Yiqi Wang. Pattern Based Information Extraction System in Business News Articles. A Master's paper for the M.S. in I.S. degree. April, 2016. 45 pages. Advisor: Jaime Arguello

Business news journals provide a rich resource of business events, which enable domain experts to further understand the spatio-temporal changes occur among a set of firms and people. However, extracting structured data from journal resource that is text-based and unstructured is a non-trivial challenge. This project designs and implements a Business Information Extraction System, which combines advanced natural language processing (NLP) tools and knowledge-based extraction patterns to process and extract information of target business event from news journals automatically. The performance evaluation on the proposed system suggests that IE techniques works well on business event extraction and it is promising to apply the technique to extract more types of business events.

Headings:

Natural Language Processing
Event Extraction -- Information Extraction
Named Entity Recognition -- Information Extraction

PATTERN BASED INFORMATION EXTRACTION SYSTEM IN BUSINESS NEWS ARTICLES

by
Yiqi Wang

A Master's paper submitted to the faculty
of the School of Information and Library Science
of the University of North Carolina at Chapel Hill
in partial fulfillment of the requirements
for the degree of Master of Science in
Information Science.

Chapel Hill, North Carolina

April, 2010

Approved by:

_____

Jaime Arguello

1. Introduction

1.1 Overview

In Gartner's analytics report (2013), the phrase "dark data" refers to the masses of unstructured information (around 80% of the total by volume) that organizations retain and store but have no meaningful way of analyzing or using for other purposes. Because of several modifiers in sentences and ever-changing ways of expression, noises and redundancy regularly occur within text information. Enterprises and analytics require great effort and time to process text information in order to find essential entities and relations between entities.

The proposed project is a particular application of information extraction, whose goal is to design and to implement a text mining system to extract structured information of business events from business news. Structured data extracted from news articles keeps track of specific types of events in the business domain and they can be utilized in further analysis such as event prediction or correlation analysis. The following example simply demonstrates the function of the analytics system. Given news:

> *Morrisville-based* ***MaxPoint Interactive*** *has priced its* ***initial public offering*** *at* ***$11.50 per share****. The company is expected to* ***debut Friday on the New York Stock Exchange****, under the* ***ticker MXPT****. — Mar 6, 2015, 8:07am EST*

The system is expected to extract structured event information as follows:

Company	Event Type
Maxpoint Interactive	IPO

Table 1.1 Company_Event

IPO ticker	IPO filed price	IPO Date
MXPT	$ 11.50	Mar/6/2015

Table 1.2 Event_Information

We will discuss several aspects of the proposed project in the next sections.

1.2 Information Extraction

Information Extraction (IE) is a technique of automatically extracting structured information from unstructured documents (Liu, 2009). As part of natural language processing, IE requires inter-disciplinary knowledge in computer science, artificial intelligence and linguistics. Liu (2009) summarized three levels of information extraction: named entity, (binary) relation and event (n-ary relation). In early stages, most of the extraction tasks were concentrated around the first two levels — identification of named entities, like people and company names and relationships among them from natural language text. (Sunita, 2008).

As mentioned in the previous section, the proposed system aims at extracting event information, the third level of IE. However, although the system focuses on the third level, the performance of the proposed systems heavily relies on the efficiency and accuracy of the entity level extraction. A good result of named entity recognition serves as ground foundation for the system to find underlying relations between different named entities. We will review relevant research in more details in Chapter 2.

1.3 Co-reference Resolution

Information extraction tasks can be further differentiated based on the length of documents. If the document consists of just several snippets or text of one-sentence length it is simpler than articles for event extraction, because the occurrence of pronouns

or definite articles, which makes text ambiguous and complex for program to process, increases with the length the document. A comparison demonstrating the complexity gap between documents with various lengths is below.

One-sentence length document:

Maxpoint Interactive plans to go public.

Document with two sentences:

Maxpoint Interactive is a technical company. ***It*** *plans to go public.*

or

IBM was founded in 1911. ***The company*** *went public in 1962.*

In the latter case, we should first let the system know which entity the pronoun 'it' (or the definite article 'the company') refers to so that the system can determine which company is going to IPO. Unfortunately, the documents our system expects to process are news articles. In news articles, pronouns and definite articles such as 'the company' are commonly used, which means the system should solve co-reference issues with the document.

Co-reference resolution is the process of determining whether two expressions in natural language refer to the same entity in the world (Soon, 2001). Most of the algorithms use syntax features such as part of speech, parse tree, and dependency to solve the problem. For instance, Hobbs (1978) proposed an approach using parse tree to search the directly dominated noun phrase for the pronoun.

One thing worth mentioning is that since co-reference resolution is quite complex and in this project, the scope of the project only focused on co-reference issues of noun phrases and particularly for two kind entities, 'ORGANIZATION' and 'PERSON'.

1.4 Event Extraction

Event extraction is a challenging task in information extraction, which aims at identifying instances of specific types of events in text and extracting event related arguments from text. The applications of event extraction varies from biological abnormal event detection in medical domain to stock analysis in finial industry. We would like to walk through several recent related work in literature review section.

2. Literature Review

This chapter reviews and summarizes prior research in the field of Information Extraction, particularly in event extraction and its preliminary subtasks. The organization of the chapter is as follows: We firstly define the scope of the proposed information extraction task in section 2.1; section 2.2 and section 2.3 introduce the definitions of Named Entity Recognition and Co-reference Resolution and make a summary of common approaches which implement these tasks. Section 2.4 discusses the scope of event extraction and makes a comparison between some of the popular methods of event extraction. Section 2.5 briefly reviews several popular information extraction systems and at the end of this chapter, we will describe the proposed system in details, including several basic functions of the system.

2.1 Information Extraction

Sarawagi (2008) defined Information Extraction to be the automatic extraction of structured information such as entities, relationships between entities, and attributes describing entities from unstructured sources.

Approaches to solve information extraction tasks and related subtasks can be generally divided into two categories: rule-based approach and learning-based approach. Rule-based approaches usually need experts in a particular domain, who browse text documents to find out any commonly used patterns for text and manually write rules to extract information. While the handcraft approach achieves high precision, the procedure

is time consuming and tedious and the recall of results is low. Additionally, sometimes it's hard to find domain experts. As for learning-based approaches, statistics and machine learning models such as generative models based on Hidden Markov Models and conditional models based on maximum entropy are commonly deployed.

2.2 Named Entity Recognition

Named Entity Recognition (NER) refers to the task which seeks to locate and to classify atomic elements in text into predefined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc. (Krishnan and Ganapathy, 2005). In the introduction to the CoNLL-2003 Shared Task, Sang and Meulder (2003) gave an example to illustrate NER analysis. Given the sentence:

U.N official Ekeus heads for Baghdad.

The NER system is supposed to recognize entities such as 'PERSON', 'LOCATION' , 'ORGANIZATION'. The result is expected to be:

[ORG U.N.] official [PER Ekeus] heads for [LOC Baghdad].

The 6th Message Understanding Conference (MUC6, 1995) first defined the scope of named entities, which includes names of person, organization, location and temporal expression, currency as well as percentage.

According to the work of Bast (2007), there are two kind of approaches to implementing NER: List Lookup Approach and Shallow Parsing Approach. The former is simply based on pre-defined gazetteers to recognize the text which is exactly matched to the word in the gazetteers. The approach is quite fast but it heavily relies on the integrity of the gazetteers and it cannot resolve ambiguity. The Shallow Parsing Approach made use of

the co-occurrence between certain entity types and certain word classes or syntactic constructs and it is more flexible since it utilizes several characteristics of entity names as evidence. For instance, capital word followed by 'City' is a case of location name. However, as the development of the NER technology progresses, most of the popular NER systems deployed both approaches and they also enable users to add and customize their own gazetteers to improve the performance of the system. Therefore, we gradually do not compare approaches based on this criteria. Instead, Lau & Zhang(2011) generally divided them into two common categories, namely rule-based approach and learning-based approach. For rule-based approach, manually pre-defined heuristic or linguistic rules are applied to identify specific types of entities such as people, organizations, places, etc. Based on this criteria, we can divide the NER system into handcrafted system, automatic system or hybrid one.

