Rule Based Event Extraction System from Newswires and Social Media Text in Indian Languages (EventXtract-IL) for English and Hindi data

Anita Saroj, Rajesh kumar Munodtiya, and Sukomal Pal

Indian Institute of Technology (BHU),
Uttar Pradesh 221005, India
{anitas.rs.cse16,rajeshkm.rs.cse16,spal.cse}@iitbhu.ac.in

Abstract. Due to today's information overload, the user is particularly finding it difficult to access the right information through the World Wide Web. The situation becomes worse when this information is in multiple languages. In this paper we present a model for information extraction. Our model mainly works on the concept of speech tagging and named entity recognization. We represent each word with the POS tag and the entity identified for that term. We assume that the event exists in the first line of the document. If we do not find it in the first line, then we take the help of emotion analysis. If it has negative polarity, then it is associated with an unexpected event which has negative meaning. We use NLTK for emotion analysis.

1 Introduction

The overload of today's information, throws enormous difficulty to access the right information especially through the World Wide Web. The situation becomes worse when the information is in multiple languages. When a user accesses a document written in multiple languages and the user faces difficulty in finding facts, it is important to remove all information from the data except only the facts that match the user's interest. To extract various types of information from document pertaining to specific languages and domains, Information Extraction (IE) systems are primarily used [2, 3]. Existing IE techniques, however, sometimes, remove specific facts from docements that match the user's interest. Also, sometimes, IE techniques return keywords that are irrelevant to the user's interests. On the other hand, users can manually find more relevant information according to their domain of interest than a system can provide [4].

Most IE systems process texts in sequential phases (or "steps") like lexical and morphological processing, identifying and typing appropriate names, analyzing large syntactic components, final extraction of anaphora and co-referent, and relationships with domain-release events and lessons [2]. Information Extraction systems need to be easily optimized for new domains in order to increase the

suitability for end-user applications [4]. Rapid growth in information technology in the last two decades has increased the amount of information available online. There is a new style social media to share information. Social media is a platform of communication between people in which public share and exchange information in virtual communities and networks (like Facebook and Twitter) [6].

2 Related Work

In Event detection, Topic Detection and tracking are one of the major component of a broad initiative. Written and spoken news stories are primarily belong to the category of Topic Detection and tracking interest based broadcasting news [1]. Driven by the MUC contest, work on Information Extraction, and especially on named entity recognition (NE), mostly concentrated on narrow subdomain, like newspaper about terrorist attacks (MUC-3 and MUC-4), and report on air vehicle launch (MUC-7) [7, 8]. Process different types of documents without involving much tuning or type of document a system is required. To adjust manually or semi-automatically new domains and application has been successfully implemented in many exsiting IE system - but there has been some progress in dealing with the problem of strengthening the system to overcome this requirement [5].

Recent research in this area starts with the notion that statistical machine learning is the best way to solve information extraction problems. To find structured information with uncontrolled or semi-structured text are the primary objective of information extraction. [6].

3 System overview

Our model primarily works on the concepts of Part of Speech tagging and Named-entity recognition. We represented each word with POS tag and identified entity for that word. We assumed that the event exists in the first line of the document. If we do not find it in the first line, then we took help of sentiment analysis. If it has negative polarity, it was found to be associated with an unexpected event that has a negative sense. We used NLTK for sentiment analysis.

Most of the information is easily extracted by Named Entity Recognition such as Time-argument(Date), and Place-argument(Location). The Speed-argument and Casualty extraction are first tagged by “CD” and further distinguished with the help of NER tag.

4 Methdology

We developed a modular method for information extraction from the Indian language. The dataset provided by organisers is collected from various social

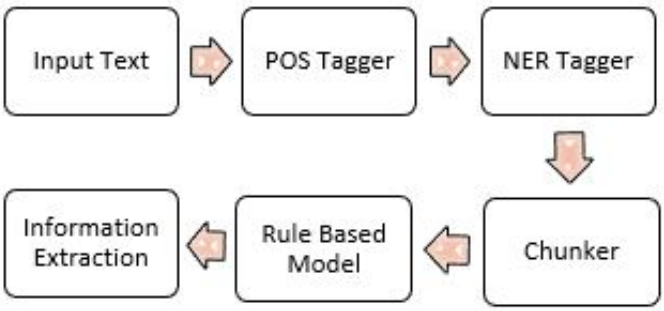

Fig. 1. Rule Based Information Extraction Framework

sources such as blogs, microblogs, social media and newswires that are either in Roman script or codemixed. Here, Indian language is mixed with English. We have worked on English and Hindi languages. The key components of proposed work (Figure 1) are described as follows.
Dataset the task organiser provide contains three languages: English, Hindi and Tamil. Training dataset is in the xml format and testing data in raw text. The statistics about given dataset are shown Table(1).

Table 1. Training dataset statistics respective to each language (Training dataset shows number of files)

Languages	Training Dataset
English	100
Hindi	107
Tamil	64

4.1 POS Tagging

We have used Stanford POS Tagger for the part of speech tagging to English dataset. We have used Hindi treebank dataset to make POS Tagger applying conditional random field algorithm. Our entire system relies on POS tagger. The part of speech tagging is the first step to perform extraction task.

