Text Analysis Technology for Big Data Utilization

TSUCHIDA Masaaki, ISHIKAWA Kai, KUSUI Dai, KUSUMURA Yukitaka, NAKAO Toshiyasu

Abstract

Since a huge amount of the texts included in big data consists of data created by humans for communicating information or expressing intentions to other humans, it is an important information source containing valuable information. NEC is tackling the development of technology for extracting “customers’ voices” and “rumors” from large amounts of text data and for utilizing them in marketing, corporate risk management and customer management. This paper introduces some of the recent research results of NEC. Included are: the recognizing textual entailment technology for recognizing included relationships of semantic content between texts, the technology for rumor detection from cyber information and the semantic search technology for improving the operation efficiency of contact centers.

Keywords

textual analysis, semantic search, rumor detection, recognizing textual entailment

1. Introduction

A text is data created by humans for communicating information or intentions to other humans. Thus, among big data, the large amount of text data existing in society, such as data contained in newspapers, magazines, web pages, corporate documents and e-mails is particularly important. This is because it is the source of information that includes information that is valuable for humans. On the other hand, utilization of the large amount of text data requires techniques for processing the data by; a) accurately and quickly finding the required data, b) disambiguating expressions, and c) understanding connotations and relationships with other information. At NEC, we are conducting R&D of the following technologies aiming at marketing, corporate risk management and customer management by extracting “customers’ voices” and “rumors” related to products, services and persons (**Fig. 1**).

(1) Language analysis technology capable of free expressions including spoken languages.

(2) Synonymy/entailment relation recognition technology for judging if two words or texts have the same meaning.

(3) Semantics/relationship extraction technology for extracting semantics and relationships.

(4) Search and monitoring technologies for the integrated utilization of various information sources based on the extracted semantic content.

In the present paper, we introduce some of our recent research results, including the recognizing textual entailment technology that relates to (2), the technology for rumor information detection from cyber information including web texts, and the semantic search technology for improving contact center operation efficiency that relate to (4).

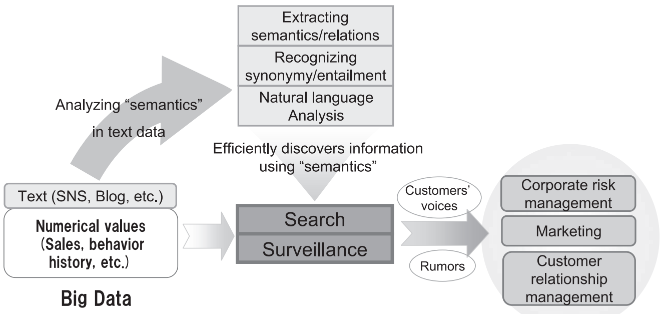

Extracting semantics/relations
Recognizing synonymy/entailment
Natural language Analysis
Analyzing “semantics” in text data
Efficiently discovers information using “semantics”
Text (SNS, Blog, etc.)
Numerical values (Sales, behavior history, etc.)
Big Data
Search
Surveillance
Customers’ voices
Rumors
Corporate risk management
Marketing
Customer relationship management

Fig. 1 Technology for extracting values for large amount of texts.

2. Recognizing Textual Entailment Technology for Recognizing the Included Relationships of Semantic Content

Regarding text data, there are many cases in which content of the same meaning is found in different expressions. The technology for recognizing that different expressions have the same meaning is called recognizing textual entailment (RTE). Our RTE technology won the first prizes both in main task and sub task of RTE-7, TAC2011, which is a world leading evaluation workshop of text analysis technologies held by the U.S. National Institute of Standards and Technology (NIST) in

2011.

RTE inputs two texts and recognizes whether one piece of text entails another. For example, the text "the president of firm A, traveled to New York City for business" contains semantically the meaning "the president of firm A visited the USA." This is because the fact of "traveling to NYC for business" always means "going to the USA."

RTE makes it possible not only to search for texts containing specific meanings. On the other hand, if a text containing a specific meaning does not exist, it is possible to detect the content of the specific meaning as "novel information," as well as to clustering texts of the same meaning.

Our technology recognizes entailment relationships with high accuracy by considering the importance of each word both in the text and in the textual structure such as in the subject and the object of the predicate. Specifically, it makes a judgment in two steps, in the first step a rough judgment is made by considering differences between the words expressed in two texts. The second step eliminates false entailment relationships by considering the linguistic and semantic structure of the texts. The technology can handle both the case in which different words express the same meaning and the case in which a word expresses different meanings and thereby prevents erroneous recognitions.

