EXTENDING ONTOLOGIES FOR ANNOTATING BUSINESS NEWS

Inna Novalija, Dunja Mladenić
Department of Knowledge Technologies
Jozef Stefan Institute
Jamova 39, 1000 Ljubljana, Slovenia
Tel: +386 1 4773144; fax: +386 1 4773315
e-mail: inna.koval@ijs.si

ABSTRACT

Ontologies are commonly used for annotating textual data mainly based on human language technologies [1]. This research focuses on manual extensions of ontologies to support the annotation of business news. Experiments were conducted on a well known Cyc ontology and using Cyc annotator on two business news datasets. We show that the proposed extensions of ontology results in annotation with better coverage of terms that are relevant for the business domain. The results of identifying financial terms in business news using the original Cyc ontology show the average precision of 56% and recall of 41% in case of Reuters news and the average precision of 69% and the recall of 57% in case of Yahoo financial news. Using the proposed extension results with increased performance, the average precision of 82% and average recall of 73% for Yahoo financial news and average precision of 84% and average recall of 63% for Reuters news.

1 INTRODUCTION

News reports are considered to be one of the largest sources of information about society. The analysis of news allows to make the important conclusions about trends in the society life. Since the news domain has several characteristics that difference it from other domains, the semantic technologies might be a good choice for news analysis [3]. One of the goals of this research is to contribute to the analysis of the financial news by the means of semantic technologies - in particular by creating and extending the Financial ontology in Cyc, which is known to have one of the largest knowledge base in the world.

There exists several challenges while using semantic technologies in the news analysis. News are dynamic, interactive and socially biased. News agencies produce huge amounts of content.

According to Jarrar [6], the following challenges should be dealt with while building any kind of ontology: Ontology reusability, Ontology application/task-independence and Ontology evolution.

The challenges of Ontology reusability and Ontology application/task-independence can be efficiently handled by Cyc, given the fact that Cyc allows using its extensive knowledge base for different tasks and purposes. Ontology evolution can also be followed in Cyc and in case of Financial ontology it should be considered as an extremely important challenge.

The creation of the Financial Ontology might be difficult due to several reasons. According to Zhang, Zhang and San Ong [12], in the financial environment the tasks are dynamic, distributed, global, and heterogeneous in nature. They are characterized by the large amount of continually changing, and generally unorganized, information available, the variety of all kinds of information (like market data, financial report data, breaking news, etc.) and many sources of uncertainty in the environment.

Mónica Martínez Montes et al. [9] as well mention several reasons explaining why the creation of the ontologies in the financial domain is difficult. Slow standardization efforts and high complexity of the financial standards, high competition and dynamics of the financial sector influence the implementation of the new technologies. Consequently, there exists a very few number of ontologies connected to the financial sphere of life. At the same time, there is a high necessity in the creation of the extensive financial ontologies which could be effectively used and reused by the financial institutions.

2 METHODOLOGY

In order to create a coherent and relevant ontology a number of methodological principles or criteria should be considered.

2.1 Design criteria for Ontology Development

Gruber [4] defines the following design criteria for ontology developing: Clarity, Coherence, Extendibility, Minimal encoding bias and Minimal ontological commitment.

Jarrar [6] states two additional methodological principles: Ontology double articulation principle and Ontology modularization principle. Since financial news are dynamic

and heterogeneous, Clarity is one of the most important principles in case of Financial ontology.
In view of the fact that there exist a very limited number of ontologies in the financial domain, Extendibility and Ontology modularization principles can be considered essential as well.

2.2 Overview of Methodologies

There exist several methods and methodologies of the ontology creation: Cyc method [7], Uschold and King's method [11], Grüninger and Fox's methodology [5], METHONTOLOGY [2], On-To-Knowledge [10] etc.
Uschold and King's methodology for developing ontologies includes the following stages:
- Indentify Purpose.
- Building the Ontology.
    - Ontology capture.
    - Ontology coding.
    - Integrating Existing Ontologies.
- Evaluation.
- Documentation.

The methodology by Grüninger And Fox can be described by the next steps:
- Capture of motivating scenarios.
- Formulation of informal competency questions.
- Specification of the terminology of the ontology within a formal language.
    - Getting informal terminology.
    - Specification of formal terminology.
- Formulation of formal competency questions using the terminology of the ontology.
- Specification of axioms and definitions for the terms in the ontology within the formal language.
- Establish conditions for characterizing the completeness of the ontology.

One the most famous and frequently used methodologies are METHONTOLOGY and On-To-Knowledge methodology. According to the developers of METHONTOLOGY, the METHONTOLOGY framework includes:
- The identification of the ontology development process.
- A life cycle based on evolving prototypes and
- the methodology itself, which specifies the steps for performing each activity, the techniques used, the products to be output, and how the ontologies are to be evaluated.

On-To-Knowledge methodology distinguishes such phases of the ontology development:
- Feasibility Study.
- Kickoff.
- Refinement.
- Evaluation and
- Application & Evolution.

