Text Assisted Defence Information Extractor

Nishant Kumar [1], Shikha Suman[2], Anubhuti Khera[3], Kanika Agarwal[4]

[1]Scientist 'C', Head (Network Services Division), DESIDOC, DRDO, Delhi, India

[2]Student, M.Tech (C.S), Jayoti Vidyapeeth Women's University, Jaipur, India

[3]Student, MCA, Guru Gobind Singh Indraprastha University, Delhi, India

[4]Student, MCA, Guru Gobind Singh Indraprastha University, Delhi, India

ABSTRACT

The huge amount of news information available on the web or in a huge textual repository requires the use of Information Extraction (IE) techniques to filter the information needed by different category of users. Such techniques differ from a normal search. The main objectives of the Information Extraction System are: reduce the time spent in reading the articles included in the corpus, avert the redundancy and addition of various annotation lists. In this context, article recognition and extraction have become key points. In this paper we initially define the various processing resources as they are the base for an efficient Information Extraction System. Following this, an information extraction tool has been discussed, which is being developed for use of Defence Research & Development Organization using open source software and focuses on extracting defence related keywords from a pool of textual data. Preliminary result is very promising, although more refinement is under process.

KEYWORDS: *Annie Transducer, Annotation sets, GATE, Information Extraction, Jape Grammar, Processing Resources, Text Mining*

I. INTRODUCTION

Development and advancement in defence domain by researchers and scientists have led to large amount of text information in form of research papers, technical reports and other textual documents. It remains a tough task for the readers or researchers to find out the relevant information from such a huge information repository in textual format. This leads to the implementation of text mining techniques on such repositories. Text mining is the process to extract expressive information from an ocean of text data by use of variety of techniques. As 'Defence Scientific Information and Documentation Centre (DESIDOC)' is the centralized information centre of DRDO, it was facing similar challenges. Therefore, text mining techniques are being implemented to enable the users in tracing out the desired information with ease and relevance. As a part of the project, a text-assisted defence information extractor has been developed using open source software. The extractor helps in identifying subject related keywords or group of keywords from domain-specific document collections. Information Extraction is a process of transforming information from unstructured data sources (e.g. .pdf, .txt files etc) to structured data. For implementing the application, an open source text mining tool has been used which is General Architecture for Text Engineering (GATE). Gate is a portable framework written in java for developing and deploying software components that describe how to process human language.

II. COMPONENT OF INFORMATION EXTRACTOR

2.1 GAZETTEER [1]

A gazetteer is a directory which consists of information about places, people etc. The role of the gazetteer is to identify entity names in the text based on lists. For example, ANNIE gazetteer lists, used in GATE, are plain text files, with one entry per line. Gazetteer list is a file with a '.lst' extension that consists of one entry at each line. A gazetteer list typically includes named entities (names, institutions, etc.), entity components such as prefixes of locations etc. Each list has a major type, minor type and a language. It is preferable to specify the major type while creating the list, while specifying the minor type and language is optional. We can modify the gazetteer lists and even create new ones. The major and minor types and the entries in the lists can also be modified. The newly created list needs to be mentioned with its major and minor type in the '.def' file of the ANNIE's gazetteer. The transducer basically uses up the entries in the gazetteer to create new annotations.

2.2 Sentence splitter

A sentence splitter divides a spawn of text into sentences. A question mark and an exclamation mark are used to end a sentence. A period succeeded by an upper case letter ends a sentence. There are some exceptions such as the abbreviations in which periods are used, but this will not end a sentence. The open source ANNIE sentence splitter produces sentences as an outcome of splitting the entire text in the corpus. To run the POS (Parts of Speech)-Tagger we need to initially implement the sentence splitter, both collectively are domain and application independent. The gazetteer lists are used up by the splitter to differentiate sentence ending full stops from abbreviations in the text. It is very essential for the working of the POS Tagger.

2.3 Tokeniser

When a stream of text is broken into some basic meaningful elements like phrases, symbols or words, so that they become inputs for further processing in text mining, the process is called tokenization and the basic elements are called tokens. In general the token is a word but it is very difficult to define what exactly a "word" is. There are basically five types of tokens-word, number, symbol (+, =, etc), punctuation marks and finally the space and control tokens. The tokenization is based upon a set of rules: that is on the left hand side we have the pattern to be matched and on the right hand side we have the actions to be taken up. The different tokeniser available is Treebank Tokeniser, S-Expression, ANNIE Tokeniser etc.

