Correcting Scientific Knowledge in a General-Purpose Ontology

Michael Lipschultz and Diane Litman

Department of Computer Science, University of Pittsburgh,
Pittsburgh PA 15260, USA
{lipschultz,litman}@cs.pitt.edu

Abstract. General-purpose ontologies (e.g. WordNet) are convenient, but they are not always scientifically valid. We draw on techniques from semantic class learning to improve the scientific validity of WordNet's physics forces hyponym (IS-A) hierarchy for use in an intelligent tutoring system. We demonstrate the promise of a web-based approach which gathers web statistics used to relabel the forces as scientifically valid or scientifically invalid. Our results greatly improve the F1 for predicting scientific invalidity, with small improvements in F1 for predicting scientific validity and in overall accuracy compared to the WordNet baseline.

Keywords: Ontology, Semantic Web, Natural Language.

1 Introduction

An ontology is a formal definition of terms and the relationships between them [1]. An existing general-purpose natural language ontology called WordNet [2] has been successfully used in various tutoring systems [3,4]. We are interested in augmenting a dialog-based physics tutoring system with an ontology of physics terms so it can identify partially correct student responses and possibly offer different remediations based on the level of incorrectness (e.g. too vague or too specific). Prior work suggests that tutoring systems that detect partially correct responses and remediate differently may improve learning [5].

By using WordNet, we would not need to construct our own physics ontology. However, general-purpose ontologies, such as WordNet, contain mistakes in scientific domains [6], causing some researchers to construct their own domain-specific ontology [7]. We describe a method for automatically correcting an existing general-purpose ontology. Our scientific domain is physics forces and the general-purpose ontology is WordNet. The method identifies scientifically invalid terms contained within WordNet's hierarchy for physics forces by using information on the web, thus improving the scientific correctness of the ontology. Other work on correcting WordNet begins by specifying formal properties that an ontology should have, then considers any violations of the properties an error to fix in the existing ontology [8]. However, this method requires human effort to correct while our method does not.

V. Aleven, J. Kay, and J. Mostow (Eds.): ITS 2010, Part II, LNCS 6095, pp. 374–376, 2010.

2 Method

We work with the hyponym (IS-A) hierarchy for the physics meaning of "force". An expert tagged each of the hyponyms for scientific correctness. Of the 75 unique terms in the hyponym hierarchy, 29 were considered scientifically invalid (called *invalid* later) and 46 were considered scientifically valid (*valid*). This original WordNet is our baseline.

To improve WordNet, we need to classify whether an existing hyponym of "force" is *invalid* or *valid*. Our construction of the classifier is similar to work in semantic class learning [9], where the goal is to learn new terms for a particular topic in a domain. Learning new terms requires a pattern template containing a wildcard where the term-to-learn will go [10] and a corpus of text to search through. The corpus is then searched for instances of the pattern template and for each match, the term that replaces the wildcard is extracted. A corpus can be either domain-specific [11] or the entire web [9]; the first can be more reliable, but the second requires less effort to create and may be larger.

In our case, we already have terms, but wish to determine whether or not they are scientifically valid. Our corpus is all .edu websites to focus the search on sites we believe will use physics terminology correctly. In this paper, we found that simply searching for the term within quotes (i.e. ⟨"*term*"⟩) to be the pattern that performed best. We then use Google to search our corpus and count the number of results returned.

3 Results

We construct 13 classifiers from the data collected as described in the previous section. These classifiers differ only in their threshold, ranging from 0 to 5,000,000. We chose the highest threshold to be larger than the highest result count. Those forces having counts above the threshold for the classifer are labeled as *valid* and those below the threshold are relabeled *invalid*.

For relabeling forces as *invalid*, the classifier outperforms the baseline. The baseline does not predict any scientifically invalid forces, so precision, recall, and F1 are all 0. All classifiers with thresholds at least 1,000 outperform the baseline. Recall and F1 increase as the threshold increases, reaching maxima of 1.00 and 0.5577 respectively. Precision plateaus at around 0.70 between thresholds 1,000 and 10,000, before dropping. So, while high thresholds are best for recall and F1, if precision is of greatest importance, then a threshold of 10,000 is best.

For overall accuracy and relabeling forces as *valid*, lower thresholds provide the greatest performance. Recall drops as the threshold increases. The greatest precision (0.6615) and F1 (0.7748) occur at threshold 10,000 (recall is 0.9348). Accuracy is around baseline, with a peak (66.67%) at threshold 10,000. So, while we are also able to improve over our baseline for labeling scientific validity and overall accuracy, the improvement is not as large.

4 Conclusions and Future Work

In this paper, we modified a method from semantic class learning to correct the scientific validity of WordNet's hyponym hierarchy of physics forces. We saw that the simple pattern ⟨"*term*"⟩ was able to label scientifically invalid forces. A threshold of 10,000 provides the best F1 for labeling scientific validity and for overall accuracy, while providing a good F1 for scientific invalidity. However, higher thresholds improve F1 for scientific invalidity.

In future work, we plan to further improve our method by exploring other algorithms, patterns, corpora, and ontologies. We also plan on addressing the incompleteness of scientific knowledge in general-purpose ontologies. Finally, we want to incorporate the corrected ontology into our tutoring system to automatically detect partially correct responses. For example, if a student answered a question with "a force" (when the correct answer was "force of gravity"), the system could respond with "You're close. Which force is acting on the keys?" instead of its current response "I disagree with you. The force of gravity is acting on the keys.".

Acknowledgements. The authors thank Art Ward, Pam Jordan, Wenting Xiong, and Joanna Drummond for their input. We also thank Guangtian Zhu and Chandralekha Singh for their help in creating the gold standard.

References

1. Hendler, J.: Agents and the semantic web. IEEE Intelligent systems 16(2) (2001)
2. Fellbaum, C., et al.: WordNet: An electronic lexical database. MIT Press, Cambridge (1998)
3. Brown, J., Frishkoff, G., Eskenazi, M.: Automatic question generation for vocabulary assessment. In: Proc. of the Conference on HLT and EMNLP. Association for Computational Linguistics (2005)
4. Ward, A., Litman, D.: Semantic Cohesion and Learning. In: Proc. 9th ITS (2009)
5. Jordan, P., Litman, D., Lipschultz, M., Drummond, J.: Evidence of Misunderstandings in Tutorial Dialogue and their Impact on Learning. In: Proc. of AIED (2009)
6. McCrae, J., Collier, N.: Synonym set extraction from the biomedical literature by lexical pattern discovery. BMC bioinformatics 9(1), 159 (2008)
7. Christopher, B., Simon, J., Joanne, L., David, S., Robert, S., Ziqi, Z.: Issues in learning an ontology from text. BMC Bioinformatics 10 (2009)
8. Gangemi, A., Guarino, N., Oltramari, A.: Conceptual analysis of lexical taxonomies: The case of WordNet top-level. In: Proc. Intl. Conf. on Formal Ontology in Information Systems. ACM, New York (2001)
9. Kozareva, Z., Riloff, E., Hovy, E.: Semantic class learning from the web with hyponym pattern linkage graphs. In: Proc. ACL 2008, HLT (2008)
10. Hearst, M.: Automatic acquisition of hyponyms from large text corpora. In: Proc. of ACL (1992)
11. Cimiano, P., Pivk, A., Schmidt-Thieme, L., Staab, S.: Learning taxonomic relations from heterogeneous sources of evidence. In: Ontology Learning from Text: Methods, evaluation and applications, pp. 59–73 (2005)