ENTITY SALIENCE IN SHORT TEXTUAL PARAGRAPHS

A PREPRINT

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

May 21, 2019

ABSTRACT

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Introduction

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.

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

2 Related Work

3 Motivation

4 Parsing

4.1 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"*).

4.2 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, as mentioned in the book [1], 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.

However, due to the limited scope of the project, 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.

5 Dataset

5.1 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.

As mentioned in the article [2] "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."

5.2 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].. This dataset contains short textual paragraphs based on economic news.

5.2.1 Pre-processing

Since the data was mentioned in a NIF format, it had to go through some pre-processing to be used by various python libraries. The input data came as a single NIF file with multiple articles and their entities inside. Each chunk began with their position on web: a url appended with the article number and the character position within the article. (Example: <*http://aksw.org/N3/Reuters-128/43#char=143,146*>)

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

The articles were embedded in a chunk containing the string *nif:Context*. Therefore, it acted as a delimiter to separate the articles from entities. The articles were read and appended into an *ArrayList* and the entities were eventually added as a property of the article. For the relevance of the project, the entities with text *nif:NamedEntity* were kept while those with *nif:CommonEntity* were removed. The named entities were mentioned in three pre-ranked categories: *most_salient, less_salient, not_salient*.

The articles, along with their given named entities were written into separate files and separated by <*delim*>. The entire input filtering was done in *JAVA*.

TODO: mention all the different datasets to be used, mention their formats.

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

6 Research Methods

The task of outputting a list of salient entities from a short textual news paragraph has been divided into multiple baselines. Each baseline implements a different approach of achieving the task. They are then tested with the annotated entities provided with the corpus.

6.1 Baseline 1 - Frequency Analysis

TODO: explain how flair works with ref

7 Results

8 Conclusion

9 Future Work

References

[1] Daniel Jurafsky and James H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 2018.

[2] Milan Dojchinovski, Dinesh Reddy, Tomáš Kliegr, Tomas Vitvar, and Harald Sack. Crowdsourced corpus with entity salience annotations. page 1, 2016.