Metaphor in Textual Entailment

Rodrigo Agerri
School of Computer Science, University of Birmimgham
B15 2TT Birmingham, UK
r.agerri@cs.bham.ac.uk

Abstract

Metaphor understanding in Computational Linguistics has largely been focused on the development of stand-alone prototypes for which only small-scale evaluations are carried out. This has made difficult the inclusion of metaphor in the development of natural language processing applications. However, dealing with metaphor properly is ultimately crucial for any automated language technology that is to be truly human-friendly or able to properly appreciate utterances by humans. This paper proposes to bring metaphor into the Recognizing Textual Entailment task. By doing so, the coverage of textual entailment systems would be broadened and metaphor research would benefit from the textual entailment evaluation framework.

1 Introduction

Using metaphorical language is common in most forms of everyday language, from ordinary conversation, "having ideas in the back of the mind", through newspaper articles, "global oil prices clung near their highest levels", to scientific articles, "the variable N goes from 1 to 100". Metaphor is important in part because it is an economical and directly appealing way of talking about many sorts of subject matter in human life, such as time, money, relationships, emotions, politics, etc. Most importantly, metaphor can have major effects on what can be properly inferred from an utterance or passage.

Most of the development of natural language processing (NLP) applications has been focused on specific tasks such as Information Retrieval (IR) and Question Answering (QA), largely ignoring the question of figurative use of language. Moreover, considering the inherent difficulty in evaluating deep approaches to language in a large-scale manner, up to date there is not a common evaluation framework, corpora or other resources for metaphor processing. It certainly has not helped that most of the computational developments on metaphor processing have largely been stand-alone systems that are not empirically evaluated on a large scale (Fass, 1997; Falkenhainer et al., 1989; Hobbs, 1992; Martin, 1990; Barnden et al., 2003).

This paper proposes to address this by adapting the Recognizing Textual Entailment (RTE) framework for metaphor interpretation. RTE aims to be an abstract and generic task that captures *major semantic inference* needs across applications (Dagan et al., 2007). RTE is considered central for the development of intelligent yet robust natural language processing systems because most of the semantic inference needed in natural language applications such as QA and IE can be characterized as problems in RTE (Dagan et al., 2007). Intuitively, textual entailment consists of determining whether a hypothesis can be inferred from a given text. The textual entailment operational definition is a directional relation between two text fragments, *Text* T and *Hypothesis* H such that T entails H if humans reading T and considering H will infer that H follows from T. An illustration can be given by example 1560 of the RTE-1 dataset (which involves a metaphor in the use of 'incubate'):

T: The technological triumph known as GPS was incubated in the mind of Ivan Getting.

H: Ivan Getting invented the GPS.

As in other NLP applications, figurative language has merely been noted as problem in the RTE field. However, we believe that RTE provides a general evaluation framework for semantic processing which can be adapted for the computational testing and evaluation of theories that aim to explain the semantic inferences involved in metaphor resolution. Furthermore, including metaphor in the RTE task may improve the performance and scope of textual entailment systems, which in turn may allow to bring metaphor into the development of NLP systems.

2 The Role of Metaphor in Textual Inference

Metaphorical use of language crucially affects the inferences that can be drawn from a text. Including metaphor in textual entailment would amount to establish whether a hypothesis H can be inferred from a text T, where (at least) T contains a metaphorical expression whose processing is relevant to judge the entailment.

2.1 Re-formulating the problem

It is usual to assume a view of metaphor understanding as involving some notion of properties and relations of events that are transferred from a source domain into a target domain. In this view, a (declarative) metaphorical text conveys information about some target domain by means of a number of correspondences between entities in the source and the target domains. Lakoff and associates argue that source to target correspondences are part of more general schemes called "conceptual metaphors" (Lakoff, 2004) which we call "metaphorical views". For the GPS T-H pair in the previous section, a metaphorical view such as MIND AS PHYSICAL SPACE would capture the correspondence between *mind* in the source to *special container or incubator* in the target.

Most of the computational approaches to metaphor processing have focused on the development of reasoning systems which take a metaphorical expression as input and perform some reasoning to prove the correct output – previously given by the researcher. The difficulties in scaling-up and the lack of empirical evaluation have been the main chronic problems of metaphor understanding systems. Furthermore, a task consisting of providing an interpretation of a text such as T above – the GPS being incubated in the mind of Ivan Getting – is very complex because it needs to consider and resolve what the "correct interpretation" is from the number of possible interpretations that can be conveyed by (the use of) 'incubating'. Conversely, the task becomes easier if the task faced is to judge whether H follows from T; in the GPS example above, H sets up the context to interpret the metaphorical expression in T, which in turn would help to correctly judge that the GPS was invented by Ivan Getting ('incubate' can be used instead of 'develop', 'invent', etc.). Thus, a slightly modified H would presumably lead to connotations of the metaphorical use of 'incubate' previously not considered:

T: The technological triumph known as GPS was incubated in the mind of Ivan Getting.

