Event Detection from Business News

Ishan Verma[1(✉)], Lipika Dey[1], Ramakrishnan S. Srinivasan[2], and Lokendra Singh[1]

[1] TCS Innovation Labs, Tata Consultancy Services Ltd., Delhi, India
{ishan.verma, lipika.dey, singh.lokendral}@tcs.com
[2] TCS Innovation Labs, Tata Consultancy Services Ltd., Cincinnati, USA
ramakrishnan.srinivasan@tcs.com

Abstract. An event is usually defined as a specific happening associated with a particular location and time. Though there has been a lot of focus on detecting events from political and other general News articles, there has not been much work on detecting Business-critical events from Business News. The major difference of business events from other events is that business events are often announcements that may refer to future happenings rather than happenings that have already occurred. In this paper, we propose a method to identify business-critical events within News text and classify them into pre-defined categories using a k-NN method. We also present an event-based retrieval mechanism for business News collections.

Keywords: Event detection · Classification · Event-based news retrieval

1 Introduction

Today's business intelligence is heavily event-driven. In this context, events of interest are those that provide information necessary to strategically exploit and improve an enterprise's processes as well as to take tactical advantage of events as they occur. With a multitude of News sources on the web, it is possible to gather a lot of information about competitors, product failures, and major global economic and political events almost as soon as they occur. Analysis of business or financial news has gained popularity due to the immense potential these articles have for contributing towards effective predictive analytics for companies. Traditionally, the most exploited technology for analyzing business news is sentiment classification that assigns positive or negative polarity to an article and thereby to the entities contained within it. There is a large body of research that studies the correlation between News sentiments and market indices. However, there are many other specific events that can be extracted from business News articles that can provide useful business insights or competitive intelligence for future planning. For example an announcement like "*Hyundai has just launched the third generation of the Santa Fe which has sleeker lines and a very executive look"* may prove to be useful for Hyundai's competitors to plan demand scenarios in similar segments.

An event is usually defined as a specific happening associated with a particular location and time. Though there has been a lot of focus on detecting events from

M. Kryszkiewicz et al. (Eds.): PReMI 2015, LNCS 9124, pp. 575–585, 2015.
DOI: 10.1007/978-3-319-19941-2_55

political and other general News articles, there has not been much work on detecting Business-critical events from Business News. The major difference of business events from other events is that one is only interested in specific types of events that have impact on business intelligence, competitive intelligence and planning. Business events are also of interest to investors and financial planners who make their decisions based on factors like brand image, disclosures on governance and environmental issues etc. Typically, business events may be announcements rather than reporting.

In this paper, we present a system that can be trained to detect and classify business-critical events from business News. We propose a k nearest neighbors (k-NN) based method for classification of business-critical events. The uniqueness of the current work lies in the use of word vectors while computing the k nearest neighbors, thereby exploiting both surface-level syntactic and semantic similarities of sentences for classification. We also present an event-based News retrieval system that helps analyst retrieve News articles of specific interest.

Section 2 provides an overview of earlier work. Sections 3, 4 and 5 present the core contributions of this paper in terms of event classification and event-based news retrieval. A specific application scenario is presented in Sect. 6 for the supply chain process. Section 7 presents results from experiments with a large News collection.

2 Review of Related Work

Event extraction from large volumes of unstructured text like News collections has emerged as one of the popular sub-tasks of information extraction (IE). An overview paper on event extraction from text documents, categorized three predominant approaches to event extraction. Reference [1] Data-driven approaches, as adopted in [2–5] rely solely on quantitative methods to discover relations and thereby events from text. They require large text corpora to develop models that approximate linguistic phenomena. Knowledge-driven event-extraction on the other hand uses patterns derived from expert inputs based on linguistic and lexicographic knowledge. The patterns are either lexico-syntactic patterns [6, 7] or lexico-semantic patterns that are customized for domains like stock-market as proposed in [8]. Reference [9] reported a large-scale event extraction system named REES that extracted 61 pre-specified types of events under different categories like crime-events, business-events, financial-events, political-events etc. REES uses a declarative, lexicon-driven approach where each lexicon entry is defined for a specific type of event. Reference [10] proposed an ontology-based event extraction mechanism to extract violent and natural disaster events from online news, which were first clustered into similar groups.

