Experiments in Candidate Phrase Selection for Financial Named Entity Extraction - A Demo

Hassan Alam
BCL Technologies
San Jose, CA 95128
Hassana@bcltechnologies.com

Aman Kumar
BCL Technologies
San Jose, CA 95128
amank@bcltechnologies.com

Tina Werner
BCL Technologies
San Jose, CA 95128
twerner@bcltechnologies.com

Manan Vyas
BCL Technologies
San Jose, CA 95128
mvyas@bcltechnologies.com

Abstract

In this study we develop a system that tags and extracts financial concepts called financial named entities (FNE) along with corresponding numeric values – monetary and temporal. We employ machine learning and natural language processing methods to identify financial concepts and dates, and link them to numerical entities.

1 Introduction

We developed a baseline system called Automatic Extraction of Financial Data from Text (AEFDT) that tags and extracts financial concepts based on the natural language text from a financial document such as 10Q, 10K and analyst's reports.

Such financial entities (FNEs – Financial Named Entities, numerical entities, and semantic tags) are useful to multiple audiences. On one hand these extracted financial concepts are useful to analysts and internal users who can benefit from a simplified overview of the financial health of a company and in writing financial reports and making budgetary decisions; on the other, it will help consumers who are interested in reviews about a company for a product, job-related news and other financial aspects. In addition, such a system will help public companies meet the SEC.gov filing requirements in an automated fashion that is less prone to errors.

A snap-shot of the working architecture of the AEFDT system is given Figure 1.

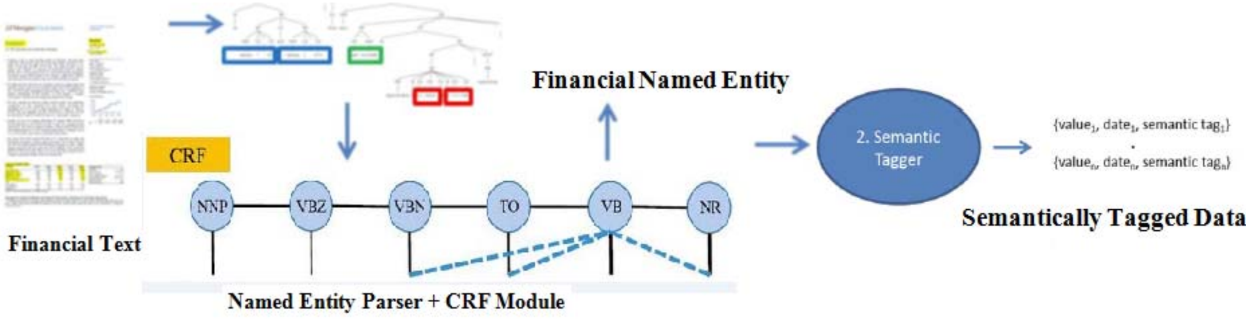
Financial Text
CRF
NNP
VBZ
VBN
TO
VB
NR
Named Entity Parser + CRF Module
Financial Named Entity
2. Semantic Tagger
{value$_1$, date$_1$, semantic tag$_1$}
{value$_n$, date$_n$, semantic tag$_n$}
Semantically Tagged Data

Figure 1. Flow chart of the AEFDT system

Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, pages 45–48, Osaka, Japan, December 11-17 2016.

2 Methodology

We have modelled our system as a Named Entity Recognition (NER) [Nadeau et al., 2007] problem. NER is an important subtask within information retrieval that locates and tags entities in text into categories such as the name, organization, location, time, quantities, etc.

Our current FNE identification system is a customized version of Stanford NER [Finkel, 2005]. Stanford NER is a Java implementation of a Named Entity Recognizer that provides a framework for training and inference of *Conditional Random Field* (CRF) model. In addition, it has the capability to incorporate other existing NLP tools like syntactic parser, Semantic Role Labelling (SRL), part-of-speech tagger, etc. We utilized our domain knowledge based on financial reports (10-Q, 10-K, etc.) to extend Stanford NER for a high precision system.

2.1 Data Collection and Annotation

We have collected around 6 MB of data in the form of 10Q/10K and non-SEC documents. We wrote programs in C++ and Python to crawl the sec.gov page. Once the sentences (from the Notes section of a 10Q document), labels, semantic tags, and numeric values are mined the training corpus is created by annotating the corpus with *labels* for training toward FNEs and creating attributes for semantic tags and numbers. Here’s a snapshot of the label annotations of a part of a 10Q document (Notes section).

