Theoretical background and practices about extracting Proper nouns in Electronically English and Arabic Texts

Marwan M. Al Omari

Lebanese University

Dr. Moustafa Al-Hajj

2018

Introduction

Before starting the research for the ways to extract proper nouns, we should ask ourselves what is proper noun. According to British council learning center, proper nouns are the names of organizations including companies, places, and people. However, what is the benefit of proper nouns? They give the reader a surface knowledge of what is going on in any given content. Sometimes, the text or sentence is worthless without the specific nouns, which include. People usually read articles for famous people, places, or organizations they speak about. Many studies in English language have been conducted in the Name recognition of proper nouns while in Arabic it is still at its early stages.

Body Background

As the search starts for ways to extract only proper nouns from texts, I come across modules that could be used in Perl language, sharpening its accuracy. To let the computer understands the language, parsing modules are important to use as like Lingua-LinkParser, Lingua-Stem, link-grammar, Lingua-EN-Tagger, Lingua-EN-NamedEntity, Lingua-Grammar, and Lingua-Sentence. Concerning proper nouns, Lingua-EN-NamedEntity is the one that would help us reach the objective of research. “"Named entities" is the NLP jargon for proper nouns which represent people, places, organisations, and so on. This module provides a very simple way of extracting these from a text” (Buvik, 2015).

Lingua-EN-NamedEntity

Therefore, I did download the previous mentioned modules to start working on the texts, testing “Lingua-EN-NamedEntity” efficiency and accuracy in extracting proper nouns. I did apply the “NamedEntity” module to “test.txt”, which is written by Berg R. (2017) under the title “The inside story of the GOP's Alabama meltdown”. Therefore, we have the following output:

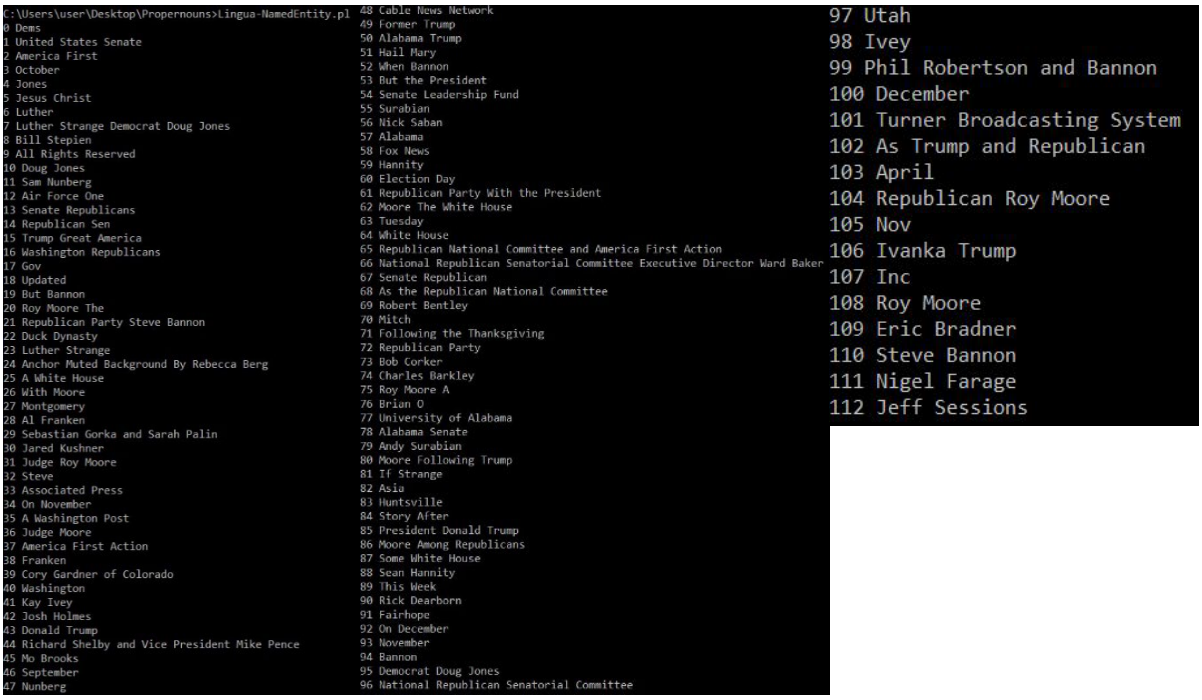

Figure (1): It shows the result of applying "Lingua-EN-NamedEntity" on "test.txt"

