Financial Events Recognition in Web News for Algorithmic Trading

Frederik Hogenboom

Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands
fhogenboom@ese.eur.nl

Abstract. Due to its high productivity at relatively low costs, algorithmic trading has become increasingly popular over the last few years. As news can improve the returns generated by algorithmic trading, there is a growing need to use online news information in algorithmic trading in order to react real-time to market events. The biggest challenge is to automate the recognition of financial events from Web news items as an important input next to stock prices for algorithmic trading. In this position paper, we propose a multi-disciplinary approach to financial events recognition in news for algorithmic trading called FERNAT, using techniques from finance, text mining, artificial intelligence, and the Semantic Web.

1 Introduction

Recently, financial markets have experienced a shift from the traditional way of trading using human brokers to the use of computer programs and algorithms for trading, i.e., algorithmic trading. Trading algorithms implemented in business tools have proven to be more efficient than conventional approaches for trading, as they provide for lower latency, larger volume, and higher market coverage degree. During the last years, it has been acknowledged the need to use Web news information in algorithmic trading in order to react real-time to market events and to enable better decision making. Responding to market events only a few milliseconds faster than the competition could mean better prices and therewith improved profitability for traders. The biggest challenge is to allow machines to identify and use the news information that is relevant for technical trading timely and accurately.

Financial markets are extremely sensitive to breaking news [22]. Financial events – phenomena that are captured in keywords pointing to specific (complex) concepts related to money and risk – like mergers and acquisitions, stock splits, dividend announcements, etc., play a crucial role in the daily decisions taken by brokers, where brokers can be of human or machine nature. Algorithmic trading enables machines to read and understand news faster than the human eye can scan them, hence allowing one to deal with larger volumes of emerging online news, and making thus better informed decisions.

The Semantic Web provides the right technologies to unambiguously identify or "tag" the semantic information in news items and represent it in a machine-understandable form. Having this information in machine-understandable form enables computers to reason and act as we humans would do. Realizing the potential the Semantic Web has to offer in making the news information semantically available, large news companies like Reuters and Dow Jones started to provide product services that offer tagged news items to be used for algorithmic trading [27].

The current annotations provided by the above vendors are *coarse-grained*, as they supply general information about the type of information available in news items, as for example company, topic, industry, etc., satisfying thus to a limited extent the information need in financial markets. For algorithmic trading, a *fine-grained* annotation [8] that allows the identification of financial events as acquisitions, stock splits, dividend announcements, etc., is needed. Additionally, most annotations are merely based on article titles instead of contents, and financial events (if any) are not linked to ontologies (hence making reasoning and knowledge inference difficult).

To our knowledge the semi-automatic recognition of financial events as a support tool for algorithmic trading has not been thoroughly investigated in previous work. Several innovative aspects play a key role here: defining a *financial ontology* for algorithmic trading, using *lexico-semantic rules* for identifying financial events in news, applying *ontology update rules* based on the previously extracted information, and employing the financial events to improve the returns generated by *trading algorithms*. In recent work, we have focused on the first two aspects. Our main contribution in the field of news analysis is the Hermes framework [12], which makes use of Natural Language Processing (NLP) techniques and Semantic Web technologies for news personalization. Additionally, we researched financial ontologies and financial event detection pipelines [4, 14] and we have introduced a lexico-semantic pattern language for Hermes [15] which is able to extract financial events from text using a financial ontology.

In light of our existing work, this paper presents the Financial Events Recognition in News for Algorithmic Trading (FERNAT) framework, which aims to automate the identification of financial events in emerging news and to apply these events to algorithmic trading. Not only does the proposed framework make use of an NLP pipeline, a financial ontology, and lexico-semantic patterns for event extraction resulting from earlier work, but it also implements a feedback loop using ontology update rules. Additionally, the discovered events are used for financial applications for risk analysis or algorithmic trading.

2 Related Work

This section discusses related work with respect to information extraction frameworks, and compares these with our proposed FERNAT framework. Additionally, we discuss work on trading in the financial markets and algorithmic trading.

2.1 Information Extraction

For the information extraction methods, we distinguish between general-purpose text processing pipelines and news-based processing frameworks. Examples of general-purpose text processing pipelines are A Nearly New Information Extraction System (ANNIE) [6] and Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations (CAFETIERE) [3]. For news-based processing frameworks we identify PlanetOnto [7] and SemNews [16].