Several automatic NER systems such as OpenNLP, Stanford NLP and LingPipe produce good results and they are widely used in text mining for various purposes. The proposed project chooses to deploy the Stanford NLP system to implement the elementary natural language processing. We would like to introduce the Stanford NER system in detail. We prefer the Stanford NER for two reasons. First, it is open source and it supports several programming languages even though it is implemented in JAVA. Second, compared to other systems, the Stanford NER system focuses on more features than models, which enables us to customize.

Stanford NER system is based on the Conditional Random Field Model, which does not assume that features are independent. The default model is trained on CoNLL, MUC and ACE datasets. Additionally, the Stanford NER system allows users to train customized

models using their training data, which is supposed to include text and the corresponding named entity tag for each word. According to the test data of CoNLL, the performance of Stanford NER system was great, which got above 90% F1 measure score.

Another potentially helpful system is *General Architecture for Text Engineering (GATE)*. GATE is a hybrid system which provides interface for users to write customized patterns to improve the system performance. Several event extraction researches built their system based on GATE (Bontcheva et al, 2002; Saggion et al., 2007).

To evaluate the NER system performance, *Precision*, *Recall*, and *F-score* are the most popular measurements to use. The definition of the *Precision* and *Recall* explained in Manning (2012) for entity category C_i is shown as *(2.2.1)* and *(2.2.2)*:

$$\text{Precision} = \frac{Number\ of\ entities\ correctly\ tagged\ with\ C_i}{Total\ number\ of\ entities\ tagged\ with\ C_i} \qquad (2.2.1)$$

$$\text{Recall} = \frac{Number\ of\ entities\ correctly\ tagged\ with\ C_i}{Total\ number\ of\ entities\ belonged\ to\ C_i} \qquad (2.2.2)$$

One thing worth noting is that for NER standard evaluation is based on per entity, instead of per token. An example demonstrating the difference is as follows:

Take the sentence "John G. Stumpf is the CEO of Wells Fargo & Co.." as example, suppose the correct NER and predicted NER for each token in the sentence is shown in Table 2.1 ('COM' and 'PER' refers to 'COMPANY' and 'PERSON', 'O' means none of entity categories matched)

Sentence:	John	G.	Stumpf	is	the	CEO	of	Wells	Fargo	&	Co.
Correct NER	PER	PER	PER	O	O	O	O	COM	COM	COM	COM
Predicted NER	PER	PER	O	O	O	COM	COM	COM	COM	COM	O

Table 2.2 Example of NER Performance Evaluation

If we use token or word as the counting unit for precision and recall, the precision and recall will be:

Person:	Precision = 2/2 = 100%	Recall = 2/3= 67.7%
Company:	Precision = 3/5 = 60%	Recall = 3/4= 75%

However, if we use entity as the counting unit, the precision and recall will be:

Person:	Precision = 0/1 = 0%	Recall = 0/1= 0%
Company:	Precision = 0/1 = 0%	Recall = 0/1= 0%

2.3 Co-reference Resolution

2.3.1 Introduction

The Concise Oxford Dictionary of Linguistics defines Co-reference to be the relation between noun phrases etc., which have the same reference. For instance, in the sentence '*David told me he wasn't at home last week.* ', '*he'* and '*David'* have the same referent - *David*. Before we discuss co-reference resolution approaches, we briefly introduce some of the terminology.

Anaphora: The use of an expression the interpretation of which depends on an antecedent expression. e.g. ***David*** *told me* ***he*** *wasn't at home last week*.

Cataphora: The use of an expression or word that co-refers with a later, more specific, expression in the discourse. e.g. *After* ***she*** *hung up the phone,* ***Mary*** *started crying*.

Co-referring noun phrases: Two or more noun phrases referring to the same reference. These phrases usually start with definite article - '*the'*. e.g. ***IBM*** *announced a re-balance of the workforce*. ***The company*** *confirms layoff*.

In this section, we generally focus on approaches to solving these three co-reference issues.

Hobbs (1978) proposed a heuristics-based approach, which deep-first traversed surface parse tree to find the first candidate to resolve pronoun references (Zheng et al., 2011). This naive algorithm laid a solid foundation for later heuristics-based pronoun resolution. Matthews (1988) conducted experiments which confirmed the efficiency of the approach by Hobbs (1978). However, this approach mainly focused on searching the antecedent noun phrases of pronouns. The performance of those heuristics-based approaches are generally 70% ~ 80% accurate.

Different from Hobbs (1978), Lappin and Leass (1994) implemented a method which performed two different operations, one for maintaining and updating salience entities in the discourse model, another for resolving each pronoun to those entities. By giving the reference candidates different values based on grammatical function, the pronoun was assigned to the candidate with highest cumulated value. Later, Ge et al. (1998) adopted a Bayesian model to calculate the probability that entity $\boldsymbol{a}$ is the antecedent of a pronoun $\boldsymbol{p}$ given a set of features $\boldsymbol{f}$. The advantage of this algorithm is that it didn't use any explicit model of disclosure.

Similar to other information extraction tasks, more and more research work focus on learning based approaches to conducting co-reference resolution. For instance, Soon et al. (2001) proposed an approach, which learned from a small, annotated corpus to solve the co-reference issues. Different from previously mentioned algorithms, this learning-based approach didn't limit to a certain type of noun phrases such as pronouns, but applied to the general noun phrases. Lee et al. (2013) proposed a deterministic approach, which

combines machine-learning models with rule-based systems. The general architecture of the implemented co-reference system served as multi-sieve to filter the search space for antecedent candidates aggressively. Lee at al. (2013) argued the architecture can keep balance of precision and recall to ensure the performance of the system.

2.3.2 Evaluation of Co-reference Resolution

Zheng et al. (2011) pointed out that the co-reference relation is reflexive, symmetric, and transitive; therefore, it's hard to define the boundary of positive prediction and negative prediction. MUC-6 proposed an evaluation scheme based on the idea of comparing equivalence classes instead of links themselves. This evaluation metric adopts precision, recall and F-score as evaluation criteria. However, this evaluation did not consider the singletons situation and tends to prefer larger chains of co-reference. Based on these shortcomings, Bagga and Baldwin (1998) proposed another evaluation named B^3, which focus on absence of entities in the equivalence classes (Zheng et al., 2011).

Barbu and Mitkov (2001) argued that several evaluation methods could not fairly compare anaphora resolution algorithms due to the difference between data and diversity of preprocessing tools. They argued the evaluation on anaphora resolution algorithms and anaphora resolution system should be conducted separately. Therefore, they implemented an evaluation workbench with four measurements, which include precision and recall as well as success rate and critical success rate. They calculated these measurements based on the following formulations:

$$\text{Precision} = \frac{Number\ of\ correctly\ resolved\ anaphor}{Number\ of\ anaphors\ attempted\ to\ be\ resolved}$$

$$\text{Recall} = \frac{Number\ of\ correctly\ resolved\ anaphor}{Number\ of\ all\ anaphors\ identified\ by\ the\ system}$$

$$\text{Success rate} = \frac{Number\ of\ correctly\ resolved\ anaphor}{Number\ of\ all\ anaphors}$$

$$\text{Critical success rate} = \frac{Number\ of\ correctly\ resolved\ anaphor}{Number\ of\ anaphors\ with\ more\ than\ one\ antecedent\ after\ a\ morphological\ filter\ was\ applied}$$

As the above formulas indicate, while precision and recall were used to evaluate the anaphora resolution system performance, the success rate and critical success rate were added to evaluate the algorithms.