Many functional approaches use a combination of the above methods to make best use of all the worlds. References [11–13] propose methods for extracting events and event templates or event schemas from large-scale text collections. While events represent a time-stamped single instance of an incident, an event schema is defined as a set of actors that play different roles in an event, such as the perpetrator, victim, and instrument in a bombing event. Reference [14] reported dynamic event discovery mechanisms based on discovery of relationships among co-bursting entities along with underlying global and local time constraints. Reference [15] proposed the use of

RelGrams which are combinations of Subject-Object-Predicates and statistical reasoning to identify event-schemas from large corpora. Reference [16] developed a probabilistic solution for template generation. The approach requires performing joint probability estimation using EM, which is not scalable to large corpora. Reference [17] worked on extracting significant events from webcast text on sports where significance of an event was computed based on people participating in reporting the event. Since an event is covered from multiple perspectives in different News articles, the focus of [18] was to generate the most compact, objective and informative headline for the event.

It may be observed that while there has been considerable work on event extraction and event schema identification, not much has been done in event-based retrieval or computing significance of an event towards analytical tasks. Consequently not much work has been done on categorizing business-critical events, their correlations and possible impact. In [20] a rule based approach was used to identify feature based sentiment and business event phrases from news documents. The relevance of news articles was decided on the basis of presence of event and sentiment polarity. This is closest to our work. However the proposed mechanisms are not rule-based, hence easily adaptable to different domains.

3 Event Extraction and Classification Architecture

In this section we present a brief overview of the proposed mechanism to extract and classify business events from News articles. The proposed system detects events of predefined types that can be used for particular business analysis tasks. The business context defines the type of events that are useful for the domain. The system is accordingly trained to extract business-critical events from News articles. An event occurrence is assumed to be detected from a single sentence within a News article. There may be multiple event occurrences of same or different classes within a single News article. Analysts can query the system for News articles that contain events of specific types associated to specific entities.

Figure 1 shows the proposed system architecture. News articles are collected from a set of pre-defined news websites using RSS feeds. Each news article is then subjected to sentence extraction, word stemming using Porter's algorithm[1] and then Named entity extraction. Named entities denote names of people, organizations, products, money or date values etc. The Stanford core NLP[2] suite has been used for extracting sentences and named-entities from the news content. All news articles along with the metadata element like source, author, date etc. and extracted sentences and named-entities are stored in a local data repository.

Each sentence is then subjected to event detection and classification. A sentence may be categorized into one of the known event classes or as "others" indicating that it does not contain any business-critical information for the given context. Subsequently

[1] http://snowball.tartarus.org/algorithms/porter/stemmer.html.

[2] http://nlp.stanford.edu/software/corenlp.shtml.

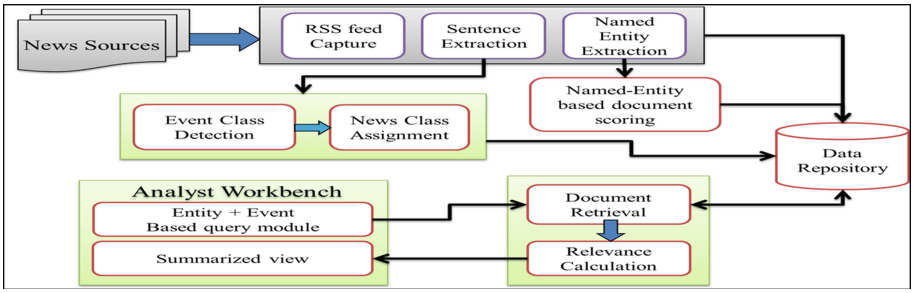

Fig. 1. Proposed event extraction and classification architecture

the event-class distribution within a News article is computed. The analyst's workbench is a search and analytics module which underlines the use of events for business analysis activities. An analyst can query the News collection for events related to specific entities. The named entities typically represent product, people or company names. The named-entities and event distributions are used for retrieving relevant News articles for a specific entity-event combination. It presents a list of articles to the user ranked by their relevance to the query. Analysts are also presented with summarized views of events extracted from the collection in association with pre-specified named entities.

