Capstone Submission

CItation analysis and extraction

Research

Review and summarization

Sehrish Iqbal, Saeed-UI Hassan, Naif Radi Aljohani, SaleemALelyan,Raheel Nawaz, Lutz Bornmann (June 23 2021),*A decade of in-text citation analysis based on natural language processing and machine learning techniques:an overview of empirical studies*, https://link.springer.com/article/10.1007%2Fs11192-021-04055-1

A literature overview of several research documents on in-text citation extraction, analysis, classification and quantification

Research summary

ML techniques used in the corpus of research documents included in the evaluation

- Support Vector Machine
- Naïve Bayes
- Multinomial Naïve Bayes
- Hidden Naïve Bayes
- Maximum Entropy
- Decision Tree
- Random Forest
- K Nearest Neighbour
- Logistic Regression

Research

Table of Literature reviewed

Table 4 Summary of the reviewed literature on citation in-text analysis

Article	Data repository	Samples size	Main results
Ritchie, Robertson, and Teufel (2008)	Association of Computational Linguistics (ACL) Anthology	9800 papers	Longer citation contexts resulted in greater retrieval effectiveness, 3sent was more effective than 1sent
Angrosh Cranefield, and Stanger (2010)	Lecture Notes in Computer Science (LNCS)	50 papers	Citation features along with the sentence features play an important role in the identification of citation context and yield an accuracy of 96.51%
Aljaber et al. (2011)	TREC Genomic	162,259 papers	Citation context is a rich source of topically related terms and many of them are semantically related to terms that are present in the original document
Angrosh et al. (2013)	Lecture Notes in Computer Science (LNCS)	20 papers	CRF with additional zero-order features identified context better than linear CRF and scored an accuracy of 91%
Zhang et al. (2013)	Social sciences –	Citations should not be treated equally as they have different reasons and functions	
Hu et al. (2013)	*Journal of Informetrics*	350 papers (11,327 citations)	In full-text articles, citations are distributed unevenly, more than 50% of the citations belong to the "introduction" section
Ding et al. (2013)	*Journal of the American Society for Information Science and Technology* (JASIST)	866 papers	Highly re-cited references (the same publication is cited multiple times in the citing paper) appear mostly in the introduction and literature review sections
Abu-Jbara et al. (2013)	ACL Anthology	30 citing papers (3500 citation contexts)	In the interpretation of citation contexts, lexical features (determiners and conjunction adverbs) are more significant than structural features (position and reference)

Research

Table of reviewed literature continued

Table 4 (continued)

Article	Data repository	Samples size	Main results
Bertin and Atanassova (2014)	Public Library of Science (PloS) Journals	9446 papers (459,834 citation sentences)	The frequency of verbs in citation contexts depends on the paper sections in which they appear and 50% of the verbs were present in the "introduction" section
Fujiwara and Yamamoto (2015)	PMC-OAS	545,147 papers	Papers cited in less than five citation contexts account for around 76% of all the cited papers in the database
Hu et al. (2015)	*Journal of Informetrics*	350 papers (11,327 citations)	In research articles, succeeding citations are more intentional and purposeful than first-time citations
Bertin et al. (2016b)	PLoS Journals	75,000 citing papers (3 million citation sentences)	The word 'show' is the most frequently occurring verb in citation contexts among all paper sections
Bertin and Atanassova (2017)	PLoS Journals	80,000 papers (3,528,514 citing sentences)	41% of the citation sentences contain MIR, most of them appear in the "Introduction" section
Small et al. (2017)	PubMed Central Open Access Subset (PMC-OAS)	1.1 million papers	Only 46% of the articles that had 'discovery' words in their citing sentences (citances) were actually scientific discoveries
Hu et al. (2017)	*Journal of Informetrics*	350 papers (16,917 citations)	25% of the references were mentioned multiple times and located in close proximity
Boyack et al. (2018)	Elsevier and PubMed Central	5 million papers	The references that are mentioned just once are typically more highly cited than references that are mentioned multiple times

Research

Summary of features reviewed

- Optimal length of citation context analysis – optimal length appears to be 4 sentences. Extended context appears to be more effective that shorter contexts up to the 4 sentence optimum.
- The use of NLP and ML techniques to analyze citation sentiment using lexicon based methods
  - Classify citation context into different polarity levels(positive , negative, neutral)
  - Feature-sets such as uni-gram. Higher levels values of n-grams(consecutive words) lead to better classification results.
  - SVM and NB classifiers are the most frequently used models for sentiment classification.
- Lexical and context features in context extraction for accuracy – lexical features(determiners and conjunction adverbs) are more significant than structural features(position and reference).
- ML Techniques for citation classification with preferred features – effective at achieving high evaluation scores.
  - Linguistic features
  - Cue words
  - Contextual features(closest noun phrase and conjunctive adverb)
  - Location feature(position and section)
- Citation based summarization.

Kaggle Competition Jupyter Notebook review

Coleridge Initiative Jupyter Notebook Authored by Prashan