Most of the presented co-reference resolution achieved around 80% accuracy. Compared to NER task, Co-reference Resolution is more complex and there is more ambiguity in co-reference resolution.

2.4 Event Extraction

2.4.1 Introduction

Several research programs in natural language processing and linguistics provide competitions or workshops on event extraction. For instance, Automatic Content Evaluation (ACE), an important program in IE research area, provide event annotation dataset for participants to design and implement IE system to extract information. Additionally, in the event extraction task of ACE 2005, 8 types and 33 subtypes of event are well defined. Table 2.4.1 summarizes the type and subtypes of events defined in ACE 2005 task (Liu, 2009).

TYPE	SUBTYPE
LIFE	Be-Born, marry, Divorce, Injure, Die
MOVEMENT	Transport
TRANSACTION	Transfer-Ownership, Transfer-Money
BUSINESS	Start-Org, Merge-Org, Declare-Bankruptcy, End-Org
CONFLICT	Attack, Demonstrate
CONTACT	Meet, Phone-Write
PERSONEEL	Start-Position, End-Position, nominate, Elect
JUSTICE	Arrest-Jail, Release-Parole, Trail-Hearing, Charge-Indict, Sue, Convict, Sentence, Fine, Execute, Extradite, Acquit, Appeal, Pardon

Table 2.4.1 Types and Subtypes of event defined in ACE 2005(Liu, 2009)

As the definition of event extraction indicates, event extraction task contains two subtasks: 1). Event detection and classification 2). Extraction of event argument such as event subject, event time etc. In actual, Text Analysis Conference (TAC) divides Event Track into two divisions: Event Nugget (EN) tasks to detect and link event and Event Argument (EA) tasks to extract event arguments and link arguments belonging to the same event. (TAC KBP, 2015). Some of the research would like to focus merely one of the subtasks (Nguyen & Grishman, 2015; Liu et al, 2016), while other proposed models to address both by first identifying event type and followed by extracting event argument (Reschke et al., 2014; Chen et al., 2015). Grishman, Westbrook and Meyers (2005) proposed a system, which applied four classifier to address the event extraction:

1) Argument classifier to decide whether the given mention is an event argument;

2) Role classifier to classify which role of the given argument plays in the event;

3) Event classifier to decide whether the given arguments constitute event mention;

4) Event co-reference classifier to decide whether two given event mentions refers to the same event.

2.4.2 Overview of Methodology

Hogeboom et al. (2009) distinguish approaches based on the field of modeling; they divided the popular approaches into three categories: Data-driven approach, Knowledge-driven approach and Hybrid event extraction approach. Data-driven approaches require large dataset to learn from and no domain knowledge is required. In contrast, knowledge-driven approach requires domain experts to lexico-syntactic and lexico-semantic patterns to extract event information. While knowledge based approach does not need large amount of data to train, several issues remain unsolved. For example, pattern based approach result in low recall of extracted information because the limited patterns generated by experts would not be applicable to the sentences whose synaptic or semantic structures different from the generated patterns. Additionally, knowledge-driven requires substantial domain experts each time the system adopt new event types. The hybrid approaches namely combine the data-driven approach and knowledge-driven approach to improve the general results. For instance, Liu and Strzalkowski (2012) proposed an information extraction system named BEAR (Bootstrapping Events And Relations) to automatically learn patterns and to extraction information from text. According to evaluation on ACE (Automatic Content Extraction) data, the performance of the BEAR system achieved great improvement on their baseline system.

2.4.3 Evaluation on Event Extraction

According to IE evaluation measurements introduced by MUC (Message Understanding Conferences), precision is and recall are defined as follow:

$$\text{Precision} = \frac{\text{Number of correctly filled template slots}}{(\text{Number of annotated slots by the system})}$$

$$\text{Recall} = \frac{\text{Number of correctly filled template slots}}{\text{Number of manually annotated slots by human}}$$

ACE 2005 provided another event evaluation scoring approach, defining the output event score to be the product of an inherent event value and the sum of the values of the event's entity participants, which evaluate the accuracy of event arguments and multiple other event attributes. Liu et al. (2015) stated that most prior evaluations on event mention did not give partial credits to partial matches, which might be important during particular setting.

2.5 Conclusions

In this chapter, we presented a general overview of research in the field of Information Extraction and its subtasks. In addition, we briefly introduced and compared current popular techniques used in this area. The prior related work lays the foundation for our proposed project and sets a benchmark for the performance of our system. Combining the previous related research in IE with the background of our project, we defined that the scope of our project is to build an IE system, which should implement the following tasks:

- Extract business related entities, mainly including: Organization, Person, Location

- Solve co-reference issues within document for ‘ORGANIZATION’ entity, basically to resolve co-reference of pronouns
- Design and implement pattern-based heuristic to extract preliminary business event information: initial public offering

3. Methodology

3.1 General Architecture

Since the proposed project's goal is to process news articles to extract structured business event information from business news, we used business news articles as data. Additionally, the system's potential stakeholders would like to be more interested in business in the RTP region. One of the potential resources is the Triangle Business Journal, which provides local business news, research and events in the Raleigh, Durham, Chapel Hill region. We will also consider using data from the Freebase database in the business domain. As far as we have learned, Freebase is one of popular databases for research in information extraction (Mintz et al., 2009; Riedel et al., 2010). Additionally, Freebase categorizes data into different categories such as Stock listing, Employer. These predefined categories will help us to create more precise IE patterns and also may potentially help train our text mining model in the future. These two data sources are available online, so we can easily conduct further analysis using natural language processing system. To fetch data from Triangle Business Journal, we use an automated program to parse html data downloaded from their official website. As for Freebase database, they provide API to end-users to query data using programming language.

Figure3.1 demonstrates the architecture of the project. We first deployed html parsing program to preprocess web data that downloaded from news website to get the clean text of news content and used the existing Natural Language Processing systems such as Stanford NLP, OpenNLP system to conduct Named Entity Recognition (NER) and Co-

reference Resolution and then we used our handcrafted rules based on our data to improve the results. Based on the improved processed data, we implemented shallow parse technique and extracted the business event information from text. The extracted structured information will serve as training examples for a future unsupervised learning model. During these procedures, we will involve a few human curators to briefly review and correct the system processed data to ensure the performance of our pattern bootstrapping model.

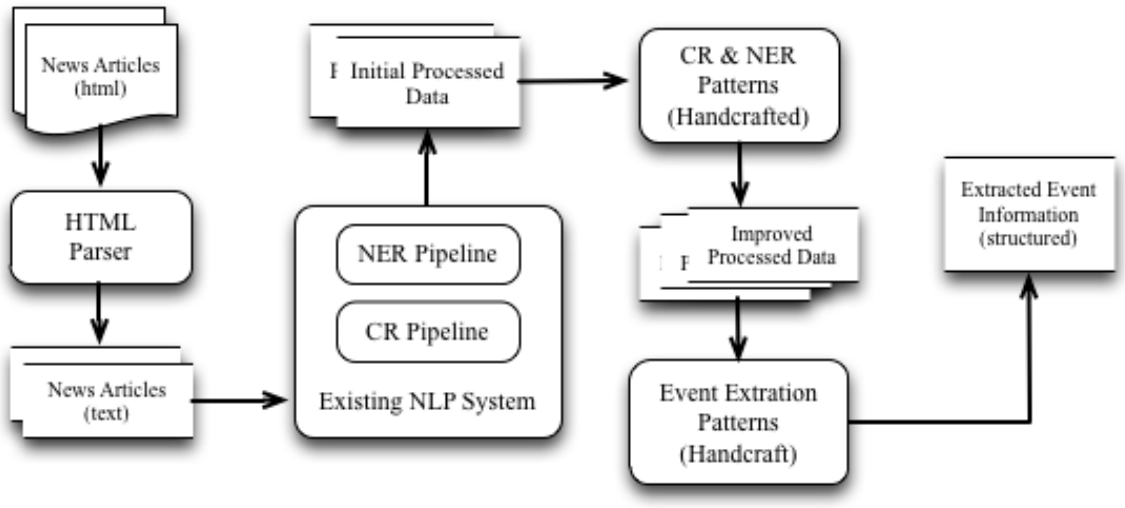

Figure 3.1 General Architecture of the Proposed Project

Our study focuses on exploiting business event information from unstructured documents. This information includes spatio-temporal transactions that occur among a changing set of firms and people, which are key sources for analyzing factors of economics of a region.

