

Toward Evidence Search

Eric Nichols[†], Junta Mizuno[‡], Yotaro Watanabe[†], and Kentaro Inui[†]

[†]Tohoku University, JAPAN [‡]Nara Institute of Science and Technology, JAPAN

{eric, junta-m, yotaro-w, inui}@is.ecei.tohoku.ac.jp

1 Evidence Search

Understanding the quantity and quality of evidence for or against a statement is essential to evaluate the trustworthiness of information on the Internet. However there is far too much information on the Web for users to cope with manually, and they need support.

It is our view 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.

The user would be told *who* holds the opinion (i.e. their qualifications); *what* the opinion is; *when* it was held (to insure its relevance); and, finally, *why* they hold that opinion (i.e. the supporting evidence for the opinion). In order to create such a solution, we would need to: (i) find evidence related to the user’s topic of interest on the Web; (ii) classify this evidence into supporting and opposing groups; and (iii) conduct quality analysis of the evidence to determine its factuality, logical soundness, and the credibility of its source.

We propose a new task to achieve these goals: *Evidence Search*. An evidence search system would find evidence related to a user’s query, organize it into supporting and opposing viewpoints, and show the user valuable meta-data like logical consistency, verifiability of information, and author credentials in the topic of interest to help the user evaluate which points of view were most supported. Evidence Search consists of two sub-tasks: (i) *detection of evidence related to the user’s query* and (ii) *evaluation of the quality of the evidence detected*.

To detect evidence related to a user query, we need to retrieve Internet texts that are related to that query, identify evidence within that text, and ensure its applicability to the query. In this paper, we present a prototype evidence search system that detects evidence related to a user query, organizes the evidence into supporting and opposing groups, and presents it to the user in an easy-to-understand manner.

To evaluate the quality of evidence, we need to determine what makes evidence good or bad. Roughly speaking, good evidence may have some of the following properties: is logically and factually consistent, is empirically verifiable, exhibits expert knowledge, or is free of bias. Likewise, bad evidence may be inconsistent, unverifiable, lacking expert knowl-

edge, or biased. In this paper, we will not address the issue of evidence quality evaluation, but propose the following hierarchy as a starting point for future research.

- unsubstantiated arguments: *evidence which cannot be confirmed or verified in any manner*
- logical argument: *relations contains a logical argument that supports a user query if its premises are sound*
- expert knowledge: *opinions from scientists, doctors, or other specialists on topic of interest*
- empirical evidence: *scientific publications, medical surveys, or journalistic reports containing results of empirical evaluation*

In this paper we present a prototype evidence search system (shown in Figure 1) where evidence applicable to a search query is automatically detected and organized into supporting or opposing viewpoints, making it easy to evaluate the evidence related to a topic of user interest. We propose an alignment-based method for determining the applicability of evidence to a user query and give preliminary experimental evaluation of its effectiveness.

2 Related Work

Evidence Search can be thought of as a combination of several areas of research.

Detection of evidence related to a user query is similar in nature to Why-QA [13, 1, 4]. 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 [1], to discourse-oriented [15] and, more recently, machine-learning [4, 16].

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 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. In addition, we do not target questions as user queries. Instead, we are more interested in handling the declarative statements that make up many online debates. This is both to eliminate the need to unravel complex interrogatives, but also to allow the possibility of moving from a user-based query system to one that actively analyzes the evidence for viewpoints expressed on the Internet without user intervention.



Figure 1: Screenshot of the Evidence Search system

The organization of evidence into groups that support or oppose the user query is a task of viewpoint detection, and it draws on the work of the Multi-Perspective Question Answering of Wiebe et al. [18], as well as later work in online stance recognition [12], and the viewpoint visualization of STATEMENT MAP [9], whose engine we make use of in our prototype system. Our goal here, again, is to supplement this cluster of viewpoints with information on the quantity and quality of evidence supporting each one.

