Statement Map: Reducing Web Information Credibility Noise through Opinion Classification

Koji Murakami† Eric Nichols‡ Junta Mizuno†‡ Yotaro Watanabe‡
Shouko Masuda§ Hayato Goto† Megumi Ohki† Chitose Sao†
Suguru Matsuyoshi† Kentaro Inui‡ Yuji Matsumoto†
†Nara Institute of Science and Technology, JAPAN
‡Tohoku University, JAPAN
§Osaka Prefecture University, JAPAN
{kmurakami,matuyosi,shouko,hayato-g,megumi-o,chitose-s,matsu}@is.naist.jp
{eric,junta-m,yotaro-w,inui}@ecei.tohoku.ac.jp

ABSTRACT

On the Internet, users often encounter noise in the form of spelling errors or unknown words, however, dishonest, unreliable, or biased information also acts as noise that makes it difficult to find credible sources of information. As people come to rely on the Internet for more and more information, reducing this credibility noise grows ever more urgent. The STATEMENT MAP project's goal is to help Internet users evaluate the credibility of information sources by mining the Web for a variety of viewpoints on their topics of interest and presenting them to users together with supporting evidence in a way that makes it clear how they are related.

In this paper, we show how a STATEMENT MAP system can be constructed by combining Information Retrieval (IR) and Natural Language Processing (NLP) technologies, focusing on the task of organizing statements retrieved from the Web by viewpoints. We frame this as a semantic relation classification task, and identify 4 semantic relations: [AGREEMENT], [CONFLICT], [CONFINEMENT], and [EVIDENCE]. The former two relations are identified by measuring semantic similarity through sentence alignment, while the latter two are identified through sentence-internal discourse processing. As a prelude to end-to-end user evaluation of STATEMENT MAP, we present a large-scale evaluation of semantic relation classification between user queries and Internet texts in Japanese and conduct detailed error analysis to identify the remaining areas of improvement.

Categories and Subject Descriptors

H.3.1 [**INFORMATION STORAGE AND RETRIEVAL**]: Content Analysis; I.2.7 [**ARTIFICIAL INTELLIGENCE**]: [Natural Language Processing]

AND'10, October 26, 2010, Toronto, Ontario, Canada.

General Terms

Human Factors, Languages, Reliability, Verification

Keywords

STATEMENT MAP, credibility analysis, opinion classification, structural alignment, discourse processing, semantic relation classification

1. INTRODUCTION

Noisy data poses challenges for a number of tasks. In information retrieval, irrelevant documents must be filtered out to produce useful results. In natural language processing, typographic errors, unknown words, and unfamiliar grammatical patterns are often encountered when dealing with Internet data like Web pages and blogs [24].

In this paper, we consider noise in a different context: that of information credibility analysis. When searching for information on the Internet, the dishonest, unreliable, or biased information users encounter constitutes another kind of noise which makes it difficult to find credible information. This is further complicated by the fact that information on users' topics of interest are often unstructured and spread over many documents, and most search engines do not aggregate this content to show users easy-to-follow summaries.

We present STATEMENT MAP, a project with the goal of helping Web users overcome the credibility and distributional noise of information on the Internet by finding documents on user topic of interest, organizing them by viewpoint, and showing supporting evidence and limitations of arguments. STATEMENT MAP combines state-of-the-art technology from IR and NLP to achieve this goal.

Several approaches have been adopted for supporting credibility analysis of online information. Services like `snopes.com` and `factcheck.org` manually debunk urban myths and fact check commonly made political claims. WikiTrust [1] identifies potentially unreliable sections of Wikipedia articles by analyzing their edit histories.

Other projects attempt to educate users about how to identify reliable information online. Meola *et al.* [15] and Metzger [16] provide good summaries of theories of user education. They criticize existing approaches as difficult because users often lack the ability or motivation to properly fact-check sources. Instead, they advocate for critical thinking, arguing that users should identify the information most