4.2 Named Entity Recognization

Similar to POS Tagging, Stanford NER Tagger was used for named entity recognization to English dataset. For Hindi language, similar to POS Tagger we use Hindi treebank dataset to create NER Tagger. By default, the Stanford NER Tagger neither use part of speech nor gazette to extract locations. The pair of POS Tag and NER Tag along with word is helpful to extract information about Person, Date,..etc.

4.3 Chunking

When we want to extract full information related to the event such as casualties, depth,..etc. then it is necessary to identify the whole phrases. Thus, the complete phrase is extracted by Stanford Parser. Sentiment analysis the extracted phrase that has tagged as a verb followed by a cardinal number but not date NER. This may have information about casualties. This extracted pattern will fail when some positive events follow the casualties' pattern. To avoid such situation we check the sentiment polarity, the threshold value for sentiment checking is 0.5.

The proposed method is a framework for information extraction from unstructured user generated contents on social media. Our information extraction systems analyse human language text as linguistics structure in order to extract information about different types of events, time, place, casualties and speed. We Selected the sentences from the dataset and perform the POS Tagging, NER Tagging and Chunking then extracted the phrases from the above-POS, NER and Chunk.

- Time extraction: We assume that if we got date tag in NER tag then that is the correct date otherwise we match the day name or day abbreviation in calender library. Apart from that we also match the word with Today, Tomorrow and yesterday strings.

- Event extraction: We took the instance from the dataset and applied the lemmatizer on the selected instance and extracted the most frequent 10 lemmas that should not be the stopword and punctuation mark and for a selection of an event among those lemma that should be the noun and belongs from first sentence of instance.

- Place extraction: Those extracted phrases, who must contain the noun, proper noun and preposition POS tag from the dataset and the selected phrases that shows location in NER tag and those words should start with a capital letter then it represents the place.

- Casualties extraction: The phrase which we selected for the extraction must hold the cardinal number, verb, noun and preposition POS tag then only we will select that phrase. After the data analysis, we found the cardinal number and verb, the window size should be from 1-5 but there is some selected

phrase which represents time and the next we check the NER tag should not be Date tag.

– Speed extraction: for the selection of phrase, the phrase which contains the cardinal number, noun and preposition then we selected those phrase. Those phrases should not match with extracted Time and Casualties and the word which hold the cardinality number's word we check that to measurement unit for the remaining phrase. The measurement unit is created by a manual dictionary which keeps the almost measurement unit.

5 Result

The evaluation matrix for the Event Extraction problem is shown in table 2. The task organiser includes the Precision(P) and Recall(R), F-measure:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$F - measure = 2 * \frac{Precision(P)*Recall(R)}{Precision(P)+Recall(R)}$$

Where TP is truly positive, FP is false positive, FN is a false negative. We have extracted five information out of seven information i.e Speed, Casualties, Time, Event, Place.

Table 2. Result of our system submissions for Hindi and English data

Language	Submissions		
	Precison%	Recall%	F-measure%
Hindi	29.65	61.39	39.90
English	34.54	64.87	45.07

We clearly saw that our Hindi model is not performing well as compare to the English model because there is no free existence of Hindi POS Tagger and NER Tagger available yet.

6 Conclusion

We have discussed our Rule-Based methodology used to solve the task of information extraction from newswires and social media text in Indian languages. Our methodology tested on Hindi and English language and have derived some insights from the achieved results. The achieved F-measure are 39.98% and 45.07% for Hindi and English respectively. We believe that the incorporation of probabilistic approach with Rule-Based will improve the results.

References

1. Allan, J., Papka, R., Lavrenko, V.: On-line new event detection and tracking. In: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 37–45. ACM (1998)
2. Appelt, D.E.: Introduction to information extraction. Ai Communications 12(3), 161–172 (1999)
3. Cowie, J., Lehnert, W.: Information extraction. Communications of the ACM 39(1), 80–91 (1996)
4. Karkaletsis, V., Spyropoulos, C.D., Petasis, G.: Named entity recognition from greek texts: the gie project. In: Advances in Intelligent Systems, pp. 131–142. Springer (1999)
5. Maynard, D., Tablan, V., Ursu, C., Cunningham, H., Wilks, Y.: Named entity recognition from diverse text types. In: Recent Advances in Natural Language Processing 2001 Conference. pp. 257–274 (2001)
6. Morgan, M.B.H., Van Keulen, M.: Information extraction for social media. In: Proceedings of the Third Workshop on Semantic Web and Information Extraction. pp. 9–16 (2014)
7. Sundheim, B.: Proceedings of the fifth message understanding conference (muc-5). Columbia, MD: ARPA, Morgan Kaufmann (1995)
8. Sundheim, B.: Proceedings of the seventh message understanding conference (muc-7). ARPA, Morgan Kaufmann (1998)