Textual Entailment as an Evaluation Framework for Metaphor Resolution: A Proposal

Rodrigo Agerri
John Barnden
Mark Lee
Alan Wallington
University of Birmingham (UK)
email: `r.agerri@cs.bham.ac.uk`

Abstract

We aim to address two complementary deficiencies in Natural Language Processing (NLP) research: (i) Despite the importance and prevalence of metaphor across many discourse genres, and metaphor's many functions, applied NLP has mostly not addressed metaphor understanding. But, conversely, (ii) difficult issues in metaphor understanding have hindered large-scale application, extensive empirical evaluation, and the handling of the true breadth of metaphor types and interactions with other language phenomena. In this paper, abstracted from a recent grant proposal, a new avenue for addressing both deficiencies and for inspiring new basic research on metaphor is investigated: namely, placing metaphor research within the "Recognizing Textual Entailment" (RTE) task framework for evaluation of semantic processing systems.

1 Introduction

The RTE task and annual Challenges (Dagan et al., 2007), starting in 2005, have arisen as an evaluation framework for applied semantics in response to the fact that in NLP applications — such as *Question Answering* (QA), *Information Retrieval/Extraction* (IR, IE), etc. — the development of semantic algorithms and models have been scattered, or tailored to specific applications, making it difficult to compare and evaluate them within one framework. RTE is interesting because QA, IE, etc. can all be cast as RTE problems. In RTE, one text fragment, the *Text* T, is said to entail another one, the *Hypothesis* H, when humans considering T and H judge that H follows from T (perhaps only plausibly/defeasibly). Thus, entailment is a commonsense matter, not a precise logic-based one. An example of a T/H pair is as follows (metaphor in italics):

(1) T: Lyon is actually the *gastronomic capital* of France.

H: Lyon is the capital of France.

Metaphor can roughly be characterized as describing something (the *target*) as if it were something else (the *source*) to which it is perceived, or set forth, as being somehow analogous. Metaphor has long been identified as being ubiquitous in language (Goatly, 1997), including ordinary conversation, newspaper articles, popular novels, popular science writing, classroom dialogue, etc. In a study (Tech. Rept. CSRP-03-05 at our School, 2003) we found one metaphorical term per 17.3 words, averaging across various discourses of different genres. This is in line with other researchers' studies though counts vary widely because of theory-relativity, researchers' aims, and marked usage differences between genres. Gedigian et al. (2006) note that 90% of uses of a set of verbs of spatial motion, manipulation, and health in a *Wall Street Journal* corpus were metaphorical. Some metaphor examples arising in past RTE datasets are the Ts in T/H pairs (1–4), with human judgments No, Yes, No and No respectively.

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

H: Ivan Getting invented the GPS.

(3) T: Convinced that pro-American officials are in the ascendancy in Tokyo, they talk about turning Japan into *the Britain of the Far East*.

H: Britain is located in the Far East.

(4) T: Even today, *within the deepest recesses of our mind, lies* a primordial fear that will not allow us to enter the sea without thinking about the possibility of being attacked by a shark.

H: A shark attacked a human being.

Importantly, metaphor is often not just a matter of particular terms with particular metaphorical senses that are entrenched (i.e., that are commonly used, default senses; and possibly listed in dictionaries). Certainly the metaphorical senses of "capital" and "incubate" used in (1) and (2) are at least moderately entrenched. For "incubate," some dictionaries list a slow, protective sense of development (in a general, possibly non-physical sense), or for “capital” a sense like “a city [or place more generally] preeminent in some special activity” [Merriam-Websters]. But (3) shows one common type of non-entrenched metaphor, of the general form “the X of Y”, where X and/or Y are often named entities. The point of such a metaphor is typically only clear with the help of context. Reference to recesses of a mind as in (4) is common, and a lexicon could reasonably include a metaphorical sense of “recess” that was directly applicable to minds (though WordNet 3.0, e.g, does not), or include “within the recesses of [X’s] mind” as a stock phrase with a sense of relative inaccessibility or unmodifiability of a thought or feeling. But the phraseology can be productively varied: e.g., “deepest” can be omitted or replaced by “dim”, “darkest”, “foulest”, “hidden”, etc. — by *any* compatible qualifier that emphasizes hiddenness or obstacles to accessibility. And the fact that such access difficulties are being *emphasized* is a matter for general semantic reasoning about the qualifier.

2 Why Metaphor in NLP?

Generally, in metaphor understanding research, a specialized system has been fed a relatively small number of metaphorical inputs, and the correct outputs have been dictated by the researcher, (e.g. Fass, 1997; Falkenhainer et al., 1989; Martin, 1990; Barnden et al., 2003). However, metaphor in particular and figurative language in general suffers a chronic lack of shared resources for proper evaluation of systems (Markert and Nissim, 2007). In using the RTE evaluation framework, computational metaphor researchers may for the first time have sizable, shared datasets, and a uniform evaluation method based on systematically and transparently collected human judgments. Also, metaphor researchers will be challenged and inspired to connect metaphor more than before to other complex linguistic phenomena.

