Semantics-Based Financial Event Detection

Frederik Hogenboom
fhogenboom@ese.eur.nl

Alexander Hogenboom
hogenboom@ese.eur.nl

Flavius Frasincar
frasincar@ese.eur.nl

Econometric Institute
Erasmus University Rotterdam
PO Box 1738, NL-3000 DR
Rotterdam, the Netherlands

ABSTRACT

Breaking news on economic events – e.g., acquisitions, profit announcements, and stock splits – has a substantial impact on financial markets. Hence, automatically identifying events in news items accurately and timely is of great importance. In this paper we propose a Semantics-based Pipeline for Economic Event Detection (SPEED), which aims at extracting financial events from news articles and annotating these events with meta-data, while retaining a speed that is high enough to make real-time use possible. In our pipeline implementation, we have reused some of the components of an existing framework and additionally, developed new ones, e.g., an Ontology Gazetteer and a Word Sense Disambiguator.

1. INTRODUCTION

Machines performing Natural Language Processing (NLP) tasks can be of great importance in today's society. Decision makers process a continuous flow of breaking news through the extraction of information. This information is subsequently used for reasoning, leading to the acquisition of knowledge on, for example, the state of markets. Most markets are extremely sensitive to breaking news on economic events like acquisitions, stock splits, dividend announcements, etc. Automating these information extraction and knowledge acquisition processes can support decision makers, as they enable faster processing of more data. An extended version of the paper that contains more details on the framework was presented at DEXA 2011 [3].

Therefore, we propose a Semantics-based Pipeline for Economic Event Detection (SPEED), i.e., a fully automated framework for processing financial news messages gathered from Really Simple Syndication (RSS) feeds. SPEED aims to extract financial events, represented in a machine-understandable way through the use of Semantic Web technologies. In our implementation, we identify in news the concepts related to economic events, which are defined in a domain ontology. These concepts are also associated to synsets from a semantic lexicon, e.g., WordNet. For concept identification, we employ lexico-semantic patterns based on ontology concepts for lexical concept matching. The identified lexical representations of relevant concepts are subject to a Word Sense Disambiguation (WSD) procedure for determining the correct senses.

DIR 2012 February 23-24, 2012, Ghent, Belgium.

2. RELATED WORK

In the field of IE, the General Architecture for Text Engineering (GATE) is a freely available general purpose framework based on which one can construct processing pipelines from components that perform specific tasks. Components are used for linguistic analysis (e.g., tokenization), syntactic analysis jobs (e.g., Part-Of-Speech (POS) tagging), and semantic analysis tasks (e.g., understanding concepts). By default, GATE loads the A Nearly-New Information Extraction (ANNIE) system, consisting of several components, i.e., *English Tokenizer*, *Sentence Splitter*, *Part-Of-Speech (POS) Tagger*, *Gazetteer*, *Named Entity (NE) Transducer*, and *OrthoMatcher*.

Examples of tools utilizing ANNIE components are the personalized news service Hermes [2], the Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations (CAFETIERE) relation extraction pipeline [1], and the Knowledge and Information Management (KIM) platform [5] for semantic annotation.

Even though ANNIE has proven to be useful in various IE applications, it lacks important features such as a WSD component, although some disambiguation can be done using JAPE rules in the *NE Transducer*. This is however a cumbersome and ineffective approach that is prone to errors. Furthermore, ANNIE lacks the ability to individually look up concepts from a large ontology within a limited amount of time. Nevertheless, GATE is easily customizable, and therefore ANNIE's components are either usable, extendible, or replaceable in order to suit our needs.

3. SPEED FRAMEWORK

While most current approaches to automated IE from news focus on annotation, our framework extracts information (events) and updates a knowledge base. SPEED consists of several components which sequentially process documents. This approach is driven by an expert-created domain ontology with information on the NASDAQ-100 companies, extracted from Yahoo! Finance. The ontology captures concepts and events from the financial domain, such as companies and competitors. Many concepts stem from a semantic lexicon like WordNet or represent named entities.

As a first processing step, the SPEED pipeline identifies individual components of the text by means of the *English Tokenizer*, which splits text into tokens (e.g., words or numbers) while taking into account rules specific to the English language. These tokens are then searched in the ontology concept lexical representations by an *Ontology Gazetteer*. Hence, tokens in the text are annotated with a reference to

their associated ontological concepts. Then, the *Sentence Splitter* groups the tokens in the text into sentences, mostly based on punctuation. These sentences are used for discovering the grammatical structure in text by determining the part-of-speech of each word token by means of the *Part-Of-Speech Tagger*. As words can have many forms that have a similar meaning, the *Morphological Analyzer* subsequently reduces the tagged words to their lemma and an affix.