Finally, our goal of classifying and evaluating the quality of evidence builds on Argumentation Theory, a branch of philosophy dedicated to diagramming and understanding the structure of arguments and the evidence and reasoning employed to support them. Walton’s Argumentation Schemes [17] provides a framework for describing rhetorical arguments and sets of critical questions that can guide the evidence quality analysis process.

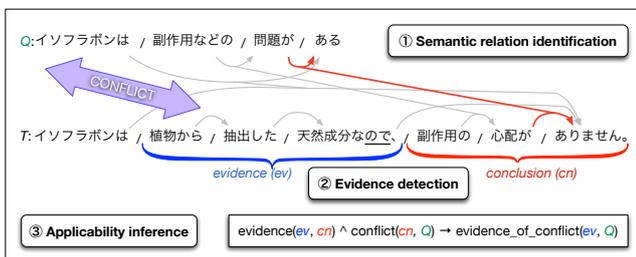


Figure 2: Overview of the evidence inference process

3 Evidence Detection

We consider evidence search to consist of two major tasks: (i) the detection of evidence that is relevant to user queries, and (ii) the evaluation of the quality of that evidence. In this paper, we address the task of evidence detection. Evaluation of evidence quality will be addressed in future research.

To successfully detect evidence that is relevant to a user query, we need to identify Internet text that is related to the query, detect evidence in that text, and make sure that the evidence applies to the query. We break this process down into the following three steps which are shown in Figure 2.

1. Identify semantic relation between *Query* and *Text*
2. Detect evidence relation in *Text*
3. Infer applicability of *Text* evidence relation to *Query*

3.1 Semantic Relation Identification

In this step, we identify text that is semantically related to the user query using the STATEMENT MAP system. STATEMENT MAP [9] helps users evaluate the credibility of online 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. It does this by classifying *Query-Text* pairs into semantic relations such as AGREEMENT and CONFLICT and organizing them to help users visualize the relationships between the viewpoints.

We use STATEMENT MAP to provide the underlying classification of evidence into SUPPORTING EVIDENCE and OPPOSING EVIDENCE show in our prototype in Figure 1. It also acts as a filter; only text that is identified as either SUPPORTING or OPPOSING makes it to the next step.

3.2 Evidence Candidate Detection

In this step, we identify evidence in Internet text that are identified as semantically relevant to the user Query in the previous step.

We follow Iida et al. [5] and frame this as a discourse parsing task like that of the Penn Discourse Treebank [10] of detecting evidence chunk and the conclusion chunk they modify. To identify evidence and conclusion chunk, we use simple pattern matching to identify explicit cues of evidence.

We currently target three explicit cues: から *kara* “since”,

| ID | Evaluation | Relation | Evidence |
|----|------------|-----------|--|
| 1 | correct | AGREEMENT | Q:牛乳は体に良い T:牛乳は栄養たっぶりだから、君たちの体にはとても大切なんだ |
| 2 | correct | CONFLICT | Q:マーガリンは体に良い T:マーガリンは悪玉コレステロールが含まれているので体に悪いんだそうです |
| 3 | correct | CONFLICT | Q:牛乳は体に良い T:つい最近話題になったアメリカの新谷弘実医師の著書病気になる生き方で牛乳の害について書かれていたため牛乳はほんとうは体に悪いのではないかとかなり議論になりました |
| 4 | correct | AGREEMENT | Q:バナナダイエットは効果がある T:バナナダイエットの効果が高いのは、やはり「誰にでもできて、そして無理をすることがないから」だと思います |
| 5 | incorrect | AGREEMENT | Q:牛乳は体に良い T:低温殺菌の「オホーツクあばしり牛乳」は、牛乳に含まれているたんぱく質や、カルシウムなどの栄養素も、熱による変性が起こっていないので、体に吸収しやすい状態を保っています |
| 6 | incorrect | AGREEMENT | Q:マーガリンは体に良い T:マーガリンは植物性ですがバターは動物性なのでバターのほうが体に悪いと思います |
| 7 | incorrect | CONFLICT | Q:マーガリンは体に良い T:バターのほうがまだマシで、マーガリンはパーム油のせいで体に悪いと聞いたのでバターしか摂取しないようにしています |
| 8 | incorrect | AGREEMENT | Q:ヨーグルトは体に良い T:でも、どっちみちヨーグルトは体に良い食べ物ですので、食べておいて損はなからう |