2.4 POS Tagger (Parts of Speech Tagger) [2]

Tagging refers to the automatic assignment of descriptors to the given tokens. The descriptor is called tag. The tag may indicate one of the parts-of-speech, semantic information etc. So tagging is a kind of classification. The process of assigning one of the parts of speech to the given word is called Parts Of Speech tagging. It is commonly referred to as POS tagging. Parts of speech include nouns, verbs, adverbs, adjectives, pronouns, conjunction and their sub-categories. Taggers use several kinds of information such as dictionaries, lexicons, rules for identification of token. A modified version of one of the best known Tagger, the Brill Tagger is ANNIE's POS Tagger. It produces a POS Tag as an annotation on each word or symbol. Default lexicon rule set is being used up in which there is alternates lexicons for all upper-case and lower case corpora. For this first the tokeniser and splitter must be run.

2.5 Transducer

A transducer refers to a technical term in physics or engineering that converts one form of energy into another. In the present context, the transducer is one that converts text from one form to another. The task of the open source 'ANNIE Name Entity Transducer' is to find terms that suggest entities. The JAPE grammar is a very important part of the transducer. We can either use the default jape rules in the main. jape (<GATEhome> / plugins/ ANNIE/ resources/ NE/ main. jape) file or we can create new transducers. These hand crafted jape grammar rules define the patterns over the annotations. The jape grammar either makes use of the gazetteer or will match the word using the 'Token String' method.

2.6 JAPE Grammar

To support our application we use an agglomeration of pattern rules popularly known as the JAPE grammar. The grammar is divided into two segments that is left side and right side. The annotation pattern that contains regular expression operators (e.g. +,*) should be present on the left hand side of the rule. The statements used to manipulate the annotations and the actions to be taken are specified on the right hand side of the rule. Labels are used in the RHS to reference the annotations matched in the LHS.

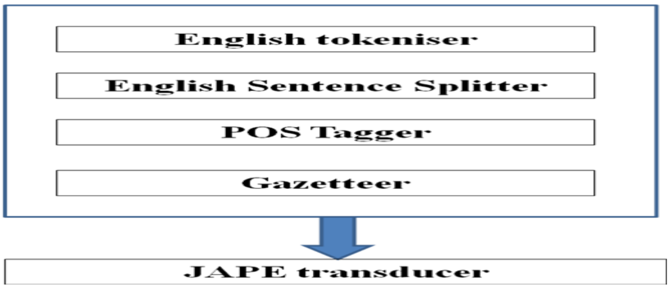


Fig 1.Components of Information Extractor

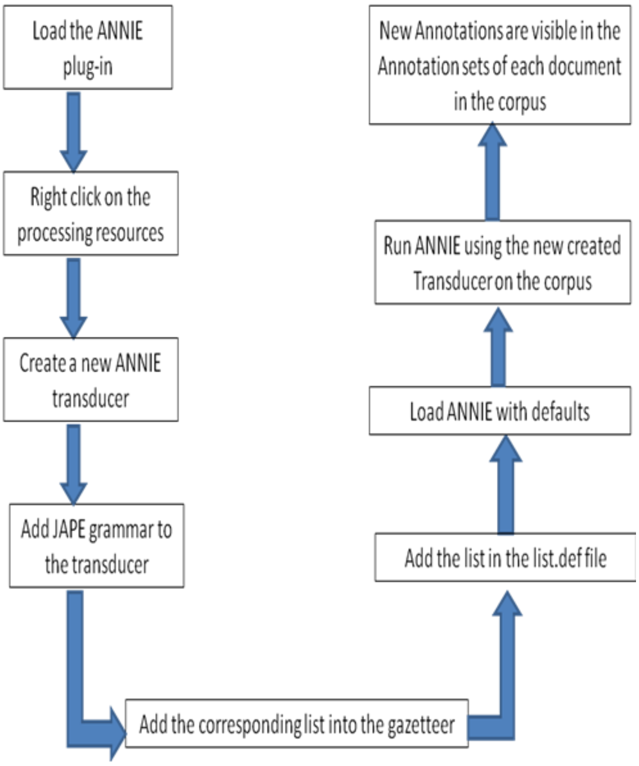

Load the ANNIE plug-in
Right click on the processing resources
Create a new ANNIE transducer
Add JAPE grammar to the transducer
Add the corresponding list into the gazetteer
Add the list in the list.def file
Load ANNIE with defaults
Run ANNIE using the new created Transducer on the corpus
New Annotations are visible in the Annotation sets of each document in the corpus

