NAIST-IS-DD0961023

Doctoral Dissertation

Organizing Information Based on Semantic Relation Recognition

Junta Mizuno

March 16, 2012

Department of Information Processing
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Junta Mizuno

Thesis Committee:

Yuji Matsumoto	Professor
Kiyohiro Shikano	Professor
Kentaro Inui	Professor
Masashi Shimbo	Associate Professor

Organizing Information Based on Semantic Relation Recognition*

Junta Mizuno

Abstract

Internet users are presented with a wealth of information on a variety of topics, but finding reliable opinions relevant to a user's query is a challenge because there are incorrect or biased information from various sources. In order to find reliable information and come to a deep understanding about the query, users need to group and arrange information. However, grouping and arranging information manually is infeasible for massive amounts of information. Given this problem, there is some research on assisting users in organizing information through automatic information classification, for instance, classification based on sentiment polarity and based on information sources. However, this approach is not suitable for classification based on the relation between a user's query and relevant sentences. For instance, to classify relevant information as agreeing and conflicting with a user's query, it is necessary to first recognize the relation between the query and the sentence.

We propose to organize information by recognizing semantic relation between information. In this dissertation, we first define the task, semantic relation recognition for the Web. Because there is no data set for the task in Japanese, we construct a development data set and analyze them to define the target semantic relations. Next, we propose a classification model which gradually relaxes a set of restrictions on the structural alignments between a query and a sentence. Then, we evaluate the model on an evaluation data set constructed separately from the development data set. Finally, we construct a Web application. In the Web application, the user can input a sentence as a query and the system organizes sentences relevant to the query and shows users a bird's-eye view. We discuss the effectiveness of information organization by comparing our system to an existing Web search engine.

*Doctoral Dissertation, Department of Information Processing, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD0961023, March 16, 2012.

The main contribution of this dissertation is the proposal of an approach to information organization based on a relation of information. Next, we defined the task of semantic relation recognition for the Web and proposed a classification model based on analyzing development data set. We confirmed that the most important technical issue for semantic relation recognition targeting Web texts is correctly detecting the region in retrieved Web texts that corresponds to the content of the user query. Finally, we evaluate the effectiveness of information organization based on relations through user evaluation.

-

Keywords:

natural language processing, world wide web, information organization, semantic relation recognition, recongnizing textual entailment

Acknowledgements

I am really indebted to many people for finishing my graduate study in natural language processing at the Nara Institute of Science and Technology and Tohoku University.

First of all, I would like to express my sincere thanks to Professor Yuji Matsumoto in Nara Institute of Science and Technology, for his valuable guidance and constant encouragement.

I would like to express my deeply grateful and sincere acknowledgment to Professor Kentaro Inui in Tohoku University. This dissertation could not have been accomplished without his well-directed advice. I was very glad I had a chance to make new laboratory in Tohoku University with him.

I also thank for the co-supervisors, Professor Kiyohiro Shikano and Associate Professor Masashi Shimbo, who gave me valuable comments and criticisms about this dissertation.

I would like to thank Statement Map Research Group in Nara Institute of Science and Technology and Tohoku University. Especially, I would like to express my thanks to Koji Murakami in Rakuten Institute of Technology and Assistant Professor Suguru Matsuyoshi in Yamanashi University who lead the group.

I had a good days in Tohoku University for the last two years. I would like to thank members in Tohoku University. Especially, I would like to express my thanks to Associate Professor Naoaki Okazaki who gave me constructive advice, Assistant Professor Yotaro Watanabe whom I talked with every time not only about research but also about private life, Researcher Eric Nichols who taught me English every time I wrote a paper in English, and Naoya Inoue who is one of the best colleagues.

In exploratory IT human resources project (MITOH Program), I would like to express my thanks to Hisashi Katsuya, Project Manager of my project. He gave me many chances to talk with a variety of people.

A lot of staff and my research group members had supported me to carry out experiments and write this dissertation at both Nara Institute of Science and Technology and

Tohoku University. I would like to express my thanks to Assistant Professor Mamoru Komachi who gave me a lot of advice, Yuko Kitagawa and Tomoko Yamaki who always help me about procedures. I appreciate to study together with all my colleagues in both laboratories. I thank Masakazu Iwatate, Erlyn Manguilimotan and Katsumasa Yoshikawa, who are Ph.D. candidates, for their valuable discussions.

Finally, I would like to thank my father Makoto Mizuno, my mother Hidemi Mizuno and my uncle Akira Mizuno who is Professor in Toyohashi University of Technology, for their supports in daily life to carry on this study.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The Web contains vast amounts of potentially useful information on a variety of topics called Information-Explosion Era [1]. Web search engines such as Google support users to search documents relevant to users' query. The size of Web data is estimated to exceed 35ZB in 2020. Therefore the importance of search engines will increase further [2]. Recently, there are many cases in which search engines retrieve a million of relevant documents from various sources. In such cases, it is impossible to check all documents. Although the searched results are ranked based on relevance, users may misunderstand that the ranking is based on credibility or trustworthiness, as can be seen in the following quotation [3]:

> A male humanities major expressed a similar understanding of the site by stating the following: "From my [experience] using Google [...] the most visited Web site is at the top so it's probably going to be the most relevant Web site and I think that's true"

Likewise, Fallows (2005) [4] reported as follows:

> 68% of users say that search engines are a fair and unbiased source of information; 19% say they don't place that trust in search engines.

By this misunderstanding, users often miss an important information buried in the ranking. For instance, there are some information which conflict to the query in the ranking. In order to understand such various information about the query, users need to group and arrange information.

However, for users to group and arrange information, they must survey and evaluate the document on their topics of interest, but there is often too much information to be manually feasible. Given this problem, a technological solution is needed to help users arrange information on the Web.

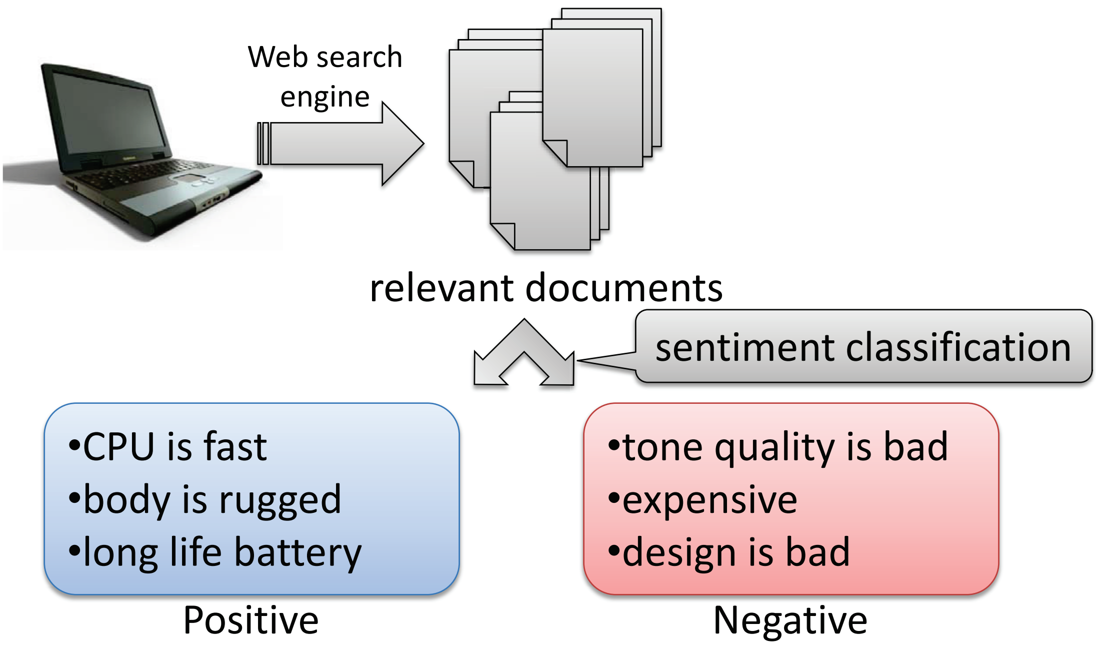

Figure 1.1: Sentiment classification example

To support users in arranging information, a typical approach is information classification. There are many researches to classify documents based on sentiment polarity [5, 6, 7]. Sentiment-based classification is effective for sentiment analysis [8, 9]. Figure 1.1 shows an example of document classification based on sentiment polarity. In the example, many documents relevant to the product are retrieved. Then, the classifier classifies the documents into positive or negative. The classification results help users make sense of the products.

Park et al. [10] classified news articles into two opposite groups on the basis of disputant relation. For given a topic, there are some disputants who claim about the topic. Then, the disputant who criticizes most frequently and the disputant who stands opposite side are specified. Finally, articles about the topic are classified into three groups: 1) the former disputant's claims, 2) the opposite disputant's claims, and 3) the other groups' claims. Their approach consists of the following three steps: 1) extracting disputants by frequency of criticisms by other disputants, 2) partitioning extracted disputants into two opposing groups, and 3) classifying the articles. They

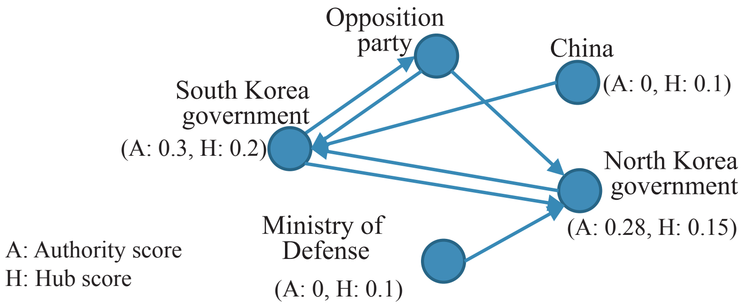

Figure 1.2: HITS graph example (Park et al.)

applied HITS algorithm in the step 2 to find major disputants (Figure 1.2). And they applied SVM in the step 3 to classify news articles, which are relevant to the main topic, into two opposing groups and one other group.

In both classification approaches, input to a classifier system is a number of documents or sentences relevant to the users' query and the system classifies them into some classes. For example, in sentiment classification task, documents or sentences are classified into three groups, positive, negative, and other. However, if we want to classify sentences by whether it agrees with a query or is conflicting to a query and conflicting to a query, such classification approach does not work well, because whether sentence agrees or conflicts with a query depends on the given query, but the existing approach does not take queries into account.

Figure 1.3 shows the previous approach to classify relevant documents. First, documents (doc 1–3) relevant to the query are retrieved by a Web search engine. Then, each document is classified into two groups A and B. In this figure, doc 1–2 are classified into group A and doc 3 is classified into group B. For this approach, relations between the query and each documents are not explicit.

We propose to classify and organize information based on semantic or logical relations. Figure 1.4 shows our approach. First, relevant documents are retrieved same as the previous approach. Then, for each documents, semantic or logical relations of the query are recognized. However, because a relation between a query (single sentence) and a document (multiple sentences) is difficult to recognize, a relation between the query and a sentence in the document is recognized instead.

Note that, if we want to classify the documents, it is achieved by grouping the relations. Hence, the proposed approach is more general than the previous approach. For

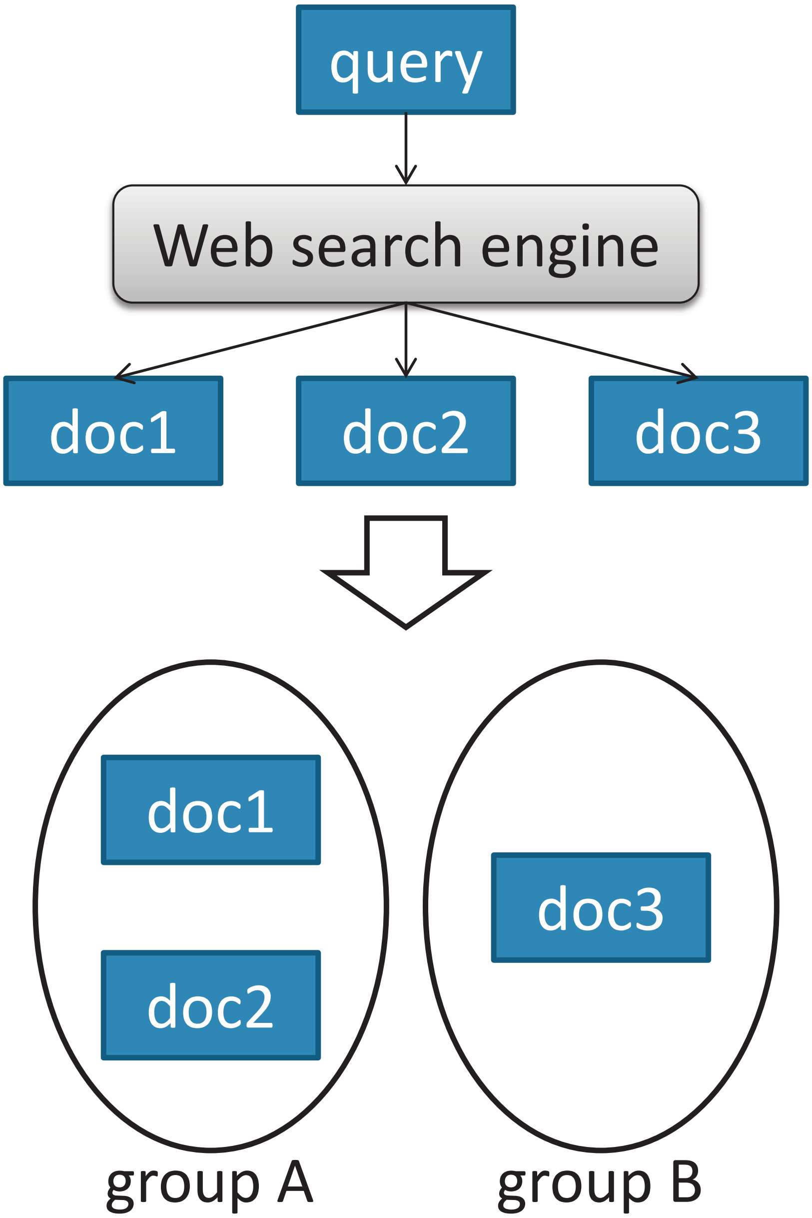

Figure 1.3: Information organization based on classification

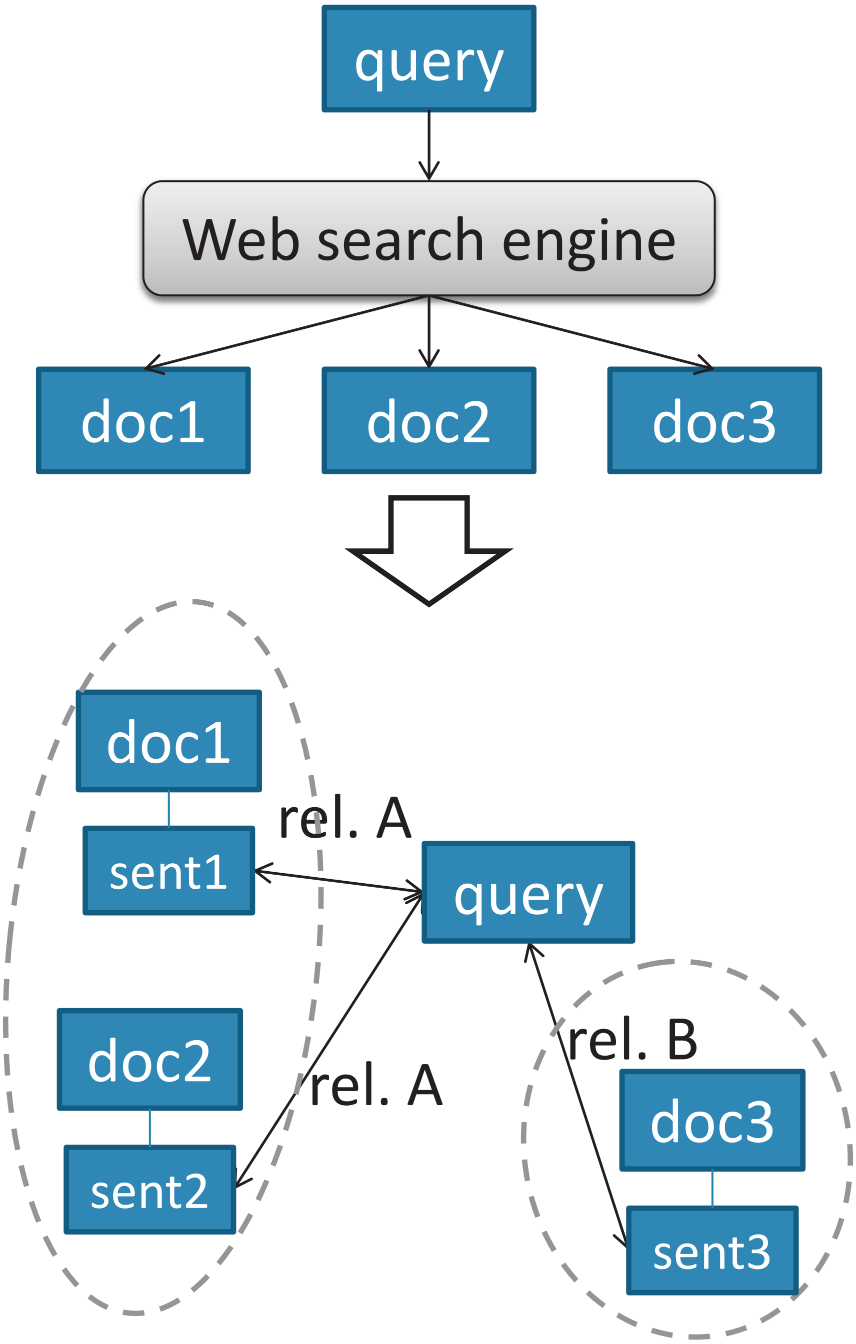

Figure 1.4: Information organization based on relation

example, in order to classify sentences by sentiment polarity, sentences which agree with a query are classified into a group whose sentiment polarity is same as the query, and sentences conflict to the query are classified into another group.

In this work, we first define the task, information organization based on semantic relation recognition for the Web. Because this is a new task, we need real data to help us identify the technical issues to overcome and the way to define the target semantic relations. Therefore, we construct a development data set and analyze them. From the analysis, the semantic relations are defined and a classification model is constructed. The classification model relaxes a set of restrictions on the structural alignments between a query and a sentence. In evaluation, we confirm that the most important technical issue for semantic relation recognition targeting Web texts is correctly detecting the region in retrieved Web texts that corresponds to the content of the user query. Finally, we construct a Web application. In this application, the user can input a sentence as a query and the system classifies sentences relevant to the query based on semantic relation between them. In user evaluation, we discuss an effectiveness of information organization by comparing to existing Web search engine.

1.1 Dissertation Outline

The rest of this dissertation is organized as follows.

In chapter 2, we propose an approach of organizing information. In the approach, the most important issue is to recognize semantic relation between sentences. Thus, we define the task, especially, a unit to organize, the set of relations.

In chapter 3, we describe how to generate STATEMENT MAP which organizes sentences relevant to user's query. We proposed the model to recognize semantic relation between sentences on the Web.

In chapter 4, we demonstrate a Web application to organize sentences relevant to user's query and evaluate the application by users.

Chapter 2

Organizing Information

In this chapter, we describe an approach to organize information based on semantic relation.

2.1 Previous Work

Many previous work indicates that information organization is required for the Web. The Semantic Web is a concept to extend the World Wide Web that enables users to share content beyond the boundaries of applications and websites. A main approach of the Semantic Web is to add machine-readable meta-data written in RDF, OWL, or XML to traditional Website written in HTML. The meta-data shows various semantic information such as a meaning of word, keywords of the Website and so on. The meta-data is useful for machine reading on the Web, however, it does not help directly to organize the number of documents, and more

2.1.1 WISDOM (Akamine et al. 2009)

The WISDOM project [11] focuses on evaluating credibility through identifying the source of information and classifying opinions by sentiment analysis. They targeted sentiment polarity to classify sentences relevant to users' query. Pang and Lee [12] also classifies opinions by sentiment analysis

However, their approach cannot be applied to a sentence which indicates fact but does not have sentiment polarity. For example, given the query sentence ⟨南極の氷は溶けている *Nankyoku-no-koori-ha-toketeiru* “The ice on the south pole is melting”⟩, the statement ⟨ 南極の氷は増えている *Nankyoku-no-koori-ha-hueteiru* "The ice on the south pole is increasing"⟩ is contradictory to the query but both do not have sentiment polarity. In order to classify such sentences, a relation between sentences is required to consider.