NLP applications that RTE serves have mostly not addressed metaphor, and neither have RTE systems themselves. Indeed, despite examples (1-4), RTE datasets have tended to avoid metaphor of more difficult types. Metaphor in general can introduce additional context-sensitivity and indeterminacy in entailment, whereas RTE Challenges have mainly concentrated on T/H pairs supporting relatively crisp, uncontroversial judgments (Zaenen et al. (2005); RTE organizers (personal communication); our own analysis of existing RTE datasets in Tech. Rept. CSRP-08-01, 2008). In fact, on examples such as (1), a system must interpret “capital” metaphorically, but as Bos and Markert (2006) reported, the inability of their system to process metaphor meant that it was incorrect on this example.

Further evidence is given by results on example (1) of the four RTE-1 systems available to us (out of 16 systems; the 4 include the 1st, 3rd and 4th best systems in terms of accuracy over whole dataset). Only one reported system run was correct. The systems mostly performed worse across (1) together with 10 other metaphor cases than on the whole dataset, with statistical significance at 0.05 level (Fisher’s independence test); only one system run gave a better performance. In RTE-2 (23 systems), analysis of 10 system runs shows that, across 9 metaphorical examples in the dataset, the systems (including the most accurate one over the whole dataset) performed worse than they did over the rest of the dataset; in 7 of the 10 RTE-2 runs compared the deficit was statistically significant at < 0.05 level (Fisher’s test, see Tech Rept. CSRP-08-01).

3 RoadMap: Datasets and Metaphor Processing

Our initiative mainly consists of **(A)**: Create a public, annotated, metaphor-focussed text dataset suitable for RTE evaluation and testing of metaphor processing systems; and **(B)**: Develop a prototype RTE system centred on processing (at least) the types of metaphor arising in (A). We will mainly address point (B) in this paper. As for A, metaphors vary along several dimensions of interest, such as: the target subject matter (e.g., Lyon's food, GPS development, Japanese foreign politics, shark fear in examples 1–4); the source subject matter; what particular, familiar metaphorical views (e.g., Idea as Living Being, in (2)) are used; whether the meaning in context is listed in dictionaries, WordNet, etc; whether the wording is (a variant of) a familiar idiom; the syntax used. Based on such dimensions, we will analyse the **types of metaphor** present in past RTE datasets and in the genres of text (e.g., newspapers) they drew from. One particular source will be the 242K-word metaphor-orientated corpus that we derived from the British National Corpus. To find metaphor examples elsewhere, we will use known metaphorical phrases and lexical/syntactic templates as seeds for automated search over general corpora or web. We will also investigate the use or adaptation of other researchers' automated detection/mining techniques (e.g. Birke and Sarkar, 2006; Gedigian et al., 2006).

4 Metaphor Processing

We will develop metaphor processing algorithms on an integrated spectrum going from relatively "shallow" forms of processing to relatively "deep" forms. The deeper are for when more inference is necessary and feasible; when less necessary or feasible, the shallower are appropriate (but they can still involve at least partial parsing, approximate semantic analysis, etc.). A significant research issue will be how to choose the attempted depth(s) of analysis and how to choose or combine different methods' results. The deeper and shallower ends of the spectrum will take as starting points our two previous metaphor-understanding approaches, from our ATT-Meta and E-Drama projects respectively (Barnden et al., 2003; Wallington et al., 2008).

For detecting metaphor, we can extend methods from our E-drama project (Wallington et al., 2008). This involves a mixture of light semantic analysis via tools such as WordNet and recognition of lexical and syntactic signals of metaphoricity (Goatly, 1997). We aim also to recognize specific idiomatic phrases and systematic variations of them. We will consider looking for semantic restriction violations, which sometimes accompany metaphor (cf. Fass (1997) and Mason (2004)), and using statistical techniques borrowed from such work as Gedigian et al. (2006).

Example (4) about "recesses" is similar to metaphor examples studied in the ATT-Meta project, and ATT-Meta-style reasoning could handle the Text. As for (2), note that the word "incubate" may have a directly usable lexicon-listed meaning (e.g., *help to develop* much as in WordNet 3.0). However, if the system's lexicon did *not* contain such a sense, ATT-Meta-style processing would apply. ATT-Meta processing would involve commonsense reasoning in source-domain terms (here, biological and other physical terms): e.g., to infer that the idea was a living being, Getting's mind was a physical region, the idea was kept warm in that region, and the idea consequently biologically developed there. Hence, there was a *relatively protracted, continuous* *process*; the idea *became more functional* as a living being; and the idea needed *protection* from physical harm (= *disenablement* of *function-inhibiting* influences). These default conclusions can then be translated into reality (target-based) terms: there was a protracted, (non-physical) continuous process of change; the idea needed protection during it; and the idea ended up being more functional. The basis for such translation is *View-Neutral Mapping Adjuncts*, a special type of mappings that cover the shape of events and processes, temporal properties and relationships, causation, functioning, mental states, emotional states, value judgments and various other matters (Agerri et al., 2007; Barnden et al., 2003).

