Lang Resources & Evaluation (2009) 43:97–104
DOI 10.1007/s10579-009-9091-2

Computational semantic analysis of language: SemEval-2007 and beyond

Introduction to the special issue

Eneko Agirre · Lluís Màrquez · Richard Wicentowski

Published online: 14 May 2009

1 Introduction

SemEval-2007, the Fourth International Workshop on Semantic Evaluations (Agirre et al. 2007) took place on June 23–24, 2007, as a co-located event with the 45th Annual Meeting of the ACL. It was the fourth semantic evaluation exercise, continuing on from the series of successful Senseval workshops.

SemEval-2007 took place over a period of about six months, including the evaluation exercise itself and the summary workshop. The exercise attracted considerable attention from the semantic processing community: 18 different evaluation tasks were organized, and more than 100 research teams and 123 systems participated in them. As a result, despite the huge effort carried out by task organizers and participant teams, time and material constraints made it virtually impossible to present thorough analyses of tasks, systems and results in the workshop proceedings. Therefore, in order to present the work and results of SemEval-2007, we assembled extended papers from the workshop as well as other contributors into this special issue of *Language Resources and Evaluation*, entitled

E. Agirre
IXA Research Group, Computer Science Department, University of the Basque Country, Manuel Lardizabal 1, 20018 Donostia, Basque Country
e-mail: e.agirre@ehu.es

L. Màrquez (✉)
GPLN Research Group, Software Department, Technical University of Catalonia, Jordi Girona Salgado 1–3, 08034 Barcelona, Catalonia, Spain
e-mail: lluism@lsi.upc.edu

R. Wicentowski
Computer Science Department, Swarthmore College, 500 College Avenue, Swarthmore, PA 19081, USA
e-mail: richardw@cs.swarthmore.edu

"Special Issue on Computational Semantic Analysis of Language: SemEval-2007 and Beyond".

The call for papers for this special issue, published in Autumn 2007, invited submissions describing evaluation exercises involving computational semantics. Although the natural candidates were papers detailing evaluation tasks from SemEval-2007, the call was also open to anyone who could report on substantial experimental evaluation of natural language semantics. The call attracted twenty high-quality papers, from which five were selected to comprise this issue of *LRE*.

This introductory article provides a brief overview of the history of Senseval and SemEval (Sect. 2), as well as other important evaluation exercises on semantic analysis of language. Section 3 then summarizes the papers included in the special issue.

2 Past, present, and future of evaluation exercises on semantic analysis of language

Evaluations for applications of language technology such as information retrieval (TREC),[1] information extraction (MUC)[2] and text summarization (DUC)[3] have been very successful in stimulating rapid scientific progress. They have brought the research community to consensus on appropriate tasks for evaluation, enabled the design of metrics for measuring comparative performance and diagnosing system strengths and weaknesses, and often led to the development of common, open, resources.

The semantic processing community quickly embraced evaluation exercises. A discussion at a workshop sponsored by the Association for Computational Linguistics Special Interest Group on the Lexicon (SIGLEX) on "Evaluating Automatic Semantic Taggers" (Resnik and Yarowsky 1997) sparked the formation of an evaluation effort for Word Sense Disambiguation (WSD), which was later named "Senseval".

2.1 Past: the Senseval series

The first Senseval evaluation exercise[4] was run by a small elected committee under the auspices of SIGLEX. Unlike the aforementioned evaluation exercises, Senseval was a grassroots enterprise, initiated and organized by WSD researchers themselves. The main goal of the first exercise was to establish the viability of WSD as a separate task, with its own evaluation methods and standards, and with the goal of paving the way for a better understanding of lexical semantics and polysemy.

The first Senseval exercise (Kilgarriff and Palmer 2000) took place in 1998, including tasks for English, French and Italian, in which 23 groups participated. Participants were provided with hand-annotated training and test data, as well as a predefined metric for evaluation. Senseval-1 produced a set of benchmarks for WSD

[1] Text REtrieval Conference: http://trec.nist.gov

[2] Message Understanding Conference: http://www-nlpir.nist.gov/related_projects/muc

[3] Document Understanding Conference: http://duc.nist.gov/

[4] See the Senseval official website for complete information on the three editions: http://www.senseval.org/

system performance, and it was followed by a workshop in Herstmonceux, Sussex, UK. The exercise was a success in terms of participation and interest, and it provided convincing evidence that the task could be evaluated consistently.

Senseval-2 (Edmonds and Kilgarriff 2002) was organized in 2001, followed by an ACL workshop held soon after as well as another ACL workshop in 2002. The second Senseval's goals were to encourage the creation of tasks in new languages, increase the number of participants and systems, and broaden the range of languages for existing tasks. A new kind of task was defined, where the word senses were defined according to possible translations into other languages. Overall, datasets for 10 languages were produced, including Basque, Czech, Dutch, English, Estonian, Italian, Japanese, Korean, Spanish and Swedish. Thirty-five research teams and 94 systems participated.