2.1.2 Concept Map (Novak 1990)

Concept Map developed by Novak [13] is an approach to organize concept such as knowledge or idea. It organizes concepts based on a relation between them. Figure 2.1 shows an example of concept map of Novak's ideas on the nature of Concept Maps [14]. For example, the relation of "Concept Maps" to "Organized Knowledge" is "represent", which shows that "Concept Maps - represent - Organized Knowledge". Concept Maps are quite widely known and used in an educational field. It shows that the importance and effectiveness to organize information based on a relation between them.

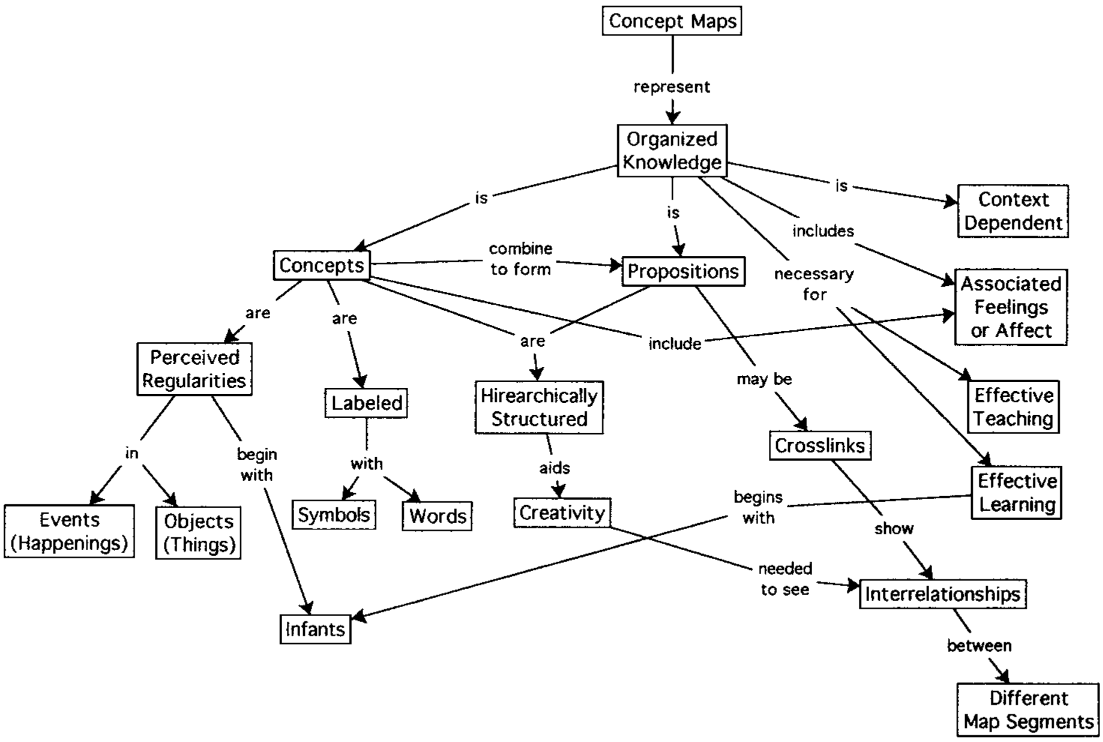

Figure 2.1: Example Concept Map (Novak, 2002)

Our proposed approach is similar to Concept Map at a fundamental level. But the most important difference between our approach and Concept Maps is the unit to organize, where Concept Map organize concepts and our approach targets sentences. When a unit to organize is a concept, a technology of information extraction and relation extraction in a natural language processing filed can be applied. However, when a unit to organize is a sentence, recognizing a relation between them is quite difficult. There is one further difference that we must not ignore. It is that we must recognize a relation between sentences automatically. In Concept Maps, because its purpose is to organize users' idea or knowledge, it is not considered to recognize relations automatically.

2.1.3 CST (Radev 2000) and RST (Mann and Thompson 1998)

Cross-document Structure Theory (CST), developed by Radev [15], is a task of recognizing semantic relations between sentences. CST is an expanded rhetorical structure analysis based on Rhetorical Structure Theory (RST: Mann and Thompson [16]), and attempts to describe the semantic relations that exist between two or more sentences from different source documents that are related to the same topic, as well as those that come from a single source document. A corpus of cross-document sentences annotated with CST relations has also been constructed (The CSTBank Corpus: Radev et al. [17]). CSTBank is organized into clusters of topically-related articles. There are 18 kinds of semantic relations in this corpus, not limited to ⟨EQUIVALENCE⟩ or ⟨CONTRADICTION⟩, but also including ⟨JUDGMENT⟩, ⟨ELABORATION⟩, and ⟨REFINEMENT⟩. Etoh et al. [18]) constructed a Japanese Cross-document Relation Corpus, and they redefined 14 kinds of semantic relations in their corpus.

CST was designed for objective expressions because its target data is newspaper articles related to the same topic. Facts, which can be extracted from newspaper articles, have been used in conventional NLP research, such as Information Extraction or Factoid Question Answering. However, there are a lot of opinions on the Web, and it is important to survey opinions in addition to facts to give Internet users a comprehensive view of the discussions on topics of interest.

2.1.4 Recognizing Textual Entailment (Dagan, Glickman and Magnini 2005)

Identifying logical relations between texts is the focus of Recognizing Textual Entailment, the task of deciding whether the meaning of one text is entailed from another text. A major task in the RTE Challenge (Recognizing Textual Entailment Challenge) [19] is classifying the semantic relation between a Text (T) and a Hypothesis (H) into ⟨ENTAILMENT⟩, ⟨CONTRADICTION⟩, or ⟨OTHER⟩. Over the last several years, several corpora annotated with thousands of (T,H) pairs have been constructed for this task. In these corpora, each pair is tagged indicating its related task (e.g. Information Extraction, Question Answering, Information Retrieval or Summarization).

The RTE Challenge has successfully employed a variety of techniques in order to recognize instances of textual entailment, including methods based on: measuring the degree of lexical overlap between bag of words [20, 21], the alignment of graphs created from syntactic or semantic dependencies [22, 23], statistical classifiers which leverage a wide range of features [24], or reference rule generation [25]. These approaches have shown great promise in RTE for entailment pairs in the corpus, but more robust models of recognizing logical relations are still desirable. The definition of contradiction in RTE is that T contradicts H if it is very unlikely that both T and H can be true at the same time. However, in real documents on the Web, there are many pairs of examples which are contradictory in part, or where one statement confines the applicability of another.

2.2 Statement Map

In this section, we describe how to organize information based on semantic relation.

We call this information organization as STATEMENT MAP generation. STATEMENT MAP means that the system maps sentences (statements) relevant to a query based on semantic relation.

When given a sentence and another sentence, and they share the same topic, there is some semantic relation such as entailment or contradiction. And given a number of such sentence pairs, the ultimate goal of the research is to connect them with an edge labeled with the semantic relation. However, it is difficult to recognize semantic relation between real sentences, so that we assume one sentence as a query which is short simple sentence, as described below. In any case, recognizing semantic relation between sentences is the most important technology in our research.

Recently, recognizing textual entailment is one of the famous tasks in natural language processing area. The task stands a part of semantic relation recognition task. In the third PASCAL RTE Challenge [26], most of the systems obtained a score in between 59% and 66%, especially, Hickl and Bensley achieved 80% accuracy [27]. Same as the RTE3, in the RTE-4 [28], the highest accuracy is 68.50% [29], and in the RTE-5 [30], the highest accuracy is 68.33% [31]. In spite of high accuracy of the RTE Challenge, as De Marneffe, Rafferty, and Manning (2008) [32] found in their RTE experiments with Web texts, real world tests are more difficult to classify. They reported as follows:

> In a real world setting, it is likely that the contradiction rate is extremely low; rather than overwhelming true positives with false positives, rendering the system impractical, we mark contradictions conservatively.

Therefore, to generate STATEMENT MAP, we first assume *query* as a core sentence of information organization where other sentences which are relevant to the *query* are retrieved. Then, a semantic relation is recognized between the *query* and the retrieved sentence. Finally, the sentences are grouped along with the semantic relation and shown users a bird's-eye view.

In these three phases, relevant sentences are retrieved with an existing information retrieval engine, and it is feasible to generate a simple bird's-eye view. Therefore, the most important technical issue is a semantic relation recognition between a query and a relevant sentence.

2.2.1 Query and Relevant Sentence

As de Marneffe, Rafferty, and Manning (2008) [32] reported the difficulty of recognizing CONTRADICTION between real sentences, in order to analyze the difficulty, we first create some queries and retrieve relevant sentences.

The queries are created by an annotator who does not concern the development of STATEMENT MAP generation system. A criterion of constructing query is described as following.

- query is a simple sentence in order to avoid the difficulty of semantic relation recognition between real sentences like Wang, Zhang and Neumann [33]

- when considering query as interrogative, we can answer it by Yes or No, in other words, agreeing opinion says Yes to the query and conflicting opinion says No to the query
- there are agreeing and conflicting opinions for the query in the Web
- there are scientific evidences for both agreeing and conflicting opinions, in other words, queries which do not supported scientifically such as superstition are excluded
- the truth of query is not cared

Relevant sentences are retrieved by the passage retrieval engine as described in Section 3.3.1. The goal of RTE is to recognize logical and factual relations between sentences in a pair, while CST targets objective expressions because newspaper articles related to the same topic are utilized. Facts, which can be extracted from newspaper articles, have been used in conventional NLP research, such as Information Extraction or Factoid Question Answering. However, there are a lot of opinions on the Web, and it is important to fully survey the opinions related to a user's topic of interest to generate a STATEMENT MAP. The task specifications of both RTE and CST do not cover opinions and their relations as illustrated below.

(1) a. There is absolutely no connection between vaccines and autism.

b. I do believe that there is a correlation between vaccinations and autism.

Subjective statements, such as opinions, have recently been the focus of various NLP research, such as review analysis, opinion extraction, opinion QA, and sentiment analysis. In the corpus conducted by the MPQA Project (Multi-Perspective Question Answering) [34], individual expressions corresponding to explicit mentions of private states, speech events, and expressive subjective elements are tagged.

2.2.2 Entailment and Agreement

In information organization task on the Web, because not every piece of information can be covered by ENTAILMENT and CONFLICT, we need to adopt a broader set of semantic relations tailored to handle statements which can be either facts or opinions.

Given the query ⟨ うがいは風邪を予防する *ugai-ha-kaze-wo-yobousuru* "Gargling prevents colds"⟩, the left side of the Figure 2.2 shows sentences which entail the query and the right side shows sentences which agree the query. For example, the upper-right sentence indicating that "gargling" is partially effective does not entail the query, but does not contradict at the same time. Therefore, we adopt AGREEMENT which is broader relation than entailment. We describe the definition in Section 2.2.3. The figure indicates that when considering only entailment, only 22% of relevant sentences can be captured.

2.2.3 Semantic Relation

In this section, we define the semantic relations that we will classify in Japanese Internet texts. Our goal is to define semantic relations that are applicable over both fact and opinions, making them more appropriate for handling Internet texts. See Table 2.1 for real examples. Hereafter, we use the term "*Q*" to refer to a query sentence and "*T*" to refer to a sentence relevant to the query.

[AGREEMENT] While ENTAILMENT is an asymmetric relation, AGREEMENT is a symmetric relation where a query and a relevant sentence have equivalent semantic content. For example, reverse ENTAILMENT, which is considered as OTHER in the RTE, is considered as AGREEMENT. Only fact is considered in ENTAILMENT, both fact and opinions are considered in AGREEMENT. The following is an example of AGREEMENT which the relevant sentence is not fact but just opinion.

(2) *Q* Gargling prevents colds.

T I think gargling can prevent colds.

Even if a relevant sentence is more informative than a query, when the region in the relevant sentence corresponds to the query agrees to a content of the query, their semantic relation is regarded as AGREEMENT. In the following example, *T* is more informative than *Q*.

(3) *Q* Bio-ethanol is good for the environment.

T Bio-ethanol is a high-quality fuel, and it has the power to deal with the environment problems that we are facing.

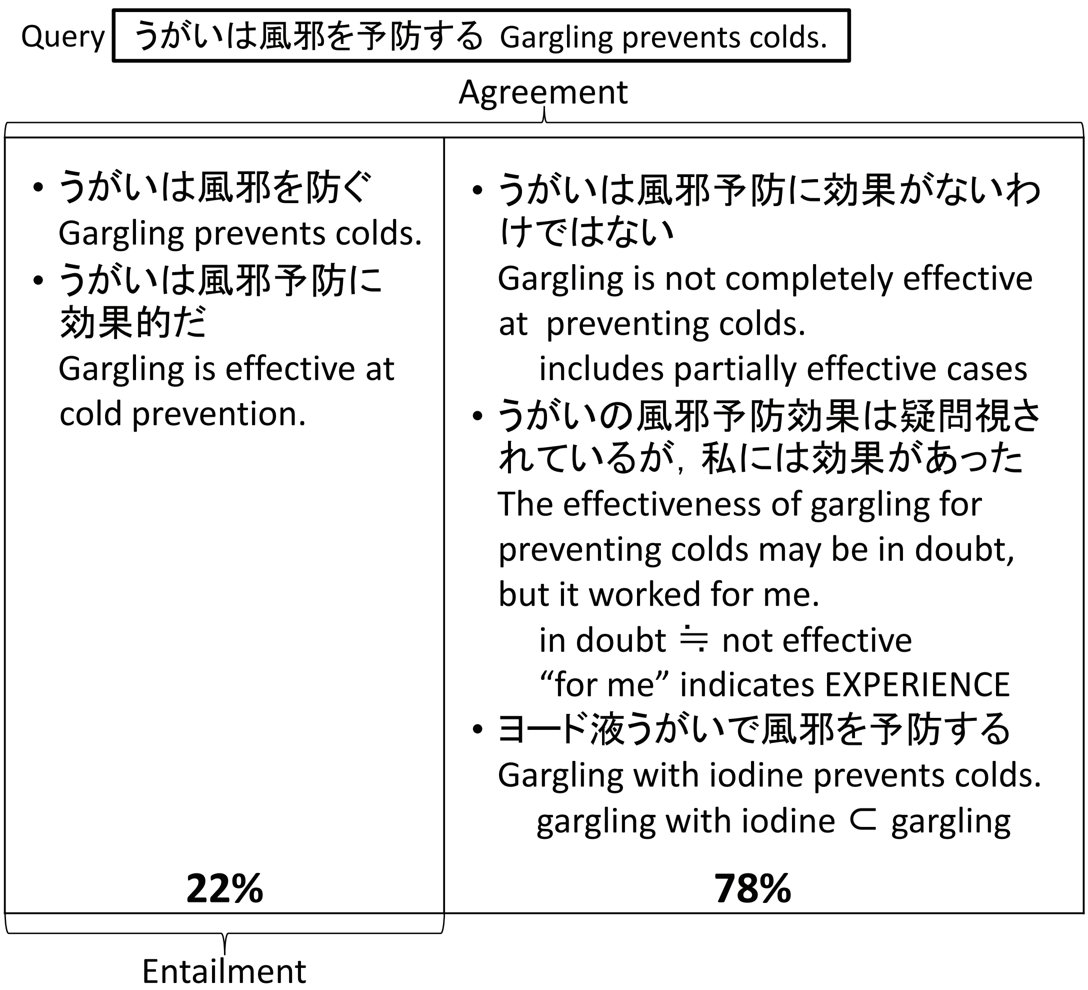

Figure 2.2: Entailment and Agreement

Given the query “Gargling prevents colds.”, the left side shows sentences which entail the query and the right side shows sentences which do not entail but agree the query. The numbers of each groups indicate the proportion of each relations in the development data set.

Table 2.1: Example semantic relation classification

Query	Relevant sentence	Semantic relation
キシリトールは虫歯予防に効果がある	キシリトールの含まれている量が多いほどむし歯予防の効果は高いようです	同意
	The cavity-prevention effects are greater the more Xylitol is included.	AGREEMENT
	キシリトールがお口の健康維持や虫歯予防にも効果を発揮します	同意
Xylitol is effective at preventing cavities.	Xylitol shows effectiveness at maintaining good oral hygiene and preventing cavities.	AGREEMENT
	キシリトールの虫歯抑制効果についてはいろいろな意見がありますが実際は効果があるわけではありません	対立
	There are many opinions about the cavity-prevention effectiveness of Xylitol, but it is not really effective.	CONFLICT
還元水は健康に良い	弱アルカリ性のアルカリイオン還元水があなたと家族の健康を支えます	同意
	Reduced water, which has weak alkaline ions, supports the health of you and your family.	AGREEMENT
	還元水は活性酸素を除去すると言われ健康を維持してくれる働きをもたらす	同意
Reduced water is good for the health.	Reduced water is said to remove active oxygen from the body, making it effective at promoting good health.	AGREEMENT
	美味しくても酸化させる水は健康には役立ちません	対立
	Even if oxidized water tastes good, it does not help one's health.	CONFLICT
バイオエタノールは環境に良い	バイオエタノールは高い可能性を秘めた高品質の燃料であり現在我々が直面する環境問題に対処し得る潜在能力を持っている	同意
Bio-ethanol is good for the environment.	Bio-ethanol is a high-quality fuel with untold potential, and it has the power to deal with the environment problems that we are facing.	AGREEMENT

Given the query "Mineral water is good for health" which asks a product is whether good or not, when a relevant sentence compares with another product, it may not always show which product is better. In the following example, relevant sentence shows only healthiness by comparing with "tap water". There is a possible that neither "mineral water" nor "tap water" are "good for health". Although there is just comparison, the semantic relation is regarded as AGREEMENT.

(4) *Q* Mineral water is good for health.

T Mineral water is healthy than tap water.

In order to capture opinions, detecting the modality of a predicate is important. Only existence or non-existence are required for capturing facts. However, when considering opinions, uncertain modality is needed. Saurí and Pustejovsky [35] reported that the modality axis distinguishes among *certain*, *probable*, *possible*, and *underspecified* for English. In Japanese, there are various expressions which indicate the modality. Matsuyoshi et al. [36] defined nine steps of modality (actuality). We use the term "modality" to refer to mean "modality" that Saurí and Pustejovsky defined and "actuality" that Matsuyoshi et al. defined.

In the following example, the modality of the predicate *prevents* in *Q* is "certain" while the modality of the predicate *prevent* in *T* is "possible". In spite of the difference in modality, the semantic relation is recognized as AGREEMENT.

(5) *Q* Gargling prevents colds.

T Gargling may prevent colds.

[CONFLICT] A symmetric relation where a relevant sentence has negative or contradicting semantic content to a query. This can range from strict logical contradiction to opposite polarity of opinions. The following pair is a CONFLICT example.

(6) *Q* Bio-ethanol is good for our earth.

T There is a fact that bio-ethanol further the destruction of the environment.

Same as AGREEMENT, we capture not only facts but also opinions.

(7) *Q* Collagen is good for the skin.

T Speaking as a researcher, just because collagen enters the body does not mean it will moisturize the skin.

[EVIDENCE] An asymmetric relation where a relevant sentence provides justification or supporting evidence for a query. The following is a typical example. The relevant sentence agrees the query. EVIDENCE is marked by underline. It provides factual information supporting the query.

(8) *Q* Xylitol is effective at preventing cavities.

T Xylitol is effective at preventing cavities <u>because the cavity-causing ba cteria streptococcus mutans cannot metabolize it</u>.

The following is another example. The relevant sentence agrees the query. EVIDENCE provides explanatory information supporting the query.

(9) *Q* Gargling prevents colds.

T The effectiveness of gargling for preventing colds is in truth, <u>because it worked for me</u>.

2.2.4 Mapping Arguments

Our goal is to help users come to a deep understanding about a user's query (e.g. "Do vaccines cause autism?"). To do this, we propose an approach to information organization based on relation between a user's query and relevant sentences. It leads to present the users with a comprehensive survey of a user's query. To be concrete, the system classifies relevant sentences as agreeing or conflicting with a user's query.

Figure 2.3 shows the results of an information organization about the example query. The group in the upper-left is labeled AGREEMENT, and it contains statements that are closest to the user's query. In this case these are opinions that support a causal link between vaccines and autism. An example is the claim "Mercury-based vaccine preservatives actually have caused autism."

The group in the upper-right is labeled CONFLICT, and it contains statements that are in opposition to the statements of focus. This includes the counter-claim "There is no valid scientific evidence that vaccines cause autism."

