World Applied Sciences Journal 32 (3): 399-402, 2014
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2014.32.03.762

Survey on Various Natural Language Processing Toolkits

L.B. Krithika and Kalyana Vasanth Akondi

School of Information Technology and Engineering,
VIT University, Vellore-632014, Tamil Nadu, India

Abstract: Natural language processing (NLP) is a sub-area of artificial intelligence that deals with human interaction with the machine. There are many tools available for natural language processing in many platforms. Each toolkit can be used with a different programming language. So users can choose different toolkits to work on NLP depending on their familiarity with a particular programming language. Most of the NLP toolkits like NLTK, Apertium, Stanford NLP etc., are open source and are easy to use. The context of many websites in the present world depends on the natural language processing. Hence, natural language processing has become an important field in computer science and hence these toolkits gained attention. This paper gives a survey about various Natural language processing toolkits and highlights the unique feature of every toolkit.

Key words: Natural Language Processing · Toolkits · Open Source · Platforms

INTRODUCTION

Natural language processing (NLP) has been noticed very often in the field of computer science in the past few decades. Enormous attempts have also been made in the field of NLP to convert the natural language processing text into computer programs [1]. This conversion system has been so wonderful that the NLP identifies conditionals, loops and other aspects in the procedural programming. It then builds the program from natural language text [1].

To achieve wonderful tasks like the one mentioned above, there is a need of using one or more NLP toolkits. Earlier, the widely used NLP toolkit was GATE [2].It was designed and released in 1995 and is still being developed. There are lot of open source toolkits available like Stanford NLP, NLTK, Apertium. Most toolkits have been designed for English language, but there are also toolkits that have been designed for other languages like 'Apertium', 'IceNLP' etc., (Apertium is designed for Spanish texts and IceNLPis designed for Icelandic texts). NLP toolkits have gained popularity because of the increase in interest in the domain of Natural Language processing. In general, any Natural Language Processing toolkit consists of a tokenizer, lexica, sentence splitter and a parts of speech tagger. A tokenizer is the one which segments natural text as sentences, words, numeric etc. [3]. A lexicon is the one which establishes or identifies language patterns or finds meaningful relation among the words [3]. Sentence splitter splits the sentences as required and a parts of speech tagger mark parts of speech like noun, preposition or verb in a particular sentence.

Most of the toolkits are released under General Public License (GPL) which is a free license which allows the linking with retail software and for further distribution under any terms [4]. There are also toolkits developed for complex languages like Chinese. Georgetown experiment is a toolkit designed for Russian texts. Fundan NLP is a toolkit for Chinese language. It has many complex algorithms involved and has parts of speech tagger like other NLP toolkits [5].

Need for Toolkits: Online documents contain huge amount of information.To organize such information in a better manner, researchers are actively involved in automatic text generation. The research is focussed on sorting documents in bulk according to the category to which they belong like sports, movies etc. In the recent years, there have been online discussion in many sites

Corrsponding Auther: L.B. Krithika, School of Information Technology and Engineering,
VIT University, Vellore-632014, Tamil Nadu, India.

and these discussions will be with respect to a subject. Hence, the texts in this discussion are sorted as positive, negative or neutral depending on the sentiment they hold. This is called sentimental analysis or opinion mining [6]. All the above mentioned problems of the present day require a lot of natural language processing.

Each Natural Language Processing toolkit has a unique feature. And some developers may find a particular tool easy to use and some may find it difficult. Hence, developers use the toolkit depending on their need and also depending on their comfort with a particular programming language.

Some nations where English is less spoken need Natural Language Processing toolkits for their regional language. Hence, Natural Language Processing toolkits have been developed for many languages apart from English. And these toolkits have gained an enormous popularity. Apertium particularly has been designed for Spanish language. It has all superb features of other toolkits like POS tagging, tokenizer, sentence splitter [4]. Ice NLP on the other hand is used to process Icelandic texts. This toolkit is for regional Iceland language. Hence processing of other language texts has become very easy because of these toolkits.

Due to the increase in non-English document on the World Wide Web, there has been a need to develop CLIR (Cross Language Information Retrieval) systems. Dictionary translation is a preferred method of translation because of its simplicity and availability of machine readable bilingual dictionaries [7]. Translation from one language to another is possible only with the interrelation of one or more Natural Language processing toolkits. Hence, natural language toolkits are very important while considering translation into the account.

Types of Toolkits: In this section we will look at various toolkits using the terminologies and concepts described in the above sections.

Icenlp Toolkit: It is an open source toolkit for processing Icelandic texts (Regional language of Iceland) [8]. This toolkit consists of the following modules-Pre-processor, POS tagger and Finite state parser [9]. Word error rate is a measure of a performance of a machine translation system.Word Error rate of Apertium and IceNLP is 50.6% [8].

Apertium Toolkit: The Apertium is aimed at Spanish language. It has also been aimed at other languages such as Danish, French and Italian. It consists of the following modules-morphological analyser, Part-of-Speech (PoS) tagger, Lexical selection, Lexical transfer, Structural transfer, morphological generator [8].