Both ANNIE and CAFETIERE are able to cope with the domain semantics to a limited extent. For gazetteering, their pipelines use a list of words for which the semantics are not defined. Also, the information extraction rules are based on lexico-semantic patterns that apply only to very concrete situations. These rules are written in JAPE [6] using Java, a low-level format which makes rules development and maintenance rather tedious. The MUlti-Source Entity finder (MUSE) [21] uses the ANNIE pipeline for named entity recognition going through the rather difficult process of defining JAPE rules for information extraction.

There are a number of tools for Ontology-Based Information Extraction (OBIE) that have adapted the ANNIE pipeline to be used in combination with ontologies. Examples of such tools are the Ontology-based Corpus Annotation Tool (OCAT) [20] that has been used for annotating documents for business intelligence purposes, and the Knowledge and Information Systems Management (KIM) platform [26], a generic annotation and document management system. While these tools benefit from the information stored in ontologies for knowledge acquisition, due to the direct use of JAPE rules, they fail to deliver an easy-to-use, high-level language for information extraction rules specification.

Differently than ANNIE, CAFETIERE provides high-level information extraction rules which makes it easier to write and update rules. Although the extraction rules are defined at lexico-semantic level, CAFETIERE does not employ ontologies (knowledge bases) that are Semantic Web-based. For this, CAFETIERE uses a specific representation, i.e., Narrative Knowledge Representation Language (NKRL), a knowledge representation language which is defined before the Semantic Web era. With the advent of the Semantic Web, we believe that both gazetteering and lexico-semantic rules can benefit from an ontology-based approach based on standards and proven tool support. Also, both ANNIE and CAFETIERE do not update their knowledge bases with extracted information that is possibly helpful in the next information extraction run.

PlanetOnto represents an integral suite of tools used to create, deliver, and query internal newsletters of the Knowledge Media Institute (KMi). Similar to the approach proposed here, domain ontologies are used for identifying events in news items. While we aim at semi-automatic information extraction from news items, PlanetOnto uses a manual procedure for identifying information in news items. SemNews on the other hand uses a domain-independent ontology for semi-automatically translating Web pages and RSS feeds onto meaningful representations given as OWL facts. For this purpose it uses OntoSem [25], an NLP tool which performs lexical, syntactic, and semantic analysis of text.

OntoSem has a specific frame-based language for representing the ontology and an onomasticon for storing proper names. In our work both the input ontology and the facts extracted from news items are to be represented in OWL. Also, our approach proposes to use, instead of an onomasticon, a semantic lexicon, a richer knowledge base that can better support the semantic analysis of text.

Many of the current approaches for automating information extraction from text use rules based on lexico-syntactic patterns. As these rules do not take into account the semantics of the different constructs involved in a pattern we do find such an approach limited in expressiveness. In our previous work [15] we aim at exploiting lexico-semantic patterns, which remove some of the ambiguity inherent to the lexico-syntactic rules. In addition, the proposed rules provide a higher abstraction level than lexico-syntactic rules, making rule development and maintenance easier.

2.2 Financial Markets

Financial markets are strongly dependent on information, and thus also on emerging news messages. Traders – whether they are technical or fundamental traders – use information in their decisions on selling and buying stocks, thus influencing the financial market. Processing and interpreting relevant information in a timely manner can be of crucial importance for the profitability of trading activities. Predicting the future course of stocks within a financial market is hard, which led to the development of theories on stock prediction, such as the random walk theory [9] and the efficient market hypothesis [10], that both recognize the influence of available information on the market.

As shown in the previous section, an extensive body of literature is available on processing text to a machine-understandable format. Also, a lot of research has been done for the prediction of market reactions to news (see [23] for an extensive survey). Many existing approaches aim to forecast price trends based on emerging news and mainly employ statistical text mining approaches to classify financial events (e.g., positive or negative). Price trends based on news messages can be used in automated trading algorithms. Examples of these algorithms are the Penn-Lehman Automated Trader (PLAT) [17], the Artificial Stock Trading Agent (ASTA) [13], and genetic algorithms-based financial trading rules [1].