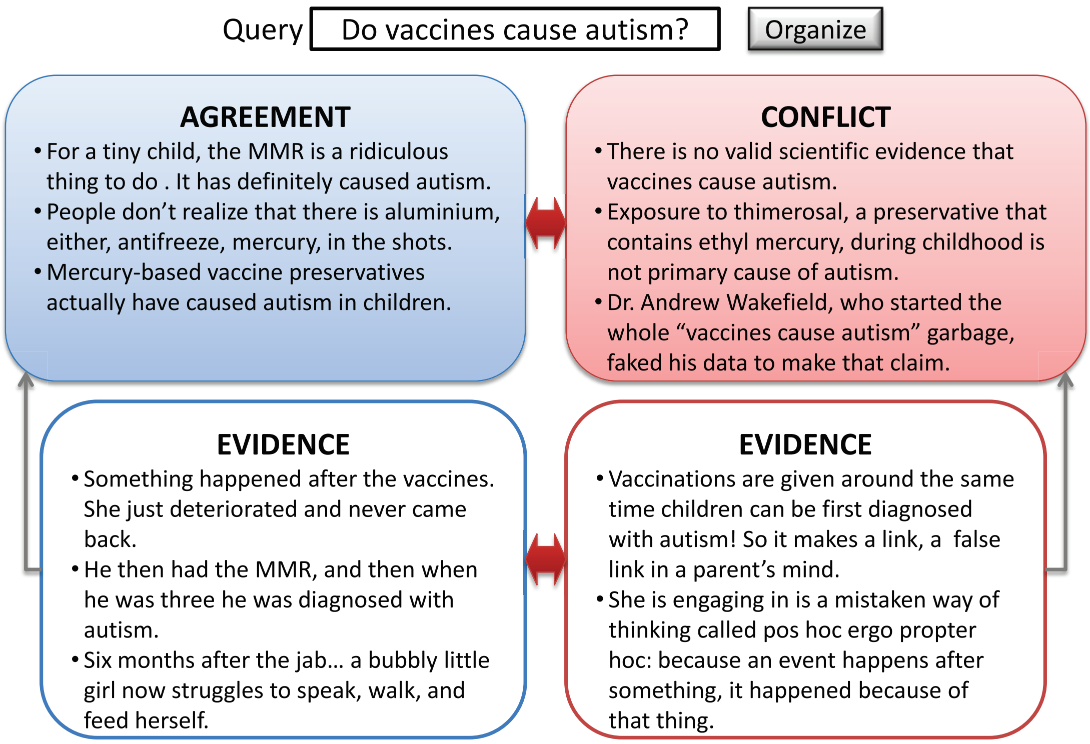

Figure 2.3: An example STATEMENT MAP for the query "Do vaccines cause autism?"

The thick, red, bi-directional arrows connecting the AGREEMENT and CONFLICT groups help that opposition in opinion stand out to the user. It is clear that these are strongly opposing opinions. The groups labeled EVIDENCE at the bottom of the figure contain supporting evidence for the AGREEMENT statements and CONFLICT statements. They are linked by thin, gray, mono-directional arrows.

Ultimately it will be up to him or her to weight the anecdotal evidence of the anti-vaxxers against the medical evidence and logical arguments of the scientific community, but by providing all of the information to the user in a way that makes it easy to see the support or lack thereof for each viewpoint, the STATEMENT MAP helps the user come to an informed conclusion.

Recognition ENTAILMENT and CONTRADICTION may be sufficient to generate STATEMENT MAPS if it were limited to factual statements, however, we also need to determine the source of experiments, opinions and their attitude in order to recognize semantic relations when dealing with opinions. We need to recognize not only logical relations, such as ENTAILMENT and CONFLICT, but also several expanded relations in order to handle opinions. The semantic relations to capture for STATEMENT MAP generation are described in Section 2.2.3. To recognize attitudinal relations, we need to combine RTE methodologies, attribution analysis for capturing the source of an opinion, and sentiment analysis to recognize the semantic relations.

2.3 Data Construction

In English, there are some corpora developed for the RTE Challenge. The corpora consist of pairs of hypothesis and text labeled whether entail or not. Also, alignment information for the RTE2 is annotated by Brockett [37]. In Japanese, Shima et al. [38] and Odani et al. [39] developed Japanese RTE development data. However, there are not currently any corpora that focus on semantic relations between both facts and opinions, and there are many challenges in constructing such a corpus. In this section, we describe the specification of the corpus we are constructing and our method of collecting samples from the Web.

2.3.1 Constructing a Japanese Corpus

Real data on the Web generally has complex sentence structures. That makes it difficult to recognize semantic relations between full sentences. For example, the following two sentences cannot be annotated with any of the semantic relations.

(10) A According to Department of Medicine, there is no link between the MMR vaccine and autism.

B The weight of the evidence indicates that vaccines are not associated with autism.

Thus, in this work, we apply two conditions to resolve difficulties.

First, query must be a simple clause: which contains just one predicate and some arguments. For example, *Mineral water is safer than tap water*（ミネラルウォーターは水道水より安全だ）, *Global warming causes rising sea levels*（地球温暖化によって海面が上昇する） (predicates are underlined). Note that, nominal predicate is not count as predicates but arguments. Then, to recognize semantic relation between a query and a relevant sentence, we just focus on semantically corresponding portions of the query and relevant sentence.

Second, relevant sentence retrieved by passage retrieval engine must share nouns with the query except nominal predicates and functional words. For example, for the query *Xylitol is effective at preventing cavities*（キシリトールは虫歯予防に効果的だ）, *xylitol* and *cavities* must be contained in retrieved sentences. Because *preventing* is nominal predicate, it must not be contained in retrieved sentences. On the other hand, no limitation is imposed for predicate in query sentence. Therefore, query sentence and retrieved sentences are relevant and various predicates are contained in retrieved sentences.

We constructed two data set: development data set and evaluation data set. Table 2.2 shows size of data set. How to generate query is described in Section 2.2.1 and how to retrieve sentences is described in Section 3.3.1.

Table 2.2: Development data and Evaluation data

	Number of queries	Number of retrieved sentences
development data	5	1,324
evaluation data	20	1,467

A part of 20 query data set is shown in Table 2.3. The queries involved in the 20 query evaluation data set had no overlap with those in the 5 query development data set. The data set is available on http://www.cl.ecei.tohoku.ac.jp/stmap/ as "STATEMENT MAP Corpus 1.0".

We explain how to construct the data set. First, 25 queries are constructed by an annotator. All queries are simple sentences described in Section 2.2.1. All queries are generated by a single annotator who does not concerned the development of STATEMENT MAP generation system.

For each queries, in order to retrieve relevant sentences on the Web, some query words are selected automatically from the query sentence. A criterion of selecting query words are described as follows.

- last phrase is excluded
- all functional words are excluded
- when a phrase contains a pattern of "Noun + Noun (nominal predicate)", behind word is excluded

- query words are remained words

Then relevant sentences are retrieved by TSUBAKI[1] [40], the information retrieval engine described in Section 3.3.1.

Semantic relation annotation was carried out by two native speakers of Japanese with an inter-annoatator agreement kappa score of 0.72.

[1]http://tsubaki.ixnlp.nii.ac.jp/

Table 2.3: Example evaluation data

Query (Query Words)	Semantic relation	Relevant sentence
リサイクルは環境に良い Recycling is good for the environment （リサイクル，環境）(recycling, environment)	同意	かけがえのない地球の環境をより良くしていく為に、全社員が環境に優しい商品づくり、リサイクル商品づくりとその技術開発に取り組んでいます
	AGREEMENT	In order to make the environment better, the whole staff works hard to create environmentally-friendly and recyclable products and related technological developments
	対立	また、ほぼ全てのリサイクルは環境によいわけではなく、「リサイクル＝環境に優しい」とは限らない
	CONFLICT	In addition, not all recycables are good for the environment , so it is not necessarily true that “recycle = good for the environment”
温暖化によって海面が上昇する Global warming causes rising sea levels （温暖化，海面）(global warming, sea levels)	同意	地球温暖化で進む海面上昇で、深刻な被害が出始めている
	AGREEMENT	Rising sea levels from global warming is starting to cause serious damage
	根拠	温暖化が原因で南極の氷が溶けて海面が上昇するなんてことも聞きますが、海面が上昇するのは、温暖化で温度が上がることで海の温度も上がって、水が膨張するのが主な原因です
	EVIDENCE	I often hear that melting of polar ice caps causes rising sea levels, however, the main reason of sea levels rising is water expansion due to rising sea tempratures, and the reason for the rising sea tempratures is global warming

ミネラルウォーターは水道水より安全だ Mineral water is safer than tap water （ミネラルウォーター，水道水） (mineral water, tap water)	同意 AGREEMENT	その当時から、外国に行くと水道の水は飲んではダメだと言われ、ホテルに泊まるとほとんどの国でミネラルウォーターのペットボトルが飲料水用に置いてあります It is often said that tap water should not be drunk in foreign countries, so, hotels in most countries provide bottles of mineral water for drinking
	対立 CONFLICT	東南アジアでは、半透明なボトルで売られている安いミネラルウォーターがありますが、これは薬品を混入させているらしく下痢を誘発する事もあります A semi-transparent bottle of mineral water is sold in south east Asia, however it may cause diarrhea
納豆ダイエットは効果がある Fermented soybeans is effective for diet （納豆，ダイエット） (fermented soybeans, diet)	同意 AGREEMENT	納豆のダイエット効果をウソだと思う人もいるかもしれませんが、ダイエット効果のある栄養素が、納豆に含まれてることは間違いありません Someone may think that the effectness of fermented soybeans diet is a lie, however, it is fact that fermented soybeans contains nutrients which are effective for dieting
血液型で性格が分かる Blood type predicts personality type （血液型，性格） (blood type, personality type)	同意 AGREEMENT	「私はA型人間です」とか「やっぱりB型の性格が出ているわね」などと、血液型によって性格を判断することがよくあります I often heard that personality type is predicted by blood type such as “I have a personality of blood type A” or “You have a personality of blood type B, don’t you?”
	根拠 EVIDENCE	自分や周りの人間の性格をみて、血液型性格学の記述があたってる、と思うこと多々あるかもしれませんが、これは心理学的トリックによるものです You may think that blood type-based characterology is true from considering the character of yourself or those around you, however, this effectiveness is really a psychological trick

Chapter 3

Statement Map Generation

In this chapter, we describe the structure of STATEMENT MAP generation system.

3.1 System Introduction

The first point to be discussed is the definition of the task to generate STATEMENT MAP.

First, it is necessary to be clear what sentences are used to classify semantic relation. While in RTE, entailment is judged between hypothesis and text, query sentence and retrieved sentence are used to classify semantic relation in STATEMENT MAP generation. The query sentence is that user input to analyze its credibility. In general, "query" is ambiguous between "sentence" and "words". However, because in STATEMENT MAP generation, "query words" are automatically selected from "query sentence" in passage retrieval phase, we discriminate them. The retrieved sentence is one of retrieved sentences in passage retrieval phrase.

Second, in order to resolve difficulties of analysis such as anaphora resolution and coreference resolution, we place a restriction that a semantic relation is recognized from only query sentence and retrieved sentence. In other words, we do not use other information such as around texts of retrieved sentence in retrieved document.

Third, in STATEMENT MAP generation, the query sentence and retrieved sentence are classified into five relations: AGREEMENT, CONFLICT, EVIDENCE and OTHER. Each relations are defined in Section 2.2.3. In RTE, the hypothesis and text are classified into three relations; ENTAILMENT, CONTRADICTION and OTHER.

STATEMENT MAP generation is the task to classify a pair of the query and a retrieved sentence into five relations. The retrieved sentence is relevant to the query sentence and gotten by passage retrieval engine described in Section 3.3.1. As our research purpose is to analyze the problem, three policies as following are premised.

1. A deep semantic analysis such as Predicate-Argument Structure Analysis and Modality Analysis are used. While the performance of such analyses is not enough, we employed the state-of-the-art systems.
2. A large scale lexical knowledge is absolutely essential to recognize the relation between words. We used existing resources aggressively.
3. A various problems are contained in this task [41]. It is important to be clear the problems, we employed the rule-based method which relaxed some constraints one by one. And we do not use machine learning technique easily because of the difficulty to analyze the machine learned models. The rule-based method is adapted to analysis of the problem.

We describe the method to realize the task in following sections.

3.2 Related Work

3.2.1 Recognizing Textual Entailment

There is some overlap between recognizing semantic relation of two sentences and recognizing textual entailment. In recognizing textual entailment task, ⟨ENTAILMENT⟩ is contained in ⟨AGREEMENT⟩, and ⟨CONTRADICTION⟩ is contained in ⟨CONFLICT⟩. Although, there are many researches of recognizing textual entailment, the performance is not achieved well.

3.2.2 Recognizing Contradiction and Conflict

Particularly, ⟨CONTRADICTION⟩ is a difficult relation to recognize. De Marneffe et al. (2008) [32] reported 23% precision and 19% recall at detecting contradictions in the RTE-3 data set [42]. However, the RTE-3 data set was constructed artificially and do not reflect real contradictions. Therefore, they collected contradictions in 'a real world'. The corpus contains 131 pairs: 19 from news wire, 51 from Wikipedia, 10 from the Lexis Nexis database, and 51 from the data prepared by LDC[1]. Their system was evaluated with the data set, however, the performance is limited.

De Marneffe, Rafferty and Manning (2011) [43] classified contradiction types in two groups, seven categories: Antonym, Negation, Numeric, Factive/Modal, Structure, Lexical, and World Knowledge. They defined feature sets according to the classification. They reported that lexical information and world knowledge are difficult to generalize as features.

Ritter et al. [44] focused on the functionality of a relation. They addressed to recognize contradictions between tuples that represent the entities in the sentences and the relationships between them (e.g., *was_born_in(Mozart, Salzburg)*). For example, *was_born_in(Mozart, Salzburg)* and *was_born_in(Mozart, Vienna)* are contradictory, and *visited(Mozart, Salzburg)* and *visited(Mozart, Vienna)* are not. They detected the contradiction by functionality of *was_born_in(x, y))*. They founded that genuine contradictions are quite rare.

3.2.3 Recognizing Evidence

In addition, only the overlap of hypothesis and text is focused in recognizing textual entailment. In other words, additional information in text are not focused. However, we focused on the additional information in retrieved sentence especially which express an evidence for formulation of query (hypothesis in RTE). Such sentence pair is classified as ⟨EVIDENCE⟩ which has never considered in RTE. Because retrieved sentence classified into ⟨EVIDENCE⟩ has useful information for user which are not contained in traditional ⟨AGREEMENT⟩ and ⟨CONFLICT⟩ sentence pair, it is important to recognize ⟨EVIDENCE⟩ and present for user.

Detection of evidence related to a user query is similar in nature to Why-QA [45, 46, 47]. The goal of Why-QA is to detect questions in *why* form and retrieve answers explaining their cause or reason. Several approaches have been proposed: from pattern-based [46], to discourse-oriented [48] and, more recently, machine-learning [47, 48]

Evidence Search shares its basic evidence detection architecture with these approaches, almost all of which simplify down to *detect semantically similar passage* and *identify explicit causal cue*. However, it differs in its approach toward the query and extracted text. In most Why-QA systems, a lot of processing is done to determine the question type and alter its search strategy accordingly. However, not much attention is paid to

[1] http://www-nlp.stanford.edu/projects/contradiction/

the contents of the answers extracted. In contrast, with Evidence Search we want to be able to distinguish between *kinds* of evidence so that their quality can be assessed.

3.3 System Architecture

The process to generate STATEMENT MAP is presented in Figure 3.1. As mentioned in Section 3.1, short length sentence written in natural language (not query words) is assumed as input of STATEMENT MAP. Then many sentences which are relevant with the input sentence are retrieved in passage retrieval phase. Finally, the system recognizes a semantic relation between query sentence and each retrieved sentences and shows users a bird's-eye view.

The STATEMENT MAP screenshot is presented in Figure 3.2. AGREEMENT are shown in a red column on the left and CONFLICT in a blue column on the right. Meta-information EVIDENCE and CONFINEMENT[2] found are displayed below the corresponding viewpoint with the important regions highlighted. A link to the page where it was found is given below each statement, and summaries of the total number of statements found for each viewpoint and type of meta-information are shown to make the amount of support clear at a glance. The order statements appear in is determined by the confidence score given by the semantic relation classifier.

As passage retrieval is quite easy to realize by using existing sentence retrieval engine, the most important problem of our research is how to recognize semantic relation between two sentences.

An approach to recognize semantic relation is similar to conventional recognizing textual entailment. They are classified into two groups: transformation-based approach and alignment-based approach [50]. In transformation-based approach, ENTAILMENT is judged by whether *text* is converted to *hypothesis*. When *text* cannot be converted to *hypothesis*, the relation is judged as non-entailment. The convertions are exchange of the passive and active, exchange of the hypernym and hyponym and so on. As an objective of transformation-based approach is to recognize ENTAILMENT or not, it is not suitable to recognize various relations. Therefore, we adopt an alignment-based approach. The alignment-based approach analyzed in three steps: 1) query sentence and retrieved sentence are deep semantic parsed, 2) when the words are identical or semantically similar then they are aligned, 3) by using information of 1) and 2).

[2]detail of CONFINEMENT recognition is reported in Ohki et al. [49].