As we see from the result, the module has provided us. It is not that accurate. It does not pick all the proper nouns in the given "test.txt" as it supposed to do. The total number of extracted proper nouns is (112) proper nouns and some of them, in fact, are not proper nouns like numbers (107)(89)(84)(18)(9). For instance, "Inc" (n.107) is not. In addition, some of the proper nouns contain words that are not considered proper nouns as in numbers (102)(99)(92)(87)(86)(81)(80)(71). For example, "Moore Following Trump" (n.80), "Following" is not a proper noun. However, it is yet giving one hundred percent accurate result.

Blacklist Method

I went back working on the code we worked on in classroom in attempt of improving its accuracy of extracting proper nouns. I made many attempts, which are of importance. Firstly, I did add a dictionary of words. "A list of 109582 English words compiled and corrected in 1991 from lists obtained from the Interociter bulletin board", and "this word list includes inflected forms, such as plural nouns and the -s, -ed and -ing forms of verbs." ("English Wordlists", n.d.). Thus, the result has been sharpened to meet the minimum requirements (proper nouns extractions). However, the number of lexical stems represented in the list is considerably smaller than the total number of words in English. The result is too narrow and it is not efficient for the task.

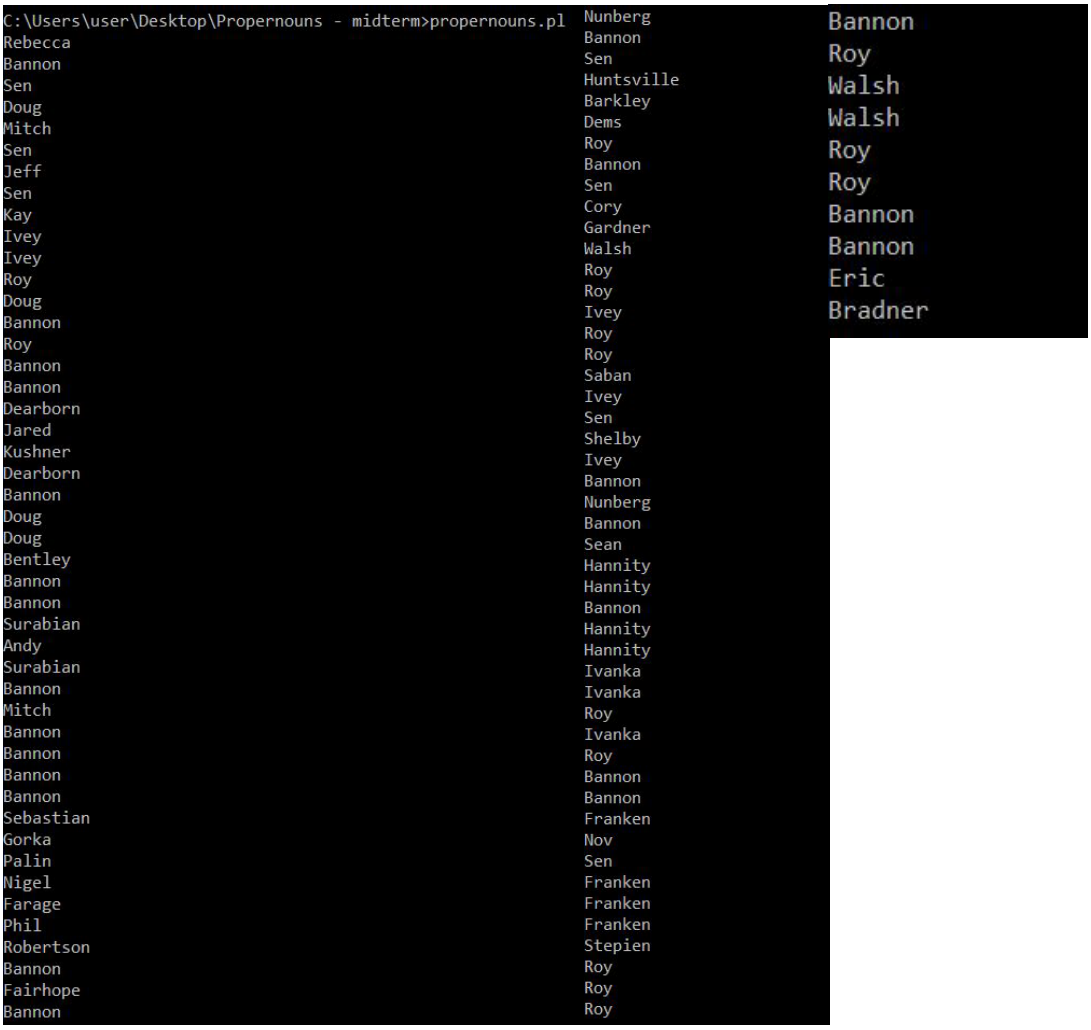

Figure (2): it shows the result obtained after adding the wordlist from “Interociter bulletin board”

The result obtained as we see in the above picture does not have any words, which are not proper nouns, but it is only the first name of the picked entity. Thus, It does not extract the proper nouns as a whole entity and even it does not meet the desirable conclusion as in Lingua-NamedEntity.

