Populating the Semantic Web by Macro-reading Internet Text

Tom M. Mitchell[1], Justin Betteridge[1], Andrew Carlson[1], Estevam Hruschka[1,2], and Richard Wang[1]

[1] Carnegie Mellon University, Pittsburgh PA 15213, USA
Tom.Mitchell@cs.cmu.edu
http://www.cs.cmu.edu/∼tom
[2] Federal University of Sao Carlos, Brazil

Abstract. A key question regarding the future of the semantic web is "how will we acquire structured information to populate the semantic web on a vast scale?" One approach is to enter this information manually. A second approach is to take advantage of pre-existing databases, and to develop common ontologies, publishing standards, and reward systems to make this data widely accessible. We consider here a third approach: developing software that automatically extracts structured information from unstructured text present on the web. We also describe preliminary results demonstrating that machine learning algorithms can learn to extract tens of thousands of facts to populate a diverse ontology, with imperfect but reasonably good accuracy.

1 The Problem

The future impact of the semantic web will depend critically on the breadth and depth of its content. One can imagine several approaches to constructing this content, including manual content entry by motivated teams of people, convincing owners of existing databases to publish them on the semantic web, and automatically extracting structured information from the vast quantity of unstructured online text. We consider here the third of these approaches, and argue both that it is feasible and that this kind of approach will be able to collect knowledge that is unlikely to be captured as easily by other approaches.

The feasibility of extracting structured information automatically from text will itself depend on the technical state-of-the-art of natural language processing (NLP) methods. We have witnessed significant progress in NLP over the past decade, on problems from sentence parsing [1] to named entity extraction [2], to question answering [3], to document classification [4]. Nevertheless, computer algorithms remain very far from being able to truly "understand" natural language text (e.g., to read and extract the full content of the paper you are currently reading). Given this shortcoming, why might we take the position that NLP algorithms offer a promising near-term approach to populating the semantic web?

We believe automatic methods offer a feasible near-term approach because the problem of automatically populating large databases from the internet can

A. Bernstein et al. (Eds.): ISWC 2009, LNCS 5823, pp. 998–1002, 2009.

be formulated so that it is much easier to solve than the problem of full natural language understanding. Our own formulation involves three key design choices:

1. Macro-reading instead of micro-reading. We use the term "micro-reading" to refer to the traditional NLP task where a single text document is input, and the desired output is the full information content of that document. In contrast, we define "macro-reading" as a task where the input is a large text collection (e.g., the web), and the desired output is a large collection of facts expressed by the text collection, without requiring that every fact be extracted. We argue that macro-reading is much easier than micro-reading, for two reasons. First, macro-reading does not require extracting every bit of information contained in the text collection. Second, in text corpora as large as the web, many important facts will be stated redundantly, thousands of times, using different wordings. A macro-reader can benefit from this redundancy by focusing on analyzing only the simple wordings of the fact, ignoring hopelessly complex sentences, and by statistically combining evidence from many text fragments in order to determine how strongly to believe a particular candidate hypothesis.

2. Ontology-driven reading. Much of the difficulty in truly understanding free-form text follows from the fact that it can say anything. In contrast, we formulate our macro-reading problem as a task of populating an ontology that is given as input, and that defines the categories (e.g., sport, person, team) and relations (e.g., plays-sport, plays-on-team) of interest. This is a natural way to frame a problem of populating some portion of the semantic web for which an ontology is available. It also makes our macro-reading problem easier in two ways. First, the system can focus only on a subset of text that is on-topic relative to the ontology. Second, the ontology itself can define meta-properties of its categories and relations that make extraction easier and more accurate (e.g., it can state that the relation 'plays-on-team' relates arguments of type 'person' and 'team').

3. Machine learning methods whose accuracy improves with ontology complexity. A third design choice is to use semi-supervised machine learning methods that automatically discover patterns of text and hypertext that support reliable fact extraction. Our machine learning approach acquires extraction patterns (e.g., "mayor of X" often implies X is a city) for each predicate (category or relation) in the input ontology. We build on earlier semi-supervised bootstrap learning methods [5,6,7,8] that learn from just a handful of labeled training examples, plus a large corpus of unlabeled text. While these earlier methods showed the feasibility of semi-supervised learning of extraction patterns, they were limited because accurate learning requires more constraints than are provided by a few dozen labeled training examples. Our algorithm achieves significantly higher accuracy by using the input ontology itself to provide additional constraints that guide the learner[9]. For example, when our algorithm learns extraction patterns for the predicates 'person', 'team' and 'plays-on-team', prior knowledge from the ontology requires that for any unlabeled sentence containing noun phrases A and B, the extractor for 'plays-on-team' can label <A, B > a positive example of the relation only if the 'person' classifier labels A positive, and the 'team' classifier

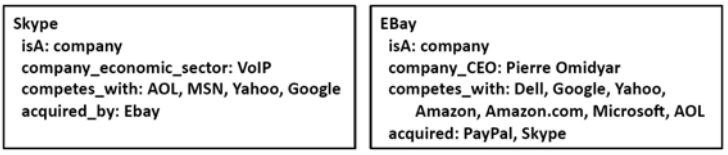

Fig. 1. Extracted facts for two companies discovered by our system. These two companies were extracted by the learned 'company' extractor, and the relations shown were extracted by learned relation extractors.