Table 1: Example detected evidence

ので *node* “so”, and *tame* “because”. As these cues are often ambiguous, having some roles that indicate evidence and others that do not, we incorporate part-of-speech, tense, and dependency information into our patterns to filter out the noise. These patterns are similar to those employed by Fukumoto [1].

While there are more sophisticated methods for detecting causal relations (e.g. [2, 3, 11]), this pattern-based approach performs sufficient enough¹ to implement our prototype Evidence Search system.

3.3 Evidence Applicability Inference

STATEMENT MAP is a combination of advanced IR and NLP technology, but at its core is the structural alignment engine of Mizuno et al. [6]. The structural alignment engine uses a combination of manually- and automatically-compiled NLP resources to align Japanese sentences on the chunk level and provide a similarity score that is used as a basis for STATEMENT MAP’s semantic classification. The structural alignment engine conducts alignments at two different levels: the chunk and the dependency tree. We use its chunk alignments to help us determine if key passages in the Internet text containing the evidence identified in the previous step are sufficiently related to the user query. Evidence applicability is inference is carried out by verifying if the chunk alignments between the *Query* and *Text* meet a series of *alignment conditions*. These conditions are described in detail in Section 4.2 and Table 2.

4 Experiment

In this section we conduct preliminary experiments to evaluate the effectiveness of alignment-based evidence detection.

4.1 Data Preparation

In order to evaluate our evidence detection approach, we need a collection of user queries and corresponding Internet texts that have been annotated with EVIDENCE relations. Why-QA is similar in nature to our task, and could provide a useful source of data. While several Why-QA corpora have been described in the literature [14, 7, 4], to the best of our

¹Our evidence detection system currently achieves a f-score of over 71 on the chunk labeling task and data presented in [5].

knowledge, there are no publicly-available resources for the Japanese language, so we need to construct our own evaluation data.

For evaluation, we manually annotate EVIDENCE relations between user queries and Internet texts. We use the evaluation data of the STATEMENT MAP project [9], which consists of a total of 1,050 *Query-Text* pairs covering 20 different that have been annotated with semantic relations. *Query-Text* pairs were gathered automatically using the Tsubaki² search engine. See [8] for an in-depth description of the evaluation data creation process.

To reduce costs, we limit annotation to *Query-Text* pairs in the STATEMENT MAP data that have been identified as having a semantic relation and that contain an explicit cue for evidence as discussed in Section 3.2. All *Query-Text* pairs satisfying these criteria are annotated as either EVIDENCE or NOT-EVIDENCE. Annotation was carried out by two native speakers of Japanese. A total of 251 *Query-Text* pairs were annotated with 105 instances of EVIDENCE.

4.2 Evaluation

In this section, we conduct experiments to determine the effectiveness of our evidence detection method, exploring several different applications of Mizuno et al.’s [6] structural alignments to infer the applicability of evidence detected using pattern-matching on Internet texts.

Our task is to determine if a given *Query-Text* pair is EVIDENCE or NOT-EVIDENCE. We do this by applying the evidence detection method described in Section 3. For evaluation, we use gold-standard STATEMENT MAP semantic relations and system output for structural alignments.

We compare the effectiveness of several different applications of Mizuno et al.’s [6] structural alignments to infer the applicability of evidence detected. Alignment conditions and results are given in Table 2.

