A Belief Revision Approach to Textual Entailment Recognition

Ralf Krestel
L3S Research Center
Universität Hannover
Germany
krestel@L3S.de

Sabine Bergler and **René Witte**
Department of Computer Science
and Software Engineering
Concordia University, Montréal, Canada
witte|bergler@cse.concordia.ca

Abstract

An artificial believer has to recognize textual entailment to categorize beliefs. We describe our system – the Fuzzy Believer system – and its application to the TAC/RTE three-way task.

1 Introduction

True understanding of natural language by computer systems is one of the main goals of artificial intelligence research. Since this goal is still out of reach, smaller, clearly defined tasks can be used to demonstrate progress towards the right direction. Recognizing textual entailment is one task a human can easily do after grasping the meaning of sentences. But for computational systems, finding out whether one statement entails another is not a trivial task. It involves coping with similarity, negation, or generalization which are concepts of natural language that have to be modeled in the system.

Recognizing textual entailment is also useful for other applications. Automatic summarization needs some kind of RTE module to avoid duplicate information in the summaries. It is also closely related to paraphrase detection as well. For opinion mining its usefulness is even more obvious. Comparing two opinions includes the identification of the polarity of both statements.

Another area where RTE can be fruitfully applied is belief revision. We developed an artificial believer (Ballim and Wilks, 1991) based on fuzzy set theory (Krestel et al., 2007a). The beliefs are extracted from reported speech in newspaper articles, which are a good source for opinionated statements. Different strategies were implemented to model a human newspaper reader, who holds certain beliefs after reading some articles. In this context it is necessary to identify opposing and supporting beliefs. We solved this task using different heuristics. In this paper we show how we employed our Fuzzy Believer system to recognize textual entailment.

2 The Three-Way RTE Task

The three-way RTE task was introduced last year and allows systems to be more precise: Instead of the possible two answers *entailment* and *contradiction* in the classic version, the new task additionally allows systems to tag a text/hypothesis pair as *unknown*. Table 1 shows examples for all three answers.

3 The Fuzzy Believer Approach

The core concept embodied in our approach (Krestel et al., 2007a; Krestel et al., 2007b) is the application of fuzzy set theory to the NLP domain. This allows for an explicit modeling of fuzziness inherent to natural languages and enables the user to control the system's behaviour by varying various runtime parameters responsible for the fuzzy processing. Reported speech statements present the basic set of beliefs for our system. These kinds of statements usually express a belief held by the source of the statement and allows a clear attribution of the statement to this source. The extracted reported speech structures are further processed and the output of external semantic parsers is utilized to identify predicate-argument structures (PAS) within the reported speech content. Each PAS defines a statement, which the system eventually either believes or rejects. They also form the foundation for the fuzzy processing and the basis for our heuristics to process beliefs.

To mirror the different processing steps, our Fuzzy Believer system consists of a set of components running consecutively. It is implemented using GATE (General Architecture for Text Engineering) (Cunningham et al., 2002), which offers a framework for developing NLP applications. For preprocessing, we use a number of standard components shipped with GATE, for high-level processing we developed our own components.

To apply our approach to the RTE task we had to change the preprocessing part and adapt some parameters. Our system runs now on the whole input document and not

ENTAILMENT

T	Boris Franz Becker, German tennis player who, on July 7, 1985, became the youngest champion in the history of the men's singles at Wimbledon.
H	Becker was a tennis champion.

CONTRADICTION

T	Set in the New York City borough of The Bronx, the show starred Ted Danson as the title character, Dr. John Becker, a doctor who operates a small practice and is constantly annoyed by his patients, co-workers, friends, and practically everything and everybody else in his world. Becker has never played tennis in his life.
H	Becker was a tennis champion.

UNKNOWN

T	Boris Becker has told a German court that he made financial mistakes 10 years ago but denied deliberately cheating his taxes.
H	Becker was a tennis champion.

Table 1: Examples for the three-way RTE task

only on reported speech sentences. The main components – predicate-argument structure extraction and belief computation – remained unaltered, requiring only adaptations to the output format. These are described in more detail in the following sections.