Due to the fact that Cyc gives an extremely powerful possibility of creating and using different ontologies, Cyc method has been chosen as a main methodology in our research. Cyc method and Cyc knowledge base are widely discussed in the following chapter.

3 CYC

Cyc Knowledge Base (Cyc KB) appears to be one of the largest knowledge bases in the contemporary IT world. It is stated as "a formalized representation of a vast quantity of fundamental human knowledge: facts, rules of thumb, and heuristics for reasoning about the objects and events of everyday life" [13] and divided into the large number of "microtheories", each of which represents the set of assumption for a particular knowledge domain.

At the present time, the Cyc KB contains nearly two hundred thousand terms and several dozen hand-entered assertions about/involving each term. New assertions are continually added to the KB by human knowledge enterers. Additionally, term-denoting functions allow for the automatic creation of millions of non-atomic terms, such as (LiquidFn Nitrogen); and Cyc adds a vast number of assertions to the KB by itself as a product of the inferencing process. [13]

According to Cyc method [7], the phases to build the Cyc ontology are following:
- Manual encoding of the explicit and implicit knowledge appearing in the knowledge sources.
- Knowledge codification that is aided by tools using knowledge already stored in the Cyc KB.
- Delegating to the tools the majority of the work.

In each phase two tasks are performed: 1. Development of a knowledge representation and top level ontology containing the most abstracts concepts. 2. Representation of the rest of the knowledge using this primitives.
There exist several reasons for using Cyc including the following. The large number of assertions currently existing in Cyc KB. Existance of several version of the system available under different licences (OpenCyc, ResearchCyc). Flexible and convenient language (CycL). Suitable interface.

The Cyc method can be also considered the most useful for news analysis due to the extensive amount of versatile integrated information in the Cyc KB. Controversially, METHONTOLOGY and OTK methodology can be more useful for the creation ontologies which are going to be used in the particular applications.

4 PRELIMINARY EXPERIMENTS

In spite of the fact that Cyc contains a very extensive knowledge base, the representation of the financial and economical information in it is far from complete. As an

experiment, two sets of the financial news - one from a well known Reuters news archive and another from Yahoo Finance news archive - have been analyzed.

4.1 Data Description

Reuters news archive contains the selected collection of 1450 news in 1996 year. News are categorized into 354 categories which enables an easy identification of a subset of business news [8] We have taken all the news that are assigned one the following categories:

I8 FINANCIAL AND BUSINESS SERVICES
I81 BANKING AND FINANCIAL SERVICES
I82 INSURANCE
I831 FINANCIAL SERVICES
I84 RENTING AND LEASING EQUIPMENT
I85 REAL ESTATE DEALING

Yahoo news for the experiment have been taken from the news archive on the Yahoo! finance website [14] It contains the raw, uncategorized, news materials for the last three months (in our case, second half of May – first half of August 2008) all together around 26000 news articles. However, all news from the archive have financial connotation.

4.2 Experimental results

Experiments were conducted on a randomly selected sample of ten documents from each dataset. The documents were manually annotated for financial terms. Then Cyc annotator was applied on them using the original Cyc ontology as well as our extension of the original Cyc ontology.

Table 1. shows the results of Cyc annotation of random samples of Reuters financial news. The first column contains the name of the news article. Then we give the total number of the words in the article, followed by the number of financial terms in the article (as selected manually). The forth column gives the number of financial terms tagged by Cyc and the fifth column contains the number of financial terms tagged by Cyc correctly. The following two columns contain the precision and recall of tagging of the financial terms by Cyc and the last two columns contain the precision and recall after adding to Cyc mis-tagged and untagged terms from Yahoo Financial Glossary. The results using the original Cyc ontology show the average precision of 56% and recall of 41% (in case of Reuters news) and average precision of 69% and the recall of 57% in case of Yahoo financial news (table with Yahoo results is omitted due to space restrictions).

Consequently, improving Cyc knowledge on finances may be a key factor in a better analysis of the financial news using Cyc. In order to estimate possible improvement of an annotator with the proposed extension of the ontology we have provisionally added to the Cyc ontology all untagged and mis-tagged financial terms that appear in the sample of the news. The experiment shows that when we add all the untagged and mis-tagged financial terms to Cyc ontology, the average precision increases up to 99% and average recall to 98%.

However, to avoid the assumption that we already know all the relevant terms that need to be annotated and thus possibly added for a particular document set, we have checked the overlap between such terms in our sample and a publicly available glossary of financial terms. We found that around 50% (52% for Yahoo financial news and 54% for Reuters) of the financial terms untagged or tagged incorrectly by Cyc, can be found in the publicly available glossary of financial terms (Yahoo Financial Glossary).

Adding mis-tagged and untagged terms from Yahoo Financial Glossary increases the average precision to 82% and average recall to 73% for Yahoo financial news and average precision to 84% and average recall to 63% for Reuters news.

