SEMI-AUTOMATIC ONTOLOGY EXTENSION USING TEXT MINING

Inna Novalija, Dunja Mladenić
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144
e-mail: inna.koval@ijs.si, dunja.mladenic@ijs.si

ABSTRACT

This paper addresses the process of the ontology extension for a selected domain of interest which is defined by keywords and possibly a glossary of relevant terms. A new methodology for semi-automatic ontology extension, aggregating the elements of text mining and user-dialog approaches for ontology extension, is proposed and evaluated. We conduct a set of ranking, tagging and illustrative question answering experiments using Cyc ontology and business news collection. The experiments show that the precision of business news tagging increases from 61% to 87% and the corresponded recall increases from 46% to 81% after the ontology extension with concepts extracted from business news.

1 INTRODUCTION

This paper explores the process of the ontology extension motivated by usage of the extended ontology for business news analysis. The main contribution of this paper is in proposing a new methodology for semi-automatic ontology extension based on text mining and user-dialogue approaches. Our research also contributes to the analysis of business news by the means of semantic technologies. The new methodology for the semi-automatic ontology extension, aggregating the elements of text mining and user-dialog approaches for ontology extension, is suggested and used for inserting the new financial knowledge into Cyc [10], which maintains one of the most extensive common-sense knowledge bases worldwide.

The experiments on ranking, business news annotation and simple question answering show that the extended financial ontology allows for a better financial news analysis.

The evaluation of the methodology of the ontology extension shows its ability to fasten the ontology extension process.

The paper is structured as follows: Section 2 presents the information about the existing approaches of ontology extension; the new methodology of ontology extension is discussed in Section 3, Sections 4 and 5 describe the experiments and the results, the discussion and future work are covered in Section 6.

2 EXISTING APPROACHES OF ONTOLOGY EXTENSION

The automatic and semi-automatic ontology extension processes are usually composed of several phases. Most approaches include defining the set of the relevant ontology extension sources, preprocessing the input material, ontology augmentation according to the chosen methodology and ontology evaluating and revision phases. The notable approaches of ontology extension include natural language processing based approach [2] [7], networks/graphs based approach [5] [6], user-dialogue based approach [8] and pattern based approach [1].

3 METHODOLOGY

As a part of the research, we propose a new methodology for semi-automatic ontology extension, which combines text mining methods with user-oriented approach and supports the extension of multi-domain ontologies. The proposed methodology embodies three main modules: the Domain Information Module (DIM), the Domain Subset Extraction Module (DSEM) and the Ontology Extension Module (OEM) and deals with an extendable multi-domain ontology.

The proposed methodology for semi-automatic ontology extension accounts for the following phases, displayed with numbers in Figure 1:

1. *Domain information identification.* The domain experts identify the appropriate Domain Keywords. As well, in DIM a Domain Relevant Glossary, containing concepts with descriptions is determined.
2. *Extraction of the relevant domain ontology subset from multi-domain ontology.* In DSEM the Keywords are used by the Upper-Level Domain Extractor to restrict the multi-domain Ontology to the specific domains of interest. Afterwards, the Domain Knowledge Extractor uses a Domain Relevant Glossary to obtain the Ontology Subset for the particular domains of interest.
3. *Domain relevant information preprocessing.* The information from the Domain Relevant Glossary and the extracted relevant Ontology Subset are linguistically preprocessed in OEM. The preprocessing phase

includes tokenization, stop-word removal and stemming.

4. *Composing the list of potential concepts and relationships for ontology extension.* The ranked list of the relevant concepts and possible relationships suitable for ontology extension is composed in OEM.
5. *User validation.* Furthermore, in OEM the user validates the initial results and the final list of ontology extension concepts and relationships in the relevant format is created.
6. *Ontology extension.* The Ontology extension is taking place in the Ontology Extension Module. It represents adding the new concepts and relationships between concepts into the Ontology.
7. *Ontology reuse* As a part of the new extension process, we reuse the previously extended Ontology in DSEM and in OEM.