In the RTE-7, TAC2011, we gained first prizes both in the main task that recognizes entailment relationships between two given texts, and in the subtask that detects whether a piece of text contains novel information or not based on entailment relationships recognized among the text and other given texts.

We are currently advancing R&D with the aim of applying our RTE technology in the fields of marketing and corporate risk management. Conventional systems of the same purposes have limited ability of processing the meaning of text that caused users to assist the system in many aspects, for example, by extracting characteristic keywords for investigating the trend of customer's voices or selecting keywords for detecting rumors or reputations. On the other hand, RTE can offer a fundamental ability to text processing systems to handle the meaning of texts on a computer so that users can use the systems with less limitation.

3. Rumor Detection Technology from Cyber Information

The recent increase in ease of communication and dispatching information via the Internet such as blogs and social networking services has resulted in issues caused by the rapid spread of rumors and damage to reputations. For example, a Japanese bank suffered with withdrawals amounted to several ten billion yen because of the spreading of a rumor of bank failure via chain mails. To prevent damage due to such harmful rumors, it is important to detect risk information that might be a source of rumor at an early stage.

NEC has developed a system for detecting information with a risk of becoming rumor from information on the Internet (hereafter "cyber information") (**Fig. 2**). The rumor detection system is composed of the "dictionary building support technology" for supporting compilation of rumor detection dictionaries and the "rumor detection" block for detecting articles containing expressions registered in the dictionaries as rumor risk information, from a large amount of text data on the Web. Among the dictionaries, the "organization name dictionary" and "product name dictionary" are used to detect the targets of rumors and the risk expression dictionary" and the "negative expression dictionary" are used to detect the contents of rumors. For example, if the "corporate name dictionary" contains "A Bank of Commerce" and the "risk expression dictionary" contains "failure," it is possible to detect texts such as "there is a rumor of failure of A Bank of Commerce" or "bad loans may bring A Bank of Commerce to failure" from a large amount of text data.

When detecting risk information without omissions, it is important to cover all the possible risk expressions to be detected

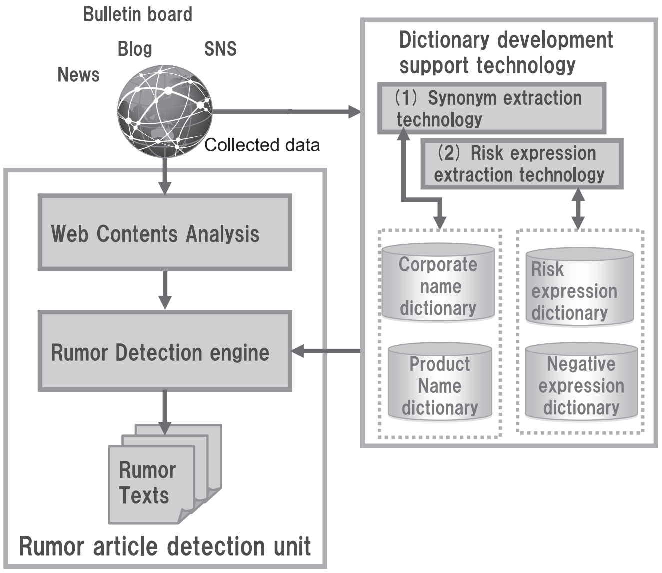

Bulletin board
Blog
SNS
News
Collected data
Dictionary development support technology
(1) Synonym extraction technology
(2) Risk expression extraction technology
Web Contents Analysis
Rumor Detection engine
Corporate name dictionary
Risk expression dictionary
Product Name dictionary
Negative expression dictionary
Rumor Texts
Rumor article detection unit

Fig. 2 Configuration of rumor detection system.

with their variations in the dictionaries. However, it is not realistic to carry this out in fully manual way because the huge labor cost is required for reading many texts and selecting applicable expressions.

Therefore, we developed dictionary building support technology to reduce the labor cost of the development of dictionaries. This technology is composed of the "synonym extraction technology" and the "risk expression extraction technology."

