The Role of Semantics in Recognizing Textual Entailment

Catherine Blake, Wu Zheng, Kyle Painter, Walker Weyerhaeuser
Graduate School of Library and Information Science
University of Illinois at Urbana Champaign
Illinois, NC 27599-3360
clblake@illinois.edu

Abstract

Systems designed to recognize textual entailment typically employ both syntax and semantics. Our goal in this paper is to explore the degree to which semantics alone can be used to accurately detect entailment so that we can gain a better understanding of this single component within an entailment system. This paper reports the knowledge-bases considered and selected for person names, locations and organizations and the results of the system when used on the Recognizing Textual Entailment (RTE) Track of the Text Analysis Conference (TAC).

1 Introduction

Systems designed to recognize textual entailment typically employ both syntax and semantics. There has been a recent trend toward the use of syntax in that fewer than half of the submissions in the first Recognizing Textual Entailment challenge (RTE1) employed syntax (13/28, 46%) (Dagan, Glickman, & Magnini, 2005), but more than two-thirds (28/ 41, 68%) of the second RTE challenge (RTE2) submissions employed syntax (Bar-Haim et al., 2006). Moreover the RTE2 challenge results suggested that systems which employed deep language features, such as syntactic or logical representations of text, could outperform the purely semantic overlap approach typified by BOW. Earlier findings such as (Vanderwende, Coughlin, & Dolan, 2005) also suggest that sentence structure plays an important role in recognizing textual entailment and paraphrasing accurately.

In a previous participation of RTE (RTE3), we too designed a system that employed syntactic features to identify entailments (Blake, 2007). The resulting system, which did not use background knowledge, had an accuracy of 60.50 and 65.87% and average precision of 58.97 and 60.96% in $RTE3_{Test}$ collection, suggesting that sentence structure alone can improve entailment accuracy by 9.25 to 14.62% over the baseline majority class. For this year's challenge, our goal was to develop an entirely different system based on semantics only which would enable us to assess to the extent to which semantics alone can help to identify entailment. The longer term goal is to combine what we learn from this system with the syntax only system.

One way to incorporate semantics into an entailment system is to unify morphological variants of the same word. In some cases, such as in Latin it is "… possible to enumerate all of the 262 suffixes that are needed to stem correctly verbs of all tenses and moods, giving a total of no less than 346 suffixes that should be removed from Latin words in order to stem them correctly." For English, systems can use information resources that enumerate each morphological variant, such as the SPECIALIST lexicon (National Library of Medicine, 2000), or algorithms that remove prefixes and suffixes (also called stemming). Although morphology adequately accounts for surface level lexical differences between words, we ultimately discarded this strategy due to the limitations with respect to resolving synonymous terms.

The proposed system uses morphological comparisons for non-stop words only when the term does not exist in the knowledge-bases described in section 3. The system first identifies entities (people, places, and organizations) from the entailment sentences and the full-text of the news articles in RTE6. For example the system should identify George Bush from the sentence 'Sheehan knows nothing can bring back her son, but she wants to talk to *President George W Bush* (document APW_ENG_20050811.0720, sentence 4). The system also unifies alternative representations of the same entity. For example, the system would unify references to Bush in the following sentences, which are made in the same article as the previous sentence, 'Nearly 40 Democratic members of Congress have asked *Bush* to talk to her.' (sentence 35) and 'On Wednesday, a coalition of anti-war groups in Washington also called on *Bush* to speak with Sheehan, who they say has helped to unify the peace movement.' (sentence 36). In addition to nouns, the system uses semantics to identify verbs and unify verb phrases from an entailment sentence pair. For example, *checkout* may entail *to purchase*. Lastly, the system resolves pronoun references and numbers.

2 Related work

In RTE4, 18 out of the 26 teams used WordNet (Giampiccolo, Dang, Magnini, Dagan, & Dolan, 2008). In RTE5, based on our observation, all participants used WordNet in their entailment systems. Wikipedia, DIRT (Lin & Pantel, 2001), and VerbOcean are among the resources that are used most in RTE systems. Although these resources were usually used together with other lexical-syntactic features, the ablation test introduced in RTE5 makes it possible to see to what extent does external knowledge help. Clark and Harrison (Clark & Harrison, 2009) reported that using only DIRT and WordNet, their system performed almost as good as the full system which also include syntactic features. WordNet helps substantially in recognizing entailments. Breck (Breck, 2009) also reported that WordNet greatly improved his system's performance. However, WordNet does not help in all systems. In (Malakasiotis, 2009), the author reported that with WordNet ablated, the system performance even improved. For most of the cases, only marginal improvements were reported using external knowledge bases (Ferrández, Munoz, & Palomar, 2009; Ren, Ji, & Wan, 2009; Varma et al., 2009).