Algorithmic trading encompasses the use of computer programs for trading purposes, which is of interest to traders as this greatly enhances trading speed, and thus increases profit expectations. Algorithms are employed for instance for correlation analyses and the identification of opportunities and risks. These algorithms are based on inputs such as statistics on the financial market, but also price trends. These price trends can be calculated based on both historical and real-time market data. However, real-time market data as for example news information is often inaccurate or too coarse to be of great value. Thus, improving processing speed and accuracy of real-time information would be beneficial for algorithmic trading.

3 FERNAT Framework

The Financial Events Recognition in News for Algorithmic Trading (FERNAT) framework proposes a pipeline to extract financial evnets from news to be exploited in algorithmic trading. First, news messages (i.e., written text in natural language that originate from RSS sources) are parsed to tokens. These tokens are then used to match patterns that identify (extract) financial events. Then, these events are used in decision making, i.e., trading in financial markets. This section continues with discussing the proposed model in more detail.

3.1 Processing Pipeline

The first part of our framework, i.e., news extraction trough a processing pipeline, is depicted in Fig. 1. The pipeline contains two parts: the lexico-syntactic analysis and the semantic analysis. The cornerstone of the pipeline is a domain ontology for financial events and their related facts. This information defines the expert view on the financial world at a certain moment in time useful for algorithmic trading. The concepts defined in the ontology are anchored to the synsets defined in a semantic lexicon, if such synsets exist. The purpose of the semantic lexicon is twofold: to help define the meaning of the domain concepts and to have access to more lexical representations (lexons) for the ontology concepts.

The lexico-syntactic analysis comprises the following processing units: text tokenizer, sentence splitter, Part-of-Speech (POS) tagger, morphological analyzer, and lexon recognizer. The text tokenizer recognizes the basic text units (tokens) such as words and punctuation. Then, the sentence splitter identifies the sentences present in the news items. After that, the morphological analyzer determines the lemma associated with each word in the text. The POS tagger associates to each word its grammatical type (e.g., noun, verb, pronoun, etc.). The lexon recognizer identifies using gazetteers lexical representations of concepts from both the domain ontology and the semantic lexicon present in news items. The lexons found outside the ontology are useful for defining the contextual meaning of a sentence, a feature exploited in the next processing unit.

The semantic analysis consists of the following processing units: lexon disambiguator, event recognizer, event decorator, ontology instantiator, and ontology updater. The lexon disambiguator uses word sense disambiguation techniques, as for example Structural Semantic Interconnections (SSI) [24], for computing the senses of the found lexons. The lexons which correspond to the financial events stored in the ontology are used for building event instances in the event recognizer. For example, the word "acquire" in the sentence "Google acquires Appjet for Word Processing Collaboration and Teracent to Beef Up Display Ad" is recognized as instance of the ontology referred to by prefix `kb`, i.e., `kb:BuyEvent`.

The event decorator uses *lexico-semantic patterns* to mine facts relevant for event description. An illustration of such a rule is a pattern that mines texts for company acquisitions, i.e.,

```
$sub:=[kb:Company] $prd:=kb:BuyEvent $obj:=([kb:Company])+
```