4 Business Event Classification

In this section we present a k-nearest neighbor based classification algorithm for detecting and classifying business events within an article. The novelty of the proposed method lies in sentence representation and similarity computation mechanism within a classical k-NN framework. The system is trained to classify each sentence of an incoming News article based on its similarity to each classified sentence in a training set. The assigned classes of the top k nearest neighbors of the sentence are then used to determine the class of the sentence based on majority voting. Multiple values of k have been used for experiments.

4.1 Word-Vector Based Sentence Representation

We have used the *word2vec*[3] tool to compute continuous distributed representations of words. Each word is represented as a 500 dimensional numerical vector. This tool learns word representations that are not sparse and semantically similar or grammatically closer words are placed closer to each other. One billion word benchmark dataset [19] was used to obtain the word vectors. A window size of 10 words was chosen and words with lesser than 10 instances were ignored.

[3] https://code.google.com/p/word2vec/.

The proposed system first takes a text corpus as input and produces the word vectors as output. Continuous bag-of-words algorithm is used to construct a vocabulary from the training text data and then learn vector representation of words. The word-vector store is indexed for fast retrieval of a vector for a given word. Using the word-vector representation ensures that use of semantically similar words also does not affect event recognition and classification accuracy.

4.2 k-NN Based Event Classification

The word vectors generated in the earlier step are used for detecting events from News articles, by classifying each sentence into a known event class or a separate class called OTHERS. The vector representation of each word in an incoming sentence is retrieved from the word-vector store. Similarity between a pair of sentences is computed in terms of the word-vectors of the contained words.

For each word in a given sentence, its most similar neighbor in another sentence is located using the cosine similarity of word vectors. This ensures position of words do not affect accuracy of classification.

To find the closest neighbor of a word 'w' in another sentence S_i, the word vector of w is compared with the word vectors of each word in S_i using the cosine similarity measure. The word of S_i which yields the highest similarity with w, is accepted as its neighbor in S_i. It may be noted that the same word of S_i may be returned as neighbor for two different words.

The function Similarity(S_X, S_i) computes similarity of two sentences using the above measure. The function kNN(S_X) assigns the class label for a new sentence S_X. Finally, the two functions are used to compute the event distribution for each News article. Stop-words are ignored while computing similarity. Since words have been already stemmed so different morphological forms of the same word have the same representation and hence the same word vector. This reduces the size of the vocabulary.

Function Similarity(S_X, S_i)

1. Let W_x and W_i denote the sets of words in S_x and S_i respectively
2. For each $\omega_p \in W_x$, find semantically similar word $\omega'_p \in W_i$ as follows:
   (a) Let $\overrightarrow{V_{\omega_p}}$ denote the word-vector for word ω_p and $\overrightarrow{V_{\omega_j}}$ denote the word-vector for word ω_j where $\omega_j \in W_i$
   (b) $\omega'_p = argmax_{\omega_j}\left[cosine\_similarity\left(\overrightarrow{V_{\omega_p}}, \overrightarrow{V_{\omega_j}}\right)\right]$
3. Calculate $Similarity(S_X, S_i) = \frac{\sum_{\omega_p}\left[cosine\_similarity\left(\overrightarrow{V_{\omega_p}}, \overrightarrow{V_{\omega'_p}}\right)\right]}{|W_x|}$

Function kNN(S_X)

1. Let T = {Set of labeled sentences where the label denotes an event class}
2. For each sentence S_i in T
   (a) Calculate $Similarity(S_X, S_i)$

3. Select $T^{'} \subset T$ such that $T^{'}$ contains k most similar sentences for S_X based on [Similarity(S_X, S_i)]
4. Obtain labels L_k from sentences in $T^{'}$
5. Count occurrences for each label l_i in L_k
6. Assign class $l_i$to S_X if count $(l_i) > k/2$ else assign class "Others"

Event distribution for each News articles is computed as follows.