We have adapted the methodology in order to obtain an exhaustive specific methodology for Cyc knowledge base extension. The main adaptations are based on microtheories (Mt) that Cyc is using to represent thematic subsets of the ontology. Usage of Knowledge Entry (KE) concept templates helps to incorporate the user feedback. Since our motivation is in business news annotation, we have chosen Business and Finances as the domains of primary interest. Given the fact that Cyc knowledge base contains common sense knowledge [4], we assume that Cyc KB includes some financial knowledge - a financial knowledge base (Cyc FKB).

4 EXPERIMENTS

In order to evaluate the proposed methodology we conducted a series of ranking, news tagging and illustrative question answering experiments on the data sources, described below.

For the data evaluation we have used the RSS feeds data from CNN [9], Reuters [14], Forbes [12], Financial Times (FT) [13] and Yahoo! Finance [11] websites.

The news collection used in the current research accounts for 1455 Reuters news, 4584 CNN news, 5812 Yahoo! Finance news, 15374 Forbes news and 34311 Financial Times news.

Following the first phase of the proposed methodology, domain knowledge identification should be made in the initial phase. For these purposes we have selected the Harvey [3] financial glossary which contains around 6000 hyperlinked financial terms.

In order to evaluate the suggested methodology, we have conducted ranking experiments on the subset of 500 random Yahoo! Finance news. The most frequent financial terms have been extracted and 100 random financial terms have been chosen. Cyc Financial Knowledge Base is then extended, using the proposed methodology, with concepts corresponding to the chosen financial terms. The efficiency of the automatic concept ranking is measured afterwards.

Tagging experiments show how the business news tagging improves after ontology extension with the domain relevant glossary. The tagging/annotation experiments provide testing on a random subset of 100 Yahoo! Finance news. We have identified the financial terms, occurring most frequently in the selected news, tagged the terms with Cyc Concept Tagger and checked the precision and recall of news tagging. Furthermore, we have added the simplest assertions about the missing financial terms into Cyc and again found the precision and recall of news tagging.

Illustrative question answering demonstrates the capacity of Cyc to answer simple financial questions before and after the extension of Cyc Financial Knowledge Base. Let us assume that we have a simple question and we want to get an answer using an unextended and extended Cyc Knowledge Base. The experimental results are given in Section 5.

5 RESULTS

The results of the experiments suggest that the financial ontology extension leads to better business news annotation and confirm the applicability of the suggested methodology for ontology extension to Cyc Knowledge Base augmentation.

5.1 Ranking

The proposed methodology results by the user geting a ranked list of relevant Cyc financial concepts for each glossary concept and confirms the relationships between the glossary and Cyc concepts. The comparison of the automated and manual rankings shows that in 60 out of 100 cases the correct concept can be found among the top 10 automatically suggested concepts for each glossary term. More precisely, for 23 out of 100 glossary terms the automatically suggested concept occupied the first position in manual ranking. It means that using the proposed methodology the user is able to compare Cyc and glossary concepts and establish the equivalence and parent-child relationships much faster than just using the manual search for the relevant concepts in Cyc.

5.2 Business News Tagging

We have found 231 financial terms in the random sample of 100 Yahoo! Finance news. The precision of business news tagging increases from 61% to 87% and the corresponded recall - from 46% to 81% after adding the simplest assertions about the missing terms into Cyc. This is confirming our hypothesis that the Cyc ontology has still space for extension as for financial domain with terms relevant for financial news analysis.

5.3 Illustrative Question Answering

In this section we illustrate relevance of the proposed Cyc ontology extension for question answering in the financial domain.