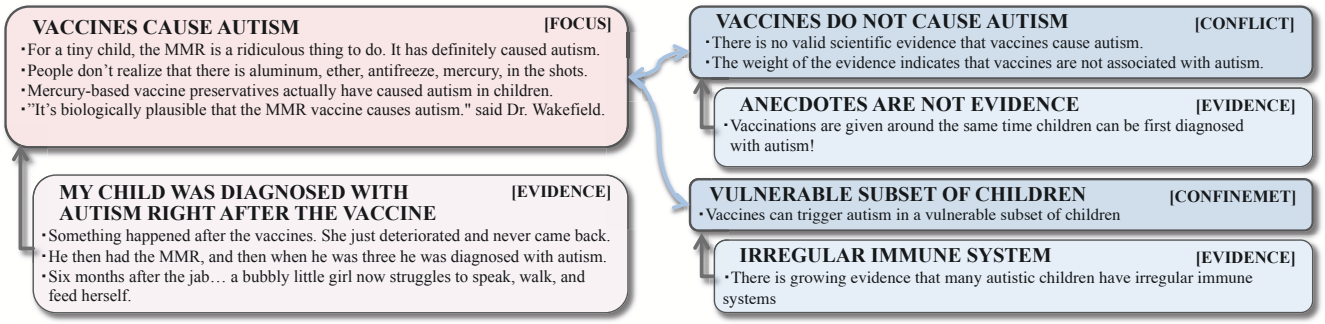

Figure 1: Overview of the STATEMENT MAP system

in need of credibility analysis and compare with multiple references, giving the greatest weight to sources that have undergone peer or editorial review.

We agree that critical thinking is essential to evaluating the credibility of information on the Web. Rather than telling users what information to trust, our goal is to make it easier for them to compare the evidence for each viewpoint on a topic of interest by applying natural language processing and information retrieval technology to automatically gather and summarize relevant sentences, organize their opinions on a topic into the different viewpoints, and show users the evidence supporting each one.

Our project is closest in spirit to Dispute Finder [4] - an extension to the Firefox Web browser, which informs a user when a web page makes a claim that is in its database of known disputes - except that we focus on aggregating information on topics of user interest rather than identifying disputes in claims found during passive Web browsing.

The rest of this paper is organized as follows. In Section 2, we outline the STATEMENT MAP project and its goals. In Section 3, we introduce a set of cross-sentential semantic relations for use in the opinion classification needed to support information credibility analysis on the Web. In Section 4, we discuss related work in the field of Natural Language Processing. In Section 5, we describe how STATEMENT MAPS are generated through relevant passage retrieval from Internet texts, structural alignment, discourse processing, and semantic relation classification. In Section 6, we evaluate our system in a semantic relation classification task. In Section 7, we discuss our findings and conduct error analysis. Finally, we conclude the paper in Section 8.

2. STATEMENT MAP

The goal of the STATEMENT MAP project is to assist Internet users with evaluating the credibility of online information by presenting them with a comprehensive survey of opinions on a topic by applying deep NLP technology, such as semantic analysis to show how they relate to each other. However, because real text on the Web is often complex in nature, we target a simpler and more fundamental unit of meaning which we call the *statement*. To summarize opinions for STATEMENT MAP users, we first convert all sentences into statements and, then, organize them into groups of agreeing and conflicting opinions that show the support and limitations for each group.

Figure 1 shows the results of a similar query "Do vaccines cause autism?" would produce with STATEMENT MAP. The group in the upper-left is labeled [FOCUS], 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."

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

When the concerned user in our example looks at this STATEMENT MAP, he or she will see that some opinions support the query "Do vaccines cause autism?" while other opinions do not, but it will also show what support there is for each of these viewpoints. In this way STATEMENT MAP helps user come to an informed conclusion.

In [20], we discussed the importance of information credibility evaluation on the Web and proposed the development of a STATEMENT MAP system. In this paper, we show how STATEMENT MAPS can be automatically generated by combining information retrieval, linguistic analysis, alignment, and classification tasks, and we present a proof-of-concept Japanese prototype system. Our system is able to detect the semantic relations that are important to STATEMENT MAP by leveraging sophisticated syntactic and semantic information, such as extended modality, to conduct accurate local structural alignment that acts as a basis for semantic relation detection.

3. SEMANTIC RELATIONS

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 1 for real examples.

[AGREEMENT] A bi-directional relation where statements have equivalent semantic content on a shared theme. Here we use *theme* in a narrow sense to mean that the semantic contents of both statements are relevant to each other. The following is an example of [AGREEMENT] on the theme of *bio-ethanol environmental impact*.

(1) a. Bio-ethanol is good for the environment.

Query	Matching sentences	Output
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	限定
	The cavity-prevention effects are greater the more Xylitol is included.	[Confinement].
	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]
	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
還元水は健康に良い	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]
イソフラボンは健康維持に効果がある	大豆イソフラボンをサプリメントで過剰摂取すると健康維持には負の影響を与える結果となります	限定
Isoflavone is effective at maintaining good health.	Taking too much soy isoflavone as a supplement will have a negative effect on one's health	[Confinement]

Table 1: Example semantic relation classification.

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

Once relevance has been established, [Agreement] can range from strict logical entailment or identical polarity of opinions. Here is an example of two statements that share a broad theme, but that are not classified as [Agreement] because *preventing cavities* and *tooth calcification* are not intuitively relevant.

(2) a. Xylitol is effective at preventing cavities.
b. Xylitol advances tooth calcification.