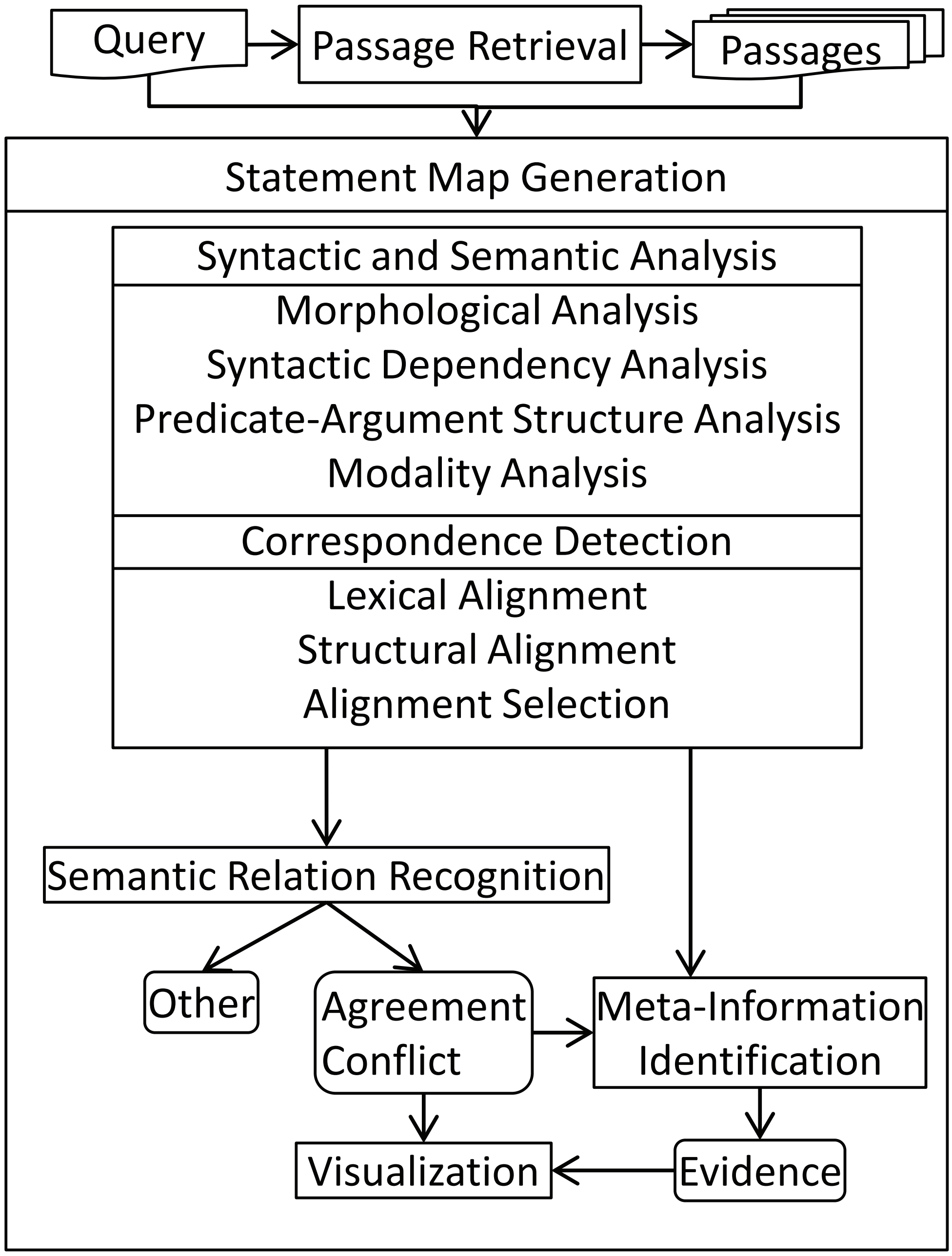

Figure 3.1: STATEMENT MAP Generation System

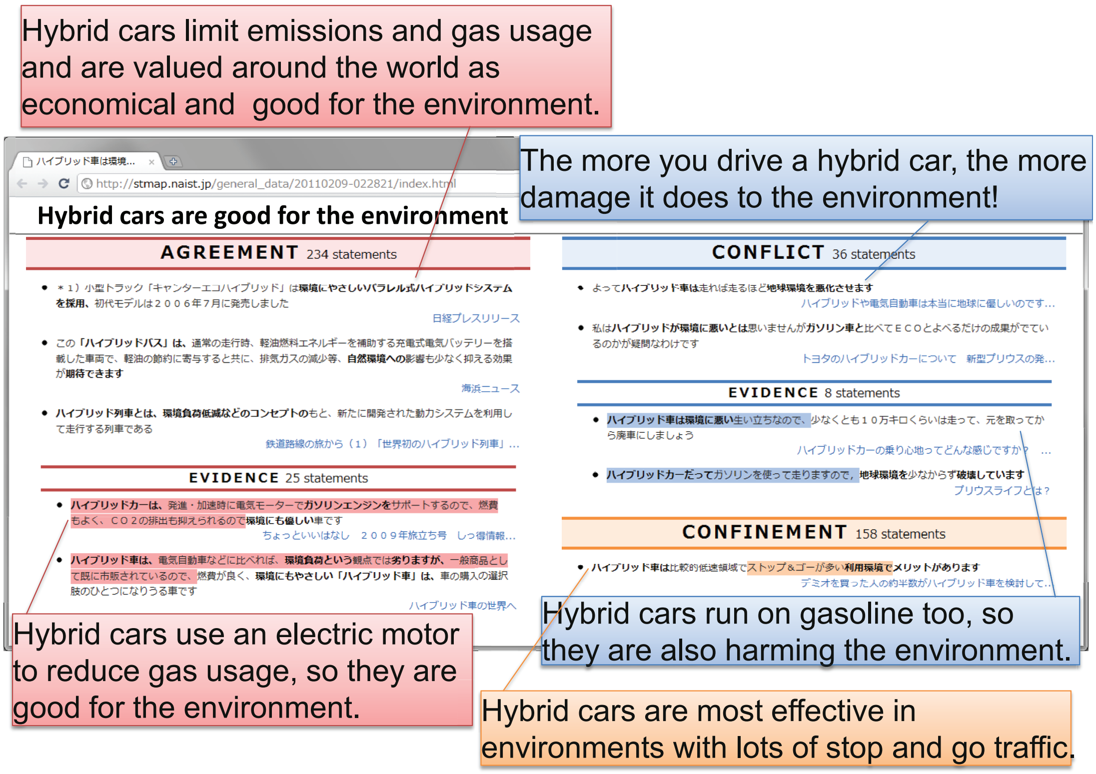

Figure 3.2: STATEMENT MAP Screenshot

Both *hypothesis* and *text* are some length sentence in previous recognizing textual entailment task such as TAC RTE [51]. Therefore, as not only predicate-argument structure analysis, but also high level analysis such as coreference resolution and anaphora resolution are required, the problem tends to be more difficult and diverged. In this work, as described in Section 2.3.1, we have two conditions to resolve difficulties: 1) a query must be simple clause, and 2) retrieved sentence must share content words with the query.

A noun alignment and a predicate alignment are considered as the same problem in previous work [52]. However, while the main problem of noun alignment is whether the knowledge base contained the relation between the words, the knowledge base coverage is not the main problem for predicate alignment. There are various types of predicate alignment, such as an alignment of single predicate and multiple predicates (E.g. 効果が_ある *kouka ga aru* "have effect" - 効果的だ *koukateki da* "is effective"), content of predicate is contained in noun (E.g. Milk is good for health - Milk is a healthy drink —— *good* is semantically contained in *healthy drink*). Therefore, it is considered that noun alignment and predicate alignment are different tasks. In this paper, we addressed predicate alignment.

In this section, we describe the STATEMENT MAP generation system with taking Figure 3.4 for instance.

3.3.1 Passage Retrieval

We retrieve passages that are relevant to the user query using the TSUBAKI search engine [40] and applying heuristics to filter out noise such as page title or marshaling of noun. The filters are described as follows[3].

- contains context noun words in query sentence
- sentence length are larger than 20 words and less than 150
- sentence must be normal sentence
- part of speech of last word must be verb, auxiliary verb, or adjective
- contains less than three post-positional particles

[3]Parameters were decided heuristically.

3.3.2 Linguistic Analysis

In order to identify semantic relations between the user Query (Q) and the sentence extracted from Web Text (T), we first conduct syntactic and semantic linguistic analysis to provide a basis for alignment and relation classification.

For syntactic analysis, we use the Japanese dependency parser CaboCha [53] and the predicate-argument structure analyzer ChaPAS [54]. CaboCha splits the Japanese text into phrase-like units called *chunks* and represents syntactic dependencies between the *chunks* as edges in a graph. ChaPAS identifies predicate-argument structures in the dependency graph produced by CaboCha.

We also conduct extended modality analysis using the resources provided by Matsuyoshi et al. [55], focusing on source, time, modality and polarity because such information provides important clues for the recognition of semantic relations between *statements*.

3.3.3 Sentiment Analysis

In order to detect strings with implicit sentiment and expressive subjective elements in a given statement, we perform sentiment analysis.

Expressions of emotion, evaluation and reputation, each of which has a sentiment orientation (i.e. positive or negative), have been collected in existing sentiment lexicons such as SentiWordNet [56] for English and Kobayashi's sentiment lexicon [57] as well as Higashiyama's sentiment lexicon [58] for Japanese. We extracted and manually checked 5,500 predicates and 13,312 compound nouns from Web documents using the methods in Kobayashi et al. [57] and Higashiyama et al. [58], respectively.

Our sentiment lexicon includes "*zenkai* (complete recovery)" and "*seiseki ga agaru* (raise one's grade)" that represent positive sentiment, and "*byoki* (disease)" and "*kosho-suru* (breakdown)" that represent negative sentiment. Our sentiment analyzer detects strings with implicit sentiment and expressive subjective elements using this lexicon, and marks them with their sentiment orientations.

3.3.4 Lexical Knowledge

According to premise 2 described in Section 3.1, we use various large scale lexical knowledge. These knowledge are used for knowledge-based lexical alignment. In this section, we describe how to use the knowledge.

Ontologies

We use the Japanese WordNet [59] to check for hypernymy and synonymy between words. E.g. ⟨ 効果 *kouka* "good effect" - 作用 *sayou* "effect"⟩.

In addition, we use Wikipedia. Recently, Wikipedia is considered as the Collective Intelligence which has a massive amount of various information such as sports, history and so on. However the entries in Wikipedia are unstructured texts. To extract knowledge which can be used in computer easily is one of the important tasks in NLP [60]. We use Wikipedia to check hypernymy [60] and synonymy. Synonymy is checked automatically based on the redirect database in Wikipedia. In Wikipedia, some words are hyper-linked to another word as "redirect". The word linked from and to are considered to synonym or paraphrase words [61].

Predicate databases

To determine if two predicates are semantically related, we consult a database of predicate relations [62] and a database of predicate entailment [63] using the predicates' default case frames. E.g. ⟨ 維持する *iji-suru* "to preserve" - 守る *mamoru* "to maintain"⟩ and ⟨ 予防する *yobou-suru* "to prevent" - 気をつける *ki-wo-tsukeru* "to be careful"⟩

Both predicate database and ontology have similar set of relation such as synonymy or hypernymy. In STATEMENT MAP generation, query is considered as hypothesis and retrieved sentence is considered as text of RTE, so that it is enough to check only hypernymy and do not check hyponymy. Because query can be hypernym of retrieved sentence but the inverse is not applied. Other relations such as part-of are also same as hypernymy.

Other Knowledge

In this section, we describe other knowledge used in STATEMENT MAP generation.

Japanese Allographic Database [64] is a database to check allographic ambiguity between two words. E.g. ⟨ 排気ガス *Haiki-gasu* "exhaust gas"⟩ and ⟨ 排ガス *Hai-gasu* "exhaust gas"⟩. When the word in query is allographic form of the word in retrieved sentence, they are aligned as synonym relation.

Distributional similarity database [65] is used to check two words based on coordination [66]. The database contains one hundred million headwords and five hundred words for each of them. The coverage of the database is strong, however, there are many unrelated word pairs. In addition, though the system recognizes CONFLICT based on aligned predicates whose relation is antonym, distributional similarity indicates just similarity but not relation. In other words, there are both synonym and antonym words in distributional similar words. Therefore, we do not use the database in the evaluation.

3.3.5 Corresponding Portions Detection

In order to recognize semantic relation between query and retrieved sentence, identifying the part of retrieved sentence which is semantically similar with query is needed. Then in the following phase, a semantic relation is recognized by observed the part. In previous work, identifying the part is done by word alignment that aligns words between sentences which are semantically similar or related [67, 52]. However, there are the cases that a words are not to be aligned in spite of that they are identical or semantically similar. Harabagiu et al. [68] proposed to consider feature of sentence structure such as syntactic dependency and semantic dependency in recognizing semantic relation phase for the problem. However, we think that the problem should be separated from the recognizing semantic relation.

The first question to be discussed is a unit of alignment. While in English, alignment is done for each words called "word alignment", in Japanese, popular unit of alignment are either a morpheme or a phrase. Though it is not clear that which units are suitable for alignment, in this dissertation, we use phrase for alignment unit called "phrase alignment". In "phrase alignment", there is a problem to align multiple phrases, E.g. alignment for 効果が_ある *kouka ga aru* "have effect" and 効果的だ *koukateki da* "is effective". Therefore, we allow to align single phrase and multiple phrases like multiword expressions. In "morpheme alignment" because there is also similar problem to align multiple morphemes, E.g. 予防 *yobou* "prevent" and 予め_防ぐ *arakajime fusegu* "protect beforehand", it is not clear that which alignment is better and we will discuss the problem in future work.

Then, the part of retrieved sentence which is related with the query in following three processes.

(1) Lexical alignment align chunks between the query and retrieved sentence based on contained words' relatedness

(2) Structural alignment align relations between the two chunks in query and the two chunks lexical aligned in retrieved sentence based on syntactic and semantic structural similarity

(3) Alignment selection select lexical alignment which are passed both (1) and (2)

Lexical Alignment

First, we conduct lexical alignment at the *chunk* level. When the content words in corresponding *chunks* are identical or semantically similar then they are aligned. We use the following resources to determine semantic similarity.

1. **Surface-based Alignment**

   When all of the content words in a phrase in t_2 are all contained in a phrase in t_1, they are aligned. Even if the number of words in the phrase of t_2 is greater than it of t_1, they are aligned.

2. **Knowledge-based Alignment**

   We use the following resources to determine semantic similarity. During the alignment phase, when a pair of phrases, one from t_1 and the other from t_2, is found in one of the resources described in Section 3.3.4, the phrases are aligned. Phrases are matched against the resources using a word-level bi-gram cosine-based similarity measure [69].

   We only use directly related words and do not expand relation through other entry. In other words, when we have two relations A - B and B - C and do not have a relation A - C, we may consider to have relation between A and C through B. We may use this inference when both relations are synonym, however, when they are not, such heuristic inference is not always concluded.

3. **Structure-based Alignment**

   In spite of using massive amounts of knowledge, there are many uncovered words. Especially, domain specific Knowledge are not covered. For instance,

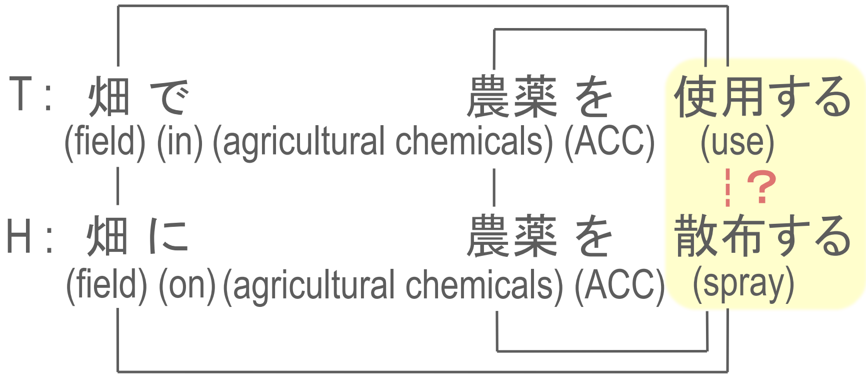

Figure 3.3: Determining the compatibility of semantic structures

considering ⟨ 畑で農薬を使用する *hatake-de-nouyaku-wo-shiyou-suru* "Agricultural chemicals are used in the field. "⟩ and ⟨ 畑に農薬を散布する *hatake-ni-nouyaku-wo-sanpu-suru* "Agricultural chemicals are sprayed on the field. "⟩, ⟨ 使用する *shiyou-suru* "used"⟩ entails ⟨ 散布する *sanpu-suru* "sprayed"⟩ but any knowledge do not contain this relation.

For the problem, we focus on a relevance of query and retrieved sentence. If the query and retrieved sentence are different of surface-level, there is topic-level relevancy. Therefore, it is estimated that the predicate in query and the predicate in retrieved sentence are semantically related, when more than two arguments are lexical aligned. In other words, the predicates are aligned because they share the same argument structures. In this way, we can align predicates which we lack lexical semantic resources for.

Figure 3.3 illustrates example of the estimation. First, a phrase set of ⟨ 畑で *hatake-de* "in the field"⟩ in Q and ⟨ 畑に *hatake-ni* "on the field"⟩ in T and a set of ⟨ 農薬を *nouyaku-wo* "agricultural chemicals"⟩ in Q and ⟨ 農薬を *nouyaku-wo* "agricultural chemicals"⟩ in T are both lexical aligned based on surface similarity. Next, focusing on ⟨ 使用する ⟨[⟩used]shiyou-suru⟩ in Q and ⟨ 散布する *sanpu-suru* "sprayed"⟩ in T, each arguments are ⟨ 畑で *hatake-de* "in the field"⟩ and ⟨ 農薬を *nouyaku-wo* "agricultural chemicals"⟩ in Q, and ⟨ 畑に *hatake-ni* "on the field"⟩ and ⟨ 農薬を *nouyaku-wo* "agricultural chemicals"⟩ in T. Note that, they are both lexical aligned respectively. Finally, ⟨ 使用する *shiyou-suru* "used"⟩ in Q and ⟨ 散布する *sanpu-suru* "sprayed"⟩ in T are judged to semantically similar based on the similarity of argument structures.

When two predicates are related on the levels described as follows, they are tend to align than the other case. So that, they are aligned even if they share only one argument.

(a) **predicates indicate existence or non-existence**

When both predicates indicate existence or non-existence such as ⟨ ある *aru* "exist"⟩ or ⟨ 少ない *sukunai* "low"⟩, the predicates are aligned even if they share only one argument. We listed manually that which predicate indicates existence or non-existence.

(b) **predicates have sentiment polarity**

When both predicates have sentiment polarity, in other words, when both predicates are same sentiment polarity (both are positive or negative) or different sentiment polarity (positive and negative), the predicates are aligned even if they share only one argument.

In these methods, (1) is the most reliable restriction to align two chunks, however, the coverage is quite low. Then we use various lexical knowledge to align phrases based on semantic similarity such as synonymy or hypernymy in (2). However, in spite of large amount of lexical knowledge, the coverage of predicate knowledge and domain specific knowledge are not enough. Therefore, in order to improve coverage of predicate alignment (3) is the method to estimate phrase alignment based on the compatibility of semantic structure. However, because an accuracy of the method is rely on predicate-argument structure analysis and the performance is not sufficient, the false positive alignments are to be contained.

Figure 3.4 shows an example of semantic relation recognition. First, *T* is retrieved by passage retrieval engine. Second, linguistic analysis is applied. For each sentences, a syntactic dependency structure is shown by above arrows, and a predicate-argument structure is shown by below arrows and semantic relation is labeled. An extended modality is given for all predicates (the figure shows only last predicates' one). Third, ⟨ うがいは ⟩ in Q and ⟨ うがいを ⟩ in T is lexical aligned by surface-based alignment. Also, ⟨風邪予防に⟩ in Q and ⟨風邪の_予防には⟩ in T are lexical aligned. ⟨効果的だ⟩ in Q cannot be aligned by neither surface-based and knowledge-based alignments, but can be aligned by structure-based alignment: i.e. 1) ⟨ 効果的だ ⟩ in Q shows positive sentiment and ⟨ ならない ⟩ in T shows negative sentiment, 2) ⟨ 風邪予防に ⟩ modifies ⟨ 効果的だ ⟩ and ⟨ 予防には ⟩modifies ⟨ ならない ⟩, 3) lexical alignment of ⟨ 風邪予

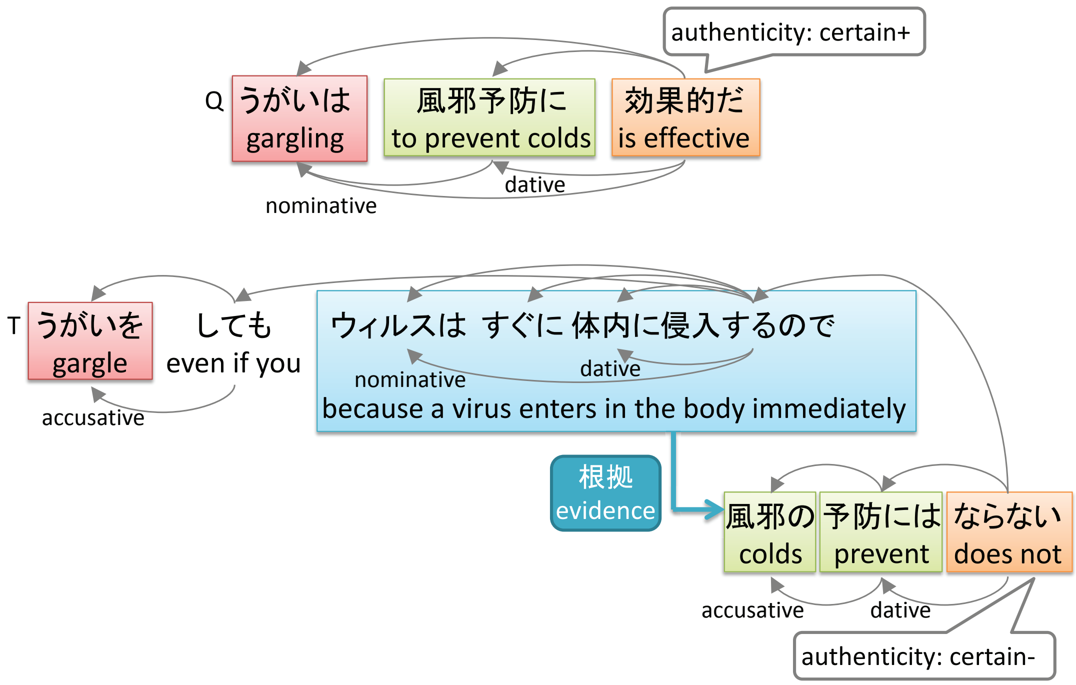

Figure 3.4: An example of semantic relation classification

Q: Gargling is effective to prevent colds.
T: Gargling does not prevent colds even if you gargle, because a virus enters in the body immediately.
The above arrows show syntactic dependency structure, the below arrows and labels show predicate-argument structure, and the colored boxes show alignments.

防に〉and〈予防には〉suggests a structural similarity between〈効果的だ〉and〈ならない〉.

Structural Alignment

(11) Q Black bass$_{\mathrm{a}}$ destroy$_{\mathrm{b}}$ the ecosystem$_{\mathrm{c}}$

T_1 Carp destroy$_{\mathrm{b}}$ the ecosystem$_{\mathrm{c}}$ of black bass$_{\mathrm{a}}$

T_2 Aggressive fish like black bass$_{\mathrm{a}}$ destroy$_{\mathrm{b}}$ the ecosystem$_{\mathrm{c}}$

T_3 Increasing a black bass$_{\mathrm{a}}$ destroys$_{\mathrm{b}}$ the ecosystems$_{\mathrm{c}}$

When the content words in corresponding *chunks* are identical or semantically similar, they are aligned by lexical alignment described in Section 3.3.5. However, this is not enough for corresponding portions detection, because lexical alignments can occur even when the words are not in syntactically and semantically corresponding portions of the query and relevant sentence. Therefore, we adapted a dependency-based approach in the spirit of Das and Smith [70] and Chan et al. [71]. Their approach cannot be applied when there is no syntactic dependency such as $T_1a - T_1c$. In T_1, by considering predicate-argument structure "black bass - nominative - destroys", the structure $T_1a - T_1c$ is structural aligned to the syntactic dependency structure $Qa - Qc$.