H': Ivan Getting accidentally invented the GPS.

A reasonable judgment is that H is not entailed by T, since an incubation process in this particular example would seem to indicate a careful nurturing of ideas that were brought slowly into life and so on, within Ivan Getting's mind. The modified H brings extra connotations of the metaphorical use of 'incubate' that are crucial to establish the negative entailment judgment.

2.2 Metaphor in RTE Challenges

Even though annotators aimed to filter out metaphorical uses of language from the RTE datasets (Zaenen et al., 2005), some metaphorical texts have eluded the annotators' selection policies (Bos and Markert, 2006). Our study of RTE datasets looking for pairs in which resolving a metaphorical expression was relevant for the entailment judgment uncovered few and mostly conventional metaphors. We have focused on 10 pairs in RTE-1 and 9 in RTE-2. Some of them are listed here:

T1: Lyon is actually the gastronomic capital of France.

H1: Lyon is the capital of France.

T2: The upper house of the Russian parliament has approved a controversial bill to tighten state control over non-governmental organisations (NGOs).

H2: Russian parliament closes NGOs.

T3: Convinced that pro-American officials are in the ascendancy in Tokyo, they talk about turning Japan into "the Britain of the Far East."

H3: Britain is located in the Far East.

T4: Stocks rallied for a second session Thursday, boosted by falling oil prices and ongoing relief that the presidential election has passed without incident.

H4: The falling oil prices had a positive impact on stocks.

An evaluation of the systems' accuracy for the pairs involving metaphor was performed to test if there was any significant difference with respect to the overall accuracy results reported in the official RTE challenges. The RTE-1 results are not publicly available, so the study is restricted to the 7 runs which were made available – including 4 of the best 5 systems. Table 1 shows the official overall accuracy results and the results of the evaluation over the 10 pairs involving metaphor:

Author (Group)	Overall	Metaphor
Bayer (MITRE)	0.586	0.4
Herrera (UNED)	0.566	0.2
	0.558	0.2
Bos (Rome/Leeds)	0.563	0.2
	0.555	0.1
Newman (Dublin)	0.563	0.1
	0.565	0.6

Table 1: RTE-1 Accuracy Comparison.

Although the sample of metaphor pairs is fairly small, table 1 shows that there is a trend for the accuracy to be significantly lower when metaphor is involved than for the overall results (which agrees with Bos and Markert's (2006) diagnostic).

RTE-2 results are publicly available and for this study 8 runs of the best scoring systems (only those which also submitted the average precision results are considered) and 2 with lower accuracy were chosen. Table 2 confirms the trend suggested by table 1, namely, that the accuracy score is lower when the judgement depends on processing metaphorical uses of language. How significant are these results? For the RTE-1 pairs, a Fisher's test of independence establishes that for 5 out of the 7 runs the difference in performance is statistically significant at the 0.05 level. The same results were obtained for 7 of the 10 RTE-2 runs compared in table 2.

Author (Group)	Overall	Metaphor
Hickl (LCC)	0.7538	0.4444
Tatu (LCC)	0.7375	0.5555
Zanzotto (Milan/Rome)	0.6388	0.2222
Adams (Dallas)	0.6262	0.3333
Bos (Rome/Leeds)	0.6162	0.1111
Kouylekov (Trento)	0.6050	0.1111
Vanderwende (Stanford)	0.6025	0.1111
Herrera (UNED)	0.5975	0.1111
Clarke (Sussex)	0.5275	0.1111
Newman (Dublin)	0.5250	0.4444

Table 2: RTE-2 Accuracy Comparison.

3 Discussion

Although only few pairs containing fairly conventional metaphors were uncovered from the RTE datasets, the results obtained confirm the hypothesis that the ability to process metaphor would broaden the coverage of textual entailment systems, thereby improving their overall performance. It should also be considered that achieving statistical significance is harder when the overall results are not that high, as shown by the fact that we get statistical significance for Hickl's system and not for Newman's and Adam's.

Moreover, it is envisaged that the relatively good performance of some of the systems (e.g, Hickl and Tatu) is due to the relative lack of open-ended metaphors in the pairs used for the analysis. This also shows that shallow techniques can be fruitful for processing conventional metaphor. However, open-ended metaphors may pose more complex problems. For example, a fairly deep analysis may presumably be needed to extract the metaphorical connotations conveyed by 'incubate' (about the source to target transfer of *carefully growing and nurturing*) to correctly judged the lack of entailment for the modified hypothesis 'Ivan Getting accidentally invented the GPS'. This is also true for metaphors about "deepest recesses of the mind" (in RTE-1 dataset), etc. This type of open-ended metaphors have been subjected to a in-depth analysis (both formal and computational) within the ATT-Meta system and approach for metaphor interpretation (Agerri et al., 2007; Barnden et al., 2003). Adapting it for textual entailment may facilitate the processing of open-ended metaphor in a textual entailment task.