[CONFLICT] A bi-directional relation where statements have negative or contradicting semantic content on a shared theme. This can range from strict logical contradiction to opposite polarity of opinions. The next pair is a [Conflict] example.

(3) a. Bio-ethanol is good for our earth.
b. There is a fact that bio-ethanol further the destruction of the environment.

[CONFINEMENT] A uni-directional relation where one statement provides more specific information about the other or quantifies the situations in which it applies. The pair below is an example, in which one *statement* gives a condition under which the other can be true.

(4) a. Steroids have side-effects.
b. There is almost no need to worry about side-effects when steroids are used for local treatment.

[EVIDENCE] A uni-directional relation where one statement provides justification or supporting evidence for the other. Both statements can be either facts or opinions. The following is a typical example:

(5) a. I believe that applying the technology of cloning must be controlled by law.
b. There is a need to regulate cloning, because it can be open to abuse.

The *statement* containing the evidence consists of two parts: one part has a [Agreement] or [Conflict] with the other *statement*, the other part provides support or justification for it.

4. RELATED WORK

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) is classifying the semantic relation between a Text (T) and a Hypothesis (H) into [Entailment], [Contradiction], or [Unknown].

The RTE Challenge has successfully employed a variety of techniques in order to recognize instances of textual entailment [12, 7, 26]. These approaches have shown great promise for current RTE corpora, but, as de Marneffe *et al.* [3] found in their RTE experiments with Web data, real world data is more difficult to classify. Broader semantic relations that can handle both facts and opinions are needed.

Cross-document Structure Theory (CST), developed by Radev [23], is another approach to recognizing semantic relations between sentences. CST is an expanded rhetorical structure analysis based on Rhetorical Structure Theory (RST: [29]), and it attempts to describe the semantic relations 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: [22]). 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 [Judgement] and [Refinement]. Etoh *et al.* [5] constructed a Japanese Cross-document Relation Corpus and redefined 14 kinds of semantic relations.

Zhang and Radev [30] attempted to classify CST relations between sentence pairs extracted from topically related documents. However, they used a vector space model and tried multi-class classification. The results were not satisfactory. This observation may indicate that the recognition methods for each relation should be developed separately. Miyabe *et al.* [17] attempted to recognize relations that were defined in a Japanese cross-document relation corpus [5]. However, their target relations were limited to [Equivalence] and

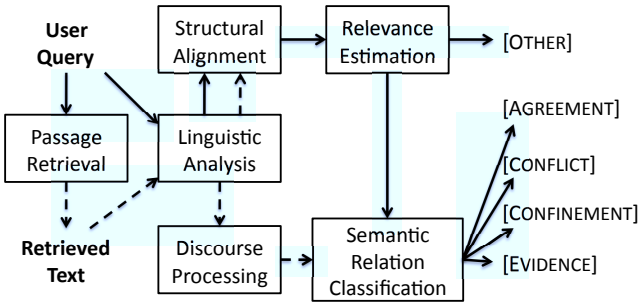

Figure 2: An overview of STATEMENT MAP generation

[TRANSITION]; other relations were not targeted. We also target [CONFINEMENT] and [EVIDENCE] because their recognition is indispensable to STATEMENT MAP.

Subjective statements, such as opinions, have recently been the focus of much NLP research including review analysis, opinion extraction, opinion question answering, and sentiment analysis. In the corpus constructed in the Multi-Perspective Question Answering (MPQA) Project [28], individual expressions are tagged that correspond to explicit mentions of private states, speech event, and expressive subjective elements.

Our task differs from opinion mining and sentiment analysis in two respects. First, the semantic relations we identify between *statements* are broader than the positive/negative classes used in sentiment analysis. Second, it is often not possible to determine if a *statement* on the Internet is a fact or an opinion. Consider the query "Xylitol is effective at preventing cavities." in Table 1. What ultimately determines whether it should be trusted or not is the quantity and quality of evidence in its favor. Our semantic relations are designed to take this ambiguity between facts and opinions into account.

5. STATEMENT MAP GENERATION

In order to generate STATEMENT MAPS, we need to identify [AGREEMENT], [CONFLICT], [CONFINEMENT], and [EVIDENCE] semantic relations between *statements* from multiple documents or user queries. Because identification of [AGREEMENT] and [CONFLICT] is a problem of measuring semantic similarity between two *statements*, it can be cast as a sentence alignment problem and solved using an RTE framework. The two *statements* need not share a source.

However, the identification of [CONFINEMENT] and [EVIDENCE] relations differs because they are fundamentally discourse relations that depend on contextual information in the sentence. For example, conditional statements or specific discourse markers like "because" act as important cues for their identification. Thus, to identify these two relations across documents, we must first identify [AGREEMENT] or [CONFLICT] between *statements* in different documents, then determine if there is a [CONFINEMENT] or [EVIDENCE] relation in one of the *statements*, and finally infer the applicability of the detected relation to a user's query or a *statement* from another document. In the discourse processing stage, we detect these relations within the Internet texts, then in the semantic relation classification stage, we determine if the relations apply to the user query as well.