This means that by extending Cyc ontology by the terms from the glossary we can considerably improve precision and recall of the annotator. Based on the experimental results we conclude that the conducted experiment shows the insufficient representation of financial domain in Cyc and the ways to effectively improve it by the means of extension of Cyc financial knowledge base.

5 DISCUSSION AND CONCLUSIONS

This research focuses on manual extensions of ontologies to support the annotation of business news. Experiments were conducted on Cyc ontology on two business news datasets (Reuters and Yahoo). We show that the proposed extensions of ontology results in news annotation with better coverage of terms that are relevant for business domain. However, we belive that the proposed approach can be applied to other domains beyond business news.

In this research we are proposing extending ontology for better coverage of business terminology by adding terms from financial glossary. However, other relevant terms from different sources of financial information can be added, such as, terms describing stock market mechanisms and stock exchange instances. Extension of the Cyc ontology in direction of stock exchange is ongoing and its evaluation is part of our future work.

6 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] K. Bontcheva, H. Cunningham, A. Kiryakov and V. Tablan. Semantic Annotation and Human Language Technology. In *Semantic Web Technology: Trends and Research*., J. Davies, R. Studer, P. Warren (eds.). John Wiley and Sons Ltd. 2006.

[2] O. Corcho O, M. Fernández-López, A. Gómez-Pérez, A. López-Cima. Building legal ontologies with METHONTOLOGY and WebODE. Springer-Verlag, LNAI. 2005.

[3] M. Grobelnik, D. Mladenic. Visualisation of News Articles. *Informatica journal, vol. 28, No. 4*. 2004.

[4] R. Gruber Thomas. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int. J. Hum.-Comput. Stud., Vol. 43, No. 5-6.* 1995.

[5] M. Gruninger, M. Fox. The role of competency in enterprise engineering. *IFIP WG5.7 Workshop on Benchmarking- Theory and Practice.* IFIP, Norway, 1994.

[6] M. Jarrar. Towards Methodological Principles for Ontology Engineering. *PhD thesis*, Vrije Universiteit Brussel, 2005.

[7] D.B. Lenat, R.V. Guha, Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project, Addison-Wesley, Boston, 1990.

[8] D. Lewis, Y. Yang, T. Rose and F. Li. RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research, 5:361-397*. 2004.

[9] M. Martínez Montes, J. Bas, S. Bellido, O. Corcho, S. Losada, R. Benjamins, J. Contreras. *WP10: Case study eBanking D10.3 Financial Ontology*. 2005.

[10] Y. Sure, R. Studer. On-To-Knowledge Methodology - Final Version. *On-To-Knowledge deliverable D-18,* Institute AIFB. University of Karlsruhe, 2002.

[11] M. Uschold, M. King. Towards a methodology for building ontologies. *Proceedings of Workshop on Basic Ontological Issues in Knowledge Sharing*, IJCAI-95. Canada, 1995.

[12] Z. Zhang, C. Zhang and S. San Ong. Building an Ontology for Financial Investment. *Proceedings of the Second International Conference on Intelligent Data Engineering and Automated Learning, Data Mining, Financial Engineering, and Intelligent Agents.* Lecture Notes In Computer Science; Vol. 198,. 2000.

[13] http://www.cyc.com/cyc/technology/whatiscyc_dir/whatsincyc

[14] http://biz.yahoo.com/top.html

[15] http://biz.yahoo.com/f/g/

Table 1. Financial news tagged by Cyc (Reuters)

Article name	Total words	Fin. Terms	Fin. Terms tagged	Fin. Terms tagged correctly	Precision, %	Recall, %	Precision after adding terms from Yahoo glossary, %	Recall after adding terms from Yahoo glossary, %
Doc. 1 Lloyd's of London serves notice of emergency stay.	648	14	13	8	62%	57%	93%	93%
Doc. 2 UK Lloyd's moves to ward off doubts on recovery.	489	13	9	6	67%	46%	75%	69%
Doc. 3 CANADA: Canadian banks poised for higher third-quarter profits.	612	45	30	17	57%	38%	69%	40%
Doc. 4 Malaysia's Intria buys into two construction firms.	432	24	15	9	60%	38%	80%	50%
Doc. 5 Slovak PM sees banks releasing funds for bad debts.	156	10	10	2	20%	20%	90%	90%
Doc. 6 ArgentBank to buy Assumption Bank & Trust.	64	9	7	6	86%	67%	100%	78%
Doc. 7 Nationwide, Halifax bid for UK defense sale - paper.	123	9	5	4	80%	44%	80%	44%
Doc. 8 Australia: Current Australian Takeovers (A to E) - Aug 26.	481	19	17	8	47%	42%	82%	74%
Doc. 9 AUSTRALIA: RTRS-Australia's COAL in A$300 mln float - paper.	365	18	13	4	31%	22%	100%	83%
Doc. 10 China state firms form new insurance company.	67	11	7	6	86%	55%	100%	64%
Avrg.	344	17	13	7	56%	41%	84%	63%