Introduction to the Shared Task on Comparing Semantic Representations

Johan Bos
University of Rome "La Sapienza" (Italy)
email: bos@di.uniroma1.it

Abstract

Seven groups participated in the STEP 2008 shared task on comparing semantic representations as output by practical wide-coverage NLP systems. Each of this groups developed their own system for producing semantic representations for texts, each in their own semantic formalism. Each group was requested to provide a short sample text, producing a shared task set of seven texts, allowing participants to challenge each other. Following this, each group was asked to provide the raw system output for all texts, which are made available on http://www.sigsem.org. Two groups were extremely inspired by the shared task and also provided gold-standard semantic representations for the seven texts, together with evaluation measures. The STEP 2008 workshop itself will continue the discussion, focusing on the feasibility of a theory-neutral gold standard for deep semantic representations.

1 Introduction

Following advances made in computational syntax in the last years, we have recently witnessed progress in computational semantics too. Thanks to the availability of wide-coverage parsers, most of them implementing statistical approaches with models trained on the Penn Treebank, we now have at our disposal tools that are able to produce formal, semantic representations on the basis of the output of the aforementioned parsers, achieving high coverage. Computational semantics isn't anymore limited to tedious paper and pencil exercise, nor to implementations of tiny fragments of natural language, and has genuinely matured to a level useful for real applications.

As a direct consequence, the question as to how to measure the quality of semantic representations output by these systems pops up. This is an important issue for the sake of the field, but difficult to answer. On the one hand one might think that the quality of semantic representations, because they are more abstract than surface and syntactic representations, should be easy to evaluate. On the other hand, however, because there are several "competing" semantic formalisms, and the depth of analysis is arbitrary, it is hard to define a universal theory-neutral gold standard for semantic representations (see, e.g. Bos, 2008a).

Partly in response to this situation in the field, a "shared task" was organised as a special event on the STEP 2008 conference. The aim of this shared task was primarily to *compare* semantic representations for texts as output by state-of-the-art NLP systems. This was seen as a first step for designing evaluation methodologies in computational semantics, with a practical bottom-up strategy: rather than defining theoretical gold standard representations, we look what current systems can actually produce and start working from that.

2 Participants

In response to the call for participation seven groups were accepted to take part in the shared task. Table 1 gives an overview of the participants, the systems they have developed, and the semantic formalism they adopted. This volume contains full descriptions of these systems (please follow the page numbers in Table 1).

Table 1: Overview of shared task participants at STEP 2008

	System	**Type of Formalism**	**Authors**	**Pages**
1	BLUE	Logical Form	Clark and Harrison	263–276
2	Boxer	Discourse Representation Theory	Bos	277–286
3	GETARUNS	Situation Semantics	Delmonte	287–298
4	LXGram	Minimal Recursion Semantics	Branco and Costa	299–314
5	OntoSem	Ontological Semantics	Nirenburg et al.	315–326
6	TextCap	Semantic Triples	Callaway	327–342
7	Trips	Logical Form	Allen et al.	343–354

All but one group have NLP systems developed to deal with the English language. One group has an NLP system for Portuguese (LXGram). This made it more difficult to organise the task (the English text had to be translated, Branco and Costa (2008)), but also more interesting. After all, it is a reasonable assumption that semantic representations ought to be independent of the source language.

Also note that basically all participants adopt different semantic formalisms (Table 1), even though they all claim to do more or less the same thing: computing semantic representations for text. These differences in (formal) background make the shared task only more interesting.

3 The Shared Task Texts

All participants were asked to submit an authentic small text, not exceeding five sentences and 120 tokens. The pool of test data for the shared task is composed out of all the texts submitted by the seven participants. This procedure allowed the participants to "challenge" each other. Below are the original texts as submitted by the participants — the numbering follows the numbering of the participants in Table 1.

Text 1
An object is thrown with a horizontal speed of 20 m/s from a cliff that is 125 m high. The object falls for the height of the cliff. If air resistance is negligible, how long does it take the object to fall to the ground? What is the duration of the fall?

Text 2
Cervical cancer is caused by a virus. That has been known for some time and it has led to a vaccine that seems to prevent it. Researchers have been looking for other cancers that may be caused by viruses.

Text 3
John went into a restaurant. There was a table in the corner. The waiter took the order. The atmosphere was warm and friendly. He began to read his book.

Text 4
The first school for the training of leader dogs in the country is going to be created in Mortagua and will train 22 leader dogs per year. In Mortagua, Joao Pedro Fonseca and Marta Gomes coordinate the project that seven people develop in this school. They visited several similar places in England and in France, and two future trainers are already doing internship in one of the French Schools. The communitarian funding ensures the operation of the school until 1999. We would like our school to work similarly to the French ones, which live from donations, from the merchandising and even from the raffles that children sell in school.

Text 5
As the 3 guns of Turret 2 were being loaded, a crewman who was operating the center gun yelled into the phone, "I have a problem here. I am not ready yet." Then the propellant exploded. When the gun crew was killed they were crouching unnaturally, which suggested that they knew that an explosion would happen. The propellant that was used was made

from nitrocellulose chunks that were produced during World War II and were repackaged in 1987 in bags that were made in 1945. Initially it was suspected that this storage might have reduced the powder's stability.