The synonym extraction technology can increase the variation of detected expressions by extracting words of the same meaning. For example, "ABC may collapse" cannot be detected if the dictionary includes "A Bank of Commerce" and "failure." However, it can be detected if the dictionary includes "ABC" and "collapse" that are their synonyms. Our developed technology can recognize whether two words are in synonymy relation based on multiple criteria including acronym transformation rules (i.e. "ABC" is generated from "A Bank of Commerce"), coincidence of translations ("failure" and "collapse" has the same Japanese translation " *hatan* ") and the similarity of the expressions in character string ("foreign currency deposit" and "foreign currency saving"). It is also capable of multi-step conversion, for example translating the input Japanese word into English and then generating the abbreviation (Japanese " *gaika yokin* " → English "foreign currency deposit" → abbreviation "FCD"). By all the above functions, our technology enables extracting wide variety of synonym candidates including completely different synonyms in character strings.

The risk expression extraction technology supports the development of the risk expression dictionary by extracting expressions representing risks. For example, organizations such as banks are conscious about the risks caused by "fictitious claims" and "phishing" as well as "bankruptcy." The developed technology inputs the risk articles related to the risks and the reference articles irrelevant of the risks of the same domain. The seed articles typically consist of announcements for customers by the other organizations of the same business domain, news articles general articles, etc. This technology extracts the expressions characteristic to the risk articles comparing with the reference articles based on the statistical differences among the frequencies of their appearances in the risk articles and the reference articles.

Other than the above extraction process, the developed technology ranks the risk expressions in the order of appropriateness. For instance, in the extraction of risk expressions related to "phishing," the expression "suspicious mail" can be regarded as being appropriate as a risk expression because it appears particularly often in risk articles. In contrast, "mail" is not extracted as a risk expression because it appears very often in articles other than the risk articles. On the other hand, while "a suspicious mail is sent" is appropriate as well, however it only covers very limited expressions and it fails to detect even a slightly different expression such as "a suspicious mail from the bank." The developed technology outputs "suspicious mail" at a higher rank compared to "suspicious mail is sent."

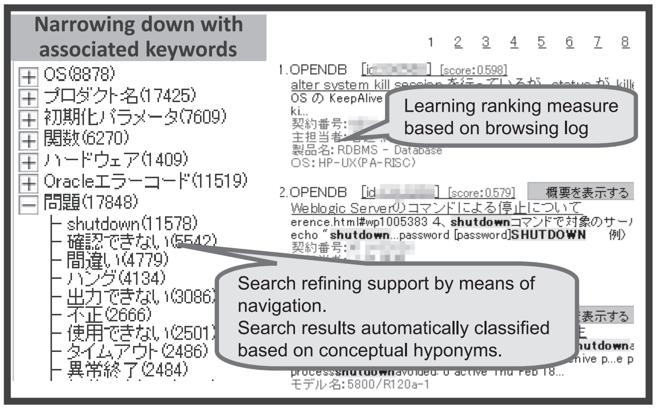
The user can create a dictionary by checking the expressions output from the dictionary building support technology and by adding appropriate expressions to the dictionary. Using our technology, the labor cost of dictionary development can be drastically lowered because the cost of checking of the expressions is much lower than reading all the texts manually to find expressions. We successfully confirmed that the accuracy of rumor detection task with a dictionary developed by using our method is equivalent to the accuracy with a dictionary developed by reading all the text in an evaluation. The time required for the development with our method is about 60% of the time required for the latter method.

Thus we developed the rumor detection system with a dictionary containing about 6,000 expressions on a server unit witch can process about 18 million Twitter texts (tweets) in about 2-1/2 hours. Introducing our dictionary development technology leads to reduction of dictionary development cost by about 30% compared to the case of manual reading of all the tweets with target organization names.

4. Semantic Search Technology for Improving the Operational Efficiencies of Contact Centers

We have developed a search technology that enables an accurate and quick response for inquiries to a contact center by utilizing the enormous inquiry response logs stored at the contact centers (**Fig. 3**).

In general, at a contact center intended for technical support, each operator should identify the subject of inquiry from each customer and find the cause and solution by referring to the large amount of stored knowledge documents (past cases, manuals, technical information, etc.). However, expansion of supported products and diversification of product functions have made the inquiries from customers more complicated and technically advanced, and some cases have rather lengthy response times in performing analysis of information and acquisition



Narrowing down with associated keywords

1.OPENDB [id[redacted]] [score:0.598]
alter system kill session を行っているが、status が kill
OS の KeepAlive
ki...
契約番号:[redacted]
主担当者:
製品名:RDBMS - Database
OS:HP-UX(PA-RISC)

2.OPENDB [id[redacted]] [score:0.579]
Weblogic Serverのコマンドによる停止について
erence.html#wp1005383 4、shutdownコマンドで対象のサー
echo "shutdown...password [password]SHUTDOWN 例)
契約番号:[redacted]

processshutdownavoided: 0 active Thu Feb 18...
モデル名:5800/R120a-1

Learning ranking measure based on browsing log

Search refining support by means of navigation.
Search results automatically classified based on conceptual hyponyms.