Next, in T_2, by considering predicate-argument structure "fish - nominative - destroys" and the "black bass" is an elaboration of "fish", it is suggested that a relation between "black bass" and "destroys" is pseudo nominative relation. Then, $T_2a - T_2c$ is structural aligned to $Qa - Qc$.

In many cases, like T_1 and T_2, there is semantically correspondence between lexical aligned chunks. However, in T_3, as there is no nominative relation between "black bass" and "destroys", $T_3a - T_3c$ should not be aligned to $Qa - Qc$. In structural alignment phase, $Qa - Qc$ is aligned to both $T_1a - T_1c$ and $T_2a - T_2$, but is not aligned to $T_3a - T_3c$.

To do such alignment, we analyzed the development data set by annotating correct structural alignment manually. Note that, a query is restricted to simple sentence. Therefore, for the query, it seems reasonable to suppose that targeting a pair of chunks which directly modified is enough to align. Thus, it is important to analyze what structures exist in relevant sentences.

The result of the analysis was that there are multiple structures to align not only syntactic dependency structure and predicate-argument structure, but also various structures such as elaboration and so on. For the problem, we applied the following four restrictions to realize relaxing a set of restrictions gradually.

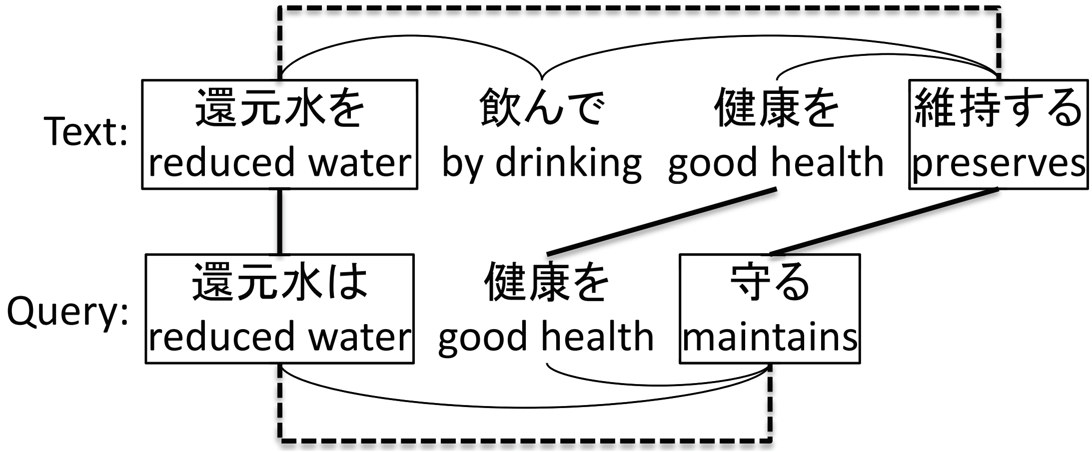

Figure 3.5: An example of structural alignment

Q: Reduced water maintains good health.
T: Reduced water preserves good health by drinking.
A curve shows syntactic dependency structure. A bold solid line shows lexical alignment. A dotted line shows structural alignment.

1. In relevant sentence, two chunks are in a relation of directly modification.

2. There is an arbitrary predicate-argument structure between two chunks. This restriction can be applied to $Qa - Qc$ and $T_1a - T_1c$.

3. Two chunks are in a relation of not directly modification but linked modification. In other words, the chunk modified the other chunk through another chunk. However, it is no allowed to link any number of chunks. By analysis of the development data set, it is limited four chunks forward and backward. For example, in Figure 3.5, structural alignment is applied through two linked-dependency structure.

Alignment Selection

In structural alignment, corresponded alignments are detected based on structural similarity. In next phase, it is necessary to judge which alignments are effective for recognizing semantic relation.

To begin with, we regarded alignments passed structural alignment as corresponded alignments. Then, a main predicate in retrieved sentence is detected. In this step, The main predicate is a last predicate in query and for retrieved sentence, corresponded predicate is detected. Finally, the main predicate and the arguments of the predicate are the part of retrieved sentence which are relevant with the query and effective for recognizing semantic relation.

We describe detail of selection with taking Figure 3.6 for example. To begin with, a main predicate of query is ⟨ 良い *yoi* "good"⟩ which is last and only predicate of the sentence. There are two predicates which correspond to ⟨ 良い *yoi* "good"⟩: ⟨ 悪いので *warui-node* "because ··· bad"⟩ and ⟨ 良い *yoi* "good"⟩. At this point, ⟨ 良い *yoi* "good"⟩ is depended from two phrases ⟨ マーガリンは *ma-garinn ha* "margarine is"⟩ and ⟨ 体に *karada-ni* "for health"⟩. A predicate which has same structures of these dependency structures as structural alignment is just ⟨ 悪いので *warui-node* "because ··· bad"⟩. Because ⟨ 良い *yoi* "good"⟩ has only one same structure with ⟨ 体に *karada-ni* "for health"⟩, it is not selected for main predicate. Finally, phrases which are structural aligned to ⟨ 悪いので *warui-node* "because ··· bad"⟩ are selected as corresponded phrases: ⟨ マーガリンが *ma-garin-ga* "margarine is"⟩ and ⟨ 体に *karada-ni* "for health"⟩. In conclusion, three lexical alignments "1", "2", "4" are selected as corresponded part to query.

There are some difficult cases to select main predicate. For example, when both agreement and conflict opinions for query are written in retrieved sentence, a main predicate can not be detected by the selecting method. In this case, we use heuristic rule that selects last predicate based on the intuition that later predicate is to be indicate conclusion of the sentence.

3.3.6 Semantic Relation Recognition

We employ two strategies to recognize semantic relation: rule-based approach and machine learning-based approach. The former approach is used the alignment results directly than the letter approach. For example, when all phrases in a query are aligned with some phrases in a retrieved text the relation between them is AGREEMENT or

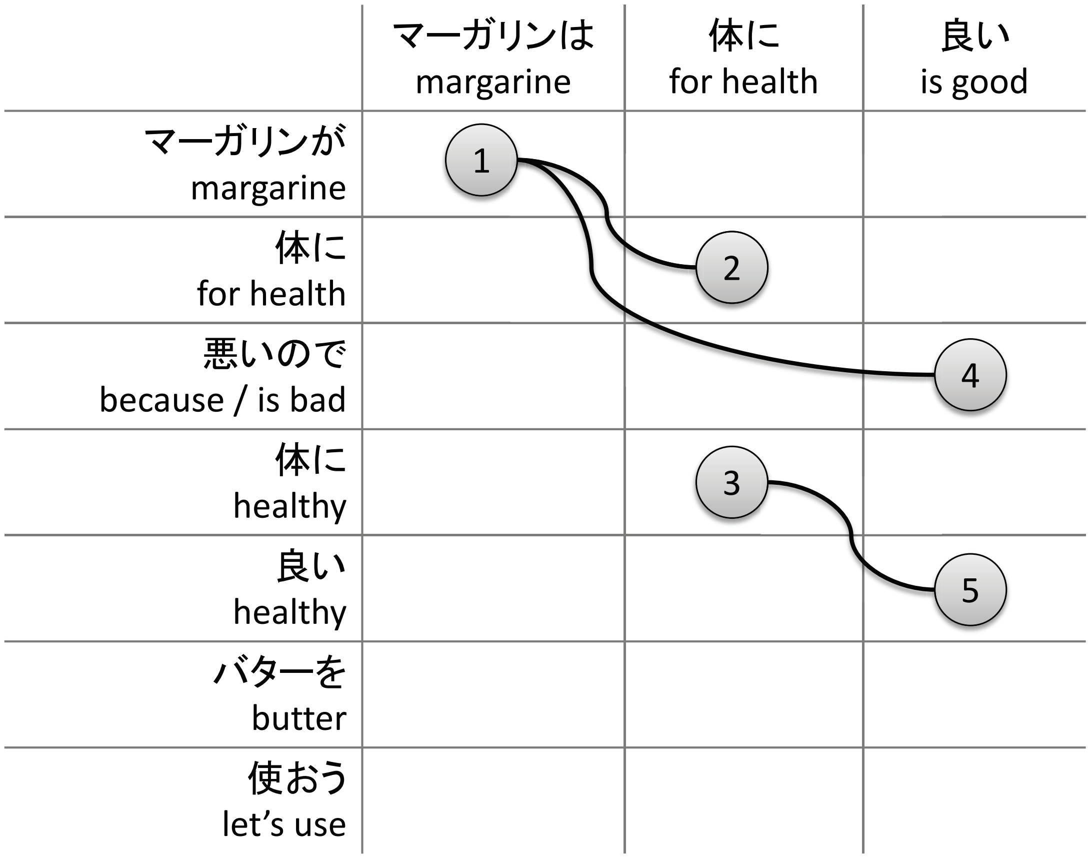

Figure 3.6: Alignment selection

Q: Margarine is good for health.

T: Because margarine is bad for health, let’s use healthy butter.

The circles show lexical alignment and the lines show structural alignment.

CONFLICT. The latter approach is used to confirm the claim of previous work that the machine learning-based approach is difficult to classify CONTRADICTION.

Rule-based Approach

The rule-based approach is based on following two assumptions obtained in analysis of development data.

Assumption 1 Given a pair of a query and a sentence, if there is some semantic relation between the query and the sentence, all phrases in the query are lexical and structural aligned to the phrases in the sentence.

Assumption 2 When existence of some semantic relation between the query and the sentence is detected by the assumption 1, the detailed semantic relation can be recognized based on combination of local semantic relations of the aligned phrases.

According to the assumptions, the strategy of semantic relation recognition consists of following two phases. The phase 1 is from the assumption 1 and the phase 2 is from the assumption 2.

relevance recognition A pair of a query and a retrieved sentences is classified to "relevant" if all of the phrases in the query are aligned to phrases in the retrieved sentence, and "irrelevant" otherwise. However we made exceptions in the above condition: when the headwords of the phrases contain *light verbs*, the phrases in the retrieved sentence are allowed to be unaligned. If the pair is classified as "irrelevant" then the system outputs ⟨OTHER⟩.

semantic relation recognition The relevant pairs are classified into two relations: AGREEMENT, CONFLICT. CONFLICT is determined by considering the semantic relation of an alignment (e.g., if the aligned predicates have an antonym relation: ⟨減少する *gensyo-suru* "decrease"⟩ and ⟨増加する *zoka-suru* "increase"⟩), factualities (e.g. factive - counter-factive: ⟨効果的だ *kokateki-da* "effective"⟩-Positive and ⟨無駄だ *muda-da* "waste"⟩-Negative), and sentiment polarities (e.g. ⟨効果的だ *kokateki-da* "effective"⟩ and ⟨非効果的だ *hikokateki-da* "noneffective"⟩).

In order to adopt opinions, we use two modifications for semantic relation recognition.

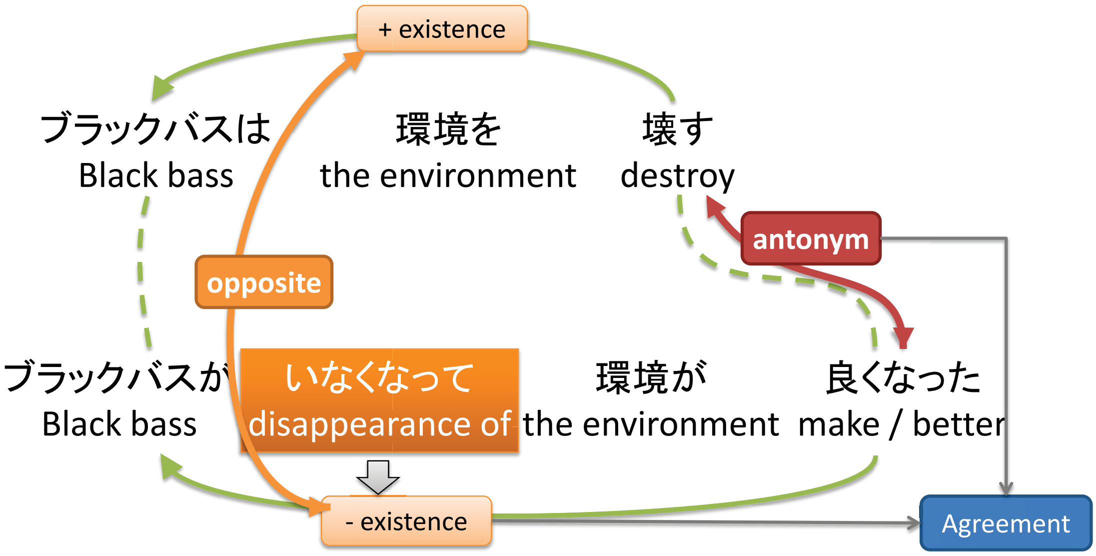

Figure 3.7: Semantic relation recognition with considering existence / non-existence

Q: Black bass destroy the environment.
T: Disappearance of black bass make the environment better.
The dotted lines show lexical alignment and orange bi-directional arrow shows structural alignment.

(1) Modality Because the strength of certainty of the event is over-classification to recognize AGREEMENT and CONFLICT, strength of modality is ignored. Thus, nine types of modality (actuality) that Matsuyoshi et al. [36] defined are mapped to three types: certain+, certain−, and unknown.

certain+ certain+, probable+, probable− → +, and certain− → +

certain− certain−, probable−, probable+ → −, and certain+ → −

unknown unknown

(2) Considering existence For given two sentences Q and T, T entails Q when a human who reads T would infer that Q is most likely also true. However, considering opinions, there are some cases that it is not clear to recognize ENTAILMENT of T and Q. Therefore, we defined AGREEMENT to catch up such pairs. For example, in Figure 3.7, the relation between "black bass" and "destroy" in Q is nominative, however the relation between "black bass" and "make / better" in T is not directly recognized. In T, "disapearance" which shows non-existence of "black bass" is required to consider. Then, the relation between non-existence of "black bass" and "make / better" is pseudo nominative.

Machine Learning-based Approach

Previous work show that recognizing CONTRADICTION is difficult based on machine learning-based approach [32]. However, we defined CONFLICT like CONTRADICTION but more wider relation, therefore, we use machine learning-based approach to confirm the performance of CONFLICT recognition. Once the structural alignment is successfully identified, the task of semantic relation classification is straightforward. We solve this problem with machine learning by training linear classifier [72]. We used L2-regularized logistic regression model. As features, we draw on a combination of lexical, syntactic, and semantic information including the structural alignments from the previous section. The feature set is:

alignments We define two binary function, $ALIGN_{word}(q_i,t_m)$ for the lexical alignment and $ALIGN_{struct}((q_i,q_j),(t_m,t_k))$ for the structural alignment to be true if and only if the node $q_i,q_j \in$Q has been semantically and structurally aligned to the node $t_m,t_k \in T$. Q and T are the Query and the Text, respectively. We also use a separate feature for a score representing the likelihood of the alignment.

modality We have a feature that encodes all of the possible polarities of a predicate node from modality analysis, which indicates the utterance type, and can be *assertive*, *volitional*, *wish*, *imperative*, *permissive*, or *interrogative*. Modalities that do not represent opinions (i.e. *imperative*, *permissive* and *interrogative*) often indicate ⟨OTHER⟩ relations.

antonym We define a binary function $ANTONYM(q_i, t_m)$ that indicates if the pair is identified as an antonym. This information helps identify ⟨CONFLICT⟩.

negation To identify negations, we primarily rely on a predicate 's *Actuality* value, which represents epistemic modality and existential negation. If a predicate pair $ALIGN_{word}(q_i, t_m)$ has mismatching actuality labels, the pair is likely a ⟨OTHER⟩.

Evidence Relation Recognition

EVIDENCE is a relation where the relevant sentence fully agrees or disagrees with the query and contains evidence supporting the conclusion. Consider the following example. Two clauses are linked by the discourse marker *because*. The clause preceding *because* is identical to the query. The clause following *because* provides support for its sibling in the discourse relation and often indicates the presence of EVIDENCE.

(12) *Q* Xylitol is effective at preventing cavities.

T Xylitol is effective at preventing cavities <u>because</u> the cavity-causing bacteria streptococcus mutans cannot metabolize it.

Because EVIDENCE relation provides useful meta-information for the user that is not reflected in simple agreeing and disagreeing classification, it is important to detect EVIDENCE relation and include them in information organization.

Our strategy for identifying EVIDENCE relation is illustrated in Figure 3.8. Because the relation provides information about a particular semantic relation AGREEMENT or CONFLICT, we must first recognize the correct semantic relation. Once that is done, we need to identify the meta-information. This is done using discourse markers like *so* and *because*. The presence of discourse markers indicating evidence are used to identify potential EVIDENCE, while discourse markers and modifiers indicating condition or degree expressions indicate potential EVIDENCE meta-information. In the Penn Discourse Treebank [73], such discourse relations are composed of *nucleus* and *satellite*,

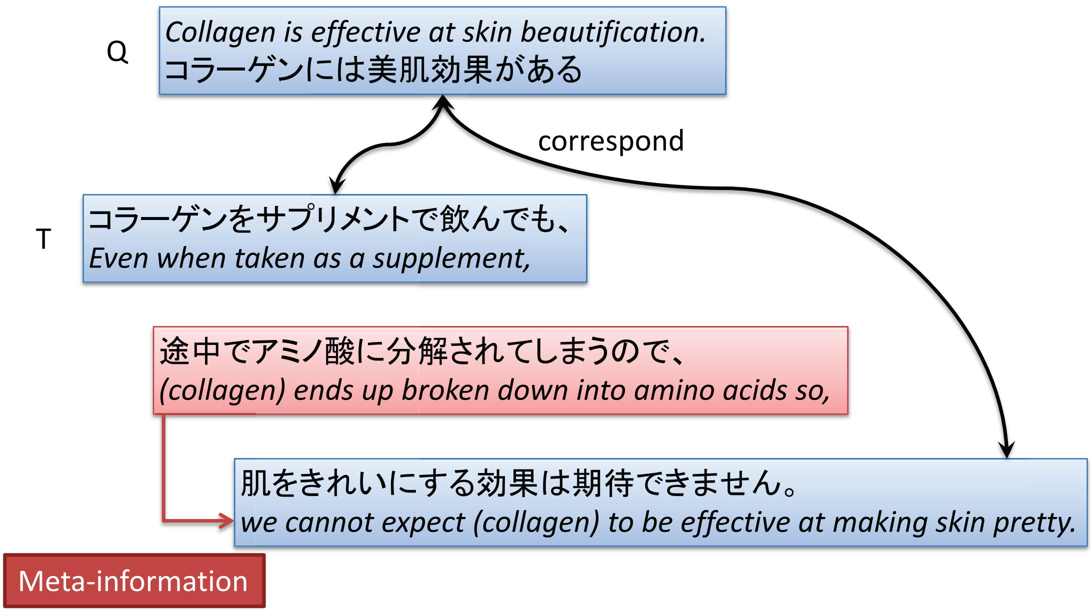

Figure 3.8: Evidence relation recognition

where *satellite* expressions modify the *nucleus*. When we identify a discourse relation, we treat the *satellite* as meta-information. Finally, we need to determine that the meta-information is relevant to the query. We do this by consulting the alignment information between the query and the relevant sentence. If sufficient alignments are found between the *nucleus* and the query, then we identify the *satellite* as meta-information.

3.4 Evaluation

In experiments, we evaluate the performance of semantic relation classification. Especially, we discuss the effect of alignment restriction. The baseline method is machine learning-based strategy used in a number of previous work.

3.4.1 Experimental Settings

We used the 5 query data set for the system development and 20 query data set for the system evaluation as described in Section 2.3.1. The data set included 1467 instances consisting of 532 AGREEMENT instances (i.e. 532 pairs of query and sentence), 238 CONFLICT, and 45 EVIDENCE instances.

In EVIDENCE recognition, while there are 532 AGREEMENT sentences and 238 CONFLICT sentences in the evaluation data, there are only 45 EVIDENCE sentences in it. This indicates that there are a small number of sentences agreeing or conflicting with showing evidence.

In the performance evaluation, we discuss an effect of alignment restriction for semantic relation classification. The restrictions used for the evaluation are defined as follows.

Lexical Alignment

exact: **exact** is the strongest restriction which aligns the phrases when the all content words are identical.

exact+dic: **exact+dic** is a weaker restriction than **exact** which uses lexical knowledge and aligns the phrases based on semantic similarity along with **exact**.

exact+dic+estimation: **exact+dic+estimation** is the weakest restriction which aligns the phrases based on structural similarity along with **exact+dic**.