Fig 2. Steps to create a new mark-up

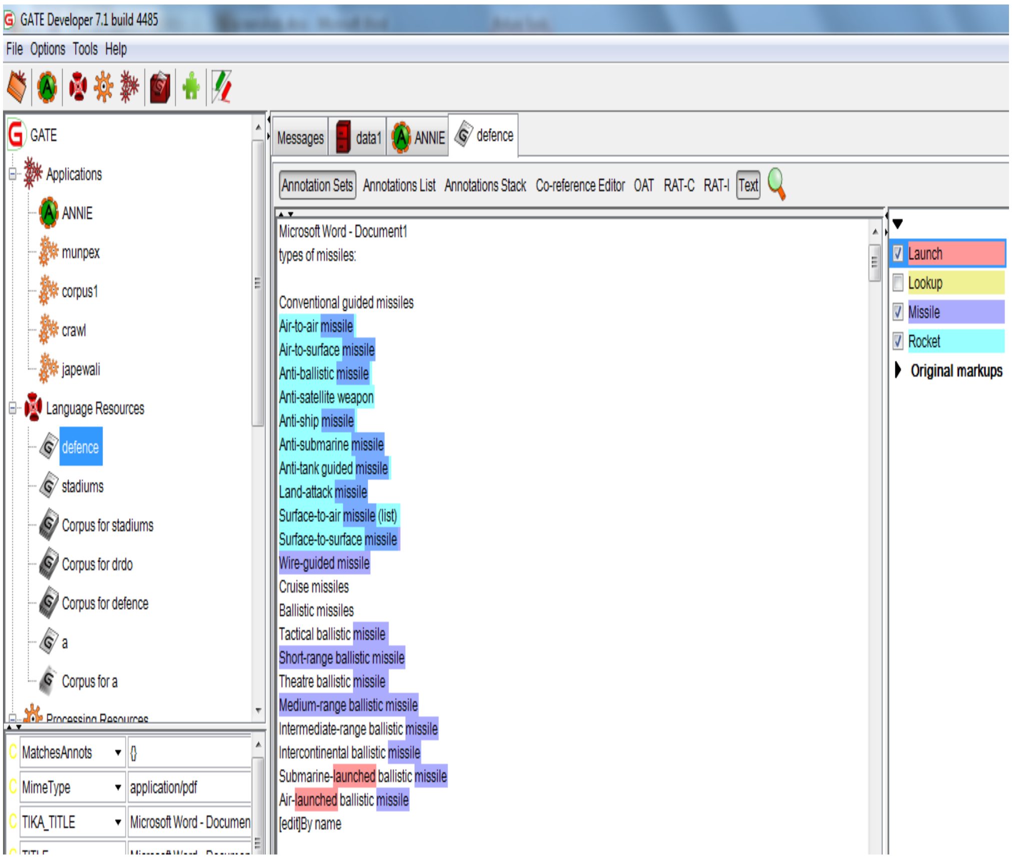
III. THE 'INFORMATION EXTRACTOR' FOR DESIDOC

As mentioned in section 1, DESIDOC has developed an information extractor, using open source software GATE, which helps in identifying subject related keywords or group of keywords from domain-specific document collections. Under this section, the methodology of this development has been explained in brief.

3.1Steps in Development of Information Extractor

Fig 2 depicts the protocol followed in development of the information extractor tool. Initially after starting the gate application, a new corpus is created and saved into a data store. Next, the newly created corpus is populated with text documents containing domain specific information. Next, the ANNIE plug-in is loaded to activate all its processing resources. A new annotation is formulated by creating a JAPE grammar based on our requirements. Now, JAPE grammar is incorporated into the newly created ANNIE transducer. Then, a new list is created in the ANNIE Gazetteer using the extension .lst. Next, the list is populated with the information about the newly created annotation. The file is listed in the .def file with its major type and minor type. Reinitialize the ANNIE Gazetteer so that the list is incorporated in it. Now load ANNIE with default, change the default transducer with the newly created one and run it on the corpus. At last for each document in the corpus the annotations are visible on the right hand side of the gate GUI.

Fig 3 presents the output of a domain specific extractor. The above mentioned idea is implemented through creating a corpus which contains the document about defence related warfare. The tree like structure on the left in the screenshot shows the Gate Application, Language Processing and Processing Resources. The central view displays the document from which the specific information is extracted. The right side tree structure depicts domain specific mark-ups. The ticked mark-ups highlight the desired keywords in the central portion of the screenshot.