Words and meanings, denoted often as synsets (set of synonyms) have a many-to-many relationship. Hence, the next step in interpreting a text is disambiguation of its words' meaning. For this, a *Word Group Look-Up* component first combines words into meaningful word groups that are maximal with respect to semantic lexicon entities. The *Word Sense Disambiguator* then determines the word sense of each word group by exploring the mutual relations between senses (as defined in the semantic lexicon and the ontology) of word groups. We propose an adaptation of the Structural Semantic Interconnections (SSI) [4] algorithm for WSD. The SSI approach uses graphs to describe word groups and their context (word senses), as derived from a semantic lexicon. The senses are determined based on the number and type of detected semantic interconnections in a labeled directed graph representation of all senses of the considered word groups. As opposed to the original algorithm, we consider the two most likely senses for each word group, and we default to the statistically most likely sense in our semantic lexicon based on the difference between the similarity of these two senses with the sentence context.

The last processing steps comprise text interpretation by introducing semantics, linking word groups to an ontology. The *Event Phrase Gazetteer* scans the text for (financial) events by using a list of phrases or concepts that are likely to represent some part of a relevant event. Additional information is added to identified events by the *Event Pattern Recognition* component, which use domain-specific lexico-semantic patterns appointed to events. Last, the knowledge base is updated by the *Ontology Instantiator*.

We have implemented the SPEED framework as a Java-based application. SPEED uses some default GATE components, i.e., the *English Tokenizer*, *Sentence Splitter*, *Part-Of-Speech Tagger*, and the *Morphological Analyzer*. Furthermore, we have extended the functionality of other GATE components for ontology gazetteering, and we also implemented new components for WSD and event detection.

The *Ontology Gazetteer* and *Word Group Look-Up* components match ontology concepts and WordNet word groups, respectively, with lexical representations stored in a look-up tree. Individual tokens are represented by nodes and a concept's lexical representation is viewed as the path from the root node to an arbitrary leaf node. Word groups associated with a path are annotated with their associated concepts. In order to reduce the tree scanning time, ontology concepts and word groups have been organized using hash maps.

4. SPEED EVALUATION

For evaluating SPEED's performance, we assess the precision and recall, as well as latency. In case of performance comparisons, we determine the statistical relevance of differences in performance by means of a paired t-test.

We use 200 manually annotated news messages extracted from the Yahoo! Business and Technology news feeds. We distinguish between ten different financial events and observe a precision for the concept identification in news items of 86% and a recall of 81%. Precision and recall of fully decorated events result in lower values of 62% and 53%. Despite using only WordNet as a semantic lexicon, we obtain high precision as many of our concepts' lexical representations are mostly monosemous named entities. The high recall scores are caused by our focus on detecting ontology concepts in the text, rather than on identifying all appearing concepts.

Our *Word Sense Disambiguator* with an adapted SSI algorithm shows a precision and recall of 59%, compared to a precision and recall of 53% and 31%, respectively, for the original algorithm, implying an overall improvement of 12% and 90% at a significance level of 0.001.

As for processing speeds, SPEED measures a latency of 632 milliseconds per document on a Intel Core i7 920 PC with 6 GB RAM. Roughly 30% is allocated to performing linguistic and syntactic analysis tasks. The subsequent WSD task on average takes up about 60% of the execution time, whereas the remaining tasks are typically performed in about 10% of the execution time.

5. CONCLUSIONS

We have proposed a semantics-based framework for economic event detection (SPEED), which extracts financial events from news articles. For our implementation, we have reused existing components and developed new ones such as gazetteers and a word sense disambiguator. Also, we make use of semantic lexicons and ontologies. The framework shows high precision and recall scores in our evaluation on news feeds. We have used specific data structures based on hash maps to reduce the latency of the event seeking process.

6. REFERENCES

[1] W. J. Black, J. M^{c}Naught, A. Vasilakopoulos, K. Zervanou, B. Theodoulidis, and F. Rinaldi. CAFETIERE: Conceptual Annotations for Facts, Events, Terms, Individual Entities, and RElations. Technical Report TR–U4.3.1, Department of Computation, UMIST, Manchester, 2005.

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

[3] A. Hogenboom, F. Hogenboom, F. Frasincar, U. Kaymak, O. van der Meer, and K. Schouten. Detecting Economic Events Using a Semantics-Based Pipeline. In A. Hameurlain, S. W. Liddle, K.-D. Schewe, and X. Zhou, editors, *Twenty-Second International Conference on Database and Expert Systems Applications (DEXA 2011)*, volume 6860 of *Lecture Notes in Computer Science*, pages 440–447. Springer, 2011.

[4] R. Navigli and P. Velardi. Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(7):1075–1086, July 2005.

[5] B. Popov, A. Kiryakov, D. Ognyanoff, D. Manov, and A. Kirilov. KIM - A Semantic Platform For Information Extraction and Retrieval. *Journal of Natural Language Engineering*, 10(3–4):375–392, September 2004.