> **<FNE>** *Amortization expense* **</FNE>** *for the three* months ended **<Date1>***January 31, 2013* **</Date1>** *and***<Date2>** *2012* **</Date2>** *was* **<NV1>***$5.0 million* **<NV1>** *and* **<NV2>** *$5.2 million***</NV2>**, *respectively.*

Our training dataset has 8000 and 2000 annotated sentences for training and testing, respectively. These datasets were obtained by randomly shuffling entire corpus and partitioning them in 80:20 ratios.

The three methods that we implemented are: (1) *CRF+No Dictionary*; (2) *CRF+Dictionary*; (3) *CRF+Dictionary+Features*. *CRF+No Dictionary* refers to the method where we just have the conditional random field model that tags the tokens (words) in a sentence with FNE or not (o). *CRF+Dictionary* is applied when we have dictionary built and we do basic statistical analysis of the dictionary items, and *CRF+Dictionary+Features* method uses post-processing linguistic rules for tagging.

3 Evaluation and Results

Three trained subject matter experts manually evaluated the results for accuracy. We tested the models on unseen 2000 sentences from a 10Q file crawled from the sec.gov webpage. The results of the preliminary system are given in Table 1.

	CRF+No Dictionary	**CRF+Dictionary**	**CRF+Dictionary+Features**
FNE % accuracy	61.33	77.82	88

Table 1. Preliminary results of the FNE identification system with three models using CRF parser

3.1 Discussion

The CRF+Dictionary+Features model gives the best results. That tells is that if we refine our heuristics and feature selection, we are likely to getter better results in future revisions. For the current set of heuristics, we have used the following feature rules for clustering and classification.

Surface Feature Selection for FNE identification system:

- Current word
- Next word
- Previous word
- Current POS (part-of-speech) Tag
- Previous POS Tag
- Next POS Tag
- Base POS tags
- Custer of “related” words
- Relative positions with numerical entity etc.
- Trigger word

In addition, we extracted features from parse tree like extracted NP sub-tree etc. As semantic features, we used semantic tree annotations extracted from Stanford parsed tree (De Marneffe et al. (2006)). We extracted a dictionary of FNEs from filed 10Qs at sec.gov and employed that as trigger words.

4 Conclusions

In this study we performed a feasibility study to tag and extract financial concepts called *financial named entities* (FNE). We employed machine learning and natural language processing methods to identify financial concepts and link them to numerical entities. The best model records an accuracy of 88% in 10Q/K files from the sec.gov webpages.

References

1. Ananiadou, S. and McNaught, J.(eds) Text Mining for Biology and Biomedicine, 2006, Artech House.
2. Bishop, Christopher M. Pattern recognition and machine learning. springer New York, 2006.

3. Cohn, Trevor. "Efficient inference in large conditional random fields." Springer, 2006.

4. De Marneffe, Marie-Catherine, MacCartney, Bill, Manning, Christopher D and others. "Generating typed dependency parses from phrase structure parses." Proceedings of LREC. 2006. 449-454.

5. Finkel, Jenny, Dingare, Shipra, Nguyen, Huy, Nissim, Malvina, Manning, Christopher and Sinclair, Gail. "Exploiting context for biomedical entity recognition: From syntax to the web." Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications. Association for Computational Linguistics, 2004. 88-91.

6. Finkel, Jenny Rose, Grenager, Trond and Manning, Christopher. "Incorporating non-local information into information extraction systems by gibbs sampling." Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2005. 363-370.

7. Lafferty, John, McCallum, Andrew and Pereira and Fernando CN. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." International Conference of Machine Learning. 2001.

8. McCallum, Andrew and Li, Wei. "Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons." Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4. 2003.

9. Minkov, Einat, Wang, Richard C, Tomasic, Anthony and Cohen, William W. "NER systems that suit user's preferences: adjusting the recall-precision trade-off for entity Extraction." Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers. Association for Computational Linguistics, 2006. 93-96.

10. Nadeau, David and Sekine, Satoshi. "A survey of named entity recognition and classification." Lingvisticae Investigationes, 2007: 3-26.

11. Ratinov, Lev and Roth, Dan. "Design challenges and misconceptions in named entity recognition." Proceedings of the Thirteenth Conference on Computational Natural Language Learning. Association for Computational Linguistics, 2009. 147-155.

12. Silla Jr, Carlos N and Freitas, Alex A. "A survey of hierarchical classification across different application domains." Data Mining and Knowledge Discovery, 2011: 31-72.

13. Vapnik, Vladimir and Cortes, Corinna. "Support vector machine." Machine learning, 1995: 273-297.