3.2 Named Entity Recognition

In this section, we generally describe the process of named entity recognition pipeline in the proposed system. We first used the Stanford Named Entity Recognizer (Stanford NER) to generate the NER labels for each token in the article. The preliminary results indicated that, the Stanford NER performed well on recognizing 'PERSON' entity, following by 'LOCATION', while they had difficulty to identifying 'ORGANIZATION' entity, owing to the difference between our dataset and its default training dataset. While Stanford NER provides the interface for user to train their own NER model by using customized training dataset, it is time and labor consuming to generate adequate quantity of data for the model to learn. Therefore, we generated several curate patterns to correct label and further improve the preliminary result of Stanford NER.

In general, some of the curate patterns are generated to correct some common mistakes made by Stanford classifier and we also create a few high precision rules to identify organizations ignored by Stanford NER. We will explain those patterns in details in the rest of the paragraph. We summarized two common situations that Stanford NER generated false negative 'ORGANIZATION' instance. One is the 'ORGANIZATION' entity mistakenly recognized to be 'PERSON' and the other is the 'ORGANIZATION' entity not recognized to be any of predefined named entity category. We generated following rules to improve the performance by solving the two problems stated above.

a. Utilize html metadata to identify organization entity. When we parsed the html data, we saved the text with hyperlink to company profile webpage into database for later organization entity correction.

b. Use external gazetteer of common position titles in organization such as Chief Executive Officer (CEO), President and Spokesman etc., to label sentences and applying several common patterns to sentences with the job position labeled:
    - *[Person]*, the *[Title]* of *[Organization]*
    - *[Organization]*'s *[Title] [Person]* or *[Organization] [Title] [Person]*
c. Correct label of token following by corporate identifier such as 'LLC.', 'Inc.','Ltd.'. We consider a sequence of tokens to be an organization entity if it starts with capital letter and ends with corporate identifier.
d. Search organization candidates near keywords such as company, group, startup
e. Generate potential alternatives 'ORGANIZATION' entity name for existing 'ORGANIZATION' entity. In a document, with existing organization entity 'Wells Fargo & Company', we generated alternative organization entity 'Wells Fargo' and label all 'Wells Fargo' appeared in the document to be an organization entity by using string match approach.
f. Resolve potential 'ORGANIZATION' entity labeled with 'PERSON' by Stanford NER. If a sequence of tokens matched with any patterns described above, we considered it to be an organization entity. However, some of them already be labeled with 'PERSON' by Stanford NER during the preprocess stage. We defined several rules to resolve such conflict, including searching any pronouns appearing near the entity referring to person like who, she, his.

By implementing these patterns, we achieved a great improvement on Name Entity Recognition on our dataset. Evaluation and results will be demonstrated in later section.

3.3 Co-reference Resolution

In this section, we would like to introduce our naïve approach to resolve co-reference of organization entity and person entity. Since the co-reference resolution is quite complex problem in natural language processing and most of existing approaches achieve 70%-80% accuracy. In our information extraction system, co-reference resolution aims to help resolve the situation when a sentence mentioned any business event but the subject entity of event was not found by NER because those entities could be replaced by any pronouns or definite articles like 'the company'. Therefore, different from traditional definition of co-reference resolution problem, we defined the scope of our resolution is to connect pronouns, possessive and some particular definite articles, such as 'the company', with any organization mentioned in previous sentences.

We limited the window of searching the antecedent of a pronoun (possessive) or a definite article in the *ith* sentence, to [i-1, i], that is, we only searched the antecedent in the *ith* sentence and the *(i-1)th* sentence.

Based on the approached proposed by Hobbs (1978), we searched the antecedent in the parse tree of the sentence and the general strategy of our searching method is as follow:

First, search any organization entity in the immediate dominated noun phrases (NP) of the pronoun (possessive) or the definite article in the same sentence. If the pronoun is part of the dominated NP of the sentence, then search any organization entity in the dominated NP of previous sentence. To find the immediate dominated NP, we first search node where the pronoun (possessive) located, and return to its ancestors to find out the first ancestor whose tag is 'VP' and then search the sibling of that ancestor to find the node with 'NP' tag and returned it, if not, continue searching ancestor until it reaches the

root. If the immediate dominated NP was not found, we consider the pronoun (possessive) or the definite article locate in the Subject of sentence, we further search the entity in dominated NP of previous sentence. To find the dominated NP of the sentence, we use breath-first-search strategy to find the first NP node and return.

Figure 3.3.1 and Figure 3.3.2 (a) and Figure 3.3.2 (b) demonstrate the searching strategy described above.

Example 1: *Since Editas Medicine first filed its intention to go public on Jan. 4, the Nasdaq Biotechnology Index has lost more than 11 percent of its value.*

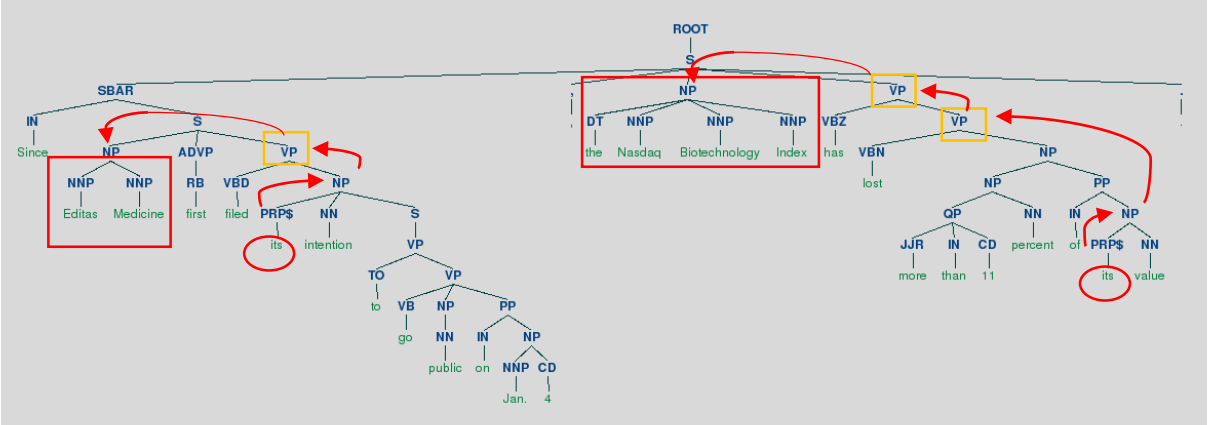

Figure 3.3.1 Parse Tree for Example 1

Example 2: *MaxPoint's technology predicts what people will buy down to the ZIP code. The company makes 13 trillion calculations per day on everything from demographic information to purchasing behavior.*