Besides measuring similarities between hypothesis and text, knowledge bases are also used in co-reference resolution (Harabagiu, Bunescu, & Maiorano, 2001; Ponzetto & Strube, 2006), word sense disambiguation (Mihalcea, 2007), and entailment rule learning (Aharon, Szpektor, & Dagan, 2010; Tatu & Moldovan, 2006) which can potentially help improve entailment performance.

3 System architecture

The proposed system uses a knowledge-based approach to identify people, places and organizations from both the entailment sentence pairs and the original news stories in the RTE6 corpora.

3.1 Information resources

Several Knowledge-Based KB's were considered, including the Stanford NER (nlp. stanford.edu/software/CRF-NER.shtml), tools from GATE (gate.ac.uk) and the Illinois Named Entity Tagger (cogcomp.cs.illinois.edu/download/ software/28), but none of those resources were used in the final system. FrameNet was also considered (Burchardt, Pennacchiotti, Thater, & Pinkal, 2009) as was a custom-built knowledge-base approach (Hickl et al., 2006). The latter strategies were not selected due to time constraints. We targeted knowledge sources that would provide good coverage for people, places and organization entities.

To identify people, both wordnet (Version 3.0 http://wordnet.princeton.edu/) (Miller, 1995), and YAGO (http://www.mpi-inf.mpg.de/yago-naga/ yago/) were considered. YAGO, which is part of the YAGO-NAGA project at the Max-Planck Institute for Informatics in Saarbrücken, Germany was selected because initial experiments suggested that the KB was more comprehensive than Wordnet with respect to people. YAGO is essentially a bootstrapped ontology, where information in wordnet is used to infer new from the web; thus YAGO and WordNet are not mutually exclusive. However, in constrast to

WordNet which is manually constructed, the bootstrapping strategy can lead is errors. Although the accuracy was reported at 95% (Suchanek, Kasneci, & Weikum, 2008), our preliminary experiments using the entire collection failed as there were too many spurious relations. URL entries within YAGO, which would not match the RTE texts were excluded, as were connections to WordNet (we chose instead to use WordNet directly). Such pruning drastically improved system performance.

The system identifies title information, such as Prime Minister, in addition to a person's name. Title abbreviations were collected from the Oxford English Dictionary list of abbreviations: (www.indiana.edu/~letrs/help-services/QuickGuides/oed-abbr.html) and a list of military ranks (copyediting-grammar-style.suite101.com/article.cfm/military_titles_and_abbreviations_in_ap_style).

Several location resources were considered including the US Postal Service (www.usps.gov), a world gazetteer (www.world-gazetteer.com/wg.php?men=home&lng=en&des=wg&srt=npan&col=abcdefghinoq&msz=1500&geo=0), the NGA GEONet Names Server (http://earth-info.nga.mil/gns/html/index.html) and Geonames (http://www.geonames.org). The postal service resource was discounted due to the restricted geographical coverage. Moreover, US locations are reasonably covered in WordNet and the GNS does not have country names and US locations.

The current system uses Geonames including city names, large and small towns, country names and geographic features such as streams, wells, and roads. The system includes the following WordNet synsets to include countries in Africa, European Union, Europe, Asia, North America, South America and general countries. American state and city names were used from WordNet because colloquialisms were well coveraged. Experiments are underway to determine the overlap and coverage within each of these resources.

Organizations were identified using YAGO, because preliminary experiments suggested that YAGO was more comprehensive than Freebase (www.freebase.com/) or DBpedia (dbpedia.org/). We are still in experimenting with the trade-off between YAGO coverage and accuracy and may identify additional external resources such as the relations identified by the DIRT system or TextRunner.

3.2 Person name resolution

The majority of people captured in YAGO (>99%) were associated with fewer than seven synonyms. One exception is George W. Bush, where a range of synonyms were captured including Bushists, Duhbbya, Dubbiya, Dubbya, Dubyuh, G.W.B, G.w.bush, GeorgeWBush, Baby Bush, Bush Jr, Bush Junior, Bush jnr, Dubya B, Dubya Bush, G.W Bush, G Dub, GW Bush, George Bush, George W. This example demonstrates that terms within YAGO may not uniquely identify an individual. For example President Bush may refer to Bush junior or Bush senior. Single word names is particularly problematic in news stories because journalists frequently refer to an individual using only his or her last name.

The proposed system includes the following strategy which was motivated by the observation that news journalists frequently provide the full name of a person within the news article, and then use just their last name in subsequent references. Before the entailment sentences are processed, the system identifies all names with more than one word from the original news stories in the RTE6 corpus. The system then uses this compiled list of people to identify other references in the article to the person's last name.