Our STATEMENT MAP generation approach is as follows:

1. Passage retrieval
2. Linguistic analysis
3. Structural alignment
4. Relevance estimation
5. Discourse processing
6. Semantic relation classification

This approach bears some similarity to RTE system of MacCartney *et al.* [12]. In particular, Steps 2-4 echo their dependency-based *annotate, align, and classify* approach. Our primary differences are that we handle a broader set of semantic relations than in RTE, our system makes use dependency parses in every stage of analysis instead of just during alignment, and we incorporate more detailed linguistic information including predicate-argument structures, extended modality, and discourse cues.

5.1 Passage Retrieval

In order to generate STATEMENT MAPS, we need documents that are relevant to a user's query and contain a variety of opinions. Because identifying semantic relations between complex sentences is difficult, our end goal is to extract *statements*, sub-sentential units of text that effectively summarize opinions, directly from text on the Web. However, as a starting point, we extract sentences from Web text instead using the system proposed by Nagai et. al [21].

5.2 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 [10] and the predicate-argument structure analyzer ChaPAS [27]. 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.* [13], focusing on source, time, modality and polarity because such information provides important clues for the recognition of semantic relations between *statements*.

5.3 Structural Alignment

In this section, we describe our approach to structural alignment. Structural alignment consists of the two phases: (i) lexical alignment, and (ii) structural alignment. Structural alignment is described in more detail in [18].

5.3.1 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.

Ontologies We use the Japanese WordNet [2] and Sumida *et al.*'s [25] to check for hypernymy and synonymy between words. E.g. <効果 *kouka* "good effect" - 作用 *sayou* "effect"> and <イソフラボン *isofurabon* "Isoflavone" - 健康食品 *kenkou-shouhin* "health food">

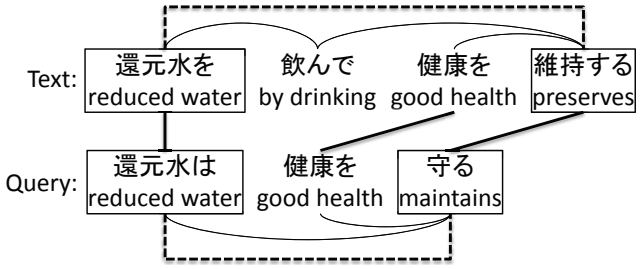

Figure 3: An example of structural alignment

Predicate databases To determine if two predicates are semantically related, we consult a database of predicate relations [14] and a database of predicate entailments [6] 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">

5.3.2 Structural Alignment

In the structural alignment stage, we align pairs of syntactically dependent chunks by determining if the words aligned during lexical alignment correspond semantically. Because lexical alignments can occur even when the words are not in syntactically and semantically corresponding portions of the Query and Text, conducting structural alignment prevents these erroneous lexical alignments from interfering with semantic classification.

For example, in Figure 3, the Query pair <還元水は *kangensui-ha* "Reduced water" → 守る *mamoru* "maintains"> and the Text pair <還元水を *kangensui-wo* "Reduced water" → 維持する *iji-suru* "preserves"> are structurally aligned because they are both semantically *means-of* relations. Similarly, <健康を *kenkou-wo* "good health" → 守る *mamoru* "maintains"> and <健康を *kenkou-wo* "good health" → 維持する *iji-suru* "preserves"> are also structurally aligned.

We treat structural alignment as a binary classification problem and train a Support Vector Machine (SVM) model[1] to decide if a pair of lexically-aligned chunks are structurally aligned. We use various features to train the model. The majority of them are related to sentence structures. Some of the main features are as follows.

- the distance in edges in the dependency graph between parent and child for both sentences
- the distance in chunks between parent and child in both sentences
- binary features indicating whether each chunk is a predicate or argument according to ChaPAS
- the parts-of-speech of the first, last, and syntactic head words in each chunk
- the lexical alignment score of each chunk pair
- the polarity of each words as determined using the resources from [9, 8]

5.4 Relevance Estimation

Murakami *et al.* [19] found that a cascaded model which first identified and excluded semantically irrelevant *statement* pairs before classifying the remaining pairs outperformed a model that identified semantically irrelevant pairs as part of a multi-class classification task.

We benefit from this finding by including a relevance estimation stage in our semantic relation identification system. The goal of relevance estimation is to eliminate Query-Text pairs that are unlikely to be semantically related, and identify if related pairs are more likely to be [AGREEMENT] or [CONFLICT]. The resulting semantic relations are passed to the semantic relation classification stage for final classification taking the discourse processing results into account.