Nltk Toolkit: The Natural Language Toolkit (NLTK), consists of ready-to-use linguistics. NLTK covers statistical NLP [10]. It is written in Python. It is currently the most preferred toolkit. It has many in-built corpora like movie reviews corpora, chat corpora etc., to test with the toolkit. It is easy to install because installation is through a simple executable file. It has many inbuilt classifiers like Naïve Bayes classifier, Maximum entropy classifier and binary tree classifier. These classifiers are easy to use as the commands to interact with them are simple.

Carabao Language Toolkit: Carabaocaptures any form of natural language entities, irrespective of spaces or complex morphology. The sequences in Carabao are viewed as regular expressions. Entity extraction in this toolkit is not limited to a fixed set. Setting up the conditions is so easy. LinguaSys (an NLP toolkit) user used Carabaoforwriting military orders in XML [11].

Ellogon: Ellogonis designed to produce language engineering systems for the End user. It has a powerful TIPSTER infrastructure. TIPSTER is the name of a text program project started in 1991 by Defence Advanced Research Projects Agency. It supported research to improve information retrieval and extraction. Ellogon's key feature gives full unicode support. It has a multilingual GUI (Graphical User Interface). It's hardware requirements are very low [12].

Monty Lingua: It was developed in MIT media labs using Python language. It covers all aspects of English text processing from raw input to summary generation. Each component of Monty Lingua is loosely coupled with each other in code and architectural level [13]. Monty Tagger (Monty Lingua Parts of Speech Tagger) is used in several contexts such as web pages. Monty Tagger has become a popular part of speech tagger [13].

PSI-Toolkit: PSI-Toolkit [4] is a tool which is also known as "Processors". It is used to process natural language processing texts. *Readers, annotators and writers* are the 3 types of processors.

- A reader creates the main data structure.
- Annotator, also known as tokenizer or parser, adds new annotations. Example: He is 6 feet tall and weighs 200 pounds. From this sentence extracting height and weight is also known as annotating height and weight.
- Writer writes to the output device (screen).

GATE (General Architecture for Text Engineering): GATE is an open source NLP toolkit. It is written in Java. So it can be used with Java applications for performing NLP tasks.
GATE has 3 main elements:

- A database for storing texts and a database schema.
- A GUI (Graphical User Interface) for viewing and processing data.
- A collection of wrappers.

GATE presents developers an environment where they can develop tools and databases with the combination of taggers or parsers [14].

Toolkits in Real-Life Contexts

Apertium in Real-life Context: Apertium is used to publish online multi-lingual versions of "La Voz de Galicia", which is a Spanish daily newspaper. It is used to generate book reviews in various languages on "Casadellibro.com", which is an online bookshop. It is also used to translate the "University of Alicante" website into Spanish and Catalan. Apertium is used by "Autodesk" which is a leading industry in the production of 3D animation software like 3DS max and Maya. Autodesk uses Apertium to translate Spanish text into Portuguese text [15].

NLTK in Real-Life Context: NLTK was developed at the "University of Pennsylvania" in combination with a Computational Linguistics course. NLTK is preferred for working on Computational Linguistics for students. This is because it is easier to perform Computational Linguistics in NLTK compared to other toolkits. Computational Linguistics is an interdisciplinary field that deals with the modelling of human language. In simple words it means, studying the language from computer's perspective. Computational Linguistics is used in many systems that are popular at present. Some of such systems are voice recognition systems, automatic voice response systems, text-to-speech converters and text editors [16].

Stanford NLP in Real-Life Context: A tool named Static UML Model Generator from Analysis of Requirements (SUGAR), generates use-case and class diagrams on analysing requirements in the form of natural language. SUGAR identifies actors, use-cases, classes, attributes and also proper associations between the classes. This tool has been implemented using the

Stanford NLP toolkit: Modified Rational Unified Process approach was used in designing SUGAR for better accuracy. This tool generates UML diagrams without the need of human interaction. Human interaction is only required to eliminate unnecessary use-cases and classes that are generated in the model [17].

Monty Lingua in Real-Life Context: A method to generate slide presentations from English text documents has been proposed using Monty Lingua toolkit. Various NLP methods like text segmentation, chunking and summarization are combined with special features like text ontology. This is done in order to build an "information extractor" that extracts information from documents and a "slide generator" that puts these extracted texts into slide presentations. Ontology means relation between a set of concepts in a specific domain. Monty Lingua helps in chunking and creating an ontology [18].

Table 1: Tabulation of Nlp Toolkits

Toolkit	Year	Language	Description
GATE (General Architecture for Text Engineering)	1995	Java	GATE has an information extraction system called as ANNIE which consists of sentence splitter, Parts of speech tagger etc.,
NLTK	2001	Python	NLTK is used to build Python programs to work with human natural language data.
MontyLingua	2004	Python and Java	MontyTagger is a MontyLingua's parts of speech tagger. MontyLingua has gained popularity only because of its POS (parts of speech tagger).
Ellogon	2004	C and C++	A power NLP tool with TIPSTER structure
IceNLP	2007	Java	It is an open source toolkit for processing Icelandic(language) texts.
Apertium	2009	C++	Apertium is a rule-based machine translation platform for Spanish language.
PSI toolkit	2011	C++	It consists of 3 processors to handle natural language processing.
Carabao Language toolkit	2004	Visual C++ and Visual Basic	Carabao captures any form of natural language entities, irrespective of spaces or complex morphology.

