Volume 5, Issue 10, October-2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Information Extraction from Text through Sequence Labelling

[1]Kavitha Raju, [2]Robert Jesuraj K, [3]Samaj Babu George, [4]P. C. Reghu Raj

[1, 4] Govt Engineering College, Sreekrishnapuram, Palakkad, Kerala

[2, 3] EY, Trivandrum, Kerala

***Abstract*— The way information is represented in natural language text is highly inconsistent, which makes it extremely complicated and practically impossible to explicitly tell a computer in which all ways a particular information can be represented in natural language texts, and enable it to read and understand it. The system presented here tries to bridge this gap by automatically extracting information from English text in web pages and representing them in a structured format so that it can be analysed efficiently.**

In order to implement the system, the task of information extraction is modelled as a sequence labelling one. It mainly uses the machine learning technique called Conditional Random Fields (CRF) to train the model and to tag the text. Further statistical and rule based techniques are also employed wherever appropriate to obtain the feature set. The system exhibits a precision of 91% and a recall of 77% that guaranties the validity of extraction techniques employed.

***Keywords*— Information extraction, machine learning, natural language processing, CRF, sequence labelling**

I. INTRODUCTION

The amount of information available to us in the web is large. A computer cannot automatically read through these numerous pages and extract the relevant information in them that are of interest to the problem at hand. If it could, then only those facts can be presented to us so that we can easily go through them, analyse and make decisions faster and more efficiently. The main barrier against this possibility is the inability of computers to read and understand natural language text. Almost all the information that the web holds is presented in web pages that are primarily intended for a human to read and understand. There is no one particular format defined for a natural language to represent a piece of information. The same thing could be stated in a number of ways, with different syntactic styles, different pragmatics, or even with different choices of words. Hence it becomes extremely complicated and practically impossible to explicitly tell a computer, in which all ways a particular information can be represented in natural language texts, and enable it to read and understand it.

The proposed system bridges this gap by automatically extracting information from English text in web pages and representing them in a structured format so that it can be analysed efficiently. It retrieves and scan the potentially relevant web documents, extract the predefined domain specific information in them, and store them in a uniform tabular pattern. This would make it possible for a human to get all the important information at one place, instead of him visiting numerous web-pages, reading through them, and writing down the information he finds in all those pages to a single document. This would make it possible for him to analyse the collected information in a more efficient manner either manually or using other data representation and analysis tools, since the data collected would be stored in a structured and machine readable format.

The news articles on the web is found to be a good choice for domain for the experiment. From them the details of various mergers and acquisitions that has taken place are extracted. That is, when one company buys another one or one dissolves into another. This information about the various mergers that happen around the world is of key importance in the business world. It enables us to keep a watch on major corporate powers and know what their current business and research focus are. It helps us to keep up-to-date with the new developments in technology, and change of trends. It also gives indications towards the distribution of territories and its expansions in the world market.

A combination of rule based and statistical approaches are found to be essential to realize the system. It is not feasible to extract all the required information in a purely rule based approach. It would require us to explicitly list all the possible patterns in which the information of interest occurs in English text, which is evidently too many. Hence the statistical techniques where the computer identifies and learns patterns from annotated corpora is more promising.

Sequence labelling is the task of assigning a single label to each token in a sequence. In NLP, according to the type of labels, the task would be called parts of speech tagging, chunking, named entity recognition etc. Our aim is to identify the pieces of information in the text. That too can be translated to a sequence labelling problem. The entities like merger or acquisition events, and details regarding them such as involved companies, date and value are all labelled accordingly whereas those words that do not constitute any entities of our interest are labelled as outside (or O). Other sequence labelling NLP tasks such as parts of speech tagging and named entity recognition are also employed in the system as part of creating the feature set for the information extraction task.

II. INFORMATION EXTRACTION FROM NATURAL LANGUAGE TEXT

Information Extraction (IE) is a technology that analyses natural language and extract snippets of information. It takes texts (and sometimes speech) as input and produces fixed-format, unambiguous data as output. The information encoded in natural language may seem unambiguous to a human reader. But when the same text is expected to be read and interpreted using a computer program, it may appear a lot more puzzling. It will be faced with a lot of issues such as word sense disambiguation, resolving structural ambiguities, etc.

Finding more efficient and effective interpretations of massive amounts of data is essential. Natural language text that is semi/un-structured should be converted to a more structured or organized format so that it can be effectively manipulated by computer programs. Specifically, most applications typically require representation that captures key events reported and the attributes of these events in a well-structured format. Thus information extraction primarily deals with text structuring.

A. Challenges

IE is often dealt as a domain-specific task; the important types of objects and events for one domain (e.g., people, companies, being hired and fired) can be quite different from those for another domain (e.g., medical science). IE for a domain can be broken down into three tasks [5] : one - determining what the important types of facts are for the domain, two - for each type of fact, determining the various ways in which it is expressed linguistically, and three - identifying instances of these expressions in text. Usually there is a fuzzy boundary between these tasks. The challenges each of these phases put forth, starting from the last are:

- Identifying instances of a linguistic expression: Instead of a word-by-word matching, the accurate analysis of the structure of sentences and entire discourses becomes necessary here. This structure exists on many levels: the structure of names; the grammatical structure of sentences; and co-reference structure across a discourse (and even across multiple discourses).
- Finding linguistic expressions of an event or relation: This is the problem of finding the myriad linguistic expressions of an event. It would include all the syntactic and semantic paraphrases of a given expression.
- Determining the domain specific facts/events: First the documents that might contain the relevant information regarding the domain are to be identified and then the required facts that need to be extracted, must be identified and clearly stated.