Relevance estimation is framed as a three-class classification problem and solved with an SVM model. We draw on a combination of lexical, syntactic, and semantic information including the syntactic alignments from the previous section to implement the following features.

alignments Binary functions representing lexical alignent of word pairs and structural alignment of Query-Text node pairs. Another feature contains a likelihood score for each alignment.

modality This feature encodes the composite polarity of a node, which is calculated as the product of its predicate and argument polarities. Modalities that do not represent opinions (e.g. *imperative*, *permissive* or *interrogative*) often indicate [OTHER] relations.

antonym This binary feature, indicating if a pair of words are antonyms, helps identify [CONFLICT] relations.

negation This binary feature indicates if a pair of words have identical *actuality* (a composite of lexical and syntactic negation) values. This value is determined using the database in [13]. Mismatching *actuality* values often indicate an [OTHER] relation.

5.5 Discourse Processing

We perform discourse processing to detect discourse markers which identify [CONFINEMENT] and [EVIDENCE] relations. Discourse processing returns a list of chunk and the cues which identify a discourse relation. This information is used in the Semantic Relation Classification stage to determine if a discourse relation discovered in the Text is applicable to the Query.

5.5.1 Confinement Detection

As cues for [CONFINEMENT] detection, we search for degree adverbs (e.g. few and some), partial negations (e.g. not all) within the alignment area and for conditional expressions (e.g. 〜ば *ba* "if 〜") in the syntactic dependencies of the alignment areas.

Consider the following example. We identify the degree adverb, *limited (effects)* and the conditional statement, *if not taken after every meal* as cues for [CONFINEMENT].

(6) T キシリトールは毎食後に摂らないと、虫歯予防の 効果は少ない
If not taken after every meal, Xylitol has limited cavity prevention effects.

5.5.2 Evidence Detection

We focus on detection of Japanese evidence relations containing one of the following explicit discourse markers: から *kara* "since", ので *node* "so", or ため *tame* "because". We identify the chunk containing the discourse marker[2] as the

[1] TinySVM `http://chasen.org/~taku/software/TinySVM/`

[2] In the case of *tame*, because it is usually isolated in its own chunk, we identify the chunk immediately preceding it.

Relation	Precision	Recall	F-Score
[AGREEMENT]	0.55 (125 / 227)	0.46 (125 / 272)	0.50
[CONFLICT]	0.54 (114 / 209)	0.66 (114 / 174)	0.59
[CONFINEMENT]	0.65 (102 / 158)	0.48 (102 / 213)	0.55
[EVIDENCE]	0.66 (21 / 32)	0.33 (21 / 64)	0.42
Total	0.58 (362 / 626)	0.50 (362 / 723)	0.54

Table 2: Semantic relation classification results

	Confl.	Agree.	Confin.	Other	Total(Corr)
[CONFLICT]	114	14	24	22	174
[AGREEMENT]	16	125	13	118	272
[CONFINEMENT]	57	14	102	40	213
[OTHER]	22	74	19	533	648
Total(System)	209	227	158	713	1,307

Table 3: Confusion matrix for semantic relations

evidence and the parent chunk of the discourse marker as the statement being supported. Detecting implicit [EVIDENCE] relations is important but remains an area of future work.

In the example below, we identify the discourse marker, *because*, linking the chunks *cannot metabolize* and *effective at preventing cavities*.

(7) T 虫歯の原因であるミュータンス菌がキシリトールを代謝できないため虫歯予防に効果的です
Xylitol is effective at preventing cavities because the cavity-causing bacteria streptococcus mutans cannot metabolize it.

5.6 Semantic Relation Classification

In this stage of processing, we combine the results of Relevance Estimation and Discourse Processing to classify semantic relations into one of the categories: [AGREEMENT], [CONFLICT], [CONFINEMENT], or [EVIDENCE].

Because relevance estimation has preliminarily classified each relation as either [AGREEMENT] or [CONFLICT], the primary task of semantic relation classification is to determine if those relations should be replaced with either [CONFINEMENT] or [EVIDENCE]. Our basic strategy is as follows:

1. Identify a [AGREEMENT] or [CONFLICT] relation between the Query and Text (*Relevance Estimation*)
2. Search the Text sentence and its surrounding context for cues that identify [CONFINEMENT] or [EVIDENCE] relations (*Discourse Processing*)
3. Infer the applicability of the [CONFINEMENT] or [EVIDENCE] relations in the Text to the Query
4. Combine Steps 1-3 to produce the final classification

Steps 1 and 2 are performed by their respective stages. To infer the applicability of discourse relations in Step 3, we judge there to be a discourse relation when the discourse cues identified fall within the portion of the Text that is aligned to the Query in the structural alignment output.

