International Journal of Soft Computing 1 (1) : 67-70, 2006

Towards Information Extraction System Based Arabic Language

[1]Mamoun, S.A.R., [2]M.B. Khaldoon and [3]H.Y. Jabar
[1]Faculty of Information Technology Al Al-Bayt University Al Mafraq-Jordan
[2]Philadelphia University
[3]Faculty of Information Science and Technology University Kebangsaan
Malaysia 43600 UKM Bangi, Selangor D.E., Malaysia

Abstract: There are vast develop in Natural Language Processing(NLP) tools such as Part Of Speech (POS) and morphology analysis, but the IE system for an Arabic texts are still not rising to comply with such progress and it is not available till now. The paper clarifies the design and implementation of a suitable information extraction system that automates input feeding of an Arabic text to derive an output template as a convenient extraction of important information. An Arabic POS tagger based on multilayer perceptron neural network (MLP) is used to achieve high accuracy and scouring for the tokenization stage. In addition, an Arabic semantic parser is suggested and addressed. The system consists of two stages, the first one is preprocessing stage, which has used to implement and model the document analysis and tokenization. The other stage is the processing stage, which has used to accomplish the morphology analysis, Pos tagger, semantic tagger and name entity extraction.

Key words: Arabic text analysis, information extraction, feature extraction, NPL, AI

INTRODUCTION

The speedy growth in information technology and great explosion of multilingual resources makes the internet and word wide web (WWW), the important source to obtain and transfer the required information. On the other side, the Arabic language is one of the weightiest languages in the world, because is the mother tongue of 300 million people, therefore, the Arabic language processing has recently becomes a focus of research and commercial development[1-3].

In general, the information extraction systems are effectively skim the text from the domain, find relevant sections and then focus only on those sections in the subsequent processing. This information is then structured and provided as output. The basic components of IE systems are (The tokenization and Tagging stage, The sentence analysis stage, extraction stage, the merging stage and Template Generation stage) which are depicted in Fig. 1,[4-9] Several systems have been developed to generate domain-specific extraction patterns automatically or semi-automatically[10,11]. There have also been efforts to automate various aspects of discourse processing[8,12]. RAPIER[13] uses inductive logic programming techniques to discover rules for extracting fields from documents. BWI[6] learns a large number of simple wrapper patterns, and combines them using boosting. LP2[5] learns symbolic rules for identifying start and end tags. EliE[14] use support

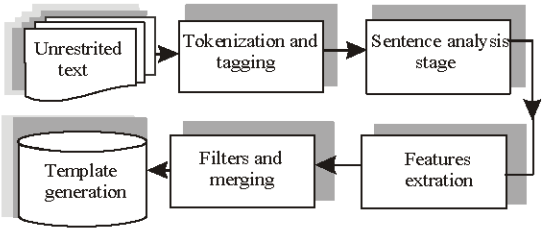

Fig. 1: IE system architecture

vector machines and several different feature-sets to build a set of classifiers for information extraction.

MOTIVATION

Arabic overview: Arabic, one of the six official languages of the United Nations[15-17]. Arabic writing differs from other languages like English and Spanish. Although there are common features that characterize such writings[18-20]. These features can be summarized by the followings:

- Writing is directed from Right to Left
- Character used are 28 in addition to special one called (ء) like in (حواء–hawaa- wich mean Eve).
- Characters within their words are connected excluding some of them which are connected to their preceding ones. Those characters are: (Alif ا ,Da د, Dah ذ, Ra ر, Za ز, Waو).

Corresponaing Author: Mamoun, S.A.R., Faculty of Information Technology Al Al-Bayt University Al Mafraq-Jordan

- Character may have more than one shape in accordance to its location in its involving word. This might be in the starting, middle or at the end of the word.