Structural Alignment

dep: **dep** is the strongest restriction. For the phrases in the query which has syntactic dependency and the phrases in the retrieved sentence which has syntactic dependency, when both pair of phrases are lexical aligned, the dependency structures are structural aligned.

dep+pat: **dep+pat** is a weaker restriction than **dep** which aligns when the pair of phrases satisfy one on the patterns defined in Section 3.3.5.

none: **none** is the weakest restriction which aligns all structures. In other words, the restriction of structural alignment is not applied.

A restriction of alignment consists of the pair of the restriction of lexical alignment and the restriction of structural alignment and declared **exact - dep**.

3.4.2 Experimental Results

Figure 3.9 and Table 3.1 show the results of semantic relation classification results of AGREEMENT, and Figure 3.10 and Table 3.2 show it of CONFLICT.

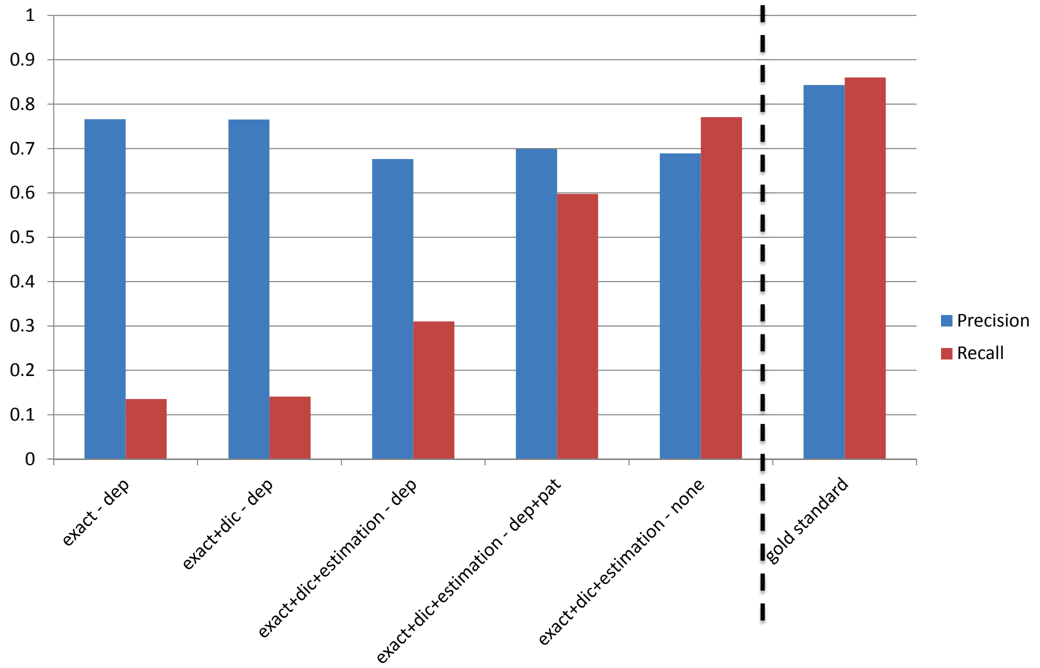

Figure 3.9: Results of semantic relation classification (AGREEMENT)

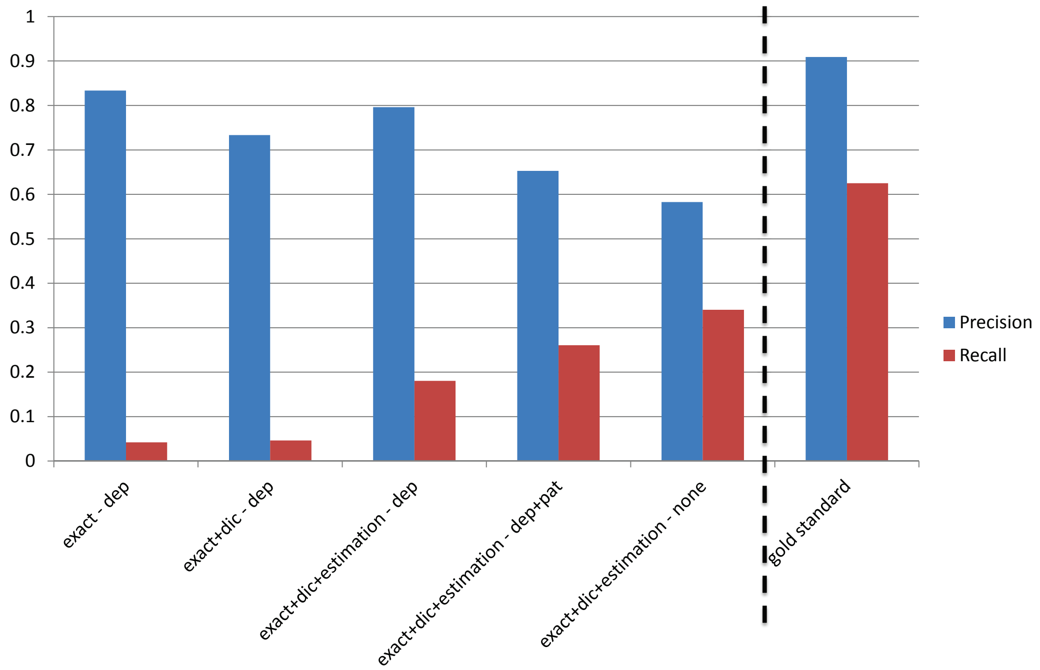

Figure 3.10: Results of semantic relation classification (CONFLICT)

Table 3.1: Results of semantic relation classification (AGREEMENT)

Restriction	Precision	Recall
exact-dep	76.6% (72/94)	13.5% (72/532)
exact+dic-dep	76.5% (75/98)	14.1% (75/532)
exact+dic+estimation-dep	67.6% (165/244)	31.0% (165/532)
exact+dic+estimation-dep+pat	69.9% (318/455)	59.8% (318/532)
exact+dic+estimation-none	68.9% (410/595)	77.1% (410/532)
gold standard	84.3% (43/51)	86.0% (43/50)
baseline (ML-based)	68.0% (374/550)	70.3% (374/532)

Table 3.2: Results of semantic relation classification (CONFLICT)

Restriction	Precision	Recall
exact-dep	83.3% (10/12)	4.2% (10/238)
exact+dic-dep	73.3% (11/15)	4.6% (11/238)
exact+dic+estimation-dep	79.6% (43/54)	18.1% (43/238)
exact+dic+estimation-dep+pat	65.3% (62/95)	26.1% (62/238)
exact+dic+estimation-none	58.3% (81/139)	34.0% (81/238)
gold standard	90.9% (20/22)	62.5% (20/32)
baseline (ML-based)	48.7% (58/119)	24.4% (58/238)

Table 3.3: Results of evidence detection

Restriction	Precision	Recall
exact+dic+estimation - dep	37.5% (6/16)	13.3% (6/45)
exact+dic+estimation - dep+pat	41.7% (20/48)	44.4% (20/45)
exact+dic+estimation - none	38.3% (23/60)	52.1% (23/45)

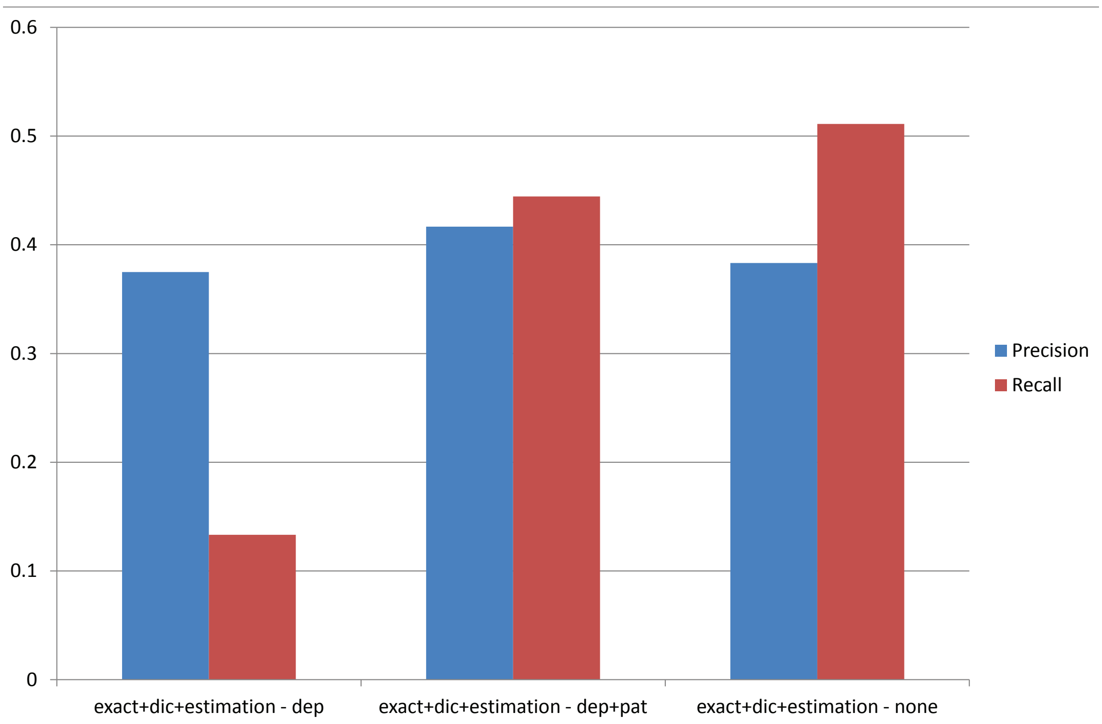

Figure 3.11: Results of evidence detection

First is the strongest restriction, **exact-dep**, where all of the phrases in the *Hypothesis* are both lexical aligned and structural aligned with phrases in the *Text*. This means that the *Hypothesis* dependency tree is found in the *Text*. This restriction performs with high precision and poor recall because of the rarity of full dependency tree matches.

There are two plans to relax the restrictions: relaxing lexical alignment restrictions and relaxing structural alignment restrictions. First, we relax the restrictions of lexical alignment gradually.

We first try relaxing the lexical alignment restrictions by allowing lexical matches using lexical resources in **exact+dic-dep**. Although we used large-scale lexical resources, there were only small gains in coverage over **exact-dep**. This is likely because *Texts* were retrieved using the nouns of *Hypothesis* as query words, making lexical knowledge about nouns less effective. We expect lexical knowledge about predicates to be more effective, but these results indicate that predicate coverage is insufficient.

Next, we further relax lexical alignment based on sentence structure similarity. This is the weakest lexical alignment restriction. This restriction improved recall from 15.6% to 35.0% for AGREEMENT and from 8.4% to 19.7% in CONFLICT in comparison to **exact+dic-dep**. These results show an importance of estimating predicate alignment correctly.

Next, we try relaxing the structural alignment restriction in two steps while retaining the highest performance lexical alignment restriction of **exact+dic+estimation**.

First, in **dep+pat** is the restriction that aligns structure more flexible than **dep** by considered not only dependency structure but also predicate-argument structure and some patterned hopped-dependency structure. This restriction achieves higher recall than **dep**. The result shows that the restriction which aligns only dependency relation is too strong for RTE. On the other hand, it is concerned that whether the weak restriction is caused to increase false alignment and lower precision. However, almost same precision for CONFLICT and higher precision for AGREEMENT.

In the second step, **none** is the weakest restriction which does not apply structural alignment at all. In this restriction, the method of predicate selection described in Section 11 is not able to apply, so that the last predicate lexical aligned with *Hypothesis* in *Text* is used as main predicate. Comparing with the **dep+pat**, the precision of CONFLICT is lower than other results. With this result, to classify semantic relation with high precision, structural alignment is important.

Comparing rule-based approach (**exact+dic+estimation-dep+pat**) with machine learning-based approach, the former approach achieved competitive results in AGREEMENT recognition. For CONFLICT recognition, rule-based approach achieved high precision (48.7$atalmostsamerecall$(24.4

However, the classification performance is not enough by automatic alignment. In addition, when both alignment and semantic relation classification are done automatically, it is difficult to analyze which are caused to the classification error. Therefore, we annotate gold alignment for a part of evaluation data to analyze the classification error of semantic relation classification phase. We also annotate which predicate alignment shows negation worked for CONFLICT recognition. By two annotations, we can separate alignment error to classification error and discuss the problem in the semantic relation classification.

However because of high cost to annotate gold alignment, we annotated 10% of evaluation data selected randomly with keeping a ratio of relations. The results shown in “gold standard”indicates that most of data is correctly classified. Through this results, the most important technical issue for semantic relation recognition targeting Web texts is detecting the region in retrieved Web texts that corresponds to the content of the query. We can see from the result that the small set of rules is enough for semantic relation classification when high performance is achieved in alignment phase. Classification errors in this evaluation setting indicate future direction of semantic relation classification. We discuss it in Section 3.4.4.

We evaluate the performance of EVIDENCE recognition. Figure 3.11 and Table 3.3 show the results. By the restriction setting “exact+dic+estimation - dep+pat”, the recall is improved and precision is also improved, comparing to “exact+dic+estimation - dep”. The improvement of AGREEMENT and CONFLICT classification mainly contributed to the improvement of EVIDENCE recognition.

(13) *Q* マーガリンは体に良い

T マーガリンは<u>生体には存在しえない油なので</u>，体に悪い影響がある

There are many cases that information of evidence is inserted between a subject and a predicate. In such cases, because there is no syntactic dependency between them, weak restriction of alignment is effective.

3.4.3 Discussion

To begin with, we analyze some pairs of query and sentence whose results are different between “exact+dic+estimation - dep” and “exact+dic+estimation - dep+pat”.

There are 176 pairs which are correctly classified with "exact+dic+estimation - dep+pat". In these pairs, 175 pairs which are wrongly classified into OTHER are classified into AGREEMENT or CONFLICT. We confirmed that structural alignment plays an important role. An example classified correctly in "exact+dic+estimation - dep+pat" is shown as follows.

(14) *Q* ヨーグルトは$_a$体に$_b$良い$_c$

T ヨーグルトを$_a$毎日欠かさず食べる習慣ができ、体調も$_b$良くなりました$_c$

Because there is no syntactic dependency between ⟨ ヨーグルトを ⟩and ⟨ 良くなりました ⟩ in T, it is not structural aligned with "exact+dic+estimation - dep". Then, lexical alignment between ⟨ ヨーグルトは ⟩ in Q and ⟨ ヨーグルトを ⟩ in T is removed in alignment selection phase and classified into OTHER. With "exact+dic+estimation - dep+pat", they are structural aligned and classified into AGREEMENT correctly.

On the other hand, there are 31 pairs which classified wrongly with "exact+dic+estimation - dep+pat". An error of structure-based alignment takes majority of wrong classifications.

(15) *Q* 酢を$_a$飲むと$_b$身体が$_c$柔らかくなる$_d$

T スポーツの後や身体の$_c$疲れたときに$_d$酢と$_a$糖分を一緒に飲んだり$_b$、食べたりすると、回復が早くなります

In the above example, the lexical alignment between ⟨ 柔らかくなる ⟩in Q and ⟨ 疲れた ⟩ in T is wrong alignment. We need to improve the conditions to apply structure-based alignment.

Next, we discuss the analysis of a difference between "gold alignment" and "exact+dic+estimation - dep+pat". There are 57 such pairs. 45 pairs in them are classified correctly with "gold alignment". Two reasons are considered for them.

1. OTHER pairs are classified wrongly into AGREEMENT or CONFLICT because of over alignment.

2. AGREEMENT or CONFLICT pairs are classified wrongly into OTHER because of lack of alignment.

While 35 pairs of 45 are caused by alignment, there are only 2 pairs whose corresponding portions are same between "gold alignment" and "exact+dic+estimation - dep+pat", but failed to recognize the relation of lexical alignment. This indicates that alignment dominates a performance of semantic relation recognition.

Finally, we discuss the influence of query difference to the classification performance. In the experiments, in order to cover various queries and reduce an annotation cost, each query has about 70 relevant sentences. This caused that there are some cases that OTHER is a majority of relations. Therefore, quantitative analysis is infeasible, we do qualitative analysis.

For CONFLICT recognition, both precision and recall tend to be high for pairs which can be classified based on sentiment polarity, such as "Fermented soybeans are good for health". On the contrary, the classification performance of pairs which can be classified based on antonymy are limited. As this tendency is particularly shown in CONFLICT recognition, we can see that there is insufficient amount of knowledge of antonymy rather than knowledge of synonymy.

3.4.4 Discussion in Semantic Relation Classification

We also analyzed the classification errors observed in the above gold alignment data set to gain more insights about what other issues should be addressed for further improvements. Through this analysis, we found that a large majority source of errors was something related to the interpretation of the factuality and attribution status of each statement in a given target text. Consider, for example, the following query-text pair:

(16) *Q* Drinking vinegar will soften your body.

T It is a common misconception that drinking vinegar will soften your body.

The system misclassified this pair as AGREEMENT because the query perfectly matches the subordinate clause of the text. However, the statement stated in the subordinate clause of the text is implicitly denied by the counter-factive noun *misconception* in the text. This sort of factuality status is currently analyzed by our extended modality analyzer [74]; however, its coverage is still limited and sometimes makes misinterpretation. Our next direction should also include addressing this factuality-related issue.

3.5 Evaluation for NTCIR-9 RITE

In this section, we describe the evaluation results that participated in the Entrance Exam Subtask of NTCIR-9 RITE. Because all the data is created from actual college-level entrance exams, we evaluate our system for another real world data.

We observe that terms are often followed by paraphrases give in brackets. We exploit this pattern to obtain additional synonym word pairs. This operation is done automatically before all other analysis. Then because bracketed expressions often cause errors in dependency parsing, the bracketed expressions are removed.

(17) t_1 16世紀に入り、海禁政策が弛緩してアメリカ大陸や日本から多くの銀（メキシコ銀、日本銀）が中国に流入した。 "In the 16th century, a lot of silver(MexicanandJapanesesilver) poured into China from America and Japan when the Haijin Policy was relaxed."

t_2 明代には、中国で日本銀が流通した。 "In the Ming era, Japanese silver circulated throughout China."

For example, in (17), ⟨ 銀 *gin* "silver"⟩ has a bracket. According to our strategy, ⟨ メキシコ銀 *Mekishiko-gin* "Mexican silver"⟩ and ⟨ 日本銀 *Nihon-gin* "Japanese silver"⟩ are synonyms of ⟨ 銀 *gin* "silver"⟩. In the alignment phrase, after removing the bracketed phrase, ⟨ 銀 *gin* "silver"⟩ of t_1 and ⟨ 日本銀 *Nihon-gin* "Japanese silver"⟩ of t_2 are aligned by this method.

3.5.1 Entailment Relation Recognition

Our approach to entailment relation recognition consists of two phases: (1) relevance recognition and (2) semantic relation recognition. Given a pair of sentences, the system at first determines relevance using a set of alignments (1). A pair is classified as "relevant" if all of the phrases in t_2 are aligned to phrases in t_1, and "irrelevant" otherwise. However we made exceptions in the above condition. Phrases in t_2 are allowed to be unaligned if the headwords of the phrases contain *light verbs*. If the pair is classified as "irrelevant" then the system outputs "non-entailment". Otherwise, the system classifies the semantic relation ("entailment" or "contradiction") of relevant pairs (2). Contradiction relations are determined by considering the semantic relation of an alignment (e.g. if the aligned predicates have an antonym relation), factualities (e.g. factive - counter-factive), and sentiment polarities.

	devel							formal run
	Prec. (Y)	Rec. (Y)	F1 (Y)	Prec. (N)	Rec. (N)	F1 (N)	Acc.	Acc.
TU1	0.733 (63/86)	0.310 (63/203)	0.436	0.659 (271/411)	0.922 (271/294)	0.769	0.672	0.649 (284/442)
TU2	0.750 (24/32)	0.667 (24/36)	0.706	0.797 (47/59)	0.855 (47/55)	0.825	0.780	0.718 (50/71)
TU3	0.767 (23/30)	0.639 (23/36)	0.697	0.787 (48/61)	0.873 (48/55)	0.828	0.780	0.718 (50/71)

Table 3.4: Results on the development data and the formal run data.

	WN	predicate relations	predicate entailments	Wikipedia	parenthesis	struct-based align.
#	817	81	414	1810	20	758

Table 3.5: The number of phrase alignments per resource/approach on the development data.

