Business Event Curation: Merging Human and Automated Approaches

Yiqi Wang[1], Huiying Ma[1], Nichola Lowe[1], Maryann Feldman[1], Charles Schmitt[1]

1 University of North Carolina at Chapel Hill (UNC), Chapel Hill, NC 27599, USA.
{wangyiqi, huiying, nlowe}@email.unc.edu; maryann.feldman@unc.edu; cschmitt@renci.org

Abstract

We present preliminary work to construct a knowledge curation system to advance research in the study of regional economics. The proposed system exploits natural language processing (NLP) techniques to automatically implement business event extraction, provides a user-facing interface to assist human curators, and a feedback loop to improve the performance of the Information Extraction Model for the automated parts of the system. Progress to date has shown that we can improve standard NLP approaches for entity and relationship extraction through heuristic means and provide indexing of extracted relationships to aid curation.

Introduction (Background)

The economic success or stagnation of a geographic region, such as the Silicon Valley, depends on a complex interplay of factors. Understanding this complexity requires collecting data on the spatio-temporal transactions that occur among a changing set of firms and people. This is a non-trivial challenge given that key data sources, i.e., business news journals, are unstructured. Human curation can ensure high quality information extraction (IE), but scales poorly. Integration with automated IE approaches are thus attractive, but must address both general IE challenges and domain-specific challenges, e.g., use of domain and regional specific terminology and the need to integrate events over news articles. We design and implement an automated IE system to help human curators extract business event information, which will be further used for local economic analysis, from large amount of text data.

Approach

Figure 1.1 illustrates the general architecture of our system. A NLP pipeline processes news articles to product articles annotated with possible business entities and events. A web interface allows searching annotated articles based on entities and events and (future) assists with curation and iterative training of IE models.

Subtasks for business event detection

Two existing approaches for conducting business event extraction are Named Entity Recognition (NER) and Co-reference Resolution (CR). NER aims at identifying phrases containing names of entities like organization and person while CR solves the complexity of implicit context owing to anaphora, cataphora and co-referring noun phrases in text documents. Although several NLP systems provide great implementation of these tasks, their performances were not ideal due to difference between their training and our testing data. A focus of our work has been to improve the performance of the NER and Co-reference Resolution System provided by Stanford NLP groups on our data.

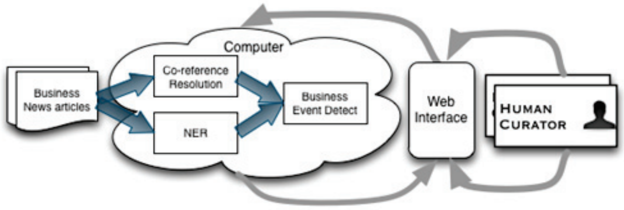

Figure 1.1 General Architecture of our business curation system

Named Entity Recognition

We developed an efficient approach to apply the Stanford NER system, particularly for the ORGANIZATION, PERSON, LOCATION classes. Our approach firstly preprocessed online news articles to extract html metadata serving as external gazetteer. Along with pre-defined gazetteer including common position titles and keywords indicating company and the external gazetteer, we exploited handcraft linguistic patterns and heuristics to correct and to improve the Stanford NER result.

Co-reference Resolution

While the Stanford CR system demonstrated good performance on our data when referent is 'PERSON', it failed

The research demonstrated in this paper is supported by NSF SMA-1439532 and UNC.

most of cases referring to ‘ORGANIZATION’ entity. We proposed a naïve approach to solve co-reference issue with ‘ORGANIZATION’ entity in particular. Similar to but simpler than the approach implemented in Hobbs (1978), the proposed approach adopted syntax features, namely Part of Speech (POS) and parse tree to solve the problem. In brief, we mainly deployed steps as follow: For definite articles such as ‘the company’, we search entity in the dominated ‘NP’ (noun phrase) of the previous sentence. As for pronouns, we first search entity in the directly dominated ‘NP’ of the pronoun in the same sentence and then follow the method we use for definite articles.

Business event extraction

We have adopted rule-based methods using both lexical and syntax features to extract information about business events, principally the concepts of IPO (initial public offering), layoff and investment. A key challenge is that these concepts are described with a variety of syntax (e.g., "going public"). The general architecture consists two steps: identifying phrases serving as event trigger and extracting potential event participants. Depending on event type, we applied either lexicon dominated patterns or syntax dominated patterns. For instance, some event information such as "investment project", did not belong to any named entity, which is hard to write surface pattern. However, syntax feature worked well in this situation. Additionally, syntax feature can partially eliminate noises brought by clause or other complex sentence structure. The form of our event patterns is similar to the patterns constructed in Liu (2009).

Curation Web Application Design

As a first step towards a curation tool for researchers, we have designed a web interface for curators to search and browse the business information processed by the Information Extraction Model.

Data Source

The web application takes the documents automatically generated by using the IE Model as CSS marked-up HTML files. All the classes detected in each documents are marked in different colors using different CSS “class”es and “id”s.

Document Indexing Method

We use Apache Solr as the full-text indexing tool for this system considering the file amounts, the retrieval speed and the powerful indexing and searching functions Apache Solr can provide, such as field matching, faceting and query re-ranking. We mapped the contents of the HTML files into different fields according to their classes detected by the IE model using a python script and generated corresponding standard indexing XML files. The XML files can easily uploaded to Solr by sending HTTP requests to the Solr server from any environment where such requests can be generated and we can also use HTTP request to retrieve information indexed on Solr.

Web Interface Design

To support category-based search, we decided to design the web interface with a main general search bar with an advanced search panel listed with filters such as location, organization and other classes detected by the IE model. There would also be a result page with all the retrieved documents and the link to open the original HTML file for curators to mark and annotate in the future.

Preliminary Result and Future Work

We used 25 news articles, 9689 words in total, as preliminary test data for our improved NER approach. As Table1.1 suggests, our approach we significantly improved the Stanford NER system performance on our data set.

Named Entity	Improved NER			Stanford NER		
	P	R	F1	P	R	F1
LOCATION	0.99	0.93	0.96	0.91	0.83	0.87
ORGANIZATION	0.94	0.88	0.91	0.97	0.50	0.66
PERSON	0.70	0.98	0.82	0.52	0.95	0.67

Table 1.1 Comparison of Stanford NER vs. Our approach

For other event extraction tasks, we conducted initial qualitative assessment since quantitative evaluation requires large labeled data sets that we have yet to generate. We plan to both increase the size of our training sets for these tasks as well as investigate the use of distant supervision for further evaluation. Finally, we will be evaluating the impact of our system on the performance of human curators as our principal assessment of utility. In future work, we plan to explore the learning-based approach for event detection without requiring large-scale labeled data, drawing upon feedback gathered from human curators as they extract entities and relationships. We also plan to investigate inference models to improve NLP.

References

Hobbs, J. R. 1976a. Pronoun Resolution. Research Report 76-1, Department of Computer Sciences, City College, City University of New York.

Liu, Ting. 2009. Bootstrapping events and relations from text. Ph.D. diss., Department of Computer Science, State University of New York at Albany, Albany, NY.

Agichtein, E., & Gravano, L. 2000, June. Snowball: Extracting relations from large plain-text collections. *In Proceedings of the fifth ACM conference on Digital libraries* (pp. 85-94). ACM.