Semantic Web-Based Knowledge Acquisition Using Key Events from News

Frederik Hogenboom Flavius Frasincar Uzay Kaymak

Erasmus School of Economics, Erasmus University Rotterdam
P.O. Box 1738, NL-3000 DR, Rotterdam, the Netherlands
{fhogenboom, frasincar, kaymak}@ese.eur.nl

The full version of this paper, entitled A SEMANTIC WEB-BASED APPROACH FOR PERSONALIZING NEWS appeared in: Proceedings of the Twenty-Fifth Symposium on Applied Computing (SAC 2010), pages 854-861, ACM, 2010

Abstract

Hermes is an ontology-based framework for building news personalization services, which focuses on news classification and knowledge base updating. The framework also allows for news querying and result presentation. In this paper, we focus on the techniques involved in keeping Hermes' internal knowledge base up-to-date. Essentially, our semi-automatic approach to knowledge acquisition from news is based on ontologies and lexico-semantic patterns.

1 Introduction

In today's information-driven world, it is beneficial to be up-to-date with emerging events. Not only regular people benefit from being updated regularly, for instance on common-day matters such as the weather, but also companies merit from being aware of the latest events in for instance their target market, as for example stock markets for financial companies. Common valuable, widely available, yet mostly unstructured sources of information are news messages. With the publishing frequency of most news sources, e.g., Web sites such as Reuters and Bloomberg, it is of utmost importance to be able to extract key events in a timely and efficient manner, and to update one's knowledge base accordingly. Reasoning with up-to-date information contributes to a valuable knowledge base that can serve many purposes.

Knowledge acquisition tasks require both proper extraction techniques, as well as adequate and easily accessible storage facilities. The Hermes news personalization framework [2] combines a Natural Language Processing (NLP) pipeline with Semantic Web domain ontologies. The framework classifies online news messages by identifying their key concepts, and updates its internal knowledge base (modeled by means of a domain ontology) based on discovered events. Also, the framework provides for news query execution and result presentation.

As updating the knowledge base is one of the most vital tasks within such frameworks, this paper focuses on the techniques involved in keeping Hermes' internal knowledge base up-to-date. Sections 2 and 3 elaborate on the Hermes framework in general, and more specifically on classification and knowledge base updating, respectively. Finally, Section 4 wraps up this paper by presenting results of an implementation of the framework, i.e., the Hermes News Portal (HNP).

2 Classification

When news items are announced through RSS feeds, the Hermes framework fetches these messages and processes them using an NLP pipeline. Text processing is done by means of an NLP pipeline based on the GATE framework [1]. The pipeline accounts for tokenization, sentence splitting, Part-Of-Speech (POS) tagging, morphological analysis, ontology gazetteering, and Word Sense Disambiguation (WSD).

Incoming news messages and their discovered concepts are stored in a domain ontology that contains the most important target domain concepts, so that future queries can be done in timely manner, avoiding superfluous text processing. Concepts are mapped to their corresponding sets of synonyms from a semantic lexicon (WordNet [3]). These sets provide domain-independent lexical representations for associated concepts, which complement domain-specific lexical representations stored in the domain ontology.

3 Knowledge Base Updating

After classification, a necessary step within the Hermes framework is updating the knowledge base. This is done by means of lexico-semantic patterns, which are designed by hand and define events using lexico-semantic arguments based on ontological classes. These patterns have one or more associated update actions that are to be executed once elements from a news message match these patterns. Within the Hermes framework, events discovered using such patterns are manually validated, to ensure a correct knowledge base.

Patterns are defined by a subject, a relation, and optionally an object. An example is `[kb:Person] kb:BecomesCEO [kb:Company]`, which identifies CEO changes in a company. Square parentheses indicate lexical representations of individuals of the enclosed type, whereas the lack of square parentheses indicates that only lexical representations of the given instance are taken into consideration.

Once the event has been manually validated, the Hermes framework updates the ontology by means of action rules that make use of SPARQL/Update [4]. The action rules are ordered, e.g., removing old CEOs before adding new CEOs to prevent incorrect updates. Furthermore, rules are executed in the order of event appearances in news. After executing all associated actions, the event effects are captured in the ontology.

4 Results

We implemented the Hermes framework in the Hermes News Portal (HNP), which allows for browsing a knowledge base, querying relevant news items, and semi-automatically updating a knowledge base. Experiments on 200 news items extracted from Yahoo! Business and Technology news feeds show 86% precision and 81% recall on concept identification, and 64% precision and 53% recall on pattern-based matching, which utilizes multiple identified concepts. Usability tests show that user interaction with the system, needed for knowledge base updating, is positively assessed by the users.

Acknowledgements

The authors are partially sponsored by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT).

References

[1] Hamish Cunningham. GATE, a General Architecture for Text Engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[2] Flavius Frasincar, Jethro Borsje, and Leonard Levering. A Semantic Web-Based Approach for Building Personalized News Services. *International Journal of E-Business Research*, 5(3):35–53, 2009.

[3] Marti A. Hearst. *Automated Discovery of WordNet Relations*, chapter 5, pages 131–151. WordNet: An Electronic Lexical Database and Some of its Applications. MIT Press, 1998.

[4] Andy Seaborne, Geetha Manjunath, Chris Bizer, John Breslin, Souripriya Das, Ian Davis, Steve Harris, Kingsley Idehen, Olivier Corby, Kjetil Kjernsmo, and Benjamin Nowack. SPARQL Update – A Language for Updating RDF Graphs, W3C Member Submission 15 July 2008, 2008.