MatchesAnnots	{}
MimeType	application/pdf
TIKA_TITLE	Microsoft Word - Documen

Microsoft Word - Document1
types of missiles:

Conventional guided missiles
Air-to-air missile
Air-to-surface missile
Anti-ballistic missile
Anti-satellite weapon
Anti-ship missile
Anti-submarine missile
Anti-tank guided missile
Land-attack missile
Surface-to-air missile (list)
Surface-to-surface missile
Wire-guided missile
Cruise missiles
Ballistic missiles
Tactical ballistic missile
Short-range ballistic missile
Theatre ballistic missile
Medium-range ballistic missile
Intermediate-range ballistic missile
Intercontinental ballistic missile
Submarine-launched ballistic missile
Air-launched ballistic missile
[edit]By name

Fig 3 Launch, Missile, Rocket Mark-up created at right hand side

IV. CONCLUSION AND FUTURE SCOPE

In this paper we present a method through which new Annie Transducer and Annotations could be created. The text assisted defence information extractor identifies desired keywords from the collection of domain specific documents. To improve the efficiency of the text assisted defence information extractor, a GUI is under development that will enable users to submit queries and search for desired information in the corpus and also a dictionary may be provided using 'WordNet' that will sense parts of speech which were earlier unsearchable.

Acknowledgement

The authors would like to thank The Director of Defence Scientific Information and Documentation Centre, Defence Research and Development Organisation, Metcalfe house, New Delhi, India.

REFERENCES

[1] http://gate.ac.uk/sale/tao/splitch6.html#x9-1330006.3

[2] http://language.worldofcomputing.net/tag/parts-of- speech

[3] http://www.kdnuggets.com/

[4] Raymond J. Mooney and Un Yong Nahm," Text Mining with Information Extraction", Multilingualism and Electronic Language Management*:* Proceedings of the 4th International MIDP Colloquium, September 2003,Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.) pp.141-160, Van Schaik Pub., South Africa, 2005

[5] Jonathan Clark,"Text Mining and Scholarly Publishing", presented at Publishing Research Consortium, Loosdrecht, The Netherlands, 2013.

[6] Fuchun Peng, Andrew McCallum, "Accurate Information Extraction from Research Papers using Conditional Random Fields"

[7] Seymore, A. McCallum, R. Rosenfeld. Learn-ing Hidden Markov Model Structure for Information Extraction. In Proceedings of AAAI'99 Workshop on Machine Learning for Information Extraction, 1999.

[8] Jian Ma, Wei Xu, Yong-hong Sun, Efraim Turban, Shouyang Wang, and Ou Liu," An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection", IEEE Transactions On Systems, Man, And Cybernetics—Part A: Systems And Humans, Vol. 42, No. 3, May 2012

[9] http://www.kdnuggets.com/software/text.html

[10] Lipika Dey, Muhammad Abulaish, Jahiruddin and Gaurav Sharma," Text Mining through Entity-Relationship Based Information Extraction", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology – Workshops, 2007.

[11] Shaidah Jusoh and Hejab M. Alfawareh,“ Techniques, Applications and Challenging Issue in Text Mining”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2 pp.1694-0814, November 2012 ISSN (Online):

[12] R. Hale, “Text mining: Getting more value from literature resources,” Drug Discovery Today, Vol. 10, No. 6, pp. 377–379, 2005.

[13] Tianxia Gong, Chew Lim Tan, Tze Yun Leong,” Text Mining in Radiology Reports ”, Eighth IEEE International Conference on Data Mining, 2008.

[14] Dhaval Thakker, Taha Osman, Phil Lakin,” GATE JAPE Grammar Tutorial”, Version 1.0**,** February 27, 2009

[15] Vishal Gupta, Gurpreet S. Lehal,”A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009

[16] A.V. Krishna Prasad, Dr. S. Ramakrishna, Dr. D. Sravan Kumar, Dr. B. Padmaja Rani, ”Extraction of Radiology Reports using Text mining”, International Journal on Computer Science and Engineering, Vol. 02, No. 05, pp. 1558-1562, 2010.

[17] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, Kalina Bontcheva, Valentin Tablan, Niraj Aswani, Ian Roberts, Genevieve Gorrell, Adam Funk, Angus RobertsDanica Damljanovic, Thomas Heitz,Mark A. Greenwood,Horacio Saggion, Johann Petrak,Yaoyong Li, Wim Peters , Developing Language Processing Components with GATE Version 7 (a User Guide), GATE version 7.2-snapshot, January 3, 2013.www. gate.ac.uk

[18] Ning Zhong, Yuefeng Li, Sheng-Tang Wu,” Effective Pattern Discovery for Text Mining” , IEEE transactions on knowledge and data engineering, Vol. 24, No. 1, January 2012