This strategy would fail if the story includes more than one person with the same last name, for example a news story that reported both Barrack and Michele Obama. We have yet to conduct a comprehensive study to measure how many articles report two individuals with the same last name, but our early tests revealed that this was uncommon, probably because a human reader of the story would also struggle with the ambiguous reference. In this implementation only the last name was used and if the article did contain two individuals with the same last name then both names would be assigned to the reference. We later observed subsequent references to first names, but this was not implemented in time for the challenge.

The people in the YAGO knowledge are usually public figures, such as politicians and entertainers. However, news story can also feature individuals who are not public figures. Thus, after the system has identified public figures from YAGO and abbreviated references to public figures, a back-up strategy is employed to ensure adequate coverage of people entities. Specifically, the system searches for trigger terms that may indicate a person's name. Trigger terms include male and female names from US Census data and titles such as Mr. and Dr. The first names of public figures were ranked by frequency and then manually inspected to supplement names in the US Census data. The system first searches for a trigger term and then proceeds to check each subsequent term for initials, terms starting with an upper case letter or name prefixes such as der.

3.3 Pronoun resolution

The system applies pronoun resolution for person names only (not locations or organizations). Pronouns can be more accurately mapped if the gender of the person is know, thus the gender of the person entities is established using either the person's first name or title. In cases where the first name may be either male or female then the person is marked as unknown. Pronouns such as he and she captures people who are already public figures, such as politicians, and entertainers. However, a news story can also feature an individual who is not a public figure.

3.4 Number resolution

In preliminary experiments revealed that inconsistencies between numeric values in the hypothesis and test sentences were highly indicative of a false entailment. The system therefore treats numbers as entities and unifies numeric and textual representations of numbers using the terms from the Unified Medical Language System as described in http://lexsrv3.nlm.nih.gov/LexSysGroup/Projects/lexicon/2003/release/LEXICON/NUMBERS/number.grammar.txt.

3.5 System submissions

The system applies the entity recognition strategies described in sections 3.1-3.4 in the following order: identify person names, resolve name references, identify locations and organizations, unify numbers and resolve pronouns from all documents in the RTE6 corpus and the hypothesis sentence (entities for the candidate test sentence were determined from the corpus). Remaining non-stopword words from the candidate sentence pairs are then compared using an exact match, a base form match, or if terms are within the same synset in WordNet. The remainder of this section will refer to the non-entities as remaining terms.

Scores were assigned for each entity depending on the number of occurrences in the hypothesis and test sentences. For example if the hypothesis sentence required more than one person entity and the candidate sentence matched more than one person entity the system would assign a score of 10 times the proportion of covered person entities; if the hypothesis sentence contained no person entities then the system assigned a score of -1; and all other cases the system assigned a score of -10 x the number of distinct entities required in the hypothesis sentence. The scoring system for persons was used for locations, organizations, numeric references and remaining terms and the over-all score for each hypothesis-candidate sentence pair was calculated.

Experiments on the RTE6 development collection revealed that a high number of remaining terms (i.e. the non-stop word terms that were not identified as entities), appeared to indicate that the candidate sentence was not an entailment. Specifically if more than 50% of the remaining terms in the hypothesis sentence were not covered by terms in candidate sentence, then the candidate sentence was unlikely to entail the hypothesis. Similarly if five or more remaining terms were not covered in the candidate sentence then it was unlikely to be entailed. Candidate sentences that did not meet both of these criteria were immediately considered as non-entailments.

Development evaluations using the RTE6 development collection revealed that several sentences were annotated as entailments even though the test sentence did not contain a reference (either directly or indirectly via pronoun), to entities in the hypothesis sentence. The description of how annotations were assigned called this

"background knowledge". To adjust for this definition of entailment, the system matches entities from the hypothesis in any part of the document rather than just the sentence under question. Although this change to the system improved performance, we posit that further discussion is warranted with respect to establishing entailment on a sentence-sentence versus sentence-document basis.

4 Results and discussion

Three RTE6 submissions were made. Thresholds were set based on the RTE6 development collection only, and then applied to the unseen test collection. The first submission used a threshold to favor precision, the second balanced precision and recall and the third favored recall. The micro averaged scores are shown in table 1.

Run	Precision	Recall	F-measure
1	46.25	23.49	31.16
2	38.11	26.46	31.23
3	31.53	33.86	32.65

Table 1. Main task results for RTE6 test set

To place these results in context, 18 teams submitted a total of 48 runs to the main task. The micro-averaged F-measure ranged between 0.1160 and 0.4801, with a median of 0.3372. Thus, the F-measure of our system was slightly below the median.