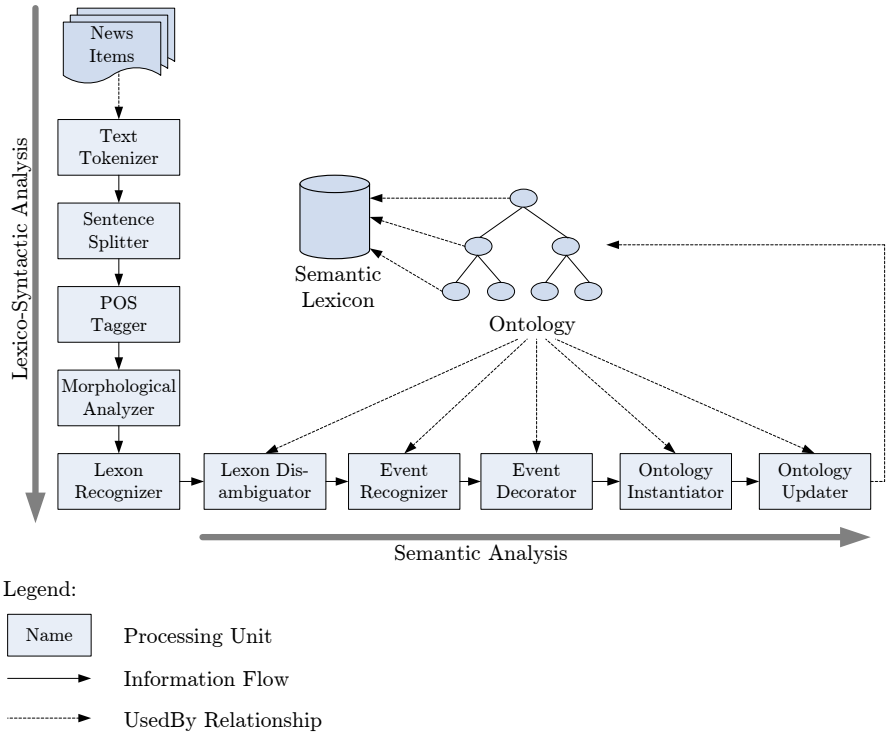

Fig. 1. The FERNAT processing pipeline

where `$sub`, `$prd`, and `$obj` are variables representing a buyer, buy event, and buyee, respectively. Furthermore, `kb:BuyEvent` and `[kb:Company]` are an instance and a class from the ontology, and `+` is the repetition operator. Based on this rule (when applied on our earlier example), `$prd` represents the previously discovered event, `$sub` is assigned to the ontology instance "Google", and `$obj` is assigned an array with the ontology instances "Appjet" and "Teracent", respectively. The lexico-semantic patterns leverage existing lexico-syntactic patterns to a higher abstraction level by using ontology concepts in the pattern construction. Also, in this processing unit, the time associated to an event instance is determined. The discovered events and their related facts need to be manually validated before the next step can proceed in order to prevent erroneous updates to cascade through the ontology, possibly causing incorrect trading decisions when used in algorithmic trading. In the ontology instantiator, the events and their associated information are inserted in the ontology.

In the last processing unit, the ontology updater uses *update rules* for implementing the effects of the discovered events in the domain ontology. An illustration of such a rule is

```
$prd:=kb:BuyEvent($sub:=[kb:Company], $obj:=[kb:Company])
-> DELETE    $sub kb:hasCompetitor $obj
   CONSTRUCT $sub kb:owns $obj
```

where `kb:hasCompetitor` and `kb:owns` are ontology relationships. In our running example, knowing that Google buys Appjet and Terracent would imply that Google and the other two companies are not anymore in the competitor relation and are now parts of the same company. While this example is about learning new relations, ontology update rules can also be used for learning new instances (e.g., a company which is not yet present in the ontology). The purpose of the ontology updates is twofold: extracting more information from subsequent news items, and providing more information in the ontology useful for algorithmic trading.

The innovation of the proposed approach stems from several issues. First, it proposes a methodology for extracting in a semi-automatic manner financial events from news items. The recognized financial events are to be used as additional input next to stock prices for trading algorithms. Second, it investigates the use of ontologies and semantic lexicons for information extraction at multiple methodological levels: domain modeling, gazetteering, word sense disambiguation, information extraction pattern construction, knowledge base update rule building, and result delivery. The envisaged ontology gazetteer is expected to go beyond state-of-the-art ontology gazetteers by allowing the automatic generation of gazetteer's lists from the ontology content. Third, it proposes the use of lexico-semantic patterns, a generalization of lexico-syntactic patterns, which makes easier the pattern development and maintenance. Last, but not least, by using update rules it implements the changes to the financial world implied by the discovered events in the domain ontology.

The implementation of the proposed methodology requires a large number of technologies like text mining tools such as GATE components, Semantic Web languages as RDF and OWL, and semantic lexicons like WordNet [18]. As most of these technologies are supported by Java libraries we develop an implementation based on the Java programming language. As input we use RSS news feeds originating from different online sources, e.g., Reuters, BBC, NYT, etc. Most of the components depicted in Fig. 1 can be implemented by reusing existing implementations. For example, the text tokenizer and sentence splitter can be implemented using ANNIE components [6], while the POS tagger can be implemented by means of the Stanford POS tagger [28]. Subsequently, the MIT Java Wordnet Interface (JWI) [19] can be employed for morphological analysis. Additionally, the lexon recognizer, lexon disambiguator, event recognizer, and ontology instantiator can be reused from our earlier work [12, 15]. Finally, although we have introduced an ontology update implementation [11], we can extend this with the more expressive update language proposed in this paper.