6. EVALUATION

6.1 Data

We constructed a corpus of sample Japanese user queries and Internet text pairs training and evaluation data for semantic relation classification. Each *query-text* pair was annotated with one of the following semantic relations: [AGREEMENT], [CONFLICT], [CONFINEMENT], [EVIDENCE], and [OTHER]. [EVIDENCE] relations were tagged only for *text* sentences containing one of the explicit discourse markers given in Section 5.5.2. The sentence pairs were manually annotated with both the correct semantic relation and correct structural alignments. Annotations were checked by two native speakers of Japanese, and any sentence pair where annotation agreement is not reached is discarded. All data, including gold standard structural alignments, will be made public at a future date.

6.2 Experiment

In this section, we present empirical evaluation the performance of our system classifying semantic relations into one of four classes: [AGREEMENT], [CONFLICT], and [CONFINEMENT], and [EVIDENCE].[3]

For this experiment, we first identify semantic relevance by using SVMs to classify semantic relations into one of three classes: [AGREEMENT], [CONFLICT], or [OTHER] as described in Section 5.6, and then identifying [CONFINEMENT] and [EVIDENCE] relations among [AGREEMENT] and [CONFLICT] candidates in a separate step based on the results of discourse processing as described in Section 5.5.

As data we use 1,307 sentence pairs from the corpus we constructed in Section 6.1. We present three evaluation measures, namely precision, recall, and f-score. We performed five-fold cross validation, training the structural alignment and relevance estimation modules on 80% of the dataset and evaluating on the held out 20%.

Table 2 shows our experimental results on the dataset described in previous section. Our system achieves moderate precision and recall for each semantic relation. The precision of [AGREEMENT], [CONFINEMENT], and [EVIDENCE] is higher than the recall, suggesting that the structural alignment and discourse processing are able to recognize the information necessary to classify those relations correctly. Examples of successfully recognized relations are given in Table 1.

7. DISCUSSION AND ERROR ANALYSIS

We constructed a prototype Japanese semantic relation classification system by combining the components described in the previous section. While the system developed is not domain-specific and capable of accepting queries on any topic, we evaluate its semantic relation classification on several user queries that are representative of our training data.

Analysis of the experimental results confirmed that semantic relation classification can perform well for real Web data. While de Marneffe [3] *et al.* reported that identifying [CONTRADICTION] in real sentences in Web data was quite difficult, we have achieved moderate performance on [CONFLICT] which includes RTE-style [CONTRADICTION].

Figure 4 shows a snapshot of the semantic relation classification system and the various semantic relations it recognized for each query. In the next example, recognized as [CONFINEMENT], our system correctly identified negation and analyzed the description "Xylitol alone can not completely" as playing a role of requirement.

[3]Although our system also classifies semantic relations as [OTHER], evaluation is omitted because it does not directly contribute to the goal of information credibility analysis.

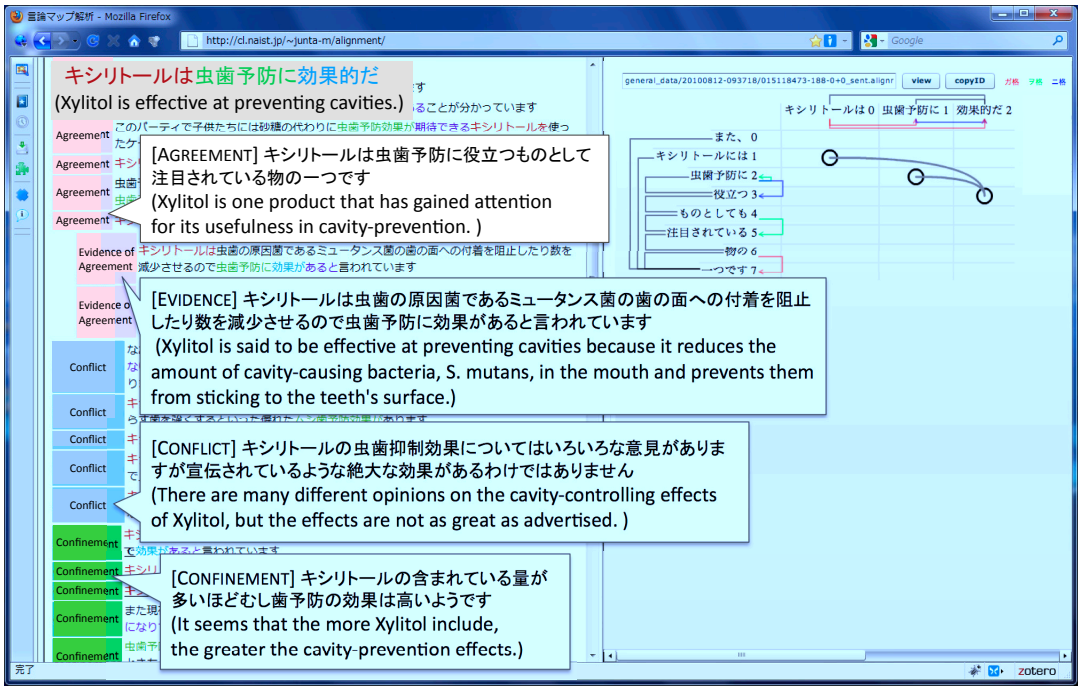

Figure 4: Alignment and classification example for the query *"Xylitol is effective at preventing cavities."*