1. For each News article D construct an event vector $\overrightarrow{E_D}$ which is initially NULL.
   (a) For each sentence $S_i \in D$. Obtain event class C_i using Function $\boldsymbol{kNN}(\boldsymbol{S_i})$
   (b) For each event class C_i
      (i) Calculate $P(C_i/D) = \left(\frac{Frequency\ of\ C_i\ in\ D}{total\ no\ of\ sentences\ in\ D}\right)$
   (c) *Event vector* $\overrightarrow{E_D}[i] = P(C_i/D)$

It may be noted that the similarity measure used here is not commutative. Since the goal of the computation is to identify for a given sentence the maximally aligned sentence from the training class and thereby its assigned label, one-way matching of the given sentences to training data was needed.

5 Event-Based News Retrieval

As stated in Sect. 3, one of the uses of event classification can be searching of a news collection using implicit or explicit event queries. An event query is of the form (E_i, C_j) where E_i denotes a named-entity of interest and C_j denotes an event class. News articles retrieved from the local repository are ranked by relevance with respect to the above query, where the total relevance score is generated by combining the two scores defined below:

Named Entity Based Relevance Score - A document D may contain multiple entities of the same or different types with different frequencies. The relevance score presented here considers the relative importance of given entity E_i within the document. The first step is to identify entity type T of E_i using lookup tables built from the named entities extracted from existing document collection. Post that the entity is represented as E_i^T.

Let E denote any entity. Let E^T denote an entity of type T

$\sum_D E_i^T$ = Total number of occurrences of E_i^T in document D

$\sum_D E$ = Total number of occurrences of entities in document D

$\sum_D E^T$ = Total number of occurrences of entities of type T in document D

$$\text{Score } (D/E_i^T) = \left[\lambda\left(\frac{\sum_D E_i^T}{\sum_D E^T}\right) + (1-\lambda)\left(\frac{\sum_D E - \sum_D E_i^T}{\sum_D E}\right)\right]$$

The first factor determines the relative importance of the given entity with respect to other entities of the same type. It takes care that a random mention of an entity name in a document does not fetch a high score for it. For example, even when a News article is predominantly about a company's performance, its competitor names are also mentioned in it. This factor makes sure that this document gets high relevance for the primary company rather than its competitors. The second factor determines whether the document is majorly talking about the relevant type of entity. This factor helps in distinguishing among different types of entities that may be related to each other. For example, when the information required is about a product, though its manufacturing company name may occur in an article, it is expected that the product name will be more frequent than its company name. News articles with fewer mentions of the product and more about the company would be ranked lower than the earlier article. We found experimentally that $\lambda = 0.8$ yields the best results.

Event Based Relevance Score – Each news article's relevance to the queried event is computed using the event distribution vector. The score is represented as:

$$\text{Score } (\text{D}/C_j) = \overrightarrow{E_D}[j]$$

Total relevance score – The total score is a linear combination of the above scores:

$$\text{Score } (\text{D}/(E_i^T, C_j) = 0.5 * \text{ Score } (\text{D}/E_i^T) + 0.5 * \text{ Score } (\text{D}/C_j)$$

6 A Use-Case Scenario for Business Event Detection for Supply Chain Analysts

In this section we present a specific application scenario for business event detection in the supply chain industry. The supply chain industry is involved in logistics and planning for each stage of product manufacturing and delivery. Manufacturing itself is preceded by procurement of raw materials or supply of components. The three stages of supply chain are

1. Sourcing of raw material or components for production - termed as Source.
2. Making or manufacturing of products - termed as Make.
3. Delivery of products are termed as source - termed as Deliver.

The supply chain is impacted by multiple factors at different stages of Source, Make or Deliver processes. Following is a detailed description of different types of events that are considered critical to the domain. Table 1 shows examples of how they occur in News articles.

1. **Market News:** Any event related to sales directly influences the Deliver process of the supply chain. Events of this type report upward or downward trends in the focus company, competitors, or the industry as a whole.
2. **Production News:** News related to manufacturing facility expansions or contractions provide insights that could lead to short-term adjustments or long-term plans.