Performance-wise, we aim to be able to rapidly and correctly identify most of the financial events present in news as required by an algorithmic trading setup. For this purpose we aim for state-of-the-art performance, i.e., precision and recall of 70-80% and sub-second performance for news processing time.

3.2 Algorithmic Trading

For the second part of the framework we investigate the use of the extracted financial events for improving the returns of technical trading rules. More precisely, we plan to associate stock price impact factors to financial events that quantify what is the effect of a financial event on a stock price. By aggregating the stock price impacts of the events found in a news item, we can determine a trading signal (e.g., buy, hold, or sell) given by the news item. Then, by combining signals generated by news items with signals obtained through technical trading, we can provide for an aggregated signal that better reflects the current situation than by using technical trading alone.

For this purpose we plan to extend a genetic programming approach which generates high performing technical trading rules. As other approaches are mostly based on ad-hoc specifications of trading rules [2, 5], by using an evolutionary algorithm we avoid the danger of ex post selection, and are able to generate rules that are in a sense optimal. Our choice for genetic programming is also motivated by the easy extension of the genetic programming solution for our current purpose, by adding news-based signals as leaves in the trading rule tree, in addition to the technical trading rules. As a last step, we will show that most of the generated (optimal) technical trading rules do make use of news and thus provide for better returns than the ones which do not make use of the news component.

High frequency trading without accounting for news is undoubtedly faster than the framework proposed in this paper, yet it does not take into account an updated knowledge base with the latest facts, generating trading decisions less informed and accurate. Lags between the publication of news and the reaction of the stock market could be substantial enough to cover for the increase in processing time caused by the usage of rather heavy Semantic Web technologies. Furthermore, one could also separate the computationally intensive event recognition tasks from algorithmic trading. This way, the knowledge base containing market facts is updated once news is processed. Trading algorithms are to be run in separate processes and make use of the (regularly updated) knowledge base, reducing the reaction time on the financial markets. Being able to reason with the financial information stored in the ontology will provide for an increased support for trading decisions.

4 Conclusions

In this position paper we have proposed the FERNAT framework for financial event recognition in news, which encompasses a news processing pipeline of which the outputs are applied to algorithmic trading. The framework builds partially on earlier work for its NLP tasks and makes use of our developed financial ontology and lexico-semantic event extraction pattern language. For ontology updating we have briefly touched upon a proposal for an update language which needs to be implemented in future work. Additional further work is related to the proposed application for algorithmic trading. We envision a genetic programming approach

that generates high-performing technical trading rules through the usage of stock price impact factors associated to financial events discovered by the proposed news processing pipeline.

Acknowledgement

The author is sponsored by the NWO Physical Sciences Free Competition project 612.001.009: Financial Events Recognition in News for Algorithmic Trading (FERNAT) and the Dutch national program COMMIT.

References

1. Allen, F., Karjalainen, R.: Using Genetic Algorithms to Find Technical Trading Rules. Journal of Financial Economics 51(2), 245–271 (1999)
2. Bessembinder, H., Chan, K.: The profitability of Technical Trading Rules in the Asian Stock Markets. Pacific-Basin Finance Journal 3(2–3), 257–284 (1995)
3. Black, W.J., M^{c}Naught, J., Vasilakopoulos, A., Zervanou, K., Theodoulidis, B., Rinaldi, F.: CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations. Technical Report TR–U4.3.1, Department of Computation, UMIST, Manchester (2005), from: `http://www.nactem.ac.uk/files/phatfile/cafetiere-report.pdf`
4. Borsje, J., Hogenboom, F., Frasincar, F.: Semi-Automatic Financial Events Discovery Based on Lexico-Semantic Patterns. International Journal of Web Engineering and Technology 6(2), 115–140 (2010)
5. Brock, W.A., Lakonishok, J., LeBaron, B.: Simple Technical Trading Rules and the Stochastic Properties of Stock Returns. Journal of Finance 47(5), 1731–1764 (1992)
6. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: 40th Anniversary Meeting of the Association for Computational Linguistics (ACL 2002). pp. 168–175. Association for Computational Linguistics (2002)
7. Domingue, J., Motta, E.: PlanetOnto: From News Publishing to Integrated Knowledge Management Support. IEEE Intelligent Systems 15(3), 26–32 (2000)
8. Drury, B., Almeida, J.J.: Identification of Fine Grained Feature Based Event and Sentiment Phrases from Business News Stories. In: Akerkar, R. (ed.) International Conference on Web Intelligence, Mining and Semantics (WIMS 2011). ACM (2011)
9. Fama, E.F.: The Behavior of Stock-Market Prices. Journal of Business 38(1), 34–105 (1965)
10. Fama, E.F.: Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance 25(2), 383–417 (1970)
11. Frasincar, F., Borsje, J., Hogenboom, F.: E-Business Applications for Product Development and Competitive Growth: Emerging Technologies, chap. Personalizing News Services Using Semantic Web Technologies, pp. 261–289. IGI Global (2011)
12. Frasincar, F., Borsje, J., Levering, L.: A Semantic Web-Based Approach for Building Personalized News Services. International Journal of E-Business Research 5(3), 35–53 (2009)
13. Hellstrom, T., Holmstrom, K.: Parameter Tuning in Trading Algorithms using ASTA. In: 6th International Conference Computational Finance (CF 1999). pp. 343–357. MIT Press (1999)

