Graph Based Automatic Domain Term Extraction

Abstract

- Domain Term, is a word or group of words, carrying a special, possibly complex, conceptual meaning, within a specific domain or subject field or community
- Because of their low ambiguity and high specificity, these words are also particularly useful to conceptualize a knowledge subject. For each domain, there is an essential need to identify the domain-specific terms as they play a vital role in many Natural Language Processing Applications
- The task of automatically extracting domain specific terms from a given text of a certain academic or technical domain, is known as Automatic Technical Domain Term Extraction
- We present a Graph Based Approach to automatically extract domain specific terms. Our approach is similar to TextRank with an extra post-processing step to reduce the noise.

Approach

- We implemented the TextRank algorithm by doing few modifications in syntactic filters, and adding a post processing step for noise removal using top 1000 common words in English from Wikipedia
- Based on syntactic filters, such as Parts of Speech (POS) Tags, words are selected as nodes and relation between the words is based on word co-occurrences , a window size (N) is assumed for word co-occurrences
- we used Noun, Proper Nouns, Adjectives as syntactic filters, window size (N = 4)

Experiments & Results

- We evaluate our approach on data provided by ICON TermTraction - 2020 shared task for four domains, Biochemistry, communication , Computer Science and Law
- In each domain we have minimum 10 files and maximum 16 files
- As our approach comes under unsupervised learning, there is no requirement of training data
- Fiind averaged precision , recall and F1 score from below table

Domain	Precision	Recall	F1
BioChemistry	0.18	0.52	0.26
Communication	0.08	0.54	0.14
Computer Science	0.13	0.56	0.20
Law	0.05	0.5	0.10

Conclusion & Future work

- Our approach showed high recall in all cases for all domains, as we used wiki pedia terms for noise removal there is no much noise in ouput domain terms
- This approach doesn't depend on any language dependant resources except POS tagger, hence we can adopt this method for any language
- We plan to use roots as nodes for morphologically rich languages like Telugu and want to use different edge relations

Authors: Hema Ala , Prof. Dipti Misra Sharma

Research Center Name : LTRC IIIT Hyderabad