For the research purposes we have selected the following simple question:

“Which stock exchanges exist in Western Europe?” which might be translated in CycL query as:

```
(#$and
  (#$isa ?X #$StockExchange)
  (#$residenceOfOrganization ?X
#$WesternEurope))
```

Using an unextended Cyc KB we get no appropriate answers because of the insufficient representation of stock exchange instances and residence relationships in Cyc. After we extended Cyc KB, using the proposed methodology, with the following new assertions about Frankfurt stock exchange:

```
Constant: FrankfurtStockExchange.
In Mt: OrganizationDataMt.
isa: StockExchange.
In Mt: OrganizationDataMt.
residenceOfOrganization:
CityOfFrankfurtGermany.
```

Extension of Cyc Knowledge Base according to the proposed methodology with European stock exchange instances allows the user to provide Cyc with new stock exchange instance and get the following answer for the asked question:

```
[Explain #0] FrankfurtStockExchange
```

6 CONCLUSION

In this paper the aspects of ontology extension and business news analysis have been explored. The new methodology of ontology extension, combining text mining methods and user-based approach, has been proposed and exposed to the preliminary evaluation.

In contrast with many other methodologies for ontology extension, our methodology deals with ontologies and knowledge bases, covering more than one domain. However, it allows restricting the area of ontology extension to a specific domain. Furthermore, the user validation helps to avoid adding to the ontology irrelevant concepts and relationships.

The future work should include further extension of Cyc Knowledge Base and using it for more sophisticated news analysis. Furthermore, the proposed methodology for ontology extension should be tested on other domains. In addition, a particular attention should be given to the problem of the disambiguation of the glossary terms and terms extracted from news sources.

7 ACKNOWLEDGMENTS

This work was supported by the Slovenian Research Agency and the IST Programme of the EC under NeOn (IST-4-027595-IP) and ACTIVE (IST-2008-215040).

References

[1] Blomqvist, E.: Pattern-based Ontology Construction. In: *KWEPSY*. 2007.

[2] Burkhardt, F., Gulla, J. A., Liu, J., Weiss, C., Zhou, J.: Semi Automatic Ontology Engineering in Business Applications. *Workshop Applications of Semantic Technologies*, INFORMATIK. 2008.

[3] Harvey, C.R.: Yahoo Financial Glossary, Fuqua School of Business, Duke University. 2003.

[4] Lenat, D.: Cyc: A Large-Scale Investment in Knowledge Infrastructure. *Communic. of the ACM 38 (11).* 1995.

[5] Liu, W., Weichselbraun, A., Scharl, A. ,Chang, E.: Semi-Automatic Ontology Extension Using Spreading Activation. *Journal of Universal Knowledge Management*, No. 1, pp. 50 – 58. 2005.

[6] McDonald, J., Plate, T., Schvaneveldt, R.: Using pathfinder to extract semantic information from text. In: *Schvaneveldt*, pp. 149–164. 1990.

[7] Sabrina T., Rosni A., Enyakong T.: Extending Ontology Tree Using NLP Technique. In: *Proceedings of National Conference on Research & Development in Computer Science REDECS 2001.* 2001.

[8] Witbrock, M., Baxter, D., Curtis, J., Schneider, D., Kahlert, R., Miraglia, P., Wagner, P., Panton, K., Matthews, G., Vizedom, A.: An Interactive Dialogue System for Knowledge Acquisition in Cyc. In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence.* 2003.