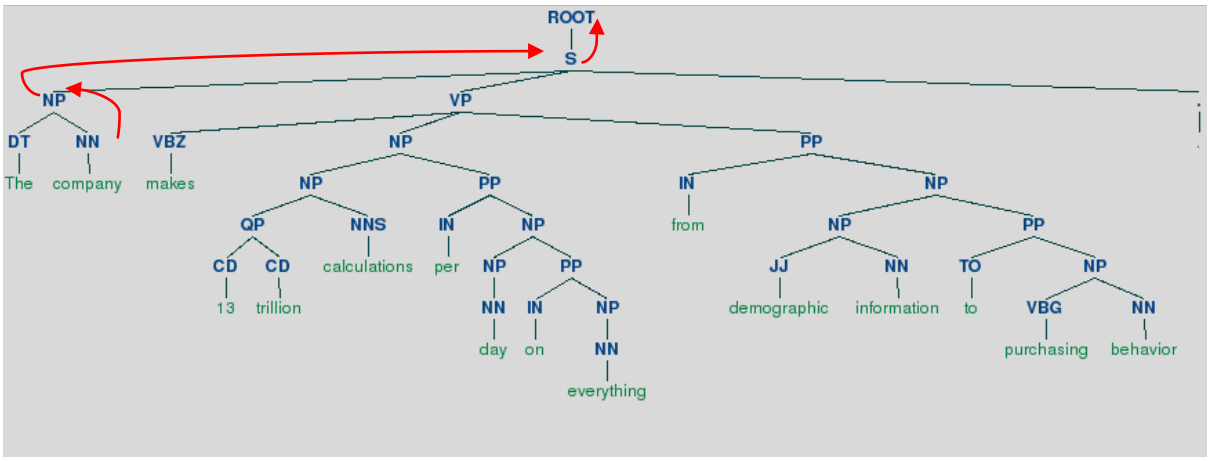

Figure 3.3.2(a) Parse Tree for Example 2

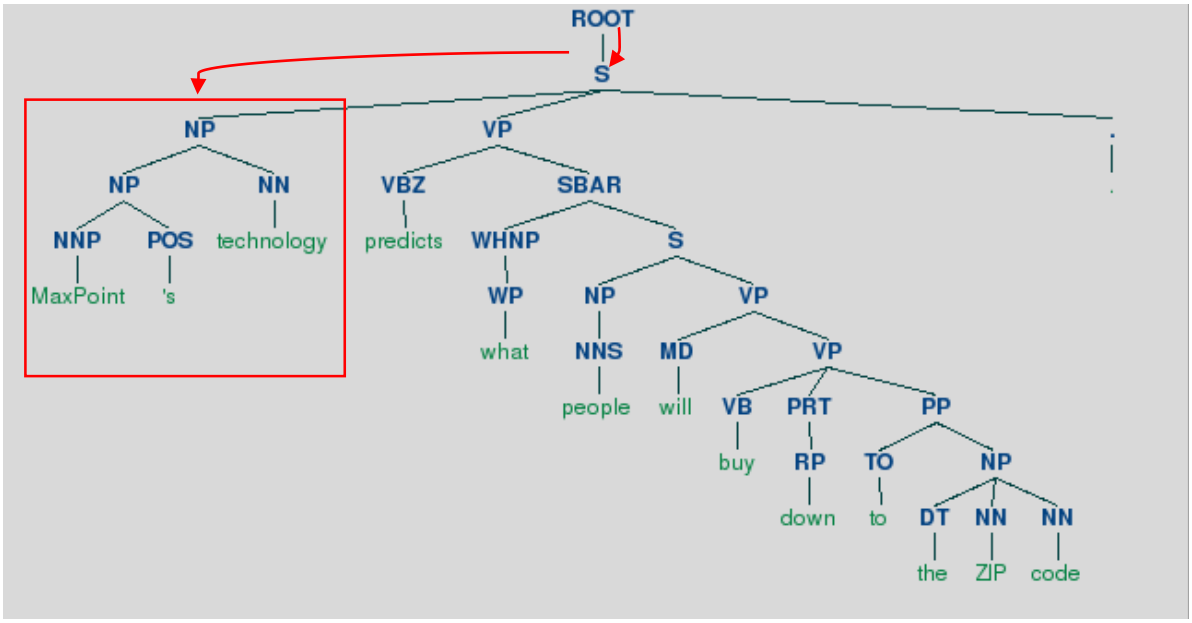

Figure 3.3.2(b) Parse Tree for Example 2

Since definite article 'The company' is in the dominated NP of the sentence, we search the dominated NP of the previous sentence instead.

While our co-reference resolution only solved part of the pronouns and in some cases, it connected some pronouns, which do not refer to any entity. For instance, the pronoun 'it' in sentence 'It is great!' does not have any referent. However, those errors were tolerated

in our system because they would only have little influence on the performance of event extraction.

3.4 Event Extraction

Similar to prior research in event extraction, our approach divided this task into event classification and argument extraction. However, different from most of prior research in event extraction, which focused on detecting and extracting event information from a sentence. The proposed information extraction system would like to aggregate and to summarize the information extracted from sentences in a document. We chose to build a model on a document instead of a sentence because of the characteristics of our data. The data we used for business event detection is news articles. Most of the news articles are narratives, where sentences connecting to each other to represent a context for readers to understand. Therefore, the information (arguments) of event are likely to spread out the articles. In addition, besides reporting core event occurring to a company, several news articles would like to provide a background about a company by reviewing a series of previous events. If we focused on sentence level to extract event, the extracted argument would be incomplete or even conflictive and would be hard to connect with other arguments.

The scope of business event we would like to limit in this research is the initial public offering (IPO) event for a company. After generally went through several news articles reporting an IPO event, we further defined several statuses (sub-events) of IPO event as follow:

a. IPO Filling: An organization filed a registration document about IPO with U.S. Securities And Exchange Commission (SEC)

b. IPO Updates: An organization updated any details such as price, the amount of shares about its planned public offering with SEC after IPO Filling
c. IPO Priced: An organization went public and its stock is available for trading on market
d. IPO Withdraw: An organization cancelled its planned IPO after filling with SEC
e. IPO Delay: An organization delayed its planned IPO and has not determined new IPO date with SEC after IPO filling
f. IPO Upcoming: An organization is expected to trade in next few days on schedule.
g. IPO Intention: An organization announced its intention to go public but had not filled any document with SEC

Based on the statuses described above, a company should be considered to hold an actual IPO event only when it becomes the IPO Priced status, otherwise, there will be a probability of canceling or delaying scheduled IPO.

After defining the scope of business event we focused in this research, we now describe the functionality that our system is expected to perform. Similar to prior approaches, we divided our event extraction task into step 1: event detection and classification and step 2: event argument extraction. For event classification stage, we consider that our system should have the following functions. First, given a news article, the proposed system should distinguish whether the news article describes IPO-related event or not. Second, if the news article is classified to be IPO-related, the system is expected to make a further classification about which IPO status the news article described. We divided the task of IPO status classification into two levels: the first level is to classify whether the given

news article described an actual IPO event and the result of classification should be a Boolean value. The advanced task is to classify the IPO-related event in the news article to be one of the IPO status summarized above. For the event argument extraction step, the proposed system should extract correct value for predefined event arguments, if any, from the given news document. We will describe the IPO event arguments in the 3.4.2 section.

We would like to introduce the general procedure of our patterned based event extraction method in the following paragraphs. After preprocessing news document by NER and co-reference resolution, we implemented the business IPO event extraction generally using Regex expression and word normalization and lemmatization. The general architecture can be described as multi-pass sieve procedure to detect sentences that are supposed to contain business event information and then extract them based on multiple handcraft patterns.