Text 6
Amid the tightly packed row houses of North Philadelphia, a pioneering urban farm is providing fresh local food for a community that often lacks it, and making money in the process. Greensgrow, a one-acre plot of raised beds and greenhouses on the site of a former steel-galvanizing factory, is turning a profit by selling its own vegetables and herbs as well as a range of produce from local growers, and by running a nursery selling plants and seedlings. The farm earned about $10,000 on revenue of $450,000 in 2007, and hopes to make a profit of 5 percent on $650,000 in revenue in this, its 10th year, so it can open another operation elsewhere in Philadelphia.

Text 7
Modern development of wind-energy technology and applications was well underway by the 1930s, when an estimated 600,000 windmills supplied rural areas with electricity and water-pumping services. Once broad-scale electricity distribution spread to farms and country towns, use of wind energy in the United States started to subside, but it picked up again after the U.S. oil shortage in the early 1970s. Over the past 30 years, research and development has fluctuated with federal government interest and tax incentives. In the mid-'80s, wind turbines had a typical maximum power rating of 150 kW. In 2006, commercial, utility-scale turbines are commonly rated at over 1 MW and are available in up to 4 MW capacity.

The first text is taken from an AP Physics exam (the fourth sentence is a simplified reformulation of the third sentence) and constitutes a multi-sentence science question (Clark and Harrison, 2008). Text 2 is taken from the *Economist*, with the third sentence slightly simplified (Bos, 2008b). Text 4 was taken from a Portuguese newspaper and translated into English (Branco and Costa, 2008). Text 6 is also a fragment of a newspaper article, namely the *New York Times* (Callaway, 2008). Text 7 is an excerpt from `http://science.howstuffworks.com`. The origin of Text 3 is unknown.

4 Preliminary Results

All groups produced semantic representations for the texts using their NLP systems. The results are, for obvious reasons of space, not all listed here, but available at the SIGSEM website `http://www.sigsem.org`. The papers that follow the current article describe the individual results in detail. It should be noted that two groups created gold standard representations for all seven texts, and already performed a self evaluation (Nirenburg et al., 2008; Allen et al., 2008).

The workshop itself (to be held in Venice, September 2008) will feature further comparison and manual evaluation of the systems' output — the system with the most complete and accurate semantic representation will receive a special STEP award. This event should naturally lead to a discussion on the feasibility of a gold standard for deep semantic representations, and furthermore identify a set of problematic and relevant issues for semantic evaluation.

References

Allen, J. F., M. Swift, and W. de Beaumont (2008). Deep Semantic Analysis of Text. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 343–354. College Publications.

Bos, J. (2008a). Let's not argue about semantics. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, Marrakech, Morocco.

Bos, J. (2008b). Wide-Coverage Semantic Analysis with Boxer. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 277–286. College Publications.

Branco, A. and F. Costa (2008). LXGram in the Shared Task "Comparing Semantic Representations" of STEP 2008. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 299–314. College Publications.

Callaway, C. B. (2008). The TextCap Semantic Interpreter. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 327–342. College Publications.

Clark, P. and P. Harrison (2008). Boeing's NLP System and the Challenges of Semantic Representation. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 263–276. College Publications.

Delmonte, R. (2008). Semantic and Pragmatic Computing with GETARUNS. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 287–298. College Publications.

Nirenburg, S., S. Beale, and M. McShane (2008). Baseline Evaluation of WSD and Semantic Dependency in OntoSem. In J. Bos and R. Delmonte (Eds.), *Semantics in Text Processing. STEP 2008 Conference Proceedings*, Volume 1 of *Research in Computational Semantics*, pp. 315–326. College Publications.

Models of semantic representation are routinely evaluated against human behavioral data. Chief among these are word association norms, priming data, and similarity ratings. Given the burgeoning interest both in semantic representation and semantic similarity, model evaluation increasingly also involves model comparisons.Â Category learning not only depends upon perceptual and semantic representations; it also leads to the generation of these representations. We describe two series of experiments that demonstrate how categorization experience alters, rather than simply uses, descriptions of objects. In the first series, participants first learned to categorize objects on the basis of particular sets of line segments. CR is thus a task which requires a form of semantic representation to cluster the mentions pointing to the same entity. We use the model proposed in (Lee et al.Â In pro-portional sampling, the probability of sampling a task is pro-portional to the relative size of each dataset compared to the cumulative size of all the datasets. Note that unlike (Subra-manian et al.Â Comparing setups A and A-GM shows how the supervision from one module (e.g. CR) can ow through the entire architecture and impact other tasksâ€™ performance: REâ€™s F1 score drops by âˆ¼1 point on A. Note that the GM setup impacts the training exit condition (the validation met-rics stop improving) and the evaluation metrics (it is well known. The aim of this shared task was primarily to compare semantic representations for texts as output by state-of-the-art NLP sys-tems. This was seen as a rst step for designing evaluation methodologies in computa-tional semantics, with a practical bottom-up strategy: rather than dening theoretical gold standard representations, we look what current systems can actually produce and start working from that. 2 Participants.Â This event should naturally lead to a discussion on the feasibility of a gold standard. Introduction to the Shared Task on Comparing Semantic Representations. 261. for deep semantic representations, and furthermore identify a set of problematic and relevant issues for semantic evaluation. References. Allen, J. F., M. Swift, and W. de Beaumont (2008).