(8) Q キシリトールは虫歯予防に効果的だ
(Xylitol is effective at preventing cavities.)
T キシリトールだけでは完全な予防は出来ません
(Xylitol alone can not completely prevent cavities.)

Our system correctly identifies [AGREEMENT] relations in other examples about reduced water in Table 1 by structurally aligning phrases like "promoting good health" and "supports the health" to "good for the health."

However, there are still various examples which the system cannot recognized correctly. The confusion matrix shown in Table 3 reveals problems in our system. The most common mistaken classifications are those where the correct answer is [AGREEMENT] but the system classify as [OTHER].

Compared to gold-standard labeled structural alignments, 110 of the 118 examples do not have enough alignment information to be classified correctly. In our estimation, the performance of the alignment is crucial, because *statement* pairs with few alignments are likely to be considered [OTHER].

The following misclassified example is telling: two alignments between the Query and Text are correctly identified, however, they are making opposite claims. In order to correctly classify examples with multiple alignment points, we need to determine the communicative goal of the Text.

(9) Q ステロイドは副作用がある
(Steroids have side-effects)
T ステロイド剤は、長期使用した場合に副作用が問題となってきますが、炎症を止める薬としては大変効果が高く、ある程度の期間なら副作用はほとんど出ません
(Steroids have side-effect in long-term use, however they significantly prevent inflammation, and they have few side-affects in short-term use.)

The confusions between [AGREEMENT] and [CONFLICT] are problematic for STATEMENT MAP generation. A large portion of the confusions is caused by lack of the information in lexical resources we used. These examples require lexical knowledge that is beyond what is currently present in our system.

We also analysed confusions where the correct answer is [CONFLICT] or [CONFINEMENT] but the system classified as [OTHER]. 23 of 62 examples can not be classified with an alignment-based approach, requiring inferential reasoning to be correctly handled. An example is shown below. While the Query describes bioethanol as good for the environment, the Text explains that bioethanol is harmful to the environment indirectly. It is quite difficult to find corresponding words or syntactic structures between these sentences. This indicates the need to adopt an inference-based approach such as [11] to correctly classify these pairs.

(10) Q バイオエタノールは地球の環境に良い
(Bioethanol is good for the environment.)
T バイオエタノールの増産が熱帯雨林を破壊する
(Increasing production of Bioethanol causes tropical rainforest destruction.)

This error analysis showed that a big cause of incorrect classification is incorrect lexical alignment. Improving lexical alignment is a serious problem that must be addressed. This entails expanding our current lexical resources and finding more effective methods of apply them in alignment. However, now that we can generate STATEMENT MAPS, we also plan to conduct an extensive usability survey of their effectiveness as a credibility analysis aid.

8. CONCLUSION

In this paper, we have described out strategy for generating STATEMENT MAPS by describing the task in terms of passage retrieval, linguistic analysis, structural alignment and semantic relation classification and presented a prototype system that identifies semantic relations in Japanese Web

texts using a combination of lexical, syntactic, and semantic information and evaluated our system against real-world data and queries. Preliminary evaluation showed that we are able to detect [AGREEMENT], [CONFLICT], [CONFINEMENT] and [EVIDENCE] with moderate levels of confidence. We discussed some of the technical issues that need to be solved in order to generate better STATEMENT MAP.

Acknowledgments

This work is supported by the National Institute of Information and Communications Technology Japan.

9. REFERENCES

[1] B. Adler, K. Chatterjee, L. de Alfaro, M. Faella, I. Pye, and V. Raman. Assigning trust to Wikipedia content. In *Proc. of WikiSym*, 2008.

[2] F. Bond, H. Isahara, S. Fujita, K. Uchimoto, T. Kuribayashi, and K. Kanzaki. Enhancing the Japanese WordNet. In *Proc. of ACL-IJCNLP*, 2009.

[3] M.-C. de Marneffe, A. N. Rafferty, and C. D. Manning. Finding contradictions in text. In *Proc. of ACL*, pages 1039–1047, 2008.

[4] R. Ennals, B. Trushkowsky, and J. M. Agosta. Highlighting disputed claims on the web. In *Proc. of WWW*, pages 341–350, 2010.