As shown in Figure 3.3.1, from top to bottom, crafted detection and extraction patterns gradually change from the word level (keyword matching) to sentence level (sentence matching), which increases the precision of event detection and extraction but decreases the recall meanwhile. However, result of each pass is individual and they are not exclusive to each other, that is, the sentence was not recognized to be IPO-related in prior stages will still be tested in the next stage because patterns defined in each stage are complimentary, which can help balance the overall recall and precision of extracted information. The procedure introduced above is applied to detect event and extract argument from each sentence in a document and we will describe the heuristics we used

to generate document level event classifier and argument extraction from those sentence level information.

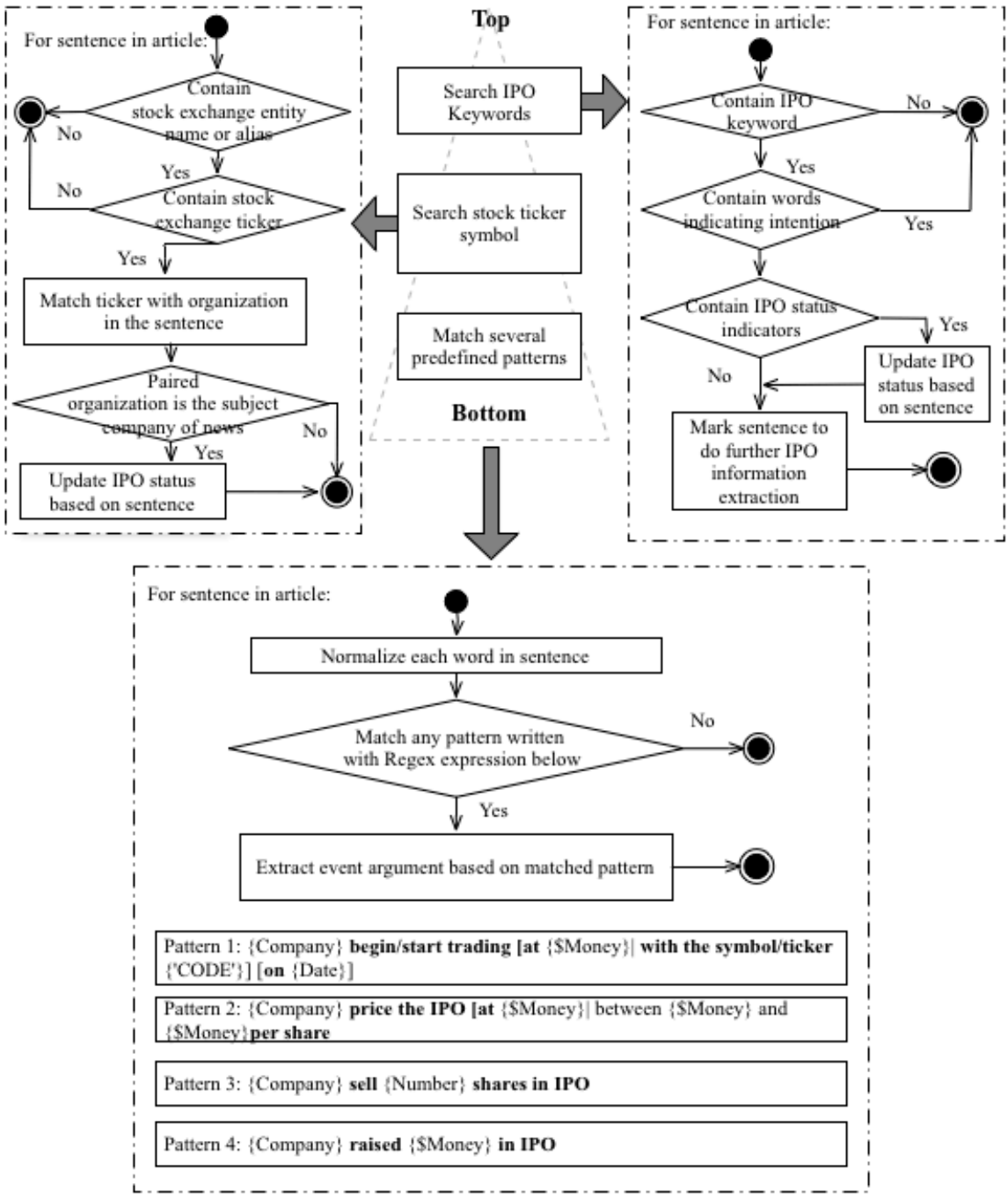

Figure 3.4.1 The general procedure of our patter based event extraction approach

3.3.1 Event Classification

After processing all sentences in a news document, we got each sentence tagged with IPO status, e.g. IPO-Filing, IPO-Priced, IPO-Intention, IPO-Updates, if a sentence is not recognized as IPO-related in any of the stage in the procedure, it is marked with 'None' tag. If all the sentences in a news document were marked with 'None' tag, this news article is considered to be irrelevant to IPO event. With those sentence-based status tags, the next step is to build a document-based IPO status classifier. We assumed one IPO-related news document only describes a particular IPO status for a particular organization. However, we could get different IPO status tags, which are mutually exclusive to each other, from different tagged sentences. Therefore, we need to design a mechanism to resolve those conflictive statuses predicted from each sentence and further to decide which status of IPO event are the news document describing. We solved this problem by giving each IPO event status a confidence value based on status indicators. The confidence value v_i for IPO status S_i, evaluates the how much the document is likely to describe IPO status S_i , if S_i appears in any sentence-based status tag regardless appearance of other status tags. The confidence value is assigned by domain experts based on their domain knowledge and the ambiguity of status keyword indicators. A keyword that potentially indicates multiple IPO statuses is ambiguous and it is not a good status indicator.

3.4.2 Event Argument Extraction

Before describing the process of argument extraction, we firstly would like to introduce the IPO-related arguments that are expected to be identified from news documents. For

IPO related event, system should identify information, if any, about IPO company, IPO status, event date, stock symbol for IPO company, IPO price per share as well as any information about the amount of shares or total amount of money offered in IPO. However, it is not rational to assume all of the arguments can be detected in the same sentence. We need to partition the candidate arguments. Considering a news article could contain multiple organizations or other annoying inference information, we cannot just combine all of information extracted from difference sentences together to represent such event mention. Therefore, we further defined several groups or pairs of arguments as follow, and these argument groups /pairs can be connected by the company name to further represent an event mention.

- IPO Company, IPO Event Status, Event Date
- Company, Stock Symbol
- IPO Company, IPO Price per share
- IPO Company, IPO Amount offered/ IPO shares offered

We considered the first group of arguments to be primary arguments of IPO-related event and they are n-ary relation. To be consistent with our assumption that one news documents only describes one core event for a particular company, the primary arguments are designed to be document level and each document should only generate one record of that group of arguments. However, the rest IPO arguments were considered to be advanced arguments and the system can extract multiple pairs of records.

As for the procedure of argument extraction, we followed the general steps demonstrated in Figure3.3.1. In the first stage, for each sentence containing IPO keywords but not containing intention words, the system would try to find out any ‘ORGANIZATION’,

‘TIME’, ’MONEY’, ’DATE’ entities or any co-referent phrase referring to such entities in the sentence. In the second stage, we only focused on matching ‘ORGANIZATION’ with its stock ticker. As for the pattern-matched stage, the proposed system would extract particular event argument defined in a pattern when it detects the sentence matched with that pattern. For instance, ‘ORGANIZATION’, ’MONEY’ and ’DATE’ are expected to be detected from the pattern -- ‘{Company} begin/start trading [at {$Money} on {Date}]’. Additionally, we derived several sub-patterns to adopt partially matched sentences. Examples are shown as follow:

Example Pattern 1: {Company} begin/start trading [at {$Money} | with the symbol/ticker {‘CODE’}] [on {Date}]

Sub-pattern: {Company} begin/start trading at {$Money}

Sub-pattern: {Company} begin/start trading on {Date}

Sub-pattern: {Company} begin/start trading with the symbol/ticker {‘Code’}

Example Pattern 2: {Company} price the IPO at {$Money} per share.

