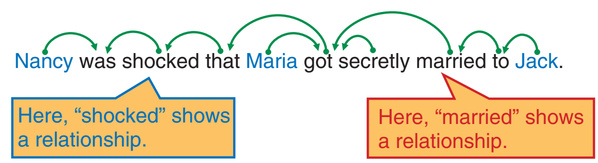

Fig. 4. Extraction of relationships between named entities.

6. Future work

In this article, we have introduced techniques for extracting knowledge from text in order to generate intelligent responses, with a particular focus on named entities. These techniques can be used in many different applications, including technology that can answer questions like the one posed in the Feature Articles entitled "Question Answering Technology for Pinpointing Answers to a Wide Range of Questions" [5]. It can also be used in search engines and the like.

Targets of knowledge extraction are not only named entities, but also relationship information (as described above), reputation information, and information to infer the attributes of blog/Twitter users. We will continue to further our study of knowledge extraction in the future in order to address the increasingly diverse needs of society and to propose new services.

References

[1] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, Vol. 3, pp. 993–1022, 2003.

[2] K. Sadamitsu, K. Saito, K. Imamura, and G. Kikui, "Entity Set Expansion Using Topic Information," Proc. of the 49th Annual Meeting of the Association for Computational Linguistics (ACL-HLT 2011): Human Language Technologies, pp. 726–731, Portland, OR, USA.

[3] R. Higashinaka, K. Sadamitsu, K. Saito, T. Makino, and Y. Matsuo, "Creating an Extended Named Entity Dictionary from Wikipedia," Proc. of the 24th International Conference on Computational Linguistics (COLING 2012), pp. 1163–1178, Mumbai, India.

[4] T. Hirano, H. Asano, Y. Matsuo, and G. Kikui, "Recognizing Relation Expression between Named Entities Based on Inherent and Context-dependent Features of Relational Words," Proc. of the 23rd International Conference on Computational Linguistics (COLING 2010), pp. 409–417, Beijing, China.

[5] R. Higashinaka, K. Sadamitsu, K. Saito, and N. Kobayashi, "Question Answering Technology for Pinpointing Answers to a Wide Range of

Questions," NTT Technical Review, Vol. 11, No. 7, 2013.
https://www.ntt-review.jp/archive/ntttechnical.php?contents=ntr201307fa4.html

Kugatsu Sadamitsu

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E., M.E., and Ph.D. degrees in engineering from Tsukuba University, Ibaraki, in 2004, 2006, and 2009, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2009. His current research interests include natural language processing and machine learning. He is a member of the Information Processing Society of Japan (IPSJ) and the Association for Natural Language Processing (NLP).

Ryuichiro Higashinaka

Senior Research Scientist, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.A. degree in environmental information, the Master of Media and Governance, and the Ph.D. degree from Keio University, Kanagawa, in 1999, 2001, and 2008, respectively. He joined NTT in 2001. His research interests include building question answering systems and spoken dialogue systems. From Nov. 2004 to Mar. 2006, he was a visiting researcher at the University of Sheffield in the UK. From 2006 to 2008, he was a part-time lecturer at Osaka Electro-Communication University. Since 2010, he has been a part-time lecturer at Keio University. He is a member of the Japanese Society for Artificial Intelligence, IPSJ, and NLP.

Toru Hirano

Research Engineer, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

He received the B.E. degree in systems engineering from Wakayama University and the M.E. and Ph.D. degrees in information science from Nara Institute of Science and Technology in 2003, 2005, and 2012, respectively. He joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2005. His current research interests include information extraction and user profile inference. He is a member of NLP.

Tomoko Izumi

Researcher, Audio, Speech, and Language Media Project, NTT Media Intelligence Laboratories.

She received the B.A. degree in global education from Hokkaido University of Education, in 2005, the M.A. degree in applied linguistics from Boston University, Massachusetts, USA, in 2007, and the M.E. degree in English Language Education from Hokkaido University of Education in 2008. She joined NTT Cyber Space Laboratories (now NTT Media Intelligence Laboratories) in 2008. Her current research focuses on automatic recognition of synonyms. She is a member of NLP.