3.1 Extracting Predicate-Argument Structures

To decide whether a sentence has the same topic as another one, we need to find a way to compare sentences with each other. To facilitate this task, we do not compare complete sentences, but rather their more fine-grained predicate-argument structures, each consisting of a "subject," "verb," and "object." Because one sentence might contain more than one statement, a correct syntactic analysis is paramount for predicate-argument structure (PAS) generation. Our experiments showed that no single parser is consistently reliable enough for PAS extraction. Thus, our PAS extraction component can work with the results of three different parsers: RASP (Briscoe et al., 2006), MiniPar (Lin, 1998), and SUPPLE (Gaizauskas et al., 2005). The best results to recognize textual entailment were achieved using RASP, thus we omit reporting results for the other parsers in this paper.

A *PAS extractor* component applies a custom rule set for each of these parsers in order to determine subject, verb, and object of a statement.

3.2 Computing Beliefs

The core of our system is the *Fuzzy Believer* component. Its tasks are:

1. Identify a topic for each statement.
2. Compute the fuzzy representation for each statement to identify polarity.
3. Process fuzzy information for each topic.

Identifying Domains. The first step is to group the statements into *domains* according to their topics. These domains constitute the basic sets for the fuzzy operations performed later on; basically, they partition the statement space into individual domains, which can be processed independently. Every domain represents one topic identified by the extracted PASs.

To determine if a statement fits into an existing domain, we use *heuristics* to measure the semantic proximity of each new statement with the statements in all existing domains. For this, the system applies two main heuristics: (1) A WordNet (Fellbaum, 1998) related heuristic, and (2) a substring heuristic.

These heuristics compare the PAS elements of one statement with the elements of the other statements in one domain and return a value representing how similar the heuristics consider the two PAS elements. A runtime option defines if *strict* matching is neccessary to include a new statement in a domain, or if a more *lenient* matching is sufficient. For a strict match, the new statement's PAS must be similar to all existing statements within a domain. In case of a lenient match, the new statement needs only to be similar to one statement of a domain, essentially implementing a transitive relation on the domain elements.

To cause a match between two statements, at least two parts of their corresponding PAS structures must be similar enough. That means, the value assigned by a heuristic must exceed the defined threshold for either subject and object, subject and verb, or verb and object.

This approach permits assigning a statement to more than one domain. If a new statement does not fit into any of the existing domains, a new domain is dynamically

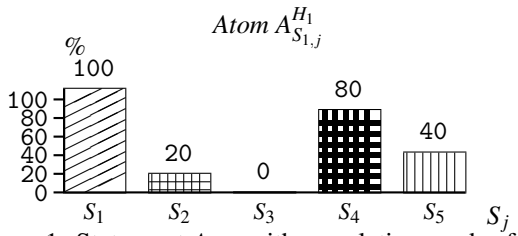

Figure 1: Statement $A_{S_{1,j}}$ with correlation grades for all statements in the domain $(S_1,\ldots,S_5)$ as computed by heuristic H_1

created, initially containing this statement.

Each domain contains all statements that have the same or opposite meaning. In other words, we try to identify each fact in the world and arrange all statements concerning this fact in one domain.

Identifying Polarity. In the next step, the statements gathered for each domain have to be evaluated by identifying their polarity. The goal is to identify opposing statements by using different fuzzy heuristics. The fuzzy representation μ_{S_i} of a statement S_i contains the degrees of similarity of this statement with all other statements within the same domain. Each degree is normalized to a fuzzy value in the $[0,1]$-interval and can be interpreted as the semantic distance between two statements. Figure 1 shows the fuzzy representation of a statement S_1 within a domain containing five statements $(S_1,\ldots,S_5)$. The fuzzy sets are interpreted in a possibilistic fashion: A fuzzy value of 0 indicates no possible semantic similarity between the two statements, while a value of 1.0 indicates the highest possibility of similarity between them. In the current implementation, only one heuristic is used. It compares the verbs of two statements using their WordNet semantic distance to find synonyms and antonyms.