Lingua::LinkParser

For Brian (n.d.), regular expressions, which used in Perl language, considered one of the strongest ones among other languages, for the ability to handle the complexity of patterns that one may find in a text. The power of regexex (regular expressions), however, falls apart when they come into the understanding of particular sentiment because firstly it lacks the knowledge of the sentence. “It's one thing to know that a phrase consists of two adjectives and two nouns -- but what you really want to know is which adjective modifies which noun. The Link Grammar does that for you” (Brian, n.d.). “The Link Grammar is based on a characteristic that its creators call *planarity*. Planarity describes a phenomenon present in most natural languages, which is that if you draw arcs between related words in a sentence (for instance, between an adjective and the noun it modifies), your sentence is ungrammatical if arcs cross one another, and grammatical if they don't.

This is an oversimplification, but it'll serve for our purposes." (Brian, n.d.). It generates, however, misleading results in conversational texts. The link grammar has achieved higher accuracy in newspaper texts (Brian, n.d.).

Moreover, according to the website **Experts Exchange**, it gives suggestions to extract proper nouns using regular expression but in PHP language:

A. A word is a proper noun if it begins with a capital letter and is NOT the first word in a sentence.
B. A word is a proper noun if it begins with a 2 or three letter title and a dot such as: Mr. Mrs. Dr. Ms.
C. Titles such as Mr. Mrs. Dr. Ms. should be included as part of the proper noun.
D. Single quotes and double quotes should NOT be included as part of the proper noun.
E. The proper noun should include all capital words in a row. For example, this is one proper noun: New York City.