3.5.2 Results

We entered three settings **TU1**,**TU2** and **TU3** in the formal run. In **TU1**, the system performs the three steps described above and classifies all of the examples in the dataset. The threshold of cosine similarity used in the alignment phase was set to 0.6. In the two settings **TU2** and **TU3**, performances of the system are evaluated with the examples in which t_2 is a simple sentence. **TU3** is the same as **TU2** except that the system uses only structure-based alignments in entailment relation recognition.

The results on the development data and formal run data are shown in Table TU2 and TU3 achieved significant improvements of performance especially on recall compared to TU1. This results suggest that our system performs well to the examples in which hypothesis has a simple syntactic structure. Although TU3 achieved a slightly higher precision compared to TU2, the performances are the same on accuracy, therefore, the structural alignment approach is less effective on this dataset.

Table 3.5 shows the number of phrase alignments on the development data for each alignment method, including different lexical resources, employed by our system. The resource making the greatest contribution was Wikpedia since there are many named entities including person names, locations and countries in the dataset. Also, Japanese WordNet and the database of predicate entailments were effective. Note that all of the alignments except for structure-based alignments may have overlaps with multiple resources. Also, there are many false positives in structure-based alignments.

3.5.3 Error Analysis

Most of the errors are due to false negatives of alignments. We show major error types with examples[4] in the following. The majority of errors are caused by lack of lexical, paraphrase, and verb entailment knowledge. The following examples are misclassified as N (Y is the correct answer) due to lack of lexical knowledge: ⟨ 征服する *seifuku-suru* "conquer" - 滅ぼす *horobosu* "destroy"⟩ and ⟨ 管轄する *kankatsu-suru* "have jurisdiction" - 統括する *toukatsu-suru* "unify"⟩.

Also, due to lack of paraphrase and entailment relation knowledge, the aligner provided false negatives: ⟨ インフレーション対策として "as a counter-inflation measures" - 物価上昇を/抑制する/ため "to curb price increases"⟩ ⟨ 自発性を/重んじる *jihatsusei-wo / omonjiru* "respect for initiative" - 自主性を/最大限に/発揮させる "exercise their autonomy in their own best"⟩.

The dataset used in the Entrance Exam Subtask contains various types of time expressions. As the time expression reasoner of the system has limited rules, it provided many false negatives: e.g. ⟨16 世紀 *16-seiki* "16th century" - 明代 *mindai* "the Ming era"⟩. Also, if there are modifiers on time expressions (e.g. beginning of), it provides 1-to-n alignments e.g. ⟨902 年 "in 902" → 10 世紀/初め "in the beginning of 10th century"⟩. Since the modifier is not aligned to any phrases in t_1, it causes incorrect entailment relation recognition.

A few examples are incorrectly classified as "entailment" due to misclassifications of factuality information.

(18) t_1 永住資格を持つ在日外国人に選挙権を付与する法案は、廃案となった。"A proposal to grant the right to vote to foreigners with permanent residency status in Japan was rejected."

t_2 永住資格を有する在日外国人も選挙権を持つ。"Foreigners with permanent residency status in Japan have the right to vote."

In this case, the factuality of the event ⟨ 選挙権を付与 "grant the right to vote"⟩ must be "counter-fact" , however, our factuality analyzer mistakenly labeled "fact" to the event.

(19) t_1 総務省が消防職員への団結権付与について検討することを決めた。"The Ministry of Internal Affairs and Communications decided to examine granting the right to organize to workers in fire departments."

[4]Some examples used in this section are slightly modified for ease of explanation.

t_2 消防職員には団結権が保証されていない。 “Workers in fire departments are not guaranteed the right to organize.”

In (19), ⟨検討する “examine”⟩ *presupposes* that the event ⟨付与 “grant”⟩ is “counter-fact”, however, the system also misclassified the factuality of this event as “fact”.

The following examples are instances of “entailment” that are misclassified by our system because t_2 contains a specific information not included in t_1:

(20) t_1 鎌倉幕府は1192年に始まったとされていたが，現在では実質的な成立は1185年であるとする説が支配的である。 “The Kamakura Shogunate had been considered to be establised in 1192, however currently the dominant theory is that it was actually established in 1185.”

t_2 12世紀に 日本では鎌倉幕府が開かれた。 “The Kamakura Shogunate was established in the 12nd century in Japan.”

(21) t_1 デイヴィッド・リヴィングストンはヨーロッパ人で初めてアフリカ大陸を横断し、現地の状況を詳細に報告した。 “David Livingstone was the first European to cross Africa. He gave a detailed report of the area.”

t_2 19世紀、リヴィングストンはアフリカ内陸部の探検を行った。 “In the 19th century, Livingstone explored inner Africa.”

These examples require additional knowledge to infer entailment relations: in (20), Kamakura shogunate was established in Japan, and in (21), David Livingstone lived from 1813 to 1873 i.e. during the 19th century.

Some examples requires more complex inference to determine the correct entailment relation.

(22) t_1 日本・イギリス・アメリカなどは、ロシア革命に対してシベリア出兵を行い、日本軍は最後までシベリアに残っていた。 “The countries including Japan, UK and USA sent troops into Siberia in response to the Russia Revolution, and only Japan remained until the end.”

t_2 日本は、ロシア革命に対してイギリスなど他の国よりも長期にわたって介入を継続した。 “Japan intervened in the Russia Revolution for a longer period than all the other countries.”

In order to infer that ⟨ 日本は最後まで残っていた “only Japan remained until the end”⟩ implies ⟨ 他の国よりも長期にわたって介入 “intervene for a longer period than all the other countries”⟩, systems are required to recognize ⟨ イギリスなどの他の国 “other countries including UK”⟩ corresponds to ⟨ イギリス・アメリカ “UK and USA”⟩ and deal with the comparative expression ⟨ イギリスなど他の国 よりも “compared to the other countries including UK”⟩.

In the following example, it is difficult to obtain the correct alignment since t_1 describes multiple and more specific events which correspond to one predicate in t_2.

(23) t_1 グスタフ・シュトレーゼマン首相はインフレ沈静化のため、ドイツ・レンテン銀行を設立し、レンテンマルクを発行した。“In order to reduce inflation, the prime minister Gustav Stresemann founded the Deutsche Rentenbank and issued the Rentenmark currency.”

t_2 シュトレーゼマンがインフレーション対策のために改革を行った。“Stresemann made reforms to reduce inflation.”

⟨ 改革を/行った “made reforms”⟩ in t_2 corresponds to multiple events, and these describes more specific level compared to t_1. How to deal with these kinds of examples is an open problem.

3.5.4 Conclusion

The resutls of the experiments show that our approach well performed for another real data. The error analysis suggest that majority of the errors still result from lack of lexical knowledge.

Chapter 4

User Evaluation

In this chapter, we describe a Web application which uses STATEMENT MAP system as one of modules. We discuss an effectiveness of information organization through user evaluation of the application.

4.1 Assisting Information Credibility Analysis

Our system is integrated into "Assisting Information Credibility Analysis System" [75] (Figure 4.1). At this time, we implemented machine learning-based approach in semantic relation recognition. Classified sentences are arranged by the classification score.

A purpose of the system is to assist users evaluate information credibility on the Web. Our system is used to organize sentences relevant to a user's query shown in Figure 4.2. By clicking on a title of original Web site, the system shows original Web site in small window (Figure 4.3).

4.1.1 Related Work

Recently, several projects have addressed problem of supporting Web users in evaluating the credibility of online information.

The WISDOM project [11] focuses on evaluating credibility through identifying the source of information and classifying opinions into viewpoints via sentiment analysis. Kawada et al. [76]'s user evaluation showed that WISDOM was effective in helping users identify bias in information and exposing them to new viewpoints, however, the

Figure 4.1: A screenshot of entrance page of assisting information credibility analysis system

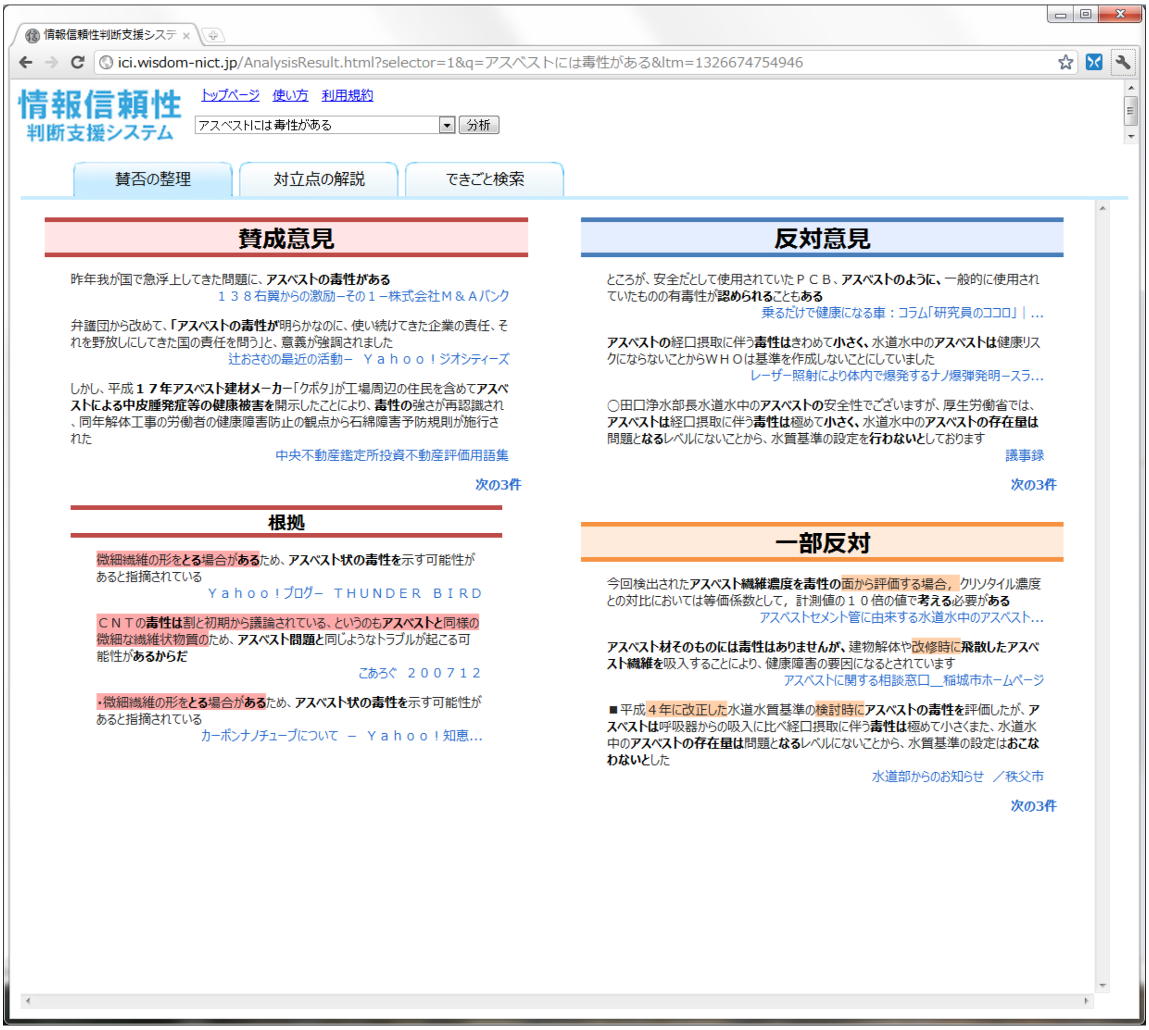

Figure 4.2: A screenshot of STATEMENT MAP in the system

Figure 4.3: A screenshot of viewing original Web site

large amount of analysis presented by WISDOM caused some users to complain of information overload.

Dispute Finder [77] is a crowd sourcing-based approach to credibility analysis. Users install a web browser plug-in that automatically highlights disputed claims in web pages they visit. When a user clicks on a disputed claim, they are shown a list of articles opposing that viewpoint. Dispute Finder builds a database of disputed claims by allowing its users to enter disputed claims and link to a trusted source of rebuttal. Ennals et al. conduct a usibility survey similar to ours, focusing on how the user personas of "skeptical reader" and "activist" that they target interact with the system.

Paul, Zhai and Girju [78] summarize contrasting viewpoints in texts, proposing an unsupervised model for extracting viewpoints and a random walk-based scoring method for detecting contrastive pairs of representative viewpoints. There is no user evaluation because their work is on a core NLP technology, not a full-fleged system.

4.1.2 User Evaluation

We conducted a usability study to gain an understanding of the issues that need consideration when deploying a viewpoint detection system to real Web users and to provide further evaluation of semantic relation classification in a real world application.

In the usability study, participants were asked to compare our system to existing web search engine[1] in investigating topics with diverse viewpoints on the Web. We prepared 54 user queries that can be categorized into the following three broad topics:

Society	裁判員になるのを拒否できる *Citizen judge duty can be refused* 血液型で性格が分かる *Blood type predicts personality type*
Health	ミネラルウォーターは水道水より安全だ *Mineral water is safer than tap water* アガリクスは健康に良い *Agaricus is healthy*
Environment	南極の氷は減っている *Polar ice caps are melting* 地球温暖化によって海面が上昇する *Global warming causes rising sea levels*

[1] We use Google as the web search engine

We recruited 112 Japanese adults (62 males and 50 females) ranging from 20 to 70 years in age. Almost all participants identified themselves as daily Internet users. In order to avoid bias, we employed participants indirectly through a recruiting agency, having the agency conduct the evaluation and distribute a survey on completion.

Before starting evaluation, the study participants freely selected 4 queries. Then, in order to avoid bias from system ordering, they alternated the order of system evaluation between each query. Upon completion of the evaluation task, participants answered a two-question survey, rating each system on the Likert scale of 1 (strongly disagree) to 5 (strongly agree).

(1) *Could you find texts on the Web that agrees and disagrees with the query?*

(2) *Could you find texts that contain evidence that supports or opposes the agree the query?*

System \ Question No.	(1)	(2)
Web search engine	3.54	3.24
STATEMENT MAP	4.06	3.85

Table 4.1: Usability study survey results

The results of the survey are given in Table 4.1. They show that satisfaction with the STATEMENT MAP system was greater for both questions. Responses to question (1) showed that over 84.8 % of users found the AGREEMENT and CONFLICT viewpoints useful, and question (2) showed that 55.4 % users found EVIDENCE detection useful.

Despite the limited performance of STATEMENT MAP in Table 3.1 and 3.2, it was still found more useful than the web search engine for identifying different viewpoints and their support. We offer two theories for this. First, the classified opinions shown to users are sorted by the system's classification confidence level, so many incorrectly classified results were likely not included in the system output. Second, users may have been able to find enough results whose they knew were correct to give STATEMENT MAP the edge in evaluation.

4.1.3 User Feedback

Study participants were also given an opportunity to give feedback during the survey. We received positive responses such as "The system is useful for finding various viewpoints on the topic." and "The system helps me organize and understand information from various sources.", showing that participants found the goal of Web information credibility analysis to be meaningful. For the question "Do you want to continue to use this system?", 70% of the participants responded in the affirmative. This percentage includes people who want to use the system together with the web search engine.

Another response showed that classify the viewpoints instead of ranking search results is better suited to analyzing information credibility: "While in Google, I only look some high-ranked results, prevent me from noticing minor viewpoints, but the [STATEMENT MAP] system is useful for objective thinking because it gives each viewpoint an equal rank."

We also received some negative responses concerning user interface problems. The majority referred to trivial issues such colors and font size, however, some raised issues about how the presentation could influence how it is perceived by users. In particular, this important response indicated that displaying information about viewpoint size could create bias toward popular ones: "The number of opinions seems to indicate that the majority viewpoint is correct." Further evaluation is needed to determine the best way of presenting information on viewpoints and their support.

Several users expressed the desire to use STATEMENT MAP to find viewpoints on arbitrary topics. This raises the question of how search queries should be generated from user input and is an area that requires investigation.

Other responses indicated the importance of identifying the author of texts: "I want only trustworthy sources." and "There are many untrustworthy sources such as weblogs." Detecting authorship is also important for recognizing EVIDENCE and needs more focus in viewpoint detection.

Chapter 5

Conclusions

This chapter summarizes this dissertation and gives future directions we intend to explore.

5.1 Summary

This dissertation has explored how to assist users organize massive amounts of information. We proposed to organize information based on recognizing relation of information on the Web.

To recognize relation of information, we applied a technology of recognizing textual entailment to semantic relation recognition which targeted more wide semantic relations targeting the Web. Then, we constructed development data set targeting Web texts and the classification model based on analysis of the data set.

In evaluation, we confirmed that it is possible to control the precision and recall of semantic relation classification by changing various restrictions. And by constructing and evaluating with gold standard alignment data we were able to detect and analyze common sources of error, identifying negation and nested sentences as posing a challenge for CONFLICT detection.

In Chapter 4, we demonstrated the Web application that organize information for users' query. The results from a survey of 112 users of our STATEMENT MAP information organization system was encouraging.

5.2 Future Directions

In performance evaluation, we confirmed the importance of detecting the region in relevant Web texts that corresponds to the content of a user's query. Especially, alignment is a core technology of the phase. In future work, there are two directions. First is to extend lexical knowledge, especially domain specific knowledge. To do this, it is important to improve conditions of structure-based alignment. Second is to improve structural alignment. In this dissertation, in spite of the approach of structural alignment is based on heuristic rule, it performed well. However, increasing lexical alignments of first direction could be caused an error of structural alignment. In future work, we plan to apply machine learning approach.

In response to user requests, in future work we plan to dynamically cluster opinions into viewpoints that go beyond simple agreement or conflict with a user query.

Currently, the system generate STATEMENT MAP when query is input and takes some minutes to generate. In future work, we plan to prepare a topic list, candidates of generating STATEMENT MAP and generate a huge number of STATEMENT MAP offline like Dispute Finder [79]. Therefore, user can access it and organize information in many directions.

Bibliography

[1] Masaru Kitsuregawa, Satoshi Matsuoka, Takashi Matsuyama, Osamu Sudoh, and Jun Adachi. Cyber Infrastructure for the Information-Explosion Era(<Special Issue>Grant in Aid for Scientific Research on Priority Areas: Cyber Infrastructure for the Information Explosion Era). *Journal of Japanese Society for Artificial Intelligence*, Vol. 22, No. 2, pp. 209–214, 2007. (in Japanese).

[2] Masaru Kitsuregawa. 1. Info-plosion: Retrospection and Outlook(<Special Section>Cyberphysical Information Processing Created in the Info-plosion Era). *The Journal of the Institute of Electronics, Information, and Communication Engineers*, Vol. 94, No. 8, pp. 662–666, 2011. (in Japanese).

[3] Eszter Hargittai, Lindsay Fullerton, Ericka Menchen-Trevino, and Kristin Yates Thomas. Trust Online: Young Adults' Evaluation of Web Content. *International Journal of Communication*, pp. 468–494, 2010.

[4] Deborah Fallows. Search engine users: Internet searchers are confident, satisfied and trusting—but they are also unaware and nav. Pew Internet & American Life Project, 2005.

[5] Peter Turney. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417–424, 2002.

[6] Yohei Seki, David Kirk Evans, Lun-Wei Ku, Le sun, Hsin-Hsi Chen, and Noriko Kando. Overview of Multilingual Opinion Analysis Task at NTCIR-7. In *Proceedings of the 7th NTCIR Workshop*, pp. 185–203, 2008.

[7] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment Classification using Machine Learning Techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing*, pp. 79–86, 2002.

[8] Hiroshi Kanayama and Tetsuya Nasukawa. Textual Demand Analysis: Detection of Users' Wants and Needs from Opinions. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pp. 409–416, 2008.

[9] Kentaro Inui, Shuya Abe, Hiraku Morita, Megumi Eguchi, Asuka Sumida, Chitose Sao, Kazuo Hara, Koji Murakami, and Suguru Matsuyoshi. Experience Mining: Building a Large-Scale Database of Personal Experiences and Opinions from Web Documents. In *Proc. of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence*, pp. 314–321, 2008.

[10] Souneil Park, Kyung Soon Lee, and Junehwa Song. Contrasting opposing views of news articles on contentious issues. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 340–349, 2011.

[11] Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. Wisdom: A web information credibility analysis systematic. In *Proceedings of the ACL-IJCNLP 2009 Software Demonstrations*, pp. 1–4. Association for Computational Linguistics, 2009.

[12] Bo Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1–135, 2007.

[13] Joseph D. Novak. Concept maps and Vee diagrams: two metacognitive tools to facilitate meaningful learning. *Instructional Science*, Vol. 19, pp. 29–52, 1990.

[14] Joseph D. Novak. Meaningful learning: The essential factor for conceptual change in limited or inappropriate propositional hierarchies leading to empowerment of learners. *Science Education*, Vol. 86, pp. 548–571, 2002.

