Open Extraction of General Knowledge through Compositional Semantics

Lenhart Schubert and Benjamin Van Durme
{*schubert, vandurme*}*@cs.rochester.edu*
Department of Computer Science, University of Rochester

Our goal is the accumulation of general knowledge, such as that a room may have windows; people may want to be rid of a dictator; when a person enters a room, it is generally through a door; and when a car crashes, the driver and passengers may well be hurt or killed. The applications we are working on are (1) improved guidance of a parser; (2) support for deep learning by reading; and (3) building a large knowledge base for our EPILOG inference engine, to support commonsense reasoning, in particular in a self-motivated, self-aware dialog agent.

Like many in the NLP community, we are attempting to exploit large text corpora for extracting the desired general knowledge. However, our target applications (particularly (2, 3)) require that the knowledge employed be expressed in a formal knowledge representation, rather than, for instance, tuples of word strings. Therefore our approach from the outset has been to derive general knowledge from fully parsed text, using compositional interpretive rules. We first reported our work in (Schubert 2002, Schubert & Tong 2003), where we describe a system called KNEXT (KNowledge EXtraction from Text), and its application to the Penn Treebank version of the Brown corpus. The system employs about 80 rules for mapping phrases to logical forms; it focuses on predicate-argument structures (dropping adjuncts) and modification structures within the logical forms, and also generalizes named entities to types (e.g., *Franco* becomes *a dictator*). In this way it derives general tentative "factoids", expressed in Episodic Logic (e.g., Schubert & Hwang 2000). KNEXT obtained 117,000 distinct factoids from the Brown corpus (more than 2 per sentence), and human judging sanctioned about 60% of these as reasonable general claims about the world. The logical factoids are also automatically rendered into approximate English, e.g., "A PERSON MAY BELIEVE A PROPOSITION", "CHILDREN MAY LIVE WITH RELATIVES", "A COMEDY CAN BE DELIGHTFUL", etc.

Since that early work we have refined KNEXT in various ways (e.g., blocking such phrases as *legal secretary* from yielding a factoid "A SECRETARY CAN BE LEGAL", while still allowing *red rose* to yield "A ROSE CAN BE RED"), and applied it to large corpora parsed with state-of-the-art statistical parsers. Our target corpora have included the 100 million word British National Corpus, yielding over 6 million factoids, and a corpus of web documents, also yielding about 6 million factoids. Human judgements for these factoids were nearly as favorable as for those derived from hand-parsed text. Our recent results are summarized in (Van Durme & Schubert 2008), where we also make comparisons with TextRunner (Banko *et al.* 2007); we have also shown the efficacy of our approach for mining class attributes from general text (Van Durme, Qian and Schubert 2008), comparing this approach with one based on Google query logs (Paşca & Van Durme 2007).

Two main goals in our current work are (i) particularizing light verbal predicates (such as HAVE) and prepositions (such as WITH) in our logical factoids; and (ii) obtaining stronger propositions either by generalization from clusters of related factoids (Van Durme, Michalak, & Schubert 2008), or by abstraction from sentences with multiple verbs, such as "*The car crashed into a tree, killing the driver*", which suggests that if a car crashes, the driver may be killed (see Van Durme 2008 for a discussion of conditional abstraction).

Acknowledgement

This work was supported by NSF grants IIS-0535105 and 0328849.

References (including relevant uncited papers)

M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni (2007), "Open information extraction from the Web", IJCAI-07, pp. 2670-2676.

F. Morbini and L.K. Schubert (2008), "Metareasoning as an integral part of commonsense and autocognitive reasoning", AAAI-08 Workshop on Metareasoning, Chicago, IL, July 13-14, pp. 155-162.

F. Morbini and L.K. Schubert (2007), "Towards realistic autocognitive inference", Logical Formalizations of Commonsense Reasoning, Papers from the AAAI Spring Symposium, Tech. Rep. SS-07-05, AAAI Press, Menlo Park, CA, March 26-28, Stanford, pp. 114-118.

M. Paşca and B. Van Durme (2007), "What you seek is what you get: Extraction of class attributes from query logs", IJCAI-07, pp. 2832-2837.

L.K. Schubert (2002), "Can we derive general world knowledge from texts?", M. Marcus (ed.), Proc. of the 2nd Int. Conf. on Human Language Technology Research (HLT 2002), March 24-27, San Diego, CA, pp. 94-97.

L.K. Schubert and M. Tong (2003), "Extracting and evaluating general world knowledge from the Brown Corpus", Proc. of the HLT-NAACL Workshop on Text Meaning, May 31, Edmonton, Alberta, pp. 7-13.

L.K. Schubert (2005), "Some knowledge representation and reasoning requirements for self-awareness", *AAAI Spring Symposium on Metacognition in Computation*, Stanford Univ., March 21-23, Vol SS05-04.

L.K. Schubert (2006), "Turing's dream and the knowledge challenge", *21st Nat. Conf. on Artificial Intelligence* (AAAI'2006), July 16-20, Boston, MA, AAAI Press, Menlo Park, CA, pp. 1534-8.

L.K. Schubert and C.H. Hwang (2000), "Episodic Logic meets Little Red Riding Hood: A comprehensive, natural representation for language understanding", in L. Iwanska and S.C. Shapiro (eds.), Natural Language Processing and Knowledge Representation: Language for Knowledge and Knowledge for Language, MIT/AAAI Press, Menlo Park, CA, and Cambridge, MA, pp. 111-174.

B. Van Durme (2008), "Notes on the acquisition of conditional knowledge", Tech. Rep. 937, Department of Computer Science, University of Rochester, Rochester, NY 14627.

B. Van Durme, P. Michalak, and L.K. Schubert (submitted 2008), "Deriving generic statements using corpus-acquired knowledge and WordNet".

B. Van Durme and L.K. Schubert (2008), "Open Knowledge Extraction through Compositional Language Processing", Symposium on Semantics in Systems for Text Processing (STEP 2008), September 22-24, Venice, Italy.

B. Van Durme, T. Qian and L.K. Schubert (2008), "Class-driven attribute extraction", COLING'08, Aug. 18-22, Manchester, UK.