Sub-pattern: {Company} price the initial public offering at {$Money} per share

Sub-pattern: {Company} price the IPO at {$Money}

4. Evaluation

4.1 Named Entity Recognition

Similar to prior evaluation approach, we used precision, recall, and F1-score as evaluation measurements and we calculated the value of those measurements by using entity instead of token as counting unit as well. Based on those criteria, we evaluated our pattern-based improving approach with the preliminary result of Stanford NER serving as baseline. We manually labeled 25 news articles, 9689 words in total, and used them as test data and the evaluation of two named entity recognizer is summarized in Table 4.1.1 and Table 4.1.2. The test data was not used in developing our improved NER approach.

Entity	P	R	F1	TP	FP	FN
DATE	1.00	1.00	1.00	172	0	0
DURATION	1.00	1.00	1.00	32	0	0
LOCATION	0.91	0.83	0.87	150	14	31
MISC	0.49	0.96	0.65	25	26	1
MONEY	1.00	1.00	1.00	94	0	0
NUMBER	1.00	1.00	1.00	108	0	0
ORDINAL	1.00	1.00	1.00	13	0	0
ORGANIZATION	0.97	0.50	0.66	178	6	175
PERCENT	1.00	1.00	1.00	30	0	0
PERSON	0.52	0.95	0.67	69	65	4
SET	1.00	1.00	1.00	3	0	0
TIME	1.00	1.00	1.00	31	0	0
TOTAL	0.89	0.81	0.85	905	111	211

Table 4.1.1 Performance Matrix for Stanford NER

Entity	P	R	F1	TP	FP	FN
DATE	1.00	1.00	1.00	172	0	0
DURATION	1.00	1.00	1.00	32	0	0
LOCATION	**0.99**	**0.93**	**0.96**	169	1	12
MISC	0.90	1.00	0.94	26	0	0
MONEY	1.00	1.00	1.00	94	0	0
NUMBER	1.00	1.00	1.00	108	0	0
ORDINAL	1.00	1.00	1.00	13	0	0
ORGANIZATION	**0.94**	**0.88**	**0.91**	310	20	43
PERCENT	1.00	1.00	1.00	30	0	0
PERSON	**0.70**	**0.98**	**0.82**	72	30	1
SET	1.00	1.00	1.00	3	0	0
TIME	1.00	1.00	1.00	31	0	0
TOTAL	0.95	0.95	0.95	1060	54	56

Table 4.1.2 Performance Matrix for our Improved NER

As table 4.1.1 and table 4.1.2 suggest, the approach we proposed has significantly improved the performance of Stanford NER for ‘ORGANIZATION’, ’PERSON’, ’LOCATION’ entity. In addition, as table 4.1.1 indicates, entities such as ‘NUMBER’, ‘TIME, ‘DATE’ obviously outperformed than ‘PERSON’ or ‘ORGANIZTAION’. The reason is that those entity types have their typical characteristics that are easy to identify. For instance, ‘NUMBER’ entity can be easily identified with digital character or the word combination of one to nine. Additionally, the date format usually is ‘MM DD, YYYY’. However, the organization entity is the entity that requires context to understand and to identify. Different from ‘DATE’, which could be enumerated by combining date (1-31), month (Jan - Dec) with year (1900-2016), new organization is created without any limitation on format or word, a company name can include both digital character and letter. As for ‘PERSON’ entity,

while we can list commonly used names, people with foreign names are difficult. Moreover, sometimes it is hard to distinguish ‘PERSON’ with ‘ORGANIZATION’ owing to the fact that several companies are named after their founders. The above situations increase the difficulty of identifying those entities.

4.2 Co-reference Resolution

Owing to the large difference between our co-reference resolution approach and prior related research in problem definition and resolution scope, we did not use exiting co-reference resolution (CR) evaluation approach to evaluate our CR method. Instead, we conducted a simple test and used the precision of connecting pronouns (possessive) or definite articles to organization entities to evaluate our approach. We used 10 documents, 310 sentences in total as test data. Our approach connected 91 pronouns (possessive) or definite articles with organization entities, 57 of which were correct.

We further analyzed the false connected instances and we found that the accuracy of connection decreased as the length of document increased. In addition, error propagation is found. Among the false examples, most of the errors are owing to the misclassified named entity, which means the antecedent was mistakenly classified to be an organization entity during NER.

4.3 Event Extraction

4.3.1 Evaluation on Event classification

We provided following criteria for evaluating our IE system:

- Whether the system can correctly identify a news article to be IPO-related (Boolean value)
- Whether the system can correctly identify an IPO-related news article describing an actual IPO event (Boolean value)
- Whether the system can correctly classify the IPO status described in an IPO-related news article (categorical value)

Among 200 news articles, the event extraction system successfully identified all the IPO-related news articles. Among the 100 IPO-related news articles, 30 news articles described actual IPO event, the proposed system recognized 21 of them, the precision and recall of IPO event classification is 0.78 and 0.70, which is same as the precision and recall of IPO_Priced status classification in Table 4.3.1. As for the status (sub-event) classification, we summarized the performance matrix as Table 4.3.1. As the evaluation result indicates, the proposed system achieved a success on event and sub-event classification. However, the low precision and recall of IPO_Upcoming draws us attention. We further analyzed the misclassified result and found that most of the misclassified IPO_Upcoming event were recognized either to be IPO_Priced or IPO_Intention. That is because sentences describing the upcoming event always contain words like ‘plan’, ‘is expected to’. These words were designed to be identifiers of IPO intention class. In addition, several upcoming event descriptions contained specific expected stock trade time, which was actually later than the time of news report, and this lead to misclassifying of the IPO status. Future work might consider improving performance for the IPO_Upcoming event category.

Event Status	P	R	F1	TP	FP	FN
IPO_Intention	0.60	0.67	0.63	6	4	3
IPO_Filing	0.81	0.96	0.88	25	6	1
IPO_Updates	0.73	0.53	0.62	8	3	7
IPO_Upcoming	0.33	0.33	0.33	2	4	4
IPO_Priced	0.78	0.70	0.74	21	6	9
IPO_Delay	0.86	0.86	0.86	6	1	1
IPO_Withdraw	0.88	1.00	0.93	7	1	0
Overall	0.75	0.75	0.75	75	25	25

Table 4.3.1 Performance Matrix for our IE system on sub-event classification

4.3.2 Evaluation on Event Argument Extraction

For argument extraction of IPO event, we only used precision as measurement for the following reason: Recall evaluation requires human efforts to read each test document carefully and then extract all of the potential arguments of event manually. This task is time consuming and tedious.

Additionally, the extracted information can sometimes be biased. We tentative asked three students with basic knowledge about IPO to extract IPO related arguments from five news documents and the extracted arguments varies from each student.

For IPO event, the primary event arguments are IPO Company, IPO Date, IPO Status. Besides, an IPO related news article could contain, if any, following extra event information: Company Ticker, IPO Price per share, Shares Offered and Offer Amount.

For the primary arguments of IPO event, our IE system extracted 122 potential event arguments from the 100 IPO-related news documents. Since the structure of the (IPO company, IPO date, IPO status) is n-ary relation. We considered an extracted record to be

correct only if all of fields in this record are correct. Based on this rule, we calculated the precision of extracted primary argument is 0.534. If we relax our evaluation criteria by dividing the n-ary relation to binary relation. In other words, we divided the record from (IPO Company, IPO Status, IPO Date) into two binary records (IPO Company, IPO Status) and (IPO Status, IPO Date) and evaluate the correctness of each record individually. In this situation, the precision of extracted primary arguments increases to 0.75.