Senseval-3 (Mihalcea and Edmonds 2004) again broadened the scope of the exercise, as shown by the subsequent ACL workshop title: "Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text". Apart from WSD tasks for various languages (Basque, Catalan, Chinese, English, Italian, Romanian, Spanish) and the cross-lingual task, Senseval-3 included tasks for the identification of semantic roles, logic forms, and sub-categorization acquisition. The evaluation exercise attracted 55 teams, which participated with over 160 systems.

The success of the Senseval series is evident from the number of benchmark datasets it produced, as well as the achievement of agreement on a common experimental design and evaluation software and hand-tagging annotation technologies. As a result of the series, explicit WSD using a fixed sense inventory attained maturity, and WSD systems were shown to be robust according to word types, frequencies, and sense distributions. The performance of WSD systems achieved accuracies close to human performance as measured by Inter-tagger Agreement measures over the course of the Senseval series. Perhaps as a consequence of this maturity, and the fact that several systems attained comparable levels of performance, the community felt that it should move on to organize evaluation exercises for other semantic processing tasks, while at the same time trying to put WSD evaluation into real application scenarios. Although this move was already evident in Senseval-3 with the inclusion of other types of semantic processing such as semantic role labeling, it was fully accomplished in the following evaluation exercise, called SemEval.

2.2 Present: SemEval-2007 and other initiatives

The broader perspective on semantic processing was made explicit in the next exercise, which was renamed as "SemEval" (Agirre et al. 2007), short for "semantic evaluations". Eighteen tasks were organized (cf. Table 1), and over 100 teams participated with over 123 unique systems.[5] Some tasks were updated versions of the WSD tasks found in Senseval-3, including lexical-sample word sense disambiguation tasks in Catalan, English, Spanish and Turkish, two all-words English word sense disambiguation tasks, and two multilingual lexical sample tasks

[5] More details are available at the official SemEval website: http://nlp.cs.swarthmore.edu/semeval/

Table 1 Tasks organized in SemEval-2007

01	Evaluating WSD on Cross-Language Information Retrieval
02	Evaluating Word Sense Induction and Discrimination Systems
04	Classification of Semantic Relations between Nominals
05	Multilingual Chinese-English Lexical Sample
06	Word-Sense Disambiguation of Prepositions
07	Coarse-Grained English All-Words Task
08	Metonymy Resolution at SemEval-2007
09	Multilevel Semantic Annotation of Catalan and Spanish
10	English Lexical Substitution Task
11	English Lexical Sample Task via English-Chinese Parallel Text
12	Turkish Lexical Sample Task
13	Web People Search
14	Affective Text
15	TempEval Temporal Relation Identification
16	Evaluation of Wide Coverage Knowledge Resources
17	English Lexical Sample, SRL and All Words
18	Arabic Semantic Labeling
19	Frame Semantic Structure Extraction

Number 03 corresponds to an accepted task that was cancelled due to the lack of participants (*Pronominal Anaphora Resolution in the Prague Dependency Treebank 2.0*). For more details on tasks and data sets consult the SemEval official website

(Chinese-English). The updates included using coarse-sense inventories or combining word sense disambiguation and semantic role classification. Only four of the WSD-related tasks were classical WSD tasks.

Some of the new tasks in SemEval were related to WSD, for example, word sense induction and lexical substitution. Others dealt with semantic properties such as metonymy, semantic relations between nominals, disambiguation of prepositions, semantic role labeling, affective text, temporal relation identification, semantic interpretation using frames, evaluation of knowledge resources, and identification of person identity over web pages. SemEval also included, for the first time, an in-vivo evaluation exercise, which explicitly measured the impact of specific NLP tasks on IR and CLIR systems. This in-vivo task was later taken over by the Cross-Lingual Evaluation Forum in 2008 and 2009.[6]

SemEval has been the primary forum for the evaluation efforts of the semantic processing community and the largest community-based evaluation effort in the NLP field. However, several other important evaluation exercises deserve mention here.

The Conference on Natural Language Learning (CoNLL),[7] yearly organized by the ACL Special Interest Group on Natural Language Learning, has organized evaluation exercises (referred to as *shared tasks*) for the last 1 year. This conference

[6] http://ixa2.si.ehu.es/clirwsd/

[7] http://www.ifarm.nl/signll/conll/

was the first to organize an SRL task (Carreras and Màrquez 2004), which was continued in 2005 (Carreras and Màrquez 2005). As a new twist, in 2008 the shared task involved a combination of parsing and SRL, using a unified dependency-based representation (Surdeanu et al. 2008). The same task, extended to multiple languages, will comprise the CoNLL 2009 shared task.