Table 1. Event examples extracted from news articles

Event class	Example	Source
Market news	Hyundai had record U.S. sales of more than 700,000 in 2012	Forbes
Production	Nissan is already building its third plant in Mexico, breaking ground in July on the $2 billion factory in the central state of Aguascalientes	Bloomberg
Supply disruption	'Toyota Motor faced another tough year in fiscal 2011 (ended March 31, 2012), hit by supply disruptions and production cuts following the Great East Japan Earthquake in March 2011'	Reuters
Launch	Hyundai has just launched the third generation of the Santa Fe which has sleeker lines and a very executive look	Irish Independent
People	Ford Motor Company today announced that Michael Boneham, president and managing director, Ford India has elected to retire effective December 31, 2012 after a successful career spanning over 27 years with Ford Motor Company	The Hindu
Product failure or recall	Ford this month recalled 16,000 Fusions for excessive engine temperatures that could lead to fire and an additional 19,000 Fusions for a defect with low-beam headlights	Reuters

These could include new equipment capabilities, labour requirements, lower cost manufacturing locations, and so on.

3. **Supply Disruption News:** Suppliers form the starting point of any supply chain. Any news regarding supply disruption related to a supplier or a region may have a ripple effect on entire business. Suppliers may supply to competing manufacturers, disruption events help manufacturers take tactical corrective action.
4. **Launch News:** Launching a new product has an indirect effect on the Deliver and Make processes. Tracking news about newly launched products will enable a business to be better prepared for positioning requisite manufacturing and warehouse capacity.
5. **People News:** Sometimes news related to people associated with a company has an impact on the demand for that company's products. Executive movements, new responsibilities, retirements, and death, can change customers' perception of a company and its products.
6. **Product Failure or Recall News:** Product failures and recall events have an indirect effect on all aspects of business. They can affect the demand for the recalled products, cast doubts about manufacturing quality, and supplier non-conformance.

Figure 2 illustrates the relationship among the different classes of events and the supply chain stage that they affect. These events are also classified as primary or secondary factors depending on their impact. Primary factors directly influence core supply chain processes, while secondary factors wield an indirect influence.

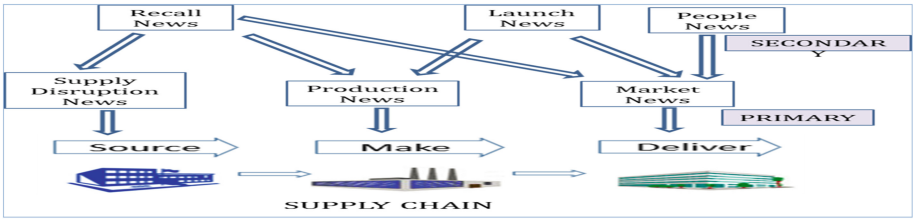

Fig. 2. Impact of events on Supply Chain – an example

7 Results for Supply Chain Use-Case

In this section we first present event detection and classification results for a collection of 30000 news articles published over a period of 15 months from Oct-2012 to Dec-2013 for major Automobile companies that trade in US. Training data for k-NN classification contained 600 manually tagged sentences for the 6 categories mentioned in Sect. 6. Table 2 shows 10-fold cross validation accuracies for different values of k from 3 to 10. It was observed that k = 7 yields best accuracy.

Table 2 10- fold cross validation k-NN classification accuracy for different k

k	Accuracy	k	Accuracy	k	Accuracy	k	Accuracy
3	0.88	5	0.89	7	0.91	9	0.89
4	0.88	6	0.88	8	0.88	10	0.86

The retrieval precision was manually verified for the most relevant article using 120 queries, generated as a combination of four entities across six event classes for five months. The precision yielded was 73.3 %. Figure 3 shows a sample screenshot of the most relevant News articles retrieved for event class 'Market News' for three automobile companies for January, 2013.

Fig. 3. Most relevant Market news for Honda, Toyota, Nissan and Hyundai (Jan,'13)

8 Conclusion

In this paper, we have presented a method for detecting and classifying business-critical events from Business News articles. It has been shown that use of word vectors can yield quite high accuracies for event classification. This paper also presents an event-based News retrieval mechanism which can be used by analysts. Rigorous experiments have been conducted with a large collection of News articles and examples from a real use case scenario have been shared. This work is being currently extended to generate automated alarms for supply chain analysts on detecting events of interest. Computing impact of events in conjunction with business data is also another area which is being explored.