As a baseline, we start with no conditions on the alignments (*has.ev*). This represents a system that relies on only the presence of explicit cues of evidence and semantic similarity as measured by STATEMENT MAP’s semantic relation classification to detect evidence for a user query. As no alignment

²<http://tsubaki.ixnlp.nii.ac.jp/>

| Precision | Recall | F-Score | Approach | Alignment Conditions |
|------------------------|-------------------------|--------------|--------------------------|---|
| 0.680 (68 / 100) | 0.648 (68 / 105) | 66.34 | <i>has.evcn</i> | no alignment conditions; evidence and conclusion are found in <i>Text</i> (baseline) |
| 0.594 (19 / 32) | 0.181 (19 / 105) | 27.74 | <i>aligned.ev</i> | an alignment exists between <i>Query</i> and the evidence in <i>Text</i> |
| 0.735 (25 / 34) | 0.238 (25 / 105) | 35.97 | <i>aligned.cn</i> | <i>Query</i> and the conclusion in <i>Text</i> are aligned |
| 0.704 (50 / 71) | 0.476 (50 / 105) | 56.82 | <i>aligned.not.ev</i> | <i>Query</i> and <i>Text</i> are aligned but the evidence is not aligned to <i>Query</i> |
| 0.722 (13 / 18) | 0.124 (13 / 105) | 21.14 | <i>aligned.cn.not.ev</i> | <i>Query</i> and the conclusion but not the evidence in <i>Text</i> are aligned |
| 0.725 (66 / 91) | 0.629 (66 / 105) | 67.35 | <i>aligned.before.ev</i> | <i>Query</i> and <i>Text</i> are aligned before the evidence |
| 0.671 (55 / 82) | 0.524 (55 / 105) | 58.82 | <i>aligned.after.ev</i> | <i>Query</i> and <i>Text</i> are aligned after the evidence |
| 0.687 (68 / 99) | 0.648 (68 / 105) | 66.67 | <i>aligned.before.cn</i> | <i>Query</i> and <i>Text</i> are aligned before the conclusion |
| 0.627 (32 / 51) | 0.305 (32 / 105) | 41.03 | <i>aligned.after.cn</i> | <i>Query</i> and <i>Text</i> are aligned after the conclusion |

Table 2: Results comparing alignment strategies for evidence detection

information is used, it is also the most lenient. In particular, it often erroneously identifies evidence that is identical in semantic content to the *Query*.

An example is given in Example 8 in Table 1. The *Query* translates to *Yogurt is good for the body*, and the *Text* translates to *But at any rate, yogurt is a food that is good for the body, so there shouldn't be any harm to eat it*. The evidence is identified as *yogurt is a food that is good for the body* and is clearly synonymous with the *Query*.

Intuitively, requiring an alignment between the conclusion chunk and the *Query* or forbidding the evidence from aligning with the *Query* should ensure that the evidence detected supports the *Query* rather than repeating it, and we test that next. As expected, the presence of a conclusion chunk aligned with the *Query* (*aligned.cn*) is the highest precision predictor, while evidence (*aligned.ev*) is the lowest, showing how the relationship between the evidence and the *Query* is essential in determining their relevance.

Given the poor performance of an aligned evidence chunk, we wonder if it is effective to forbid it entirely. We find that forbidding it while requiring another alignment (*aligned.not.ev*) or while requiring the conclusion to align with the *Query* (*aligned.cn.not.ev*) achieve high precision albeit at the cost of coverage.

We find that while alignments conditions with the conclusion are high-precision, they are too strict to achieve an acceptable level of recall. We theorize that a single chunk may be too small of a target for the alignments to hit consistently. We, thus, finally consider alignment ranges: conditions where alignments occur before or after the evidence or conclusion chunk (*aligned.before.ev*, *aligned.after.ev*, *aligned.before.cn*, *aligned.after.cn*). These conditions produce the best results, with alignments before the conclusion (*aligned.before.cn*) showing an f-score increase of 0.33 over the baseline, and alignments before the evidence (*aligned.before.ev*) as the best performing system with an f-score of 67.35: an increase in f-score of 1.01 over the baseline.