Table 1. Horn clause rules learned from extracted instances. Numbers indicate the conditional probability that the literal to the left of the ":-" will be satisfied if the literals to its right are satisfied.

```
0.84 playsSport(?x,?y) :- playsFor(?x,?z), teamPlaysSport(?z,?y)
0.70 playsSport(?x,baseball) :- playsFor(?x,yankees)
0.82 teamPlaysSport(?x,?y) :- playsFor(?x,?z), playsSport(?z,?y)
0.73 teamPlaysSport(?x,baseball) :- playsAgainst(?x,yankees)
```

labels B positive. As the ontology grows in complexity, the set of constraints on the learner also grows, resulting in even higher accuracy.

In summary, our approach uses a coupled semi-supervised learning algorithm to acquire extraction strategies for each predicate in the input ontology, and applies these to macro-read millions of web pages to populate that ontology.

2 The ReadTheWeb System

Our current system learns extraction patterns defined over free text and over HTML structure, starting from an initial ontology containing dozens of categories and relations, and 10-15 seed examples of each. The textual pattern learner, CBL [9], iteratively grows a set of extraction patterns while obeying mutual exclusion, subset, and type checking constraints given by the ontology. The HTML pattern learner, SEAL [10], learns patterns of HTML and text tokens that capture regularities such as HTML lists of predicate instances. Based on the belief that these techniques should make independent errors, our system only trusts instances that are extracted by both techniques. Such instances are added to the current beliefs at the end of each iteration, and the process repeats by invoking the subordinate techniques with the newly promoted instances. Figure 1 shows some facts extracted by a recent run of the system (see the complete results at http://rtw.ml.cmu.edu/readtheweb.html).

In a recent experiment involving 16 categories, our current system achieved an average precision of 97% while promoting 4224 category instances. In experiments with CBL which involved an additional 14 relations, it achieved an average precision of 83% for the categories and 84% for the relations while promoting 15520 category instances and 2674 relation instances.

Another component of our system mines the thousands of extracted beliefs, to learn probabilistic Horn clause rules that capture empirical regularities in this data. The resulting rules (Table 1) can then be used to infer additional beliefs to further populate the ontology. In this case the new beliefs are not extracted from text, but are instead inferred from the learned rules and other previously extracted beliefs. Note each learned rule contributes yet another constraint to couple the subsequent training of extractors for the predicates it mentions.

3 Conclusions

We argue that macro-reading the web to populate target ontologies is feasible in the near term, especially as progress continues on coupled semi-supervised learning methods, and intelligent approaches to lightly supervising them. Our preliminary results demonstrate that such an approach can successfully extract tens of thousands of beliefs to populate an input ontology, at imperfect but reasonable accuracy.

One key to our argument is that macro-reading is much easier than solving the full NLP problem. We note that micro-reading will also be important, especially for anotating individual web pages, and for extracting information that is stated only infrequently on the web (e.g., personal information that appears only on a person's home page). One direction for future work is to explore whether and how a macro-reader like ours can help train a micro-reader.

While we believe machine reading will play an important role in populating the semantic web, other approaches will be valuable too, and it is useful to understand different roles these different approaches can play. For example, publishing pre-existing databases may be especially useful for providing deep coverage over fairly narrow domains (e.g., the census data). In contrast, our approach of macro-reading the web may be better suited to populating more broad ontologies, especially with items that are mentioned frequently on the web (and hence easiest to macro-read). Because it can be retrained to fairly arbitrary ontologies, our approach might also be useful for more specialized applications where manual methods are prohibitively expensive.

Acknowledgements. This research has been supported by DARPA, Google, and the Brazilian Research Agency CAPES. Yahoo! Inc. has provided graduate fellowship support as well as access to their M45 computing cluster.

References

1. Nivre, J.: Incremental non-projective dependency parsing. HLT-NAACL, 396–403 (2007)
2. Tjong Kim Sang, E.F., De Meulder, F.: Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In: Proceedings of CoNLL 2003, pp. 142–147 (2003)
3. Vorhees, E.: Overview of TREC 2007. In: TREC (2007)

4. Nigam, K., Andrew McCallum, S.T., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Machine Learning 39, 103–134 (2000)
5. Yarowsky, D.: Unsupervised word sense disambiguation rivaling supervised methods. In: ACL, pp. 189–196 (1995)
6. Blum, A., Mitchell, T.: Combining labeled and unlabeled data with co-training. In: COLT (1998)
7. Riloff, E., Jones, R.: Learning dictionaries for information extraction by multi-level bootstrapping. In: AAAI (1999)
8. Brin, S.: Extracting patterns and relations from the world wide web. In: WebDB (1998)
9. Carlson, A., Betteridge, J., Hruschka Jr, E.R., Mitchell, T.M.: Coupling semi-supervised learning of categories and relations. In: Proceedings of the NAACL HLT 2009 Workshop on Semi-supervised Learning for NLP (2009)
10. Wang, R.C., Cohen, W.W.: Language-independent set expansion of named entities using the web. In: ICDM (2007)