Word analysis: The Arabic word is described as follows: [pre1][pre2] [stem] [suf1] [suf2] [suf3]

The word between the square brackets it is optional choice it may be appeared or absent. The word (stem) is any combination of letters that give a useful meaning. See the following example: the word

(س----------ي------ن----------ص----------نول----------مه)
seyinasroonahim , which mean " they will supporting them")

(س----------ي----ن----------ص----------نول----------مه)					
[suf3]	[suf2]	[suf1]	[stem]	[pre2]	[pre1]

Word division: The Arabic word can be divided into three types, Noun, Verb and Particle.

Noun: is a word that describes a person, thing, or idea, such as Zahra(زهرة) which means flower, madrasa(مدرسة) which means school, Ali(علي) is person name.

VERB independently means something and points to a tense e.g. (past ضامي, present حاضر, imperative املا).

Imperative	present	past
Edriss ادرس	yadress يدرس	darasa درس
You must study	He is studying	He was study

Particle means something only when preceded or succeeded by a Noun or a Verb. They could serve certain purposes like confirmation Prepositions, Adverbs, Conjunctions, and Interjections.

Preposition	Adverse	Interjections	Conjunctions
بين between	فقط only	حي to perform	ثم then
بعد After	اجل yes	اضا to nag	و and
تحت under	جدا very	اها to moan	اذا since

SYSTEM ARCHITECTURE

Overview: There has been an increasing interest in developing information extraction software's for Arabic text because of the huge available of Arabic information, newspapers and documents on internet which can be searched , browsed and summarized .The overall objective of the presented work is to setup a global input/output expression system. This system automates input feeding of an Arabic text to derive an output template as a convenient extraction of important information to be formulated and classified into a special template.

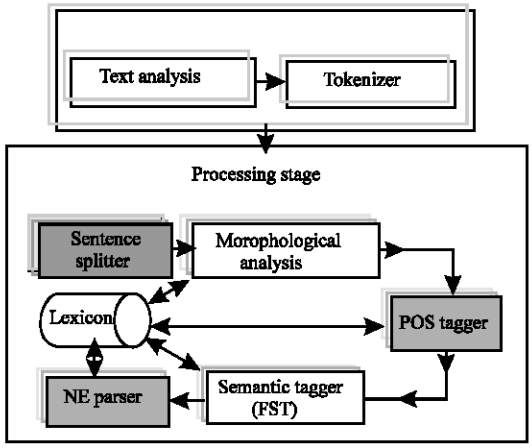

Fig. 2: System architecture

System design: The over all system components are depicted in Fig. 2.

The proposed system is involved the following constituents:

A pre-processing stage: Is implemented and utilized to achieve the following tasks:

Text analysis: Which is used for a document format detection (HTML , DOC , RTF , XML , Email ,… , etc). And The Tokenizer is used to split the text into simple tokens, such as numbers, punctuation, symbols and words of different types.

The processing stage: Is the main important mission of proposed work which it is consists of the following phases: The sentence splitter which is define as the task of assigning the exact roles to individual words, taking into consideration the grammatical rules of a language and a description (grammar) of how words can be put together in that language. Morphology Analysis which is mean the study of the way words are built up from smaller meaning-bearing units called morphemes (stem), Lexicon is a list of stems and their affixes (prefixes, suffixes, infixes), The Semantic Tagger that consists of set of grammars which used to annotations previously generated to recognize NE's. The grammars contain the rules from POS tagger and Lexicon, and combine them to produce new NE's annotations and Name Entity Parsing of Arabic sentences which is consider as a difficult task. This difficulty comes from several sources. One is that sentences are long and complex. The average length

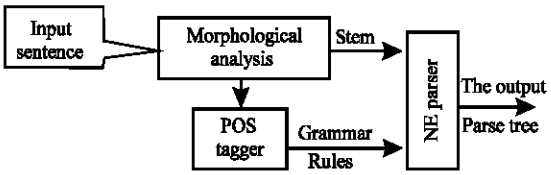

Fig. 3: The parsing architecture

of an Arabic sentence is 20 to 30 and maybe exceed to 100 words. Second, the Arabic sentence is syntactically ambiguous due to the frequent usage of grammatical relations, order of words and phrases, conjunctions and other constructions. The proposed parser is depicted in Fig. 3.