By all the humble knowledge, I have reached so far which gathered from many resources. Thus, I have started to develop my own method of extracting proper nouns in English language. I followed the contextual method to deal with the language on its surface structure away from the complexities of morphological and syntactical rules. My own method is based on the words (key words) that proceed with the nouns in most cases. For instance, the particles as in, on, the, at, to are followed by proper nouns (firs letter capitalized) as in "on November, in Washington, the White House, at Brooklyn Park, to Steve Harvey, etc). In addition, abbreviation for someone's title as "president, mister (mr.), miss (ms./mrs.), etc to indicate the Name of person that follows. Thus, the code has got a good result so far. The following pictures are depicted from the output generated from the same corpus.

```
C:\Users\user\Desktop>TestingPNE.pl
GOP
Anchor Muted Background
CNN
By Rebecca Berg
GMT
HKT
Updated
December
GOP
Alabama
GOP
Alabama
Alabama
HIGHLIGHTS
STORY
HIGHLIGHTS
STORY HIGHLIGHTS
GOP
Former Trump
Steve Bannon
Luther Strange
Sen. Luther Strange
Democrat Doug Jones
CNN
Mitch McConnell
Alabama
President Donald
Republican Sen
Jeff Sessions
President Donald Trump
McConnell
Sen. Jeff Sessions
November
Senate
November
Donald Trump's
CNN
The Alabama
Luther Strange
Sen. Luther Strange
CNN. The
Alabama
Alabama
the White House
McConnell
Kay Ivey
White House
Washington Republicans
Gov. Kay Ivey
April
Ivey
Ivey
#1ssteps
Republican Roy Moore
Democrat Doug Jones
Republican
Democrat
Democrat
Senate
Roy Moore's
the White House
White House
the White House
White House
Josh Holmes
McConnell
McConnell's
the Republican Party
Republican Party
Murphy
GOP
And GOP
Moore
the White House
Washington Republicans
McConnell
White House
Washington
CNN
the Alabama Senate
Alabama Senate
Senate Republicans
Steve Bannon
CNN
White House
White
the National Republican Senatorial
the National Republican
Judge Roy Moore
Mo Brooks
McConnell
National Republican Senatorial Committee
Rep. Mo Brooks
Rep
GOP
GOP
Dearborn
President and
Some White House
Rick Dearborn
Jared Kushner
House
Dearborn
GOP
Sessions
GOP
The President
The President's
Luther
President ultimately
The President
Doug Jones
Moore
Doug Jones
Moore
If Strange
Robert Bentley
Gov. Robert Bentley
Trump
Gov
Alabama
If Strange's
The McConnell
The McConnell
Senate Leadership Fund
Washington Republicans
Strange
Washington
the White House
When Bannon
White House
McConnell
August
McConnell's
PAC
the White House
Andy Surabian
White House
Trump Great America
Mitch McConnell
Steve Bannon
Steve
the Republican Party
Steve Bannon
Republican Party
the Republican Party
Steve Bannon
Republican Party
President still
Election Day
Sebastian Gorka
Sarah Palin
Nigel Farage
Duck Dynasty
Phil Robertson
Moore
Montgomery
Fairhope
Gorka Sarah
Robertson Bannon
Duck Dynasty's
Matt Drudge
Sam Nunberg
Bannon
President had
President indicated
Bob Corker
Sen. Bob Corker
Alabama
Sen
Corker
National Republican Senatorial Committee Executive Director Ward Baker
Alabama
President stood
September
Huntsville
September
Strange
Charles Barkley
Dems
Roy Moore
```

Figure (3): The code that I have worked on myself. However, there are more result (+).

The code still lack the accuracy because there are repeated proper nouns. Thus, the result should not mention a proper noun for more than one time and if so, it should mention the frequency. Nevertheless, the output does contain the keyword so I can some modification on the code. It is still a work for the future.

Arabic Proper Noun Recognition

On the other hand, there are also some researches have been done in extracting proper nouns in the Arabic language. However, there are no modules or snaps of such application available to use or test except by paying money for his/her inventor/researcher. As it mentioned in an article under title of "Arabic Language in the Context of Information Extraction Task" by Alruily M. et al. (2011), it summarizes the most recent studies that follows the rule-based linguistics to extract proper nouns. For example, TAGARAB, Mesfar, Abueil, NERA, Al-Shalabi, PNAES, and Traboulsi, etc. All of which use keywords and morphological knowledge to recognize proper nouns in a given text. The following table shows the differences in precision, recall, and f-measure of each system.