For the advanced IPO arguments, our IE system extracted 221 records in total, including 101 records of amount or shares offered, 55 records about stock price and the rest 65 records about ticker symbol. The overall precision of extracted advanced argument is summarized in Table 4.3.2.

Argument type	**Total**	**Correct**	**Precision**
Amount (Shares) Offered	101	61	0.60
Stock Price	55	35	0.64
Ticker Symbol	65	49	0.75
Total	221	145	0.66

Table 4.3.2 Precision of extracted advanced argument of IPO event

As the several evaluation results from different aspects suggest, the proposed Business Information Extraction System worked well on detecting and extracting IPO related information from business journals.

5. Conclusion

In our research, we designed and implemented an Information Extraction System to extract IPO event information from business news journals. We designed a heuristic approach to improve the annotated result of Stanford Named Entity Recognizer on business new articles, particularly in 'ORGANIZATION', 'PERSON' and 'LOCATION' entities. Based on the preprocessing result of NER and Co-reference Resolution, we handcrafted several high precision rules to identify IPO-related events from sentences. With preliminary sentences classification result, we applied several criteria regarding IPO status priority and ambiguity of status identifier to build an IPO status classifier on document. After conducting event classification, we used pre-generated patterns to extract corresponding event arguments from target sentences. To evaluate the performance of our proposed system, we manually labeled 25 documents for NER evaluation and 100 documents for IPO sub-events evaluation. As the precision and recall of the results indicates, the system improve the performance of existing NER system and it can successfully detect and classify IPO event and most of its sub events except IPO upcoming sub-event. We further analyzed the reason of low performance of IPO upcoming status classification and plan to refactor the generated patterns to improve the result.

Future work might consider applying similar techniques to extract other types of business events such as layoff and fund raising. Additionally, it might be helpful to use patterned based extraction results to train a learning based model, which

takes the advantage of machine learning technique so that the system would not require substantial domain expertise to craft new patterns.

Acknowledgement

This research was supported by NSF SMA-1439532 and University of North Carolina at Chapel Hill.

References

2005. The ACE 2005 (ACE05) evaluation plan. http://www.nist.gov/speech/tests/ace/ace05/doc/ace05-evalplan.v3.pdf.

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.

Bagga, A., & Baldwin, B. (1998, May). Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference* (Vol. 1, pp. 563-566).

Barbu, C., & Mitkov, R. (2001, July). Evaluation tool for rule-based anaphora resolution methods. In *Proceedings of the 39th Annual Meeting on Association for* Computational Linguistics (pp. 34-41). Association for Computational Linguistics.

Bast, H. Named Entity Recognition. Retrieved from: http://ad-teaching.informatik.uni-freiburg.de/efficient-nlp-ws1112/Efficient%20Natural%20Language%20Processing,%20WS%201112,%20Session%206,%207Dec11.pdf

Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., & Cunningham, H. (2002, June). Shallow methods for named entity coreference resolution. In *Chaınes de références et résolveurs d'anaphores*, workshop TALN.

Chen, Y., Xu, L., Liu, K., Zeng, D., & Zhao, J. (2015). Event Extraction via Dynamic Multi-Pooling Convolutional Neural Networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing* (Vol. 1, pp. 167-176).

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., & Weischedel, R. M. (2004, May). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. In *LREC* (Vol. 2, p. 1).

Gartner, Inc. Turning Dark Data into Smart Data. Retrieved from: http:／／www.federalnewsradio.com/wp-content/uploads/pdfs／031115_gartner_co_branded_newsletter_turning_dark_data_into_smart_data.pdf

Grishman, R., Westbrook, D., & Meyers, A. (2005). NYU's English ACE 2005 system description. *ACE*, 5.

Ge, N., Hale, J., & Charniak, E. (1998, August). A statistical approach to anaphora resolution. In *Proceedings of the sixth workshop on very large corpora* (Vol. 71, p. 76).

Hobbs, J. R. (1978). Resolving pronoun references. *Lingua*, *44*(4), 311-338.

Hogenboom, F., Frasincar, F., Kaymak, U., & De Jong, F. (2011, October). An overview of event extraction from text. In *Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) at Tenth International Semantic Web Conference* (ISWC 2011) (Vol. 779, pp. 48-57).

Krishnan, V., & Ganapathy, V. (2005). Named Entity Recognition.

Lappin, S., & Leass, H. J. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, 20(4), 535-561.

Lau, R. Y., & Zhang, W. Semi-supervised statistical inference for business entities extraction and business relations discovery. In *SIGIR 2011 workshop*, July (Vol. 28).

Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M., & Jurafsky, D. (2013). Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4), 885-916.

Liu, T. (2009). Bootstrapping Events and Relations from Text. Ph.D. Dissertation. State Univ. of New York at Albany, Albany, NY, USA. Advisor(s) Tomek Strzalkowski. AAI3387201.

Liu, T., & Strzalkowski, T. (2012, April). Bootstrapping events and relations from text. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 296-305). Association for Computational Linguistics.

Liu, S., Liu, K., He, S. & Zhao, J., A Probabilistic Soft Logic Based Approach to Exploiting Latent and Global Information in Event Classification, In *Proceedings of AAAI 2016*, Phoenix, USA, Phoenix, 12-17

Liu, Z., Mitamura, T., & Hovy, E. (2015, June). Evaluation Algorithms for Event Nugget Detection: A Pilot Study. In *Proceedings of the 3rd Workshop on EVENTS at the NAACL-HLT* (pp. 53-57).

Linguistic Data Consortium, 2005. *ACE (Automatic Content Extraction) English Annotation Guidelines for Events*, version 5.4.3 2005.07.01 edition.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2* (pp. 1003-1011). Association for Computational Linguistics.

Manning, Christopher. Information Extraction and Named Entity Recognition. Retrieved from: http://spark-public.s3.amazonaws.com/nlp/slides/Information_Extraction_and_Named_Entity_Recognition_v2.pdf

Nguyen, T. H., & Grishman, R. (2015). Event detection and domain adaptation with convolutional neural networks. *Volume 2: Short Papers*, 365.

Riedel, S., Yao, L., & McCallum, A. (2010). Modeling relations and their mentions without labeled text. In *Machine Learning and Knowledge Discovery in Databases* (pp. 148-163). Springer Berlin Heidelberg.

Reschke, K., Jankowiak, M., Surdeanu, M., Manning, C. D., & Jurafsky, D. (2014). Event Extraction Using Distant Supervision. In *LREC* (pp. 4527-4531).

Saggion, H., Funk, A., Maynard, D., & Bontcheva, K. (2007). *Ontology-based information extraction for business intelligence* (pp. 843-856). Springer Berlin Heidelberg.

Sarawagi, S. (2008). Information extraction. *Foundations and trends in databases*, 1(3), 261-377.

Soon, W. M., Ng, H. T., & Lim, D. C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4), 521-544.

Sunita Sarawagi (2008), "Information Extraction", *Foundations and Trends® in Databases*: *Vol. 1: No. 3*, pp 261-377. http://dx.doi.org.libproxy.lib.unc.edu/10.1561/1900000003

Strzalkowski, T., & Wang, J. (1996, August). A self-learning universal concept spotter. In *Proceedings of the 16th conference on Computational linguistics*-Volume 2 (pp. 931-936). Association for Computational Linguistics.

Tjong Kim Sang, E. F., & De Meulder, F. (2003, May). Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL* 2003-*Volume 4* (pp. 142-147). Association for Computational Linguistics.

Weld, D. S., Hoffmann, R., & Wu, F. (2009). Using wikipedia to bootstrap open information extraction. *ACM SIGMOD Record*, 37(4), 62-68.

Zheng, J., Chapman, W. W., Crowley, R. S., & Savova, G. K. (2011). Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6), 1113-1122.