5 Conclusion

Gathering and evaluating evidence is crucial to deciding what information to trust or what viewpoints to adopt. In this paper, we proposed a new task for NLP, *evidence search*, with the goal of supporting users in evaluating the evidence supporting and opposing topics of interest. We defined evidence search in terms of two sub-tasks: evidence detection and evidence quality evaluation. Our prototype evidence search system and evidence detection method show it is possible to find evidence relating to a user query with simple methods. How-

ever, there remain many technical issues that need to be addressed to provide users with relevant evidence and the analysis of its quality needed to support users in evaluating conflicting information and viewpoints.

Acknowledgments

This work was partly supported by Nifty Labs Japan, by the Japan MEXT Grant-in-Aid for Scientific Research on Priority Areas, Cyber Infrastructure for the Information-explosion Era (No. 19024033), and by Japan National Institute of Information and Communications Technology.

References

- [1] Junichi Fukumoto. Question answering system for non-factoid type questions and automatic evaluation based on BE method. In *Proc. of the Sixth NTCIR Workshop*, pages 441–447, 2007.
- [2] Roxana Girju. Automatic detection of causal relations for question answering. In *Proc. of MultiSumQA '03*, pages 76–83.
- [3] Ryuichiro Higashinaka and Hideki Isozaki. Automatically acquiring causal expression patterns from relation-annotated corpora to improve question answering for why-questions. *TALIP*, 7:1–29, April 2008.
- [4] Ryuichiro Higashinaka and Hideki Isozaki. Corpus-based question answering for why-questions. In *Proceedings of IJCNLP*, vol. 1, pages 418–425, 2008.
- [5] Ryu Iida, Kentaro Inui, and Yuji Matsumoto. The task definition of evidence-conclusion relation extraction and its preliminary empirical evaluation. In *Proc. of NLP2009*. (in Japanese).
- [6] Junta Mizuno, Hayato Goto, Yotaro Watanabe, Koji Murakami, Kentaro Inui, and Yuji Matsumoto. Local Structural Alignment for Recognizing Semantic Relations between Sentences. In *Proc. of IPSJ-NL196*, 2010. (in Japanese).
- [7] Joanna Mrozinski, Edward Whittaker, and Sadaoki Furui. Collecting a why-question corpus for development and evaluation of an automatic QA-system. In *Proceedings of ACL-08: HLT*, pages 443–451.
- [8] Koji Murakami, Shouko Masuda, Suguru Matsuyoshi, Eric Nichols, Kentaro Inui, and Yuji Matsumoto. Annotating semantic relations combining facts and opinions. In *Proc. of LAW III*, pages 150–153, 2009.
- [9] 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 *Proc. of AND 2010*.
- [10] 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.
- [11] Hiroki Sakaji, Satoshi Sekine, and Shigeru Masuyama. Extracting causal knowledge using clue phrases and syntactic patterns. In *Practical Aspects of Knowledge Management*, volume 5345 of *Lecture Notes in CS*, pages 111–122. Springer Berlin/Heidelberg, 2008.
- [12] Swapna Somasundaran and Janyce Wiebe. Recognizing stances in ideological online debates. In *Proc. of CAAGET '10*, pages 116–124.
- [13] 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*, *EACL '06*, pages 39–46, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.
- [14] Suzan Verberne, Lou Boves, and Nelleke Oostdijk. Data for question answering: The case of why. In *In Proc. of LREC 2006*.
- [15] Suzan Verberne, Lou Boves, and Nelleke Oostdijk. Discourse-based answering of why-questions, 2007.
- [16] Suzan Verberne, Stephan Raaijmakers, Daphne Theijssen, and Lou Boves. Learning to rank for why-question answering. *Information Retrieval*. DOI 10.1007/s10791-010-9136-6.
- [17] Douglas Walton, Chris Reed, and Fabrizio Macagno. *Argumentation Schemes*. Cambridge University Press, 2008.
- [18] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2/3):164–210, 2005.