As with participating in any challenge, the system design strongly reflects the task at hand. However, one of the system design characteristics in this system - specifically that the system checks entities anywhere in the document from which candidate sentences were drawn, rather than just within the candidate sentences - was inconsistent with our mental-model of how a sentence-sentence entailment system would operate. If the challenge continues to be framed as a sentence-sentence comparison then perhaps annotators should be shown the same set of hypothesis and test sentences that the system considers. We posit that further discussion is warranted with respect to the sentence to sentence versus sentence to document annotations and look forward to such discussions during the workshop.

5 Closing comments

Our goal this year was to explore the degree to which semantics alone can accuracy determine entailment. The results thus far suggest that semantics alone are inadequate. With that said, we are still in the process of evaluating the coverage of these knowledge sources and have not yet exhausted alternative entity recognition strategies.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under Grant IIS-0812522. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

References

Aharon, R. B., Szpektor, I., & Dagan, I. (2010). Generating entailment rules from framenet. In Proceedings of the ACL 2010 Conference Short Papers (pp. 241-246). Association for Computational Linguistics.

Bar-Haim, R., Dagan, I., Dolan, B., Ferro, L., Giampiccolo, D., Magnini, B., et al. (2006). In The Second PASCAL Recognising Textual Entailment Challenge. Paper presented at the In Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment, Venice, Italy.

Breck, E. (2009). A simple system for detecting non-entailment. In Proceedings of the Text Analysis Conference (TAC).

Burchardt, A., Pennacchiotti, M., Thater, S., & Pinkal, M. (2009). Assessing the impact of frame semantics on textual entailment Natural Language Engineering, 15(4), 527-550.

Clark, P., & Harrison, P. (2009). An inferencebased approach to recognizing entailment. In Preproceedings of the Text Analysis Conference (TAC).

Dagan, I., Glickman, O., & Magnini, B. (2005, 11 - 13 April 2005). In The PASCAL Recognising Textual Entailment Challenge. Paper presented at the In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Southampton, U.K.

Ferrández, O., Munoz, R., & Palomar, M. (2009). Alicante University at TAC 2009: Experiments in RTE. In Proceedings of the TAC 2009 Workshop on Textual Entailment.

Giampiccolo, D., Dang, H. T., Magnini, B., Dagan, I., & Dolan, B. (2008). The fourth pascal recognizing textual entailment challenge. In Preproceedings of the Text Analysis Conference (TAC).

Harabagiu, S. M., Bunescu, R. C., & Maiorano, S. J. (2001). Text and knowledge mining for coreference resolution. In Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001 (pp. 1-8). Association for Computational Linguistics.

Hickl, A., Williams, J., Bensley, J., Roberts, K., Rink, B., & Shi, Y. (2006). In Recognizing Textual Entailment with LCC's GROUNDHOG System (pp. 80-86). Paper presented at the The Second PASCAL Recognising Textual Entailment Challenge (RTE-2).

Lin, D., & Pantel, P. (2001). DIRT-discovery of inference rules from text. In Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining (pp. 323–328). Citeseer.

Malakasiotis, P. (2009). AUEB at TAC 2009. In Preproceedings of the Text Analysis Conference (TAC).

Mihalcea, R. (2007). Using wikipedia for automatic word sense disambiguation. In Proceedings of NAACL HLT (Vol. 2007).

Miller, G. (1995). WordNet: A Lexical Database for English. Communications of the ACM, 38(11), 39-41.

National Library of Medicine. (2000). The SPECIALIST Lexicon from www.nlm.nih.gov /pubs/factsheets/umlslex.html

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (pp. 192-199). Association for Computational Linguistics.

Ren, H., Ji, D., & Wan, J. (2009). WHU at TAC 2009: A Tri-categorization Approach to Textual Entailment Recognition. In Preproceedings of the Text Analysis Conference (TAC).

Suchanek, F.M., Kasneci, G., & Weikum, G. (2008). YAGO:A Large Ontology from Wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), 203-217.

Tatu, M., & Moldovan, D. (2006). A logic-based semantic approach to recognizing textual entailment. In Proceedings of the COLING/ACL on Main conference poster sessions (pp. 819-826). Association for Computational Linguistics.

Varma, V., Bharat, V., Kovelamudi, S., Bysani, P., Santosh, G. S. K., Kumar, K., & Maganti, N. (2009). IIIT Hyderabad at TAC 2009. In Proceedings of Text Analysis Conference 2009 (TAC 09).

Vanderwende, L., Coughlin, D., & Dolan, B. (2005). In What Syntax can Contribute in Entailment Task. Paper presented at the The PASCAL Recognising Textual Entailment Challenge.