14. Hogenboom, A., Hogenboom, F., Frasincar, F., Kaymak, U., Schouten, K., van der Meer, O.: Semantics-Based Information Extraction for Detecting Economic Events. Multimedia Tools and Applications, Special Issue on Multimedia Data Annotation and Retrieval using Web 2.0 (2012), DOI: 10.1007/s11042-012-1122-0 (to appear)
15. IJntema, W., Sangers, J., Hogenboom, F., Frasincar, F.: A Lexico-Semantic Pattern Language for Learning Ontology Instances from Text. Journal of Web Semantics: Science, Services and Agents on the World Wide Web (2012), DOI: 10.1016/j.websem.2012.01.002 (to appear)
16. Java, A., Finin, T., Nirenburg, S.: Text Understanding Agents and the Semantic Web. In: 39th Hawaii International Conference on Systems Science (HICSS 2006). vol. 3, p. 62b. IEEE Computer Society (2006)
17. Kearns, M.J., Ortiz, L.E.: The Penn-Lehman Automated Trading Project. IEEE Intelligent Systems 18(6), 22–31 (2003)
18. Laboratory, P.C.S.: A Lexical Database for the English Language (WordNet) (2008), from: `http://wordnet.princeton.edu/`
19. Mark Finlayson: JWI – the MIT Java Wordnet Interface (2012), from: `http://projects.csail.mit.edu/jwi/`
20. Maynard, D., Saggion, H., Yankova, M., Bontcheva, K., Peters, W.: Natural Language Technology for Information Integration in Business Intelligence. In: Abramowicz, W. (ed.) 10th International Conference on Business Information Systems (BIZ 2007). Lecture Notes in Computer Science, vol. 4439, pp. 366–380. Springer (2007)
21. Maynard, D., Tablan, V., Cunningham, H., Ursu, C., Saggion, H., Bontcheva, K., Wilks, Y.: Architectural Elements of Language Engineering Robustness. Journal of Natural Language Engineering – Special Issue on Robust Methods in Analysis of Natural Language Data 8(1), 257–274 (2002)
22. Mitchell, M.L., Mulherin, J.H.: The Impact of Public Information on the Stock Market. Journal of Finance 49(3), 923–950 (1994)
23. Mittermayer, M.A., Knolmayer, G.F.: Text Mining Systems for Market Response to News: A Survey. Working Paper 184, Institute of Information Systems, University of Bern (2006), from: `http://www2.ie.iwi.unibe.ch/publikationen/berichte/resource/WP-184.pdf`
24. Navigli, R., Velardi, P.: Structural Semantic Interconnections: a Knowledge-Based Approach to Word Sense Disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence 27(7), 1063–1074 (2005)
25. Nirenburg, S., Raskin, V.: Ontological Semantics. MIT Press (2004)
26. Popov, B., Kiryakov, A., Ognyanoff, D., Manov, D., Kirilov, A.: KIM - A Semantic Platform For Information Extraction and Retrieval. Journal of Natural Language Engineering 10(3–4), 375–392 (2004)
27. Reuters: Reuters NewsScope Archive (2012), from: `http://www2.reuters.com/productinfo/newsscopearchive/`
28. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network. In: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2003). pp. 252–259. Association for Computational Linguistics (2003)