Natural Language processing applications

Omar Alrassam

Outlines

- What is NLP?
- The importance of NLP
- Applications of NLP
- NPL in Kurdish language
- Future work
- Conclusion
- References

What is NLP?

- **N**atural **L**anguage **P**rocessing, or (NLP) for short is a tract of Artificial Intelligence and Linguistics, devoted to make computers understand the statements or words written in human languages.
- Natural language processing is the area of study dedicated to the automatic manipulation of speech and text by software.
- Natural Language Processing, is broadly defined as the automatic manipulation of natural language, like speech and text, by software.
- The study of natural language processing has been around for more than 50 years and grew out of the field of linguistics

The importance of NLP

- NLP is important for scientific, economic, social, and cultural reasons. NLP is experiencing rapid growth as its theories and methods are deployed in a variety of new language technologies. For this reason it is important for a wide range of people to have a working knowledge of NLP. Within industry, this includes people in human-computer interaction, business information analysis, and web software development. Within academia, it includes people in areas from humanities computing and corpus linguistics through to computer science and artificial intelligence. (To many people in academia, NLP is known by the name of “Computational Linguistics.”)

Some Applications of NLP

- Spelling checker
- Machine translation
- Text summarization
- Automatic Question Answering
- Speech Recognition
- Text-to-Speech Synthesis
- Information Retrieval

Spelling Checker

Spelling Correction is a process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. Spell Checker is an application program that flags words in a document that may not be spelled correctly. Spell Checker may be stand-alone capable of operating on a block a text

Spelling Checker cont

* Error Detection Techniques

1- Dictionary Lookup Technique

2- N-gram Analysis Technique $$P(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

3- Neural Net Techniques

Some important terms in NLP

- Corpus

1- One of the first things required for natural language processing (NLP) tasks is a corpus.

2- A large collection of written or spoken language, that is used for studying the language.

- Stop Words: A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

Some important terms in NLP

- Example of stop words

Sample text with Stop Words	Without Stop Words
GeeksforGeeks – A Computer Science Portal for Geeks	GeeksforGeeks , Computer Science, Portal ,Geeks
Can listening be exhausting?	Listening, Exhausting
I like reading, so I read	Like, Reading, read

Some important terms in NLP

- Stemming
- Exmaple
  - Cars, car's, cars' □ car
  - Am, is, are □ be

  - the boy's cars are different colors
    the boy car be differ color

Challenges in Kurdish Text Processing

According to the study of (Sheykh Esmaili, 2014)

- Despite having a large number of speakers, the Kurdish language is among the less-resourced languages.
- Spoken by more than 30 million people in western Asia.
- Has many dialects, the two major ones
  - Northern Kurdish (Kurmanji)
  - Central Kurdish (Sorani)

Challenges in Kurdish Text Processing

1- Dialect Diversity

- Grammar (F,M)