RESULTS AND DISCUSSION

The input to our system is XML file or text file. The text is processed by preprocessing stage, which is identified the type of document and then split the text into simple tokens, such as numbers, punctuation, symbols, and words of different types. The format of any XML file is as follows:

```
<?xml version = “1.0" encoding = “utf-8" ?>
<tei.2>
<file Desc>
<title Stmt>
<title> دم الشهيد قصة من الصين </title>
------
------
------
<text>
<body>
<p> هل سبق لكم أن التقيتم مدينة بكين العائلة, عاصمة البلاد
الصين؟ لا شك إذن, أنكم التقيتم الجمال العمالقة, التي ينتصب في
إحدى ساحاتها الرئيسة, ووقفتم, وقد عقدت ألسنتكم الدهشة, من
فرط الإعجاب بهذه التحفة الفنية العالية فمن هو مبدعها يا
ترى? </p>
```

Then the Tokenizer read the XML file and split the text into tokens and save the result into Excel file. The output of excel file is illustrated as follows:

?	الإعجاب	الرئيسة	أنكم	العالية
ترى	فرط	ساحاتها ,		بكين
يا	من	إحدى	إذن	مدينة
مبدعها	الدهشة	في	شك	التقيتم
هو	ألسنتكم	ينتصب	لا	أن
فمن	عقدت	الذي ?		لكم
العالية	وقد ,		الصين	سبق
الفنية ,		العمالقة	البلاد	هل
التحفة	ووقفتم	الجمال	عاصمة	
بهذه ,		التماثيل		

After the preprocessing stage is finished then we have an excel file which contains the output of Tokenizer task. This file will be use as an input to processing stage, which involves set of tasks. First, the sentence splitter is used only one time, because we used an Arabic Part Of Speech (POS) based multilayer perceptron tagger[21] . The training and learning of neural network needs a training data set, which is used in the adaptation stage. The using of tagger based neural approaches not only consummate the associations (word-to-tag mappings) from a representative training data set but it can also generalize to unseen exemplars. The experiments had achieved accuracy of 100% in tagging residuals, prepositions and verbs categories. While the accuracy that achieved in tagging of nouns category is, 98 %. The accuracy result is depicted in Fig. 4 .

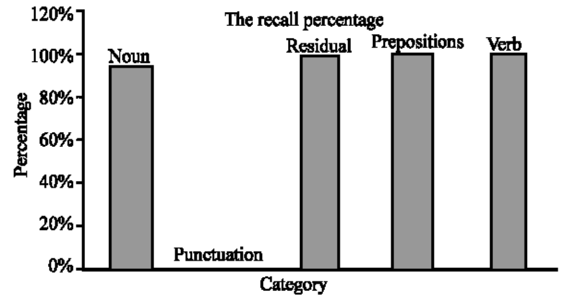

Fig.4: The recall percentage

Table 2: Best and average fitness

Optimization summary	Best fitness	Average fitness
Generation #	4	4
Minimum MSE	0.02117005	0.031124054
Final MSE	0.02117005	0.031124054

The lowest average cost reported by the Error Criterion since the beginning of the run is called Best Fitness. While the average fitness, is the average cost reported by the Error Criterion. The result of the best fitness is depicted in Table 2.

The model of[2] it will be use as an Arabic Morphological Analyzer, which is built using finite-state compilers and algorithms, and the results are stored and run as finite-state transducers. The output of Beesley Morphological Analyzer is the stem (word root) and the pos tagger output is grammar role, so we can do the parse phase easily Fig. 3.