ACE,[8] the NIST series of information extraction technology evaluation, has run Entity Detection and Recognition (EDR) evaluations throughout the years, which is a component of the detection of complex event structures.

Another recent proposal[9] at the Lexical Semantics Workshop at the European Summer School in Logic, Language and Information involves a number of tasks that focus on inducing lexical-semantic properties of words, such as free association, categorization and generation of salient properties of concepts.

The Recognizing Textual Entailment challenge (RTE) has been run yearly since 2004.[10] This challenge proposes RTE as a generic task that captures major semantic inference needs across many natural language processing applications, such as Question Answering (QA), Information Retrieval (IR), Information Extraction (IE), and (multi-)document summarization. RTE requires participant systems to recognize, given two text fragments, whether the meaning of one text is entailed (can be inferred) from the other text. The organization of the challenge for 2008 has been taken over by NIST.[11] Given the NLP modules involved, this task encompasses many of the tasks mentioned above, such us WSD, SRL, EDR, and others.

At the time of writing this introduction, two related events are upcoming. The SEW-2009 workshop,[12] aims at analyzing and discussing practical and foundational aspects of semantic processing of language, as an intermediate step between the SemEval-2007 and SemEval 2010 exercises. The organization of the next edition of SemEval in 2010 is underway.[13] As with the Senseval series, the semantic processing community itself proposed which tasks should be included. After the initial call and a competitive selection process, 18 tasks have been selected for SemEval-2010, a significant number of which are new to the SemEval series.

3 Articles in this special issue

Twenty high-quality papers were submitted to *LRE* in response to the call for papers for this special issue. After a very competitive selection process, five papers were selected for inclusion. Of the five papers accepted, four are papers that present detailed descriptions of tasks organized at SemEval-2007, while the fifth stands independent of the workshop. Each of the first four papers (Girju et al. 2009; McCarthy and Navigli 2009; Verhagen et al. 2009; Markert and Nissim 2009) detail

[8] http://www.itl.nist.gov./iad/894.01/tests/ace/2008/

[9] http://wordspace.collocations.de/doku.php/esslli:task

[10] http://www.pascal-network.org/Challenges/RTE3

[11] http://www.nist.gov/tac/tracks/2008/TAC

[12] http://www.lsi.upc.edu/∼lluism/sew2009/

[13] http://semeval2.fbk.eu/

SemEval tasks by outlining the motivation for the task, the guidelines used to create the data and resources, the participant systems from SemEval-2007, and the main contributions and lessons learned from the evaluation. The fifth paper (Chen and Palmer 2009) presents a work on robust verb sense disambiguation, which also includes a post-workshop evaluation using SemEval-2007 data.

Girju et al. (2009) present SemEval-2007 Task 4: Classification of Semantic Relations between Nominals. The task designers selected a set of seven relations between nominals, such as *X causes Y*, which were then used to form search engine queries. The results of the searches were hand-labeled by annotators to form the task data set. Participants were provided with the query and the search result, and were required to determine the latent relationship between a pair of labeled nominals. The authors found that the best systems out-performed the inter-annotator agreement rate on the task, and that systems did not benefit from including the query in the feature set.

McCarthy and Navigli (2009) describe SemEval-2007 Task 10: The English Lexical Substitution Task. Similar to a lexical sample task, participants were presented with single sentences, each containing a single word of interest. However, in the lexical substitution task, participants were asked to choose the most appropriate substitute for the word in the sentence rather than matching the word against a pre-defined sense inventory, thereby avoiding debate about coarse- or fine-grained sense distinctions. Since there can be no definitive "truth set" for this task, the authors spend considerable time discussing the formation of the data set and the post-hoc analysis of the participant systems' results.

Verhagen et al. (2009) present SemEval-2007 Task 15: TempEval Temporal Relation Identification. The TempEval task encompasses three temporal relation subtasks: specifying the relation between an event and a time expression within a sentence, specifying the relationship between an event and the document creation time, and providing an ordering of events in consecutive sentences. Participants were required to specify the temporal relations using a pre-defined subset of the TimeML annotation language (Pustejovsky et al. 2003). The performances of the six participating systems were somewhat similar, despite their architectural differences. In one subtask, the best systems were only slightly better than the baseline. This leads the authors to speculate on future evaluations using different subtasks and ways of combining the subtasks into a single evaluation metric.

Markert and Nissim (2009) discuss SemEval-2007 Task 8: Metonymy Resolution. The task was set up as a lexical sample task, where participants had to determine whether the target word was being used literally or figuratively. Teams could choose to specify the granularity of their solution, ranging from determining coarse-grained distinctions, where it was only necessary to specify if the target word is being used literally or figuratively, up through fine-grained distinctions, where it was necessary to specify the particular metonymic pattern exhibited by the target word. The target words were drawn from proper names for locations and organizations. The organizers found that relying only on information about grammatical roles resulted in a very competitive baseline.