[9] CNN News, http://www.cnn.com

[10] Cycorp, Inc., http://www.cyc.com

[11] Yahoo! Finance, http://finance.yahoo.com

[12] Forbes News, http://www.forbes.com

[13] Financial Times News, http://www.ft.com/home/europe

[14] Reuters News, http://www.reuters.com

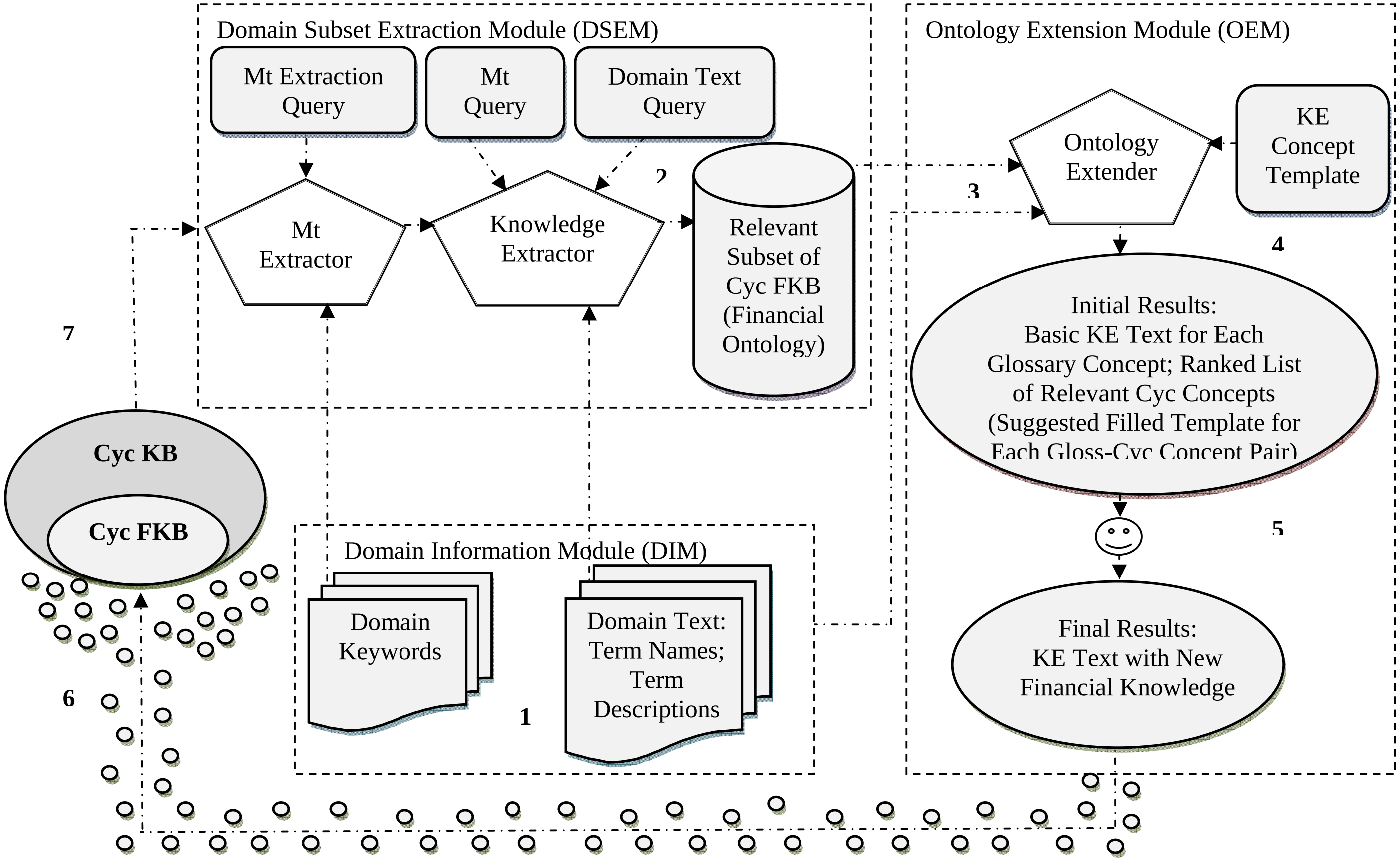
Domain Subset Extraction Module (DSEM)
Mt Extraction Query
Mt Query
Domain Text Query
Mt Extractor
Knowledge Extractor
2
Relevant Subset of Cyc FKB (Financial Ontology)
Ontology Extension Module (OEM)
Ontology Extender
KE Concept Template
3
4
Initial Results: Basic KE Text for Each Glossary Concept; Ranked List of Relevant Cyc Concepts (Suggested Filled Template for Each Gloss-Cyc Concept Pair)
5
Final Results: KE Text with New Financial Knowledge
7
Cyc KB
Cyc FKB
6
Domain Information Module (DIM)
Domain Keywords
Domain Text: Term Names; Term Descriptions
1

Figure 1: *Methodology for Semi-Automatic Ontology Extension (Cyc Adaptation).*