Metaphor understanding systems have not aimed to be empirically evaluated on a large-scale, but have chosen to focus instead on the in-depth analysis of small number of examples. As a consequence, there are not common resources such as corpora or shared task evaluation exercises for metaphor resolution. In order to make use of the RTE evaluation framework to promote empirically-based research on metaphor understanding, the first task would aim to build datasets that for the first time would allow researchers to train and (empirically) evaluate their systems. An obvious strategy would be to follow RTE guidelines with the additional requirement that at least T should contain a metaphorical expression relevant to judge the entailment.

The RTE evaluation framework has the advantage that it is theory neutral, namely, it does not depend on any semantic formalism and works on open domain data. However, the RTE evaluation framework has the disadvantage of being a "blackbox" type of evaluation. It makes very difficult to isolate the semantic task from the task of retrieving the necessary background knowledge (Zaenen et al., 2005; Bos, 2008). Furthermore, it is not designed to measure performance on specific semantic phenomena, and it is therefore difficult to know why a system is working correctly or incorrectly. For example, all but one of the RTE-1 runs studied incorrectly judged the T1-H1 pair to be true (about 'gastronomic capital'). It is difficult to be certain that this was solely due to a lack of ability to deal with metaphor instead of a problem about noun modifiers. However, there is not currently a suitable alternative to RTE semantic evaluation as trying to isolate the semantic task (e.g., metaphor) from background knowledge usually results in using artificial examples. On the bright side, the RTE framework will allow metaphor research to grapple more extensively than before with the interactions between metaphor and other language phenomena.

4 Concluding Remarks

The aim of this paper is two fold: Firstly, it provides evidence showing that the ability of processing metaphor may improve the performance of textual inference systems. Secondly, it argues that RTE may provide a much needed general semantic framework for common evaluations and computational testing of theories that aim to explain open-ended usages of metaphor in everyday text. The ATT-Meta approach and system to metaphor interpretation may be adapted for this particular task (Barnden et al., 2003). Including metaphor processing in textual entailment systems can also promote the inclusion of metaphor resolution in NLP applications such as Question Answering, Document Summarization or Information Retrieval.

Acknowledgments

Thanks to the RTE-1 participants that made available their results. Special thanks to John Barnden and Alan Wallington for their comments and suggestions.

References

Agerri, R., J.A. Barnden, M.G. Lee, and A.M. Wallington. 2007. Metaphor, inference and domain independent mappings. In *Proceedings of Research Advances in Natural Language Processing (RANLP 2007)*, pages 17–24, Borovets, Bulgaria.

Barnden, J., S. Glasbey, M. Lee, and A. Wallington. 2003. Domain-transcending mappings in a system for metaphorical reasoning. In *Companion Proceedings of the 10th Conference on the European Chapter of the Association for Computational Linguistics (EACL-03)*, pages 57–61.

Bos, J. and K. Markert. 2006. Recognizing textual entailment with robust logical inference. In Quiñonero-Candela, J., I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *MLCW 2005*, volume 3944 of *LNAI*, pages 404–426. Springer-Verlag.

Bos, J. 2008. Lets not argue about semantics. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco.

Dagan, I., O. Glickman, and B. Magnini. 2007. The PASCAL Recognising Textual Entailment challenge. In Quiñonero-Candela, J., I. Dagan, B. Magnini, and F. d'Alché Buc, editors, *MLCW 2005*, volume 3944 of *LNAI*, pages 177–190. Springer-Verlag.

Falkenhainer, B., K.D. Forbus, and D. Gentner. 1989. The structure-mapping engine: algorithm and examples. *Artificial Intelligence*, 41(1):1–63.

Fass, D. 1997. *Processing metaphor and metonymy*. Ablex, Greenwich, Connecticut.

Hobbs, J.R. 1992. Metaphor and abduction. In Ortony, A., J. Slack, and O. Stock, editors, *Communication from an Artificial Intelligence Perspective*, pages 35–58. Springer-Verlag, Berlin.

Lakoff, G. 2004. Conceptual metaphor home page. http://cogsci.berkeley.edu/lakoff/MetaphorHome.html.

Martin, J.H. 1990. *A computational model of metaphor interpretation*. Academic Press, New York.

Zaenen, A., L. Karttunen, and R. Crouch. 2005. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL 05 Workshop on Empirical Modelling of Semantic Equivalence and Entailment*, pages 31–36.