To recognize textual entailment, the only fuzzy operation necessary is *merging*. The merging is done directly on the fuzzy set representation of each statement, which has been generated as described above.

Based on the fuzzy representation, the merge operation groups all statements into one class, if a threshold of semantic similarity is reached. Usually, merging all statements leads to two classes within each domain, one containing statements about a topic and the other one containing opposing statements about this topic. If *Text* and *Hypothesis* were grouped into one domain we have *Entailment* or Contradiction. Otherwise we label the pair as *Unknown*.

4 Evaluation

Results from our system can be seen in Table 2. The runs differ only in the configuration of the parameters. Since we compare predicate-argument structures, one parameter is the fuzzy threshold defining when we consider two PAS elements similar. In the first run, this threshold is set to 0.5 whereas in the second run we used a more strict threshold of 0.7. In the third run we lowered the fuzzy threshold for merging from 0.6 to 0.4.

	Run IDs		
	1.	2.	3.
Accuracy 2-way	0.51	0.51	0.54
Accuracy 3-way	0.43	0.41	0.43

Table 2: Accuracy of our system for 3 runs with different parameter settings

Detailed results for the 3-way task can be seen in Table 3. This ranks our system in the lower 25% of all participants for this task. As can be seen the most errors occured by classifying relations as "unknown" whereas they are actually "entailed".

We also ran our system on a subset of previous RTE data (Bar-Haim et al., 2006) and achieved accuracy of 58% for the 2-way classification task using the MiniPar parser (Krestel et al., 2007b). For last years PASCAL challenge we participated with the same system in the 3-way classification task and achieved accuracy of 42% using MiniPar and 45% using Rasp.

5 Conclusions

We showed that our Fuzzy Believer system can be used to solve textual entailment tasks. The extension beyond reported speech and the generality of our approach however limit the success. We are not able to compete against highly specialized systems developed solely for RTE. Essentially, the Fuzzy Believer was designed to build knowledge bases from large amounts of complete newspaper articles, while retaining a prescribed degree of consistency through recognizing and removing inconsistent statements. These capabilities are not used at all within the RTE setting, where a statement contains typically only two sentences each.

Results of our system on the RTE data of previous years reveal an increasing level of difficulty in the task *and* an improvement of the participating systems.

References

Afzal Ballim and Yorick Wilks. 1991. *Artificial Believers: The Ascription of Belief.* Lawrence Erlbaum Associates, Inc.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proc. of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*.

		System Response			
		Entailment	Unknown	Contradition	Total
Gold Standard	Entailment	207	234	59	500
	Unknown	127	174	49	350
	Contradiction	70	47	33	150
	Total	404	455	141	1000

Table 3: Confusion matrix for the 3-way task

E. Briscoe, J. Carroll, and R. Watson. 2006. The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. 2002. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proc. of the 40th Anniversary Meeting of the ACL*.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

R. Gaizauskas, M. Hepple, H. Saggion, M. A. Greenwood, and K. Humphreys. 2005. SUPPLE: A practical parser for natural language engineering applications. In *Proceedings of the 9th International Workshop on Parsing Technologies (IWPT2005)*, Vancouver.

Ralf Krestel, René Witte, and Sabine Bergler. 2007a. Creating a Fuzzy Believer to Model Human Newspaper Readers. In Z. Kobti and D. Wu, editors, *Proc. of the 20th Canadian Conference on Artificial Intelligence (Canadian A.I. 2007)*, LNAI 4509, pages 489–501, Montréal, Québec, Canada, May 28–30. Springer.

Ralf Krestel, René Witte, and Sabine Bergler. 2007b. Processing of Beliefs extracted from Reported Speech in Newspaper Articles. In *Proc. of Recent Advances in Natural Language Processing (RANLP-2007)*, Borovets, Bulgaria, September 27–29.

Dekang Lin. 1998. Dependency Based Evaluation of MINIPAR. In *Proc. of the Workshop on the Evaluation of Parsing Systems, First Intl. Conf. on Language Resources and Evaluation*.