As the information extraction is normally dealt as a domain specific application, the steps and implementation strategies would vary according to the domain's nature and needs.

B. Proposed System

The information extraction system should collect news on company mergers and acquisitions. These would contain details like which company acquired the other, which was the company that merged, when the merger occurred etc. The information sought out are the company names involved in the merger, date of the merger and the overall monetary value of the merger.

First the system collects list of web-pages that contains news articles in web, about recent mergers and acquisitions. This is obtained by using the Google custom search API and querying for news on mergers and acquisitions. The search result would contain not only the links to the news articles, but short descriptions about them as well. Retrieving the actual web pages, and analysing the entire article to extract the required information is expected to be unnecessary and rather inefficient. Hence we analyse only the article headings. Headings can be regarded as precise summary statements about the article. It is bound to carry the important information discussed and tends to be less misleading.

Now the collected heading texts are subjected to natural language processing tasks. The aim is to feed it to a sequence labelling model so that the model would process the text and identify those information such as the companies involved, the date and value of the merger. For that we need to extract certain features of the text. These extracted features become the features that aid the model in predicting the labels. Labelling or tagging is done at word level and later aggregation is done to accommodate multi-word entities. The feature set used includes values like part-of-speech tag of the word, named entity labels, information regarding the initial capitalization of words and most importantly the context of the word. The context information is collected in terms of nearby words, their part-of-speech tag etc.

There is a simple, customizable, and open source implementation of Conditional Random Fields (CRFs) for labelling sequential data. It takes a feature set, which is defined by us according to our problem, and tag the tokens in accordance with the feature functions learned from the training corpus. The training is done a priori with the same set of features using real text drawn from the same domain that is manually annotated. The learned model will have feature function weights determined statistically for every combination of feature values.

III. OBSERVATIONS AND RESULTS

Precision and recall are the basic measures used in evaluating search strategies. They are used to measure how efficient a search system or an information retrieval system is. Our case, though an information extraction system, can be thought of as similar to a retrieval system. Where a retrieval system is evaluated on the basis of documents it has retrieved from the available collection, the extraction system can be evaluated in terms of the pieces of information it has extracted from the available set.

The precision and recall values are computed based on the roles identified from the collected news articles. The news articles collected by the Google search module of the system has been scanned manually to extract information of the type the system looks for. They were identified and recorded. This became the reference set for the evaluation. Thus the system has been evaluated against the human performance.

The evaluation is done only for the post web search part of the system. It is because, the web search has been implemented completely depending upon the Google search engine which is an external system. Precisely, the system is evaluated based on the accuracy in labelling the information snippets in the text. That is where, the intelligence is present in the system.

TABLE I SYSTEM EVALUATION

Precision Measure		
Identified Roles	**Correctly Identified**	**Precision**
169	154	91.12%

Identified roles refer to the total number of roles the system has identified from the news articles. It means the entities that the system labelled as a potential information. Correctly identified, refers to those that are correct among the identified roles. This measure reflects the correctness or relevance of the extracted information.

TABLE III SYSTEM EVALUATION

Recall Measure		
Total no. of Roles	**Correctly Identified**	**Precision**
201	154	76.62%

The total number of roles is obtained from the manually labelled output. The number of roles there actually is. This figure is similar in concept to the total number of relevant documents in the set. The correctly identified, is the same value described above. The ratio gives an insight on how much of the actually relevant information has been successfully extracted by the system.

IV. CONCLUSIONS

This paper proposes a statistical approach towards the extraction of domain specific information from natural language text by converting it into a sequence labelling problem. It described a feature set using which, we could satisfactorily identify and extract the relevant information contained in the news text. Rule based techniques were also applied in feature extraction where they were found to be more appropriate.

The feature set described here is a minimal one, just sufficient to provide a proof of concept. Adding more features would certainly increase the performance of the system.

ACKNOWLEDGMENT

This work was done as part of the internship program at EY, Trivandrum during my studies at Government Engineering College, Sreekrishnapuram. I thank the guides and advisors from both the institutions for their valuable support and guidance.

REFERENCES

[1] Douglas E. Appelt. Introduction to information extraction. *AI Commun.*, 12(3):161–172, August 1999. ISSN 0921-7126

[2] John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data, In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML ’01, pages 282–289, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-778-1.

[3] Xavier Carreras and Llu´ıs M`arquez. Introduction to the conll-2005 shared task: Semantic role labeling, In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, CONLL ’05, pages 152–164, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics.

[4] Hamish Cunningham, Information extraction, automatic, *Encyclopedia of language and linguistics*, pages 665–677, 2005.

[5] Ralph Grishman, Nlp: An information extraction perspective, In *Proceedings of the Recent Advances in Natural Language Processing*, RANLP 2005

[6] I.A. Bolshakov and A. Gelbukh, *Computational Linguistics: Models, Resources,Applications*, Ciencia de la computaci´on. Instituto Polit´ecnico Nacional, 2004. ISBN 9789703601479.

[7] Eugene Charniak. *Statistical Language Learning*. MIT Press, Cambridge, MA, USA, 1994. ISBN 0262032163.

[8] D. Jurafsky and J.H. Martin. *Speech and Language Processing: An Introduction toNatural Language Processing, Computational Linguistics, and Speech Recognition*, Prentice Hall series in artificial intelligence. Pearson Prentice Hall, 2009. ISBN9780131873216

[9] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist*, 19(2):313–330, June 1993. ISSN 0891-2017

[10] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.