System	Entity	Precision	Recall	F- measure	Year
TAGARAB	Number	82.8	97.0	97.3	1998
	Time	91.0	80.7	85.5	
	Location	94.5	85.3	89.7	
	Person	86.2	76.2	80.9	
Mesfar	Number	97.0	94.0	95,5	2007
	Time	97.0	95.0	96.0	
	Location	82.0	71.0	76.0	
	Person	92.0	79.0	85.0	
Abueil	Event	86	81	84	2007
NERA	Time	97.25	94.5	95.4	2008
	Location	77.4	96.8	85.9	
	Person	86.3	89.2	87.7	
Al-Shalabi	Time	89.4	×	×	2009
	Location	91.6	×	×	
	Person	81.1	×	×	
PNAES	Person	93	86	89	2009
Traboulsi	Person	×	×	×	2009

Furthermore, there are systems that follow and use the machine learning method to recognize Arabic proper nouns. As it mentioned in the article under title of "Arabic Language in the Context of Information Extraction Task" by Alruily M. et al.(2011), it summarizes the most recent study has provided systems following the machine learning method. For example, ANERsys

and AbdelRahman (with pattern feature), etc. The following table shows the differences in precision, recall, and f-measure of each system.

System	Entity	Precision	Recall	F- measure	Year
ANERsys (using ME) (with gazetteers)	Location	82.17	78.42	80.25	2007
	Person	54.21	41.01	46.69	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
(without gazetteers)	Location	82.41	76.90	79.56	
	Person	52.76	38.44	44.47	
	Misc.	61.54	32.65	42.67	
	Organisation	45.16	31.04	36.79	
ANERsys (using CRF)	Location	93.03	86.67	89.74	2008
	Person	80.41	67.42	73.35	
	Misc.	71.0	54.20	61.47	
	Organisation	84.23	53.94	65.76	
AbdelRahman (with pattern feature)	Location	96.05	80.86	87.80	2010
	Person	89.20	54.68	67.80	
	Organisation	84.95	60.02	70.34	
(without pattern feature)	Location	89.37	69.25	78.03	
	Person	87.01	53.23	66.05	
	Organisation	88.45	49.00	63.07	
Abdul-hamid and Darwish	Location	93	83	88	2010
	Person	90	75	81	
	Organisation	84	64	73	

Conclusion

The coding files can meet the goal of extracting proper nouns but not even close to 100% accuracy. Using modules that understand the sentence structure as Lingua::EN::NamedEntity, the result would score a higher accuracy, however, does not achieve a satisfactory result. In addition, the Arabic systems of its both methods: rule-based or machine learning are still no comparable to systems that have been developed in the English language, which opens a wide scope for researches to work on NERA (Name Entity Recognition in Arabic language).

References

"[SOLUTION] Find proper nouns using REGEX." Experts Exchange, www.experts-exchange.com/questions/23264090/Find-proper-nouns-using-REGEX.html.

Brian, D. (n.d.). Parsing Natural Language with Lingua::LinkParser. Retrieved January 17, 2018, from https://www.foo.be/docs/tpj/issues/vol5_3/tpj0503-0010.html

Alruily, M., Ayesh, A., & Zedan, H. (2011, January 01). Arabic language in the context of information extraction task. Retrieved January 25, 2018, from https://www.dora.dmu.ac.uk/handle/2086/11188

Buvik, R. (2015, October 10). Lingua::EN::NamedEntity. Retrieved January 20, 2018, from http://search.cpan.org/dist/Lingua-EN-NamedEntity/NamedEntity.pm

Berg, R. (2017, December 14). The inside story of the GOP's Alabama meltdown. Retrieved January 20, 2018, from http://edition.cnn.com/2017/12/13/politics/alabama-senate-election/index.html

English Wordlists. (n.d.). Retrieved January 01, 2018, from http://www-01.sil.org/linguistics/wordlists/english/

Proper nouns. (n.d.). Retrieved January 27, 2018, from https://learnenglish.britishcouncil.org/en/english-grammar/nouns/proper-nouns