[5] J. Etoh and M. Okumura. Cross-document relationship between sentences corpus. In *Proc. of NLP*, pages 482–485, 2005. (in Japanese).

[6] C. Hashimoto, K. Torisawa, K. Kuroda, M. Murata, and J. Kazama. Large-Scale Verb Entailment Acquisition from the Web. In *Proc. of EMNLP*, pages 1172–1181, 2009.

[7] A. Hickl, J. Williams, J. Bensley, K. Roberts, B. Rink, and Y. Shi. Recognizing textual entailment with lcc's groundhog system. In *Proc. of PASCAL*, 2005.

[8] M. Higashiyama, K. Inui, and Y. Matsumoto. Acquiring noun polarity knowledge using selectional preferences. In *Proc. of NLP*, 2008.

[9] N. Kobayashi, K. Inui, Y. Matsumoto, K. Tateishi, and T. Fukushima. Collecting evaluative expressions for opinion extraction. *Journal of NLP*, 12(3):203–222, 2005.

[10] T. Kudo and Y. Matsumoto. Japanese dependency analysis using cascaded chunking. In *Proc. of CoNLL 2002*, pages 63–69, 2002.

[11] B. MacCartney, M. Galley, and C. D. Manning. Modeling semantic containment and exclusion in natural language inference. In *Proc. of Coling*, pages 521–528, 2008.

[12] B. MacCartney, T. Grenager, M.-C. de Marneffe, D. Cer, and C. D. Manning. Learning to recognize features of valid textual entailments. In *Proc. of HLT/NAACL 2006*, 2006.

[13] S. Matsuyoshi, M. Eguchi, C. Sao, K. Murakami, K. Inui, and Y. Matsumoto. Annotating event mentions in text with modality, focus, and source information. In *Proc. of LREC*, 2010.

[14] S. Matsuyoshi, K. Murakami, Y. Matsumoto, and K. Inui. A database of relations between predicate argument structures for recognizing textual entailment and contradiction. In *Proc. of ISUC*, pages 366–373, 2008.

[15] M. Meola. Chucking the checklist: A contextual approach to teaching undergraduates web-site evaluation. *Libraries and the Academy*, 4(3):331–344, 2004.

[16] M. J. Metzger. Making sense of credibility on the web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology*, 58(13):2078–2091, 2007.

[17] Y. Miyabe, H. Takamura, and M. Okumura. Identifying cross-document relations between sentences. In *Proc. of IJCNLP*, pages 141–148, 2008.

[18] J. Mizuno, H. Goto, Y. Watanabe, K. Murakami, K. Inui, and Y. Matsumoto. Local Structural Alignment for Recognizing Semantic Relations between Sentences. In *Proc. of IPSJ-NL196*, 2010. (in Japanese).

[19] K. Murakami, E. Nichols, K. Inui, J. Mizuno, H. Goto, M. Ohki, S. Matsuyoshi, and Y. Matsumoto. Automatic classification of semantic relations between facts and opinions. In *Proc. of NLPIX*, 2010.

[20] K. Murakami, E. Nichols, S. Matsuyoshi, A. Sumida, S. Masuda, K. Inui, and Y. Matsumoto. Statement map: Assisting information credibility analysis by visualizing arguments. In *Proc. of WICOW*, pages 43–50, 2009.

[21] T. Nagai, K. Kaneko, Hideyuki, M. Nakano, R. Miyazaki, M. Ishioroshi, and T. Mori. Passage extraction based on textrank. In *Proc. of NLP*, 2010.

[22] D. Radev, J. Otterbacher, and Z. Zhang. CSTBank: Cross-document Structure Theory Bank. http://tangra.si.umich.edu/clair/CSTBank, 2003.

[23] D. 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*, pages 74–83, 2000.

[24] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi. A survey of types of text noise and techniques to handle noisy text. In *Proc. of AND*, pages 115–122, 2009.

[25] A. Sumida, N. Yoshinaga, and K. Torisawa. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proc. of LREC*, 2008.

[26] I. Szpektor, E. Shnarch, and I. Dagan. Instance-based evaluation of entailment rule acquisition. In *Proc. of ACL*, pages 456–463, 2007.

[27] Y. Watanabe, M. Asahara, and Y. Matsumoto. A structured model for joint learning of argument roles and predicate senses. In *Proc. of ACL*, 2010.

[28] J. Wiebe, T. Wilson, and C. Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3):165–210, 2005.

[29] M. William and S. Thompson. Rhetorical structure theory: towards a functional theory of text organization. *Text*, 8(3):243–281, 1988.

[30] Z. Zhang and D. Radev. Combining labeled and unlabeled data for learning cross-document structural relationships. In *Proc. of IJC-NLP*, 2004.