RTE-2 organisers claimed there has been a trend towards using more deep inference and that this has been beneficial *provided* that it is based on enough knowledge. (See also Bos and Markert (2006) and Clark et al. (2007)). The depth and knowledge needs of ATT-Meta's processing are like those of deeper parts of existing RTE systems, but ATT-Meta is currently equipped only with small, hand-constructed knowledge bases. So, our main focus in deeper processing will actually be on a shallowed, broadened form of ATT-Meta-style reasoning. In this sense, we aim to look at common-sense knowledge resources such as ConceptNet 3.0 which contains relationships such as causation, function, etc. — the types of information transferred by some of ATT-Meta's VNMAs — or modified WordNets enriched by extra, web-mined knowledge (Veale and Hao, 2008), where the extra knowledge is especially relevant to metaphorical usages.

ATT-Meta's reasoning is by backwards chaining from goals derivable from context surrounding a metaphor. This relevance-based inference-focussing will be key in other processing we develop, and is highly RTE-compatible. An RTE Hypothesis can act as (part of) a reasoning goal, with the metaphor's within-Text context supplying further information or goal parts. T's own original context is of course unavailable, but this obstacle faces human judges as well.

We will also further extend our E-Drama project's methods, based there on robust parsing using the *Rasp* system and computation over WordNet. The methods are currently largely confined to metaphors of "X is Y" form where X is a person and Y is some type of animal, supernatural being, artefact or natural physical object. We will generalize to other categories for X and Y, and to cases where the categorization is only implicit. We will make the syntactic/semantic analysis of WordNet synset-glosses and the way the system traverses the network more advanced. We will extend our associated treatment of metaphorically-used size adjectives (as in "little bully").

However, the methods are also currently confined to detecting emotional/value judgments about X (unpleasantness, etc.), and mainly exploit metaphorical information that is already implicitly in WordNet, e.g., "pig" meaning a coarse, obnoxious person, in one synset. Substantial research is needed to go beyond these limitations. One avenue will be to apply VNMAs: when a synset gloss couches a metaphorical sense, we could extract not just affect but other types of information that VNMAs handle (causation, process shape, etc.); and when a gloss couches a non-metaphorical sense, we could translate some aspects of it via VNMAs.

5 Concluding Remarks

This paper aims to provide an avenue for giving metaphor-understanding the prominence it merits in NLP applications and in RTE, and thereby also to engender basic and applied research advances on metaphor by ourselves and others.

RTE researchers and NLP applications developers will benefit, as systems will gain added accuracy and coverage by addressing metaphor. Beneficiary application areas aside from QA, IR, etc., include Knowledge Management, Information Access, and intelligent conversation agents.

References

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

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)*, pp. 57–61.

Birke, J. and A. Sarkar (2006). A clustering approach for the nearly unsupervised recognition of nonliteral language. In *Proceedings of the 11th Conference on the European Chapter of the Association for Computational Linguistics (EACL 2006)*, Trento, Italy, pp. 329–336.

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

Clark, P., W. Murray, J. Thompson, P. Harrison, J. Hobbs, and C. Fellbaum (2007). On the role of lexical and world knowledge in RTE3. In *Proceedings of the Workshop on Textual Entailment and Paraphrasing*, Prague, pp. 54–59. ACL 2007.

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

Falkenhainer, B., K. 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*. Greenwich, Connecticut: Ablex.

Gedigian, M., J. Bryant, S. Narayanan, and B. Ciric (2006). Catching metaphors. In *Proceedings of the 3rd Workshop on Scalable Natural Language Understanding*, New York, pp. 41–48.

Goatly, A. (1997). *The Language of Metaphors*. Routledge.

Markert, K. and M. Nissim (2007, June). Semeval-2007: Metonymy resolution at semeval-2007. In *International Workshop on Semantic Evaluations (SemEval-2007)*, Prague, pp. 36–41. Association for Computational Linguistics (ACL-07).

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

Mason, Z. (2004). CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics 30*(1), 23–44.

Veale, T. and Y. Hao (2008). Enriching WordNet with folk knowledge and stereotypes. In *Proceedings of the 4th Global WordNet Conference*, Szeged, Hungary.

Wallington, A., R. Agerri, J. Barnden, M. Lee, and T. Rumbell (2008, May). Affect transfer by metaphor for an intelligent conversational agent. In *Proceedings of the LREC 2008 Workshop on Sentiment Analysis: Emotion, Metaphor, Ontology and Terminology (EMOT 08)*, Marrakech, Morocco, pp. 102–107.

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*, pp. 31–36.