Knowledge Base Evaluation for Semantic Knowledge Discovery

James Mayfield, Bonnie J. Dorr, Tim Finin, Douglas W. Oard and Christine D. Piatko
Human Language Technology Center of Excellence
810 Wyman Park Drive, Baltimore MD 21211-2840 USA
hltcoe.org

james.mayfield@jhuapl.edu; bonnie@umiacs.umd.edu; finin@cs.umbc.edu;
oard@umd.edu;christine.piatko@jhuapl.edu

Semantic knowledge discovery has traditionally been evaluated at the text level. For example, evaluations such as MUC [MUC-7] and ACE [ACE 2008] evaluate the information extraction of particular types of semantic roles and relations primarily at the mention level. We suggest that evaluating at the level of a knowledge base (KB) extracted from the text has significant advantages over evaluation at the text level. By knowledge base, we mean the combination of a database, a descriptive schema for the contents of the database, a collection of background knowledge, and an inference capability.

A knowledge base and the processes that populate it might be evaluated along a variety of dimensions, each corresponding to an aspect of semantic knowledge discovery:

- **Accuracy**: What is the precision of the extracted information? What is the system's level of confidence that the information was properly extracted? Is the extracted information factual?
- **Usefulness**: Is the extracted information relevant to the target task? Was all of the relevant information discovered? Is redundancy removed? Is the granularity of the extracted information at the appropriate level?
- **Augmentation**: How does the knowledge discovery augment information already in the knowledge base? Does it exploit that information to improve its performance?
- **Explanation**: What is the provenance of the induced information? Are attribution and committed belief properly expressed? How is contradictory information handled?
- **Adaptation**: Can the techniques be applied to new languages, genres, and domains? How much tuning is required to make such a leap?
- **Temporal qualification**: Does the knowledge base support evolution of assertions over time? Are the assertions in the knowledge base accurately marked with the time span during which they are true? Is information that is currently true properly distinguished from information that was true at some time in the past but that is no longer true?

Each of these evaluation axes might be used to assess the output of a semantic knowledge discovery tool directly, without reference to a knowledge base. However, performing the evaluation over the resultant knowledge base has several advantages:

- It forces the evaluation away from linguistic constructs and fully into the realm of semantics.
- It allows inference rules to be applied in performing the evaluation.
- It enables the use of existing knowledge bases as ground truth, possibly reducing annotation costs significantly.

How might KB evaluation be carried out? A straightforward way that handles some of the above evaluations is to treat the KB as a set of assertions, and use set-oriented measures such as

precision and recall. Treating each assertion as atomic avoids the need to perform alignment between system output and ground truth. Comparing the system and ground truth KB requires encoding the assertions in compatible or mappable ontologies. Moreover, the sets of assertions to be compared should be the deductive closure of the assertions in the output and ground truth. Finally, identifying the differences should ideally take into account the logical dependencies between assertions so as not over-penalize a system for missing one assertion from which many others are derivable [Zeginis et al. 2007].

Other existing measures can also be applied here, such as those used in ACE and question answering. If a confidence has been associated with each assertion, full precision/recall curves can be calculated by ordering the assertions from highest confidence to lowest confidence. Evaluation of temporal qualification can be partially handled by treating the KB as a sequence of fixed sets of assertions over time. Augmentation can be examined by performing ablation studies over the assertions in the KB (although deciding which assertions to remove is an open question).

An alternative approach is to perform an extrinsic evaluation by using the KB in a downstream task. This approach has the advantage that the downstream task need not have any awareness of the linguistic nature of the source information. Further investigation will be required to find measures for the other aspects of evaluation listed above.

Note that a KB evaluation approach is applicable not just to situations where fixed known fields are to be extracted, but also to situations where the kinds of information to be extracted are induced from the data themselves [Sekine 2006]. In such cases, the KB approach can be quickly adapted to new kinds of information by writing inference rules that relate the new types of information to be extracted to the ontology that serves as the schema for the knowledge base.

References

[ACE 2008] *Automatic Content Extraction 2008 Evaluation.* <http://www.nist.gov/speech/tests/ace/2008/>.

[MUC-7] *Seventh Message Understanding Conference Proceedings.* <http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html>.

[Sekine 2006] Sekine, S. On-demand information extraction. In *Proceedings of the COLING/ACL Annual Meeting of the ACL*. Association for Computational Linguistics, Morristown, NJ, pp. 731-738. 2006.

[Zeginis et al. 2007] Dimitris Zeginis, Yannis Tzitzikas, Vassilis Christophides, On the Foundations of Computing Deltas between RDF models. *Proceedings of the Sixth International Semantic Web Conference*, Busan KR, November 2007.