- Writing system

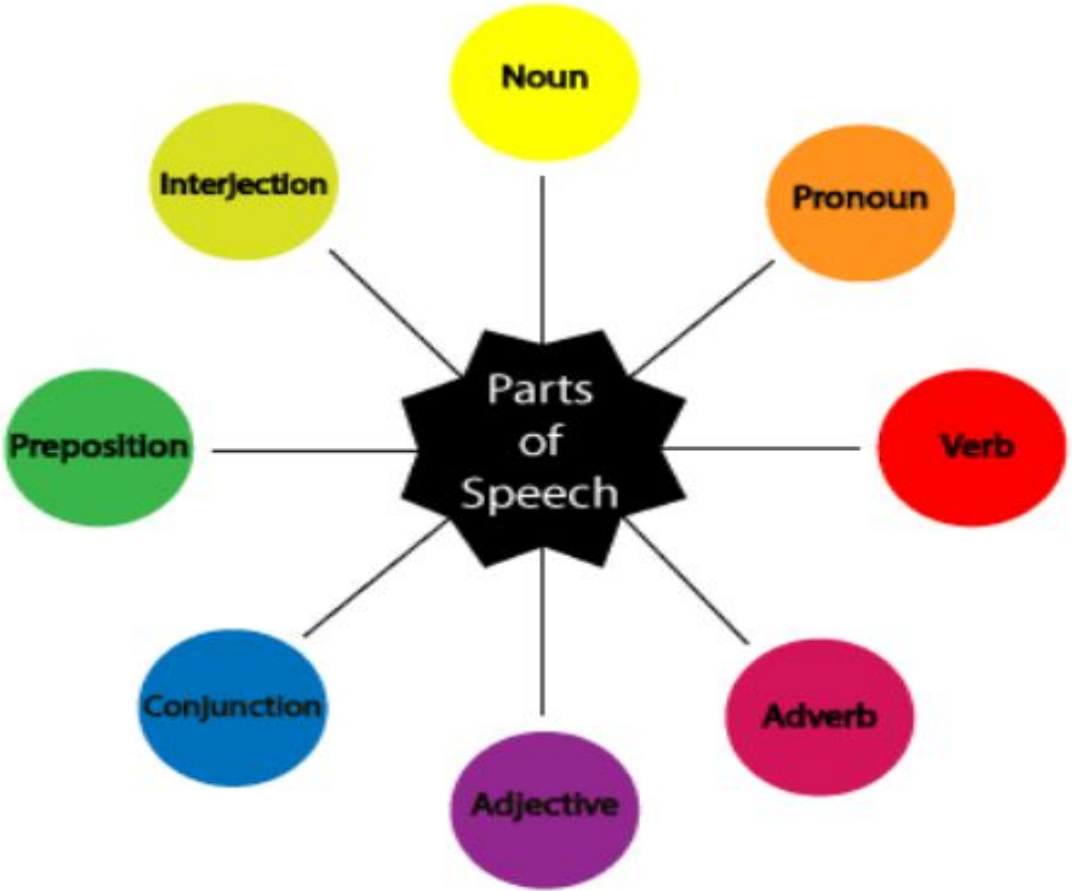
Kurmanji using Latin-based letters

Sorani using Arabic-based letters

As a result Difficulties in Segmentation and Tokenization process

Challenges in Kurdish Text Processing

2- part-of-speech detection.



Challenges in Kurdish Text Processing

3- Arabic alphabet does not have capitalization and therefore it is more difficult to recognize sentence boundaries as well as recognizing Named Entities.

Challenges in Kurdish Text Processing

4- Lack of Language Resources Kurdish is a resource-scarce language for which the only linguistic resource available on the Web is raw text. More concretely, in spite the few attempts in building corpus and lexicon, Kurdish still does not have any large-scale and reliable general/domainspecific corpus.

Furthermore, no test collection {which is essential in evaluation of Information Retrieval systems{ or stemming algorithm has been developed for Kurdish so far.

Challenges in Kurdish Text Processing

Updating words

1_کارێزما_____کەسی پیرۆز
2_ریفورم _____ چاکسازی
3_ناسیونالیزم _____نەتەوەگەری
4_ستاندار _____ پێوانەیی
5_پرۆتوکول _____ڕێکەوتن
6_پڕۆلیتار _____ چینی کرێکاران
7_کاپیتالیزم _____ سەرمایەداری
8_فیودالیزم _____دەرەبەگایەتی
9_ئەنارشیزم _____ئاژاوەچێتی
10_ئایدیالیزم _____فەلسەفەی میسالی(خەیالی)
11_ریالیزم _____فەلسەفەی واقعگەری(واقعی)
12_ یۆتۆپیا_____ ئەم زاراوایی زیاتر بۆ خەیالی فەلسەفی فەیلەسوفان بەکادێت و تایبەتە بە فەیلەسوفان
13_ئەلتەرناتیف ـ بەدیل(جیگرە وە)
14_ئەکادیمی ـ زانستی
15_ئەنتی ـ دژ

Challenges in Kurdish Text Processing

16_زیندەوەڕ ناسی ـ بایلۆجی
17_کومیدیا ـ بەزمە سات
18_جینوساید ـ قرکڕدن، قەلاچوکڕدن
19_دیمۆگڕافی ـ دانیشتوانی
20_ئەکتیڤ ـ چالاک،کارا
22_ستاندار ـ پێوانەیی
23_ئەتنۆگرافیا___ وەسفی گەلان، گەلناسی
24_ئوتوکراتی ___ حوکمی تاکەکەس
25_ئۆرگان ___ ئەندام،زمانحال
26_ئۆپزسیون ___بەڕەلستکار
27_ ئینسکلۆپیدیا ___ موسوعە(زانیاری نامە)
28_ئیکۆلوژی ___ ژینگەناسی
29_ئیکۆنومیزم___ئابوریناسی
30_هیومانیزم ___ مڕۆڤگەڕایی

Challenges in Kurdish Text Processing

31_کەپیتالیزم __ سەرمایەداری
32_کاندید ___ پالێوراو
33_کامپ ___ نیشینگە
34_کۆنکریتی___بەلگەدار
35_دوکیۆمێنت ___بەلگەنامە
36_کۆمسیون ___لیژنە
37_دیسپلین ___ڕێکخەر
38_دیموکراسی __ حوکمی گەل
39_دۆسییە___فایل
40_دیمۆگرافی ___دانیشتوویی
41_دیالیکت ___شێوەزار
42_دیفاکتۆ___ئەمری واقیع
43_پراکتیک ــ جێبەجێکردن

Challenges in Kurdish Text Processing

پرۆسه ــ کرده_46

47_ناسیونالیزم ــ نەتەوەگەری

48_کۆلۆنیالیزم ــ داگیرکەر

49_کۆلونیالکڕا ــ داگیرکڕاو

50_ماتریالیزم ــ ماددییەت

Challenges in Kurdish Text Processing

Future work
Deep learning Algorithm

- **Deep learning** (also known as **deep structured learning** or **hierarchical learning**) is part of a broader family of machine learning methods based on artificial neural networks.

Conclusion

- In this seminar the importance of NLP is discussed and the challenging of Kurdish text was explained.
- Also, focusing on the Text processing in Kurdish language was the main aim of this seminar.