CONCLUSION

This paper provides an insight of different types of NLP toolkits. Developers can process natural language texts in a systematic way.Toolkits allow them to solve many problems like opinion mining, text summarization etc. It also helps to process texts of other languages. A developer who identifies a unique feature in a particular toolkit may not be able to use it because he may not be familiar with that particular programming language. The main advantage of toolkits is that most of them like Apertium, NLTK and Stanford NLP are open source toolkits. Almost all toolkits are available for free of cost. NLTK for python developers and Stanford NLP for Java developers are the most preferred toolkits as these offer many features of NLP like tokenization, sentence splitting, POS tagging etc., Some toolkits like Apertium, Ice NLP etc., are used in integration with NLTK and Stanford NLP to perform translation tasks. Hence NLP toolkits are used to perform complex natural language tasks like sentiment extraction, translation, spell checking, word prediction etc.

REFERENCES

1. Rada Mihalcea, Hugo Liu and Henry Lieberman, 2006. NLP (Natural Language Processing) for NLP (Natural Language Programming), In Proceedings of 7th international conference on Computational Linguistics and Intelligent Text Processing, pp: 319-330.
2. Cunningham, H., 2011. Text Processing with GATE (Version 6), University of Sheffield Department of Computer Science, ISBN 0956599311,
3. Jonathan J. and Webster Chunyu Kit, 1992. Tokenizationasthe Initial Phase in NLP, In Proceedings of conference on Computational linguistics, 4: 1106-1110,
4. Filip Grali´nski, 2013. Krzyszt of Jassem, Marcin Junczys-Dowmunt, PSI Toolkit, Computational Linguistics Studies in Computational Intelligence, 58: 27-39.
5. FengJi, WenjunGao, XipengQiu and Xuanjing Huang, 2009. FudanNLP: A Toolkit for Chinese Natural Language Processing with Online Learning Algorithms.
6. Bo Pangand, Lillian Lee and Shivakumar Vaithyanathan, 2002. Thumbs up, Sentiment Classication using Machine Learning Techniques, In Proceedings of EMNLP, pp: 79-86.
7. Jianfeng Gao, Jian-Yun Nie, Endong Xun, Jian Zhang, Ming Zhou and Changning Huang, 2001. Improving Query Translation for Cross-Language Information Retrieval using Statistical Models, In Proceedings of annual international ACM SIGIR conference on Research and development in information retrieval, pp: 96-104,
8. Brandt, M.D., H. Loftsson, H. Sigurþórsson and F.M. Tyers, 2011. Apertium-icenlp: A rule-based icelandic to english machine translation system, In Proceedings of Annual Conference of the European Association of Machine Translation, pp: 217-224.
9. Hrafn Loftsson and EiríkurRögnvaldsson, 2007. Ice NLP: A Natural Language Processing Toolkit for Icelandic, In Proceedings of InterSpeech, Special session: Speech and language technology for less-resourced languages". Antwerp, Belgium.
10. Edward Loper and Steven Bird, 2009. NLTK: The Natural Language Toolkit.
11. Vadim Berman, 2012. Text Analytics with Carabao Language Kit: Linguistic Business Intelligence, Linguasys.
12. Georgios Petasis, Vangelis Karkaletsis, Georgios Paliouras, Ion Androutsopoulos and Constantine D. Spyropoulos, 2002. Ellogon: A New Text Engineering Platform.
13. Ling, Maurice H.T., 2006. An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor. The Python Papers, 1(1): 5-13.
14. Hamish Cunningham, Yorick Wilks and Robert J. Gaizauskas, 1995. GATE - a General Architecture for Text Engineering", In Proceedings of 16th International Conference on Computational Linguistics: ACL.
15. Luis Villarejo, Mireia Farr´us, Sergio Ortiz and Gema Ram´ýrez, 2010. A web-based translation service at the UOC based on Apertium, In Proceedings of the International Multiconference on Computer Science and Information Technology, pp: 525-530.
16. Mykhailo Lobur, Riy Romanyuk and Mariana Romanyshyn, 2011. Using NLTK for educational and scientific purposes, Polyana-Svalyava (Zakarpattya), UKRAINE.
17. Deeptimahanti Deva Kumar and Ratna Sanyal, 2009. Static UML Model Generator from Analysis of Requirements (SUGAR), In Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing, pp: 147-163,
18. Gokul K. Prasad, Harish Mathivanan and T.V. Geetha, 2009. Document Summarization and Information Extraction for Generation of Presentation Slides, In Proceedings of International Conference on Advances in Recent Technologies in Communication and Computing.