CONCLUSIONS

We hope to demonstrate a proposed information extracting system. The multi-layered perceptron neural network tagger has solved the problem of Arabic part of speech efficiently. The new tagger approach is highly accurate. As well as, the testing data sets involved several equivocal tags. Successfully, the MLP tagger had tagged them in correct conformation. Likewise, the experiments had achieved accuracy of 100% in tagging

residuals, prepositions and verbs categories. While the accuracy that achieved in tagging of nouns category is 98%. Recently, we need to test the proposed system and hope to obtain a good result in extracting the important information from the Arabic document.

More attention must consecrate to solve the problem of Punctuation tagging failed and we test new approach based recurrent neural network. Recently, we test the proposed system and hope to make a good result in extract the Arabic document.

REFERENCES

1. Fouad Soufiane Douzidia and Guy Lapalme, 2004. Lakhas, an Arabic summarization system. In Proceedings of DUC.
2. Beesley Kenneth, 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001, ACL, Arabic NLP Workshop, Toulouse.
3. Anne , N., De Roeck and Waleed Al-Fares, 2000. A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. In Proceedings of the 38th ACL, Hong Kong.
4. Ciravegna, F., 2000. Learning to Tag for Information Extraction from Text, In Workshop Machine Learning for Information Extraction, European Conference on Artificial Intelligence ECCAI, Berlin, Germany.
5. Ciravegna Fabio, 2001. (LP), an adaptive algorithm for information extraction from web-related texts. In Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining. Seattle, WA.
6. Freitag Dayne and Kushmerick Nicholas, 2000. Boosted wrapper induction. In Proc. of the 17th National Conference on Artificial Intelligence AAAI-2000, pages 577-583, 2000. http://citeseer.ist.psu.edu/freitag00boosted.html
7. Freitag, D., 1998. Machine Learning for Information Extraction in Informal Domains, PhD Thesis, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA.
8. Riloff, E. and W. Lehnert, 1994. Information Extraction as a Basis for High-Precision Text Classification. ACM Transactions on Information Systems, 12: 296-333.
9. De Sitter An and Walter Daelemans, 2003. Information extraction via double classification. In Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining (ATEM). Cavtat-Dubronik, Croatia.
10. Scott, B., 1995. Huffman. Learning information extraction patterns from examples. In Working Notes of the IJCAI-95Workshop on New Approaches to Learning for Natural Language Processing, pp: 127-134.
11. Riloff, E., 1996. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. Artificial Intelligence, 85: 101-134.
12. Joseph, F.M. and G. Wendy, 1995. Lehnert, Using Decision Trees for Coreference Resolution. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, pp: 1050-1055.
13. Califf, M. and R. Mooney, 2003. Bottom-up relational learning of pattern matching rules for information extraction. J. Machine Learning Research, 4: 177-210.
14. Belal Abu-ata, 2001. An Arabic Stemming Algorithm on ERA for Information Retrieval, PhD. Thesis, Kebangsaan University.
15. Nikkhou Mahtab and Khalid Choukri, 2005. Survey on Arabic Language Resources and Tools in the Mediterranean Countries. Last access on, pp: 20-02. http://www.nemlar.org/Publications/Nemlar-report-ind-needs_web.pdf .
16. Ahmed, F., 1999. Developing an Arabic Parser in multilingual machine translation system, M. Sc. Thesis, Cairo University.
17. Salah, R., 1998. Al-Najem, An Exploration of Computational Arabic Morphology, PhD Thesis, Essex University.
18. Ahmed Abedali, Jim and Cowie, 2004. Arabic Information Retrieval perspectives, JEP TALAN, Arabic Language Processing, Fez, pp: 19-22.
19. Beesley Kenneth, 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001, ACL, Arabic NLP Workshop, Toulouse.
20. Hani Al-omari, 1999. ALMAS: An Arabic Language Morphological Analyzer System, M. Sc. Thesis, Kebangsaan University.
21. Jabar, H. Yousif and M.T. Tengku, 2005. Sembok Arabic Part-Of-Speech Tagger Based Neural Networks, ACIT, Amman, Jordan.