Chen and Palmer (2009) describe a supervised word-sense disambiguation system for English verbs that make use of linguistic features such as syntactic

alternations, named entity tags, and pronoun resolution, as well as WordNet synsets and hypernyms. These features ameliorate the sparse data problems faced by WSD algorithms that use only lexical features. The system is evaluated using data drawn from the OntoNotes project (Hovy et al. 2006) where the performance matched the inter-annotator agreement rate. In addition, the system is evaluated using data from SemEval-2007 task 17 (Pradhan et al. 2007) where results matched or exceeded the best systems.

4 Conclusions

Semantic processing is at the core of language understanding. It comprises a myriad of related tasks, which need to be tackled in order to grasp the meaning of texts. The Senseval and SemEval campaigns are grassroots efforts to provide evaluation datasets for semantic tasks, in a broad sense.

This special issue presents a significant portion of the most relevant tasks in SemEval-2007, with detailed analysis of tasks on nominal relations, metonymic relations, lexical substitution and temporal relations. Those papers outline the motivation for the task, the guidelines used to create the data and resources, the participant systems, and the main contributions and lessons learned from the evaluation. In addition, this issue presents a work on robust verb sense disambiguation, which also includes a post-workshop evaluation using SemEval-2007 data.

Acknowledgements We are especially grateful to the numerous reviewers who offered their time and expertise to select the papers presented here. Also, we thank the authors of submitted papers for their interest and hard work.

References

Agirre, E., Màrquez, L., & Wicentowski, R. (Eds.). (2007, June). *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*. Association for Computational Linguistics, Prague, Czech Republic.

Carreras, X., & Màrquez, L. (2004). Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the eighth conference on computational natural language learning (CoNLL-2004)* (pp. 89–97). Boston, MA, USA.

Carreras, X., & Màrquez, L. (2005). Introduction to the CoNLL-2005 shared task: Semantic role labeling. In *Proceedings of the ninth conference on computational natural language learning (CoNLL-2005)* (pp. 152–164). Ann Arbor, MI, USA.

Chen, J., & Palmer, M. S. (2009). Improving English verb sense disambiguation performance with linguistically motivated features and clear sense distinction boundaries. *Language Resources and Evaluation, 43*(3). doi:10.1007/s10579-009-9085-0.

Edmonds, P., & Kilgarriff, A. (2002). Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering, 8*(4), 279–291.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., & Yuret, D. (2009). Classification of semantic relations between nominals. *Language Resources and Evaluation, 43*(3). doi:10.1007/s10579-009-9083-2.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., & Weischedel, R. (2006). Ontonotes: The 90% solution. In *Proceedings of HLT-NAACL06*, New York.

Kilgarriff, A., & Palmer, M. (2000). Introduction to the special issue on Senseval. *Computers and the Humanities, 34*(1–2), 1–13.

Markert, K., & Nissim, M. (2009). Data and models for metonymy resolution. *Language Resources and Evaluation, 43*(3). doi:10.1007/s10579-009-9087-y.

McCarthy, D., & Navigli, R. (2009). The English lexical substitution task. *Language Resources and Evaluation, 43*(3). doi:10.1007/s10579-009-9084-1.

Mihalcea, R., & Edmonds, P. (Eds.). (2004, July). *Proceedings of SENSEVAL-3, the third international workshop on the evaluation of systems for the semantic analysis of text*. Association for Computational Linguistics, Barcelona, Spain.

Pradhan, S., Loper, E., Dligach, D., & Palmer, M. (2007). SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th international workshop on semantic evaluations (SemEval-2007)* (pp. 87–92). Prague, Czech Republic.

Pustejovsky, J., Castaño, J., Ingri, R., Saurí, R., Gaizauskas, R., Setzer, A., et al. (2003). TimeML: Robust specification of event and temporal expressions in text. In *Proceedings of the fifth international workshop on computation semantics (IWCS-5)*.

Resnik, P., & Yarowsky, D. (1997). A perspective on word sense disambiguation methods and their evaluation. In *Proceedings of the ACL SIGLEX workshop on tagging text with lexical semantics: Why, what, and how?* (pp. 79–86). Washington, DC, USA.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th conference on computational natural language learning (CoNLL-2008)*. Manchester, UK.

Verhagen, M., Gaizauskas, R., Schilder, F., Hepple, M., Moszkowicz, J., & Pustejovsky, J. (2009). The TempEval challenge: Identifying temporal relations in text. *Language Resources and Evaluation, 43*(3). doi:10.1007/s10579-009-9086-z.