References

1. Hogenboom, F., Frasincar, F., Kaymak, U., de Jong, F.: An overview of event extraction from text. In: Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2011) (2011)
2. Lei, Z., WU, L.-D., Zhang, Y., Liu, Y.-C.: A system for detecting and tracking internet news event. In: Ho, Y.-S., Kim, H.J. (eds.) PCM 2005. LNCS, vol. 3767, pp. 754–764. Springer, Heidelberg (2005)
3. Tanev, H., Piskorski, J., Atkinson, M.: Real-time news event extraction for global crisis monitoring. In: Kapetanios, E., Sugumaran, V., Spiliopoulou, M. (eds.) NLDB 2008. LNCS, vol. 5039, pp. 207–218. Springer, Heidelberg (2008)
4. Liu, M., Liu, Y., Xiang, L., Chen, X., Yang, Q.: Extracting key entities and significant events from online daily news. In: Fyfe, C., Kim, D., Lee, S.-Y., Yin, H. (eds.) IDEAL 2008. LNCS, vol. 5326, pp. 201–209. Springer, Heidelberg (2008)
5. Okamoto, M., Kikuchi, M.: Discovering volatile events in your neighborhood: local-area topic extraction from blog entries. In: Lee, G.G., Song, D., Lin, C.-Y., Aizawa, A., Kuriyama, K., Yoshioka, M., Sakai, T. (eds.) AIRS 2009. LNCS, vol. 5839, pp. 181–192. Springer, Heidelberg (2009)
6. Nishihara, Y., Sato, K., Sunayama, W.: Event extraction and visualization for obtaining personal experiences from blogs. In: Salvendy, G., Smith, M.J. (eds.) HCH 2009. LNCS, vol. 5618, pp. 315–324. Springer, Heidelberg (2009)
7. Hung, S.-H., Lin, C.-H., Hong, J.-S.: Web mining for event-based commonsense knowledge using lexico-syntactic pattern matching and semantic role labelling. Elsevier J. Expert Syst. Appl. **37**(1), 341–347 (2010)
8. Li, F., Sheng, H., Zhang, D.: Event pattern discovery from the stock market bulletin. In: Lange, S., Satoh, K., Smith, C.H. (eds.) DS 2002. LNCS, vol. 2534, pp. 310–3115. Springer, Heidelberg (2002)
9. Chinatsu, A., Ramos Santacruz, M.: REES: a large-scale relation and event extraction system. In: ANLC 2000 Proceedings of the Sixth Conference on Applied Natural Language Processing, pp. 76–83 (2000)
10. Piskorski, J., Tanev, H., Wennerberg, P.O.: Extracting violent events from on-line news for ontology population. In: Abramowicz, W. (ed.) BIS 2007. LNCS, vol. 4439, pp. 287–300. Springer, Heidelberg (2007)

11. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative event chains. In: Proceedings of ACL-08: HLT (2008)
12. Chambers, N., Jurafsky, D.: Unsupervised learning of narrative schemas and their participants. In: Proceedings of ACL (2009)
13. Chambers, N., Jurafsky, D.: A database of narrative schemas. In Proceedings of LREC (2010)
14. Chambers, N., Jurafsky, D.: Template-based information extraction without the templates. In: Proceedings of ACL (2011)
15. Balasubramanian, N., Soderland, S., Mausam., Etzioni, O.: Generating coherent event schemas at scale. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1721–1731 (2013)
16. Cheung, J., Poon, H., Vandervende, L.: Probabilistic frame induction. In: Proceedings of NAACL, HLT (2013)
17. Chen, C.-M., Chen, L.-H.: A novel approach for semantic event extraction from sports webcast text. In: Multimedia Tools and Applications, December 2012
18. Alfonseca, E., Pighin, D., Garrido, G.: HEADY: news headline abstraction through event pattern clustering. In: ACL 1, pp. 1243–1253. The Association for Computer Linguistics (2013)
19. Chelba, C., Mikolov, T., Schuster, M., Ge, Q., Brants, T., Koehn, P., Robinson, T.: One billion word benchmark for measuring progress in statistical language modeling (2013). arXiv preprint arXiv:1312.3005
20. Drury, B., Almeida, J.J.: Identification of fine grained feature based event and sentiment phrases from business news stories. In: Proceedings of the International Conference on Web Intelligence, Mining and Semantics, p. 27. ACM, May 2011