Fig. 3 Example of knowledge search system display (Partial).

of accurate information.

The semantic search technology is a technology for reducing search oversights by exhaustive searches of texts associated with key search words. It searches a wide range of associated documents by extending the searched words to their synonyms, hypernyms and hyponyms. For example, if the search keyword is "OS," documents including semantically associated words such as the synonyms ("operating system"), hypernyms ("software") and hyponyms ("Linux") can also be searched. The primary features of our semantic search technology are as follows.

- **Compressed index management for high speed semantic searches**
  The search system compresses and saves index data that describes the relationship between the search keyword, including both their hypernyms and hyponyms, together with relevant texts. When a compact index is used to deploy the search keyword using its hypernyms and hyponyms, on-memory processing is possible and the search processing speed can be increased significantly.
- **Search refining support via navigation**
  The system classifies search results automatically using the important words contained in them based on a large-scale hypernym-hyponym dictionary that has been prepared in advance. This strategy offers an overview of the refining keyword candidates independently of the operator's knowledge, thereby allowing the narrowing down of the search results, even by a novice operator.
- **Search ranking learning based on a browsing log**
  The system learns the relationship between the search keyword and the usefulness of documents by combining the search/browsing log (data listing the combinations of the search keywords and browsed documents) recorded in the server and the evaluation data for each document input by the operator. This process contributes to an improved search accuracy.

We applied a search system incorporating the above technologies to the NEC Oracle Response Center, which is one of our contact centers for technical support, and evaluated the effects. As a result, it was confirmed that the average turn around time (average TAT) until the completion of a response decreased in spite of an increased number of inquiries. Moreover, we also confirmed a trend toward improved customer satisfaction.

Specifically, the average TAT dropped by 19.1% and the "perfect" evaluation in the customer satisfaction survey increased by 7.8%, while the number of inquiries increased by 31%. For the working time of operators, the average search time per search of middle ranked operators reduced by 14% and the number of search operations per day increased by 28%. With the novice operators, the reduction in average search time per search was only 3% but the number of search operations per day increased by 95%.

Based on the above results, we believe that the reduction in the search time and increase in the search count have enabled quick and polite responses and brought about the reduction in the average TAT and have improved the customer satisfaction simultaneously.

5. Conclusion

In the above, we introduced some of the results of our recent R&D in the field of natural language processing technology for the utilization of large amounts of text data. It is considered that, in order to utilize big data effectively, it is important to use technology capable of the collective analysis both of texts and non-texts that include numerical values. Especially, in the various decision making scenarios, the results of analysis alone are not enough, but the information on the causes leading to the results and for supporting the interpretation of the results are also necessary. The texts themselves are the important sources for obtaining such information. It is our intension in this context to continue our commitment to expanding the R&D of support technologies.

*Twitter is a registered trademark of Twitter, Inc.

*Linux is a registered trademark or trademark of Linux Torvalds in the U.S. and other countries.

Text Analysis Technology for Big Data Utilization

Authors' Profiles

TSUCHIDA Masaaki
Assistant Manager
Knowledge Discovery Research Laboratories
Central Research Laboratories

ISHIKAWA Kai
Principal Researcher
Knowledge Discovery Research Laboratories
Central Research Laboratories

KUSUI Dai
Principal Researcher
Knowledge Discovery Research Laboratories
Central Research Laboratories

KUSUMURA Yukitaka
Assistant Manager
Knowledge Discovery Research Laboratories
Central Research Laboratories

NAKAO Toshiyasu
Senior Manager
Knowledge Discovery Research Laboratories
Central Research Laboratories

Information about the NEC Technical Journal

Thank you for reading the paper.
If you are interested in the NEC Technical Journal, you can also read other papers on our website.

Link to NEC Technical Journal website

Japanese

English

Vol.7 No.2 Big Data

Vol.7 No.2

September, 2012

Special Issue TOP

◇ Papers for Special Issue

Big data processing platforms

Big data processing infrastructure

Data analysis platforms

Information collection platforms

Advanced technologies to support big data processing

◇ General Papers