ENTITY SALIENCE IN SHORT TEXTUAL PARAGRAPHS

A PREPRINT

Harsh Khandelwal
Bachelors Computer Science
Vrije UniversityAmsterdam
h.khandelwal@student.vu.nl

May 21, 2019

1 Abstract

2 Introduction

Make sure to have a short introduction of the oncoming sections. A brief agenda.

3 Problem Statement

With the advancements in technology, human attention span has been a central factor in design of interactive applications and algorithms. Paragraphs containing verbose and expressive text require effective decimation to convey the bulk information. Textual paragraphs are usually centered towards a subject which is conveyed through multiple entities. Entity is another word for the classification of word / information in a text. Common examples of entity classes include Person, Location, Organization and Time.

Entity salience refers to the act of parsing entities from a piece of text and then ranking them in order of their importance / relevance to the text. For this project, short news articles are considered as input text. The core part of the project is to develop an algorithm that can rank the parsed entities in the order of their importance. For the given news article as input text, the output should be a sorted list of entities.

4 Related Work

5 Motivation

6 Parsing

6.1 Challenges

To be able to rank the entities, it is important to parse the entirety of an entity. A simple approach to parsing words from an input string is to use space *" "* as a delimiter. However, this is limited to only single lettered entities. If the parser comes across *"New York"* in a string, it would parse *"New"* and *"York"* as two completely different entities which is wrong for two reasons. First, it is not the same as the original intended entity. Second, "New" does not readily classify as a location / person / organization or other classes of entities.

The task of entity parsing is to be dealt with more intricacy. News articles often introduce apostrophes to words. In the input text *"Microsoft organized a conference in Amsterdam's office."*, among others, *"Amsterdam's"* will be parsed as an entity. However, the entity is actually a location in this case and should be *"Amsterdam"*. As such, the parser should remove apostrophes from words.

The parser should also take into consideration other anomalies in usual news text using hyphens. In the input text *"Stock prices have gone down for PMC-Sierra"*, *"PMC-Sierra"* should be rightly classified as a company / organization. Furthermore, the parser should also handle abbreviations(*"Ph.D"*), special characters(*"Johnson & Johnson"*), amounts (*$50.00*), dates (*01/11/2018*), email addresses (*"someone@somedomain.com"*), hashtags (*"#nlp"*), and even URLs (*"https://www.somerandomwebsite.com"*).

6.2 Ranking strategies

There can be some entity ranking strategies that can be applied during parsing. The case of a word can help distinguish between abbreviations and other common words (*"US" - country vs. "us" - pronoun*). A major section of entities tend to be proper nouns - locations, people, and organization. In English, proper nouns always begin with a capital letter. Taking the case of the first letter of a word into consideration can also help in classifying an entity.

News article writers tend to emphasize certain points by quoting speech of people. In the text: *"And now for the national military news, Trump asserts, "We are going to increase budget for the Navy S.E.A.L.". Families of troops were also assured free education and non-taxable income."*, it is certain that the subject of the text is *the increment of budget for Navy S.E.A.L.*. In such situations, authors credit people by quoting them. As such, if entities belong to a sentence that was quoted, it is very likely that the entity is important.

Furthermore, if it is known that the text is **bold**, underlined or, *italicized*, the fact can be used to emphasize on the entity. Often, the presence of punctuation also helps in laying the importance of an entity. If the entity was followed by a punctuation, be it exclamation *"!"* or comma *","*, the author signifies a break in the sentence flow, indicating the importance of the entity. However, a strong metric is needed to measure such factors while parsing and reducing the chance of causing false positives.

7 Dataset

7.1 Collecting Data

For this project, the input data has been restricted to short textual news paragraphs. While there are a myriad of news articles available on the web, collecting data specific to this research purpose has been a challenge. Most of the news sources limit the usage of their data by copyright restrictions. Furthermore, if lucid data is available, then obtaining it's accurate annotation becomes a problem. A preferable dataset would be that of a huge news corpus along with annotated salience entities for testing.

The research community has explored already existing datasets for the entity salience task, e.g. the Microsoft Document Aboutness (MDA) dataset, and the New York Times (NYT) dataset. However, neither dataset provides the underlying document content due to copyright restrictions. Moreover, in these dataset the entity candidates have been generated automatically using proprietary NER systems. [1]

7.2 Reuters-128

Dojchinovski, et al. provide corpus with crowdsourced entity salience annotations by reusing the NEL Reuters–128 corpus published in he NLP Interchange Format (NIF). This dataset can be downloaded here [1].

TODO: mention all the different datasets to be used, mention their formats, mention their pre-processing, mention a way of testing their result.

TODO: plots for data. plot for avg. amounf of entities parsed in an article in a dataset.

8 Baseline 1

Parsing of the entities is not central to the project, and hence, pre-trained models from various libraries will be used. At the moment, SpaCy, Flair and Stanford CoreNLP library in python appear to be state-of-the-art.

TODO: explain how flair works with ref

[1] https://github.com/KIZI/ner-eval-collection

9 Results

10 Conclusion

11 Future Work

References

[1] Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomas Vitvar, Harald Sack *Crowdsourced Corpus with Entity Salience Annotations*

[2] Marco Ponza, Luciano Del Corro, Gerhard Weikum (2018) *Facts That Matter*

[3] Diksha Khurana, Aditya Koli, Kiran Khatter, Sukhdev Singh (2018) *Natural Language Processing: State of The Art, Current Trends and Challenges*

[4] Daniel Jurafsky (2018) *Speech and Language Processing*

[5] Kamal Sarkar *A Hybrid Approach to Extract Keyphrases from Medical Documents*

[6] H. P. Edmundson *New Methods in Automatic Extracting*

[7] Wei Shen, Jianyong Wang *Entity Linking with a Knowledge Base: Issues, Techniques, and Solutions*

[8] Ernesto D'Avanzo, Alessandro Vallin, Bernardo Magnini *Keyphrase Extraction for Summarization Purposes*

[9] Güneş Erkan, Dragomir R. Radev *LexRank: Graph-based Lexical Centrality as Salience in Text Summarization*

[10] Lunh (1958) *The Automatic Creation of Literature Abstracts*