[15] Dragomir R. Radev. Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. In *Proceedings of the 1st SIGdial workshop on Discourse and dialogue*, pp. 74–83, 2000.

[16] Mann William and Sandra Thompson. Rhetorical Structure Theory: Towards a functional theory of text organization. *Text*, Vol. 8, No. 3, pp. 243–281, 1988.

[17] Dragomir Radev, Jahna Otterbacher, and Zhu Zhang. CSTBank: Cross-document Structure Theory Bank. http://tangra.si.umich.edu/clair/CSTBank, 2003.

[18] Junji Etoh and Manabu Okumura. Cross-document relationship between sentences corpus. In *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing*, pp. 482–485, 2005. (in Japanese).

[19] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *First PASCAL Machine Learning Challenges Workshop*, Vol. 3944, pp. 177–190, 2005.

[20] Oren Glickman, Ido Dagan, and Moshe Koppel. Web based textual entailment. In *Proc. of the First PASCAL Recognizing Textual Entailment Workshop*, pp. 33–36, 2005.

[21] Jijkoun V. and M. de Rijke. Recognizing Textual Entailment Using Lexical Similarity. In *Proc. of the First PASCAL Challenges Workshop*, pp. 73–76, 2005.

[22] Erwin Marsi and Emiel Krahmer. Classification of Semantic Relations by Humans and Machines. In *Proc. of ACL-05 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pp. 1–6, 2005.

[23] Bill MacCartney, Trond Grenager, Marie-Catherine de Marneffe, Daniel Cer, and Christopher D. Manning. Learning to recognize features of valid textual entailments. In *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, pp. 41–48, 2006.

[24] Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts Bryan Rink, and Ying Shi. Recognizing Textual Entailment with LCC's Groundhog System. In *Proc. of the Second PASCAL Challenges Workshop*, 2005.

[25] Idan Szpektor, Eyal Shnarch, and Ido Dagan. Instance-based evaluation of entailment rule acquisition. In *Proc. of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 456–463, 2007.

[26] Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 1–9, 2007.

[27] Andrew Hickl and Jeremy Bensley. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pp. 171–176, 2007.

[28] Danilo Giampiccolo, Hoa Trang Dang, Bernardo Magini, Ido Dagan, Elena Cabrio, and Bill Dolan. The Fourth PASCAL Recognizing Textual Entailment Challenge. In *Text Analysis Conference (TAC 2008)*, 2008.

[29] Adrian Iftene. UAIC Participation at RTE4. In *Text Analysis Conference (TAC 2008)*, 2008.

[30] Luisa Bentivogli, Ido Dagan, Hoa Trang Dang, Danlio Giampiccolo, and Bernardo Magini. The Fifth PASCAL Recognizing Textual Entailment Challenge. In *Text Analysis Conference (TAC 2009)*, 2009.

[31] Adrian Iftene and Mihai-Alex Moruz. UAIC Participation at RTE5. In *Text Analysis Conference(TAC)*, 2009.

[32] Marie-Catherine de Marneffe, Anna N. Rafferty, and Christopher D. Manning. Finding contradictions in text. In *Proceedings of ACL-08: HLT*, pp. 1039–1047, 2008.

[33] Rui Wang, Yi Zhang, and Guenter Neumann. A joint syntactic-semantic representation for recognizing textual relatedness. In *Proc. of Recognizing Textual Entailment*, 2009.

[34] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, Vol. 39, No. 2-3, pp. 165–210, 2005.

[35] Roser Saurí and James Pustejovsky. Determining Modality and Factuality for Textual Entailment. In *Proceedings of the International Conference on Semantic Computing*, pp. 509–516, 2007.

[36] Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoyo. Annotating Event Mentions in Text with Modality, Focus, and Source Information. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2010)*, pp. 1456–1463, 2010.

[37] Chris Brockett. Aligning the rte 2006 corpus. In *Microsoft Research Technical Report MSR-TR-2007-77*, 2007.

[38] Hideki Shima, Hiroshi Kanayama, Cheng-Wei Lee, Chuan-Jie Lin, Teruko Mitamura, Yusuke Miyao, Shuming Shi, and Koichi Takeda. Overview of NTCIR-9

RITE: Recognizing Inference in TExt. In *Proceedings of the 9th NTCIR Workshop*, pp. 291–301, 2011.

[39] Michitaka Odani, Tomohide Shibata, Takayuki Nakata, and Sadao Kurohashi. Development of Textual Entailment Data for Japanese and Recognizing Inference Relation Based on Automatically Acquired Similar Expressions. In *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing*, 2008. (In Japanese).

[40] Keiji Shinzato, Tomohide Shibata, Daisuke Kawahara, Chikara Hashimoto, and Sadao Kurohashi. Tsubaki: An open search engine infrastructure for developing new information access methodology. In *Proc. the 3rd International Joint Conference on Natural Language Processing (IJCNLP2008)*, pp. 189–196, 2008.

[41] Mark Sammons, V.G.Vinod Vydiswaran, and Dan Roth. "Ask Not What Textual Entailment Can Do for You...". In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1199–1208, 2010.

[42] Ellen M. Voorhees. Contradictions and Justifications: Extensions to the Textual Entailment Task. In *Proceedings of ACL-08: HLT*, pp. 63–71, 2008.

[43] Marie-Catherine de Marneffe, Anna R. Rafferty, and Christopher D. Manning. Identifying Conflicting Information in Texts. In *Handbook of Natural Language Processing and Machine Translation: DARPA Global Autonomous Language Exploitation*. Springer, 2011.

[44] Alan Ritter, Stephen Soderland, Doug Downey, and Oren Etzioni. It's a Contradiction – no, it's not: A Case Study using Functional Relations. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pp. 11–20, 2008.

[45] Suzan Verberne. Developing an approach for why-question answering. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pp. 39–46, 2006.

[46] Junichi Fukumoto. Question answering system for non-factoid type questions and automatic evaluation based on BE method. In *Proceedings of the Sixth NTCIR Workshop*, pp. 441–447, 2007.

[47] Ryuichiro Higashinaka and Hideki Isozaki. Corpus-based question answering for why-Questions. In *Proceedings of the International Joint Conference on Natural Language Processing*, pp. 418–425, 2008.

[48] Suzan Verberne, Lou Boves, and Nelleke Oostdijk. Discourse-based answering of why-questions. *Traitement Automatique des Langues. Special issue on Computational Approaches to Discourse and Document Processing*, 2007.

[49] Megumi Ohki, Eric Nichols, Suguru Matsuyoshi, Koji Murakami, Junta Mizuno, Masuda Shouko, Kentaro Inui, and Yuji Matsumoto. Recognizing confinement in web texts. In *Proceedings of IWCS 2011*, pp. 215–224, 2011.

[50] Shachar Mirkin, Ido Dagan, and Sebastian Pado. Assessing the Role of Discourse References in Entailment Inference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1209–1219, 2010.

[51] Luisa Bentivogli, Peter Clark, Ido Dagan, Hoa Trang Dang, and Danilo Giampiccolo. The sixth pascal recognizing textual entailment challenge. In *Proceedings of Text Analysis Conference (TAC)*, 2010.

[52] Bill MacCartney, Michel Galley, and Christopher D. Manning. A phrase-based alignment model for natural language inference. In *Proceedings of 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP-08)*, pp. 802–811, 2008.

[53] Taku Kudo and Yuji Matsumoto. Japanese dependency analysis using cascaded chunking. In *CoNLL 2002: Proceedings of the 6th Conference on Natural Language Learning 2002 (COLING 2002 Post-Conference Workshops)*, pp. 63–69, 2002.

[54] Yotaro Watanabe, Masayuki Asahara, and Yuji Matsumoto. A Structured Model for Joint Learning of Argument Roles and Predicate Senses. In *Proceedings of the ACL 2010 Conference Short Papers*, pp. 98–102, 2010.

[55] Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Annotating event mentions in text with modality, focus, and source information. In *Proc. of LREC*, 2010.

[56] Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06)*, pp. 417–422, 2006.

[57] Nozomi Kobayashi, Kentaro Inui, and Yuji Matsumoto. Opinion mining from web documents: Extraction and structurization. *Journal of the Japanese Society for Artificial Intelligence*, Vol. 22, No. 2, pp. 227–238, 2007.

[58] Masahiko Higashiyama, Kentaro Inui, and Yuji Matsumoto. Acquiring noun polarity knowledge using selectional preferences. In *Proc. of the 14th Annual Meeting of the Association for Natural Language Processing*, 2008. (in Japanese).

[59] Francis Bond, Hitoshi Isahara, Sanae Fujita, Kiyotaka Uchimoto, Takayuki Kuribayashi, and Kyoko Kanzaki. Enhancing the Japanese WordNet. In *Proceedings of the 7th Workshop on Asian Language Resources*, pp. 1–8, 2009.

[60] Asuka Sumida, Naoki Yoshinaga, and Kentaro Torisawa. Boosting Precision and Recall of Hyponymy Relation Acquisition from Hierarchical Layouts in Wikipedia. In *Proc. of the 6th International Language Resources and Evaluation (LREC'08)*, pp. 2462–2469, 2008.

[61] Hideki Shima, Yuanpeng Li, Naoki Orii, and Teruko Mitamura. LTI's Textual Entailment Recognizer System at NTCIR-9 RITE. In *Proceedings of the 9th NTCIR Workshop*, pp. 386–393, 2011.

[62] Suguru Matsuyoshi, Koji Murakami, Yuji Matsumoto, , and Kentaro Inui. A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In *Proc. of the 2nd International Symposium on Universal Communication (ISUC2008)*, pp. 366–373, 2008.

[63] Chikara Hashimoto, Kentaro Torisawa, Kow Kuroda, Masaki Murata, and Jun'ichi Kazama. Large-scale verb entailment acquisition from the web. In *Conference on Empirical Methods in Natural Language Processing (EMNLP2009)*, pp. 1172–1181, 2009.

[64] Masahiro Kojima, Masaki Murata, Jun'ichi Kazama, Kow Kuroda, Atsushi Fujita, Eiji Aramaki, Masaaki Tsuchida, Yasuhiko Watanabe, and Kentaro Torisawa. Using Various Features in Machine Learning to Obtain High Levels of Performance for Recognition of Japanese Notational Variants. In *Proceedings*

of The 24th Pacific Asia Conference on Language, Information and Compuation (PACLIC 24), pp. 653–660, 2010.

[65] Jun'ichi Kazama and Kentaro Torisawa. Inducing gazetteers for named entity recognition by large-scale clustering of dependency relations. In *Proceedings of ACL-08: HLT*, pp. 407–415, 2008.

[66] Patrick Pantel, Eric Crestan, Arkady Borkovsky, Ana-Maria Popescu, and Vishnu Vyas. Web-Scale Distributional Similarity and Entity Set Expansion. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pp. 938–947, 2009.

[67] Mark Sammons, V. G. Vinod Vydiswaran, Tim Vieira, Nikhil Johri, Ming-Wei Chang, Dan Goldwasser, Vivek Srikumar, Gourab Kundu, Yuancheng Tu, Kevin Small, Joshua Rule, Quang Do, and Dan Roth. Relation Alignment for Textual Entailment Recognition. In *Proceedings of Recognizing Textual Entailment 2009*, 2009.

[68] Sanda Harabagiu, Andrew Hickl, and Finley Lacatusu. Negation, contrast and contradiction in text processing. In *Proceedings of the 21st national conference on Artificial intelligence*, pp. 755–762, 2006.

[69] Naoaki Okazaki and Jun ichi Tsujii. Simple and Efficient Algorithm for Approximate Dictionary Matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pp. 851–859, 2010.

[70] Dipanjan Das and Noah A. Smith. Paraphrase Identification as Probabilistic Quasi-Synchronous Recognition. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pp. 468–476, 2009.

[71] Ming-Wei Chang, Dan Goldwasser, Dan Roth, and Vivek Srikumar. Discriminative learning over constrained latent representations. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 429–437, 2010.

[72] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, Vol. 9, pp. 1871–1874, June 2008.

[73] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. The penn discourse treebank 2.0. In ELRA, editor, *Proc. of LREC'08*, Marrakech, Morocco, May 2008.

[74] Suguru Matsuyoshi, Megumi Eguchi, Chitose Sao, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Annotating Event Mentions in Text with Modality, Focus, and Source Information. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pp. 1456–1463, 2010.

[75] Yuzuru Okajima, Takao Kawai, Satoru Nakazawa, Koji Murakami, Suguru Matsuyoshi, Junta Mizuno, Eric Nichols, Yotaro Watanabe, Kentaro Inui, Hideyuki Shibuki, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi, and Tatsunori Mori. Development and User Evaluation of an Information Credibility Analysis Support System Using Temporal and Logical Relations from the Web. In *Proc. of the 17th Annual Meeting of the Association for Natural Language Processing*, pp. 53–56, 2011. (in Japanese).

[76] Takuya Kawada, Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Yutaka I. Leon-Suematsu, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. Web information analysis for open-domain decision support: system design and user evaluation. In *Proceedings of the 2011 Joint WICOW/AIR Web Workshop on Web Quality*, pp. 13–18, 2011.

[77] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. Highlighting disputed claims on the web. In *Proceedings of the 19th international conference on World wide web*, pp. 341–350, 2010.

[78] Michael Paul, ChengXiang Zhai, and Roxana Girju. Summarizing contrastive viewpoints in opinionated text. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 66–76, 2010.

[79] Rob Ennals, Dan Byler, John Mark Agosta, and Barbara Rosario. What is disputed on the web? In *Proceedings of the 4th workshop on Information credibility*, pp. 67–74, 2010.

List of Publications

Journal Papers

- Junta Mizuno, Yotaro Watanabe, Eric Nichols, Koji Murakami, Kentaro Inui and Yuji Matsumoto. Organizing Agreeing and Conflicting Opinions Based on Semantic Relation Recognition. In *Transactions of Information Processing Society of Japan*, Vol. 52, No. 12, pp. 3408–3422, 2011. (In Japanese).

International Conferences/Workshop Papers (refereed)

- Megumi Ohki, Eric Nichols, Suguru Matsuyoshi, Koji Murakami, Junta Mizuno, Shouko Masuda, Kentaro Inui and Yuji Matsumoto. Recognizing Confinement in Web Texts. In *Proceedings of the 9th International Conference on Computational Semantics*, pp. 215–224, 2011.

- Koji Murakami, Eric Nichols, Junta Mizuno, Yotaro Watanabe, Shouko Masuda, Hayato Goto, Megumi Ohki, Chitose Sao, Suguru Matsuyoshi, Kentaro Inui and Yuji Matsumoto. Statement Map: Reducing Web Information Credibility Noise through Opinion Classification. In *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*. pp. 59–66, 2010.

- Koji Murakami, Eric Nichols, Junta Mizuno, Yotaro Watanabe, Hayato Goto, Megumi Ohki, Suguru Matsuyoshi, Kentaro Inui and Yuji Matsumoto. Automatic Classification of Semantic Relations between Facts and Opinions. In *Proceedings of the Second International Workshop on NLP Challenges in the Information Explosion Era*. pp. 21–30, 2010.

- Junta Mizuno, Jun Ogata, Masataka Goto. Similar Content Retrieval Method for Podcast Episodes. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, pp. 297–300, 2008.

International Workshop Papers (domestic)

- Yotaro Watanabe, Junta Mizuno, Eric Nichols, Katsuma Narisawa, Keita Nabeshima and Kentaro Inui. TU Group at NTCIR9-RITE: Leveraging Diverse Lexical Resources for Recognizing Textual Entailment. In *Proceedings of the 9th NTCIR Workshop Meeting on Evaluation of Information Access Technologies*. pp. 418–421, December 2011.

Awards

- Analysis of Logical Relations in Web Texts. The 3rd place of the 1st Survey of Technology Trends in 2011. The Nikkei, 2011.
- The Best Paper Award of the 16th Annual Meeting of the Association for Natural Language Processing, 2010
- Chief Creater of the Exploratory IT Human Resources Project (MITOH Program), 2009.

Other Publications (refereed)

- Junta Mizuno, Yuichi Murata and Hisashi Katsuya. Development of Information Retrieval/Recommendation System Using Query Expansion Based on User Preferences. In *Proceedings of the 2nd Rakuten R&D Symposium*, pp. 21–24, 2009. (In Japanese).

Other Publications (domestic)

- Kazuya Narita, Junta Mizuno and Kentaro Inui. An Empirical Analysis of Issues in Japanese Factuality Analysis. In *IPSJ SIG Technical Report 2011-NL-204*, pp. 1–8, 2011. (In Japanese).
- Junta Mizuno and Kentaro Inui. Analysis of Semantic Relations in Semantic Relation Recognition. The 6th NLP Symposium for Young Researchers, 2011. (In Japanese).

- Kazuya Narita, Junta Mizuno and Kentaro Inui. An Empirical Analysis of Issues in Japanese Factuality Analysis. The 6th NLP Symposium for Young Researchers, 2011. (In Japanese).

- Junta Mizuno, Eric Nichols, Yotaro Watanabe, Koji Murakami, Suguru Matsuyoshi, Megumi Ohki, Kentaro Inui and Yuji Matsumoto. Current Status and Issues of Statement Map Generation Technology. In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pp. 49–52, 2011. (In Japanese).

- Yuzuru Okajima, Takao Kawai, Satoru Nakazawa, Koji Murakami, Suguru Matsuyoshi, Junta Mizuno, Eric Nichols, Yotaro Watanabe, Kentaro Inui, Hideyuki Shibuki, Masahiro Nakano, Rintaro Miyazaki, Madoka Ishioroshi and Tatsunori Mori. Development and User Evaluation of an Information Credibility Analysis Support System Using Temporal and Logical Relations from the Web. In *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pp. 53–56, 2011. (In Japanese).

- Eric Nichols, Junta Mizuno, Yotaro Watanabe, Kentaro Inui. Toward Evidence Search. In *Proceedings of The Seventeenth Annual Meeting of the Association for Natural Language Processing*, pp. 880–883, 2011.

- Megumi Ohki, Koji Murakami, Suguru Matsuyoshi, Junta Mizuno, Kentaro Inui and Yuji Matsumoto. Defining and Recognizing Confinement in Texts. In *IPSJ SIG Technical Report 2010-NL-199*, pp. 1–9, 2010. (In Japanese).

- Junta Mizuno, Shoko Masuda, Koji Murakami, Yotaro Watanabe and Kentaro Inui. An Effectiveness of Local Structural Alignment in Semantic Relation Recognition. The 5th NLP Symposium for Young Researchers, 2010. (In Japanese).

- Junta Mizuno, Hayato Goto, Yotaro Watanabe, Koji Murakami and Kentaro Inui. Local Structural Alignment for Recognizing Semantic Relations between Sentences. In *IPSJ SIG Technical Report 2010-NL-196*, pp. 1–8, 2010. (In Japanese).

- Koji Murakami, Junta Mizuno, Hayato Goto, Megumi Ohki, Suguru Matsuyoshi, Kentaro Inui and Yuji Matsumoto. Statement Map Generation Based on Recognizing Semantic Relation between Sentences. In *Proceedings of The Sixteenth*

Annual Meeting of the Association for Natural Language Processing, pp. 559–562, 2010. (In Japanese).

- Megumi Ohki, Koji Murakami, Junta Mizuno, Shoko Masuda, Kentaro Inui and Yuji Matsumoto. Recognizing Confinement Relation: Task Definition, Analysis and Preliminary Experiment. In *Proceedings of The Sixteenth Annual Meeting of the Association for Natural Language Processing*, pp. 788–791, 2010. (In Japanese).

- Hayato Goto, Junta Mizuno, Koji Murakami, Kentaro Inui and Yuji Matsumoto. Structural Alignment for Recognizing Semantic Relations between Sentences. In *Proceedings of the 16th Annual Meeting of the Association for Natural Language Processing*, pp. 848–851, 2010. (In Japanese).

- Junta Mizuno, Kentaro Inui and Yuji Matsumoto. Human-Computer Interaction System Using Web News. In *Proceedings of the SIG-SLUD*, pp. 1–6, 2009. (In Japanese).

- Junta Mizuno, Jun Ogata and Masataka Goto. A Similar Episode Retrieval Method for Podcast. In *IPSJ SIG Technical Report 2008-NL-185*, pp. 31–38, 2008. (In Japanese).

- Junta Mizuno, Hiromi Oyama, Tomoyuki Kobayashi, Kohei Sakata, Noah Evans, Yusaku Taniguchi and Yuji Matsumoto. Development of Example Extraction System for Assisting Japanese Reading. In *Proceedings of Workshop of the 14 the Annual Meeting of the Association for Natural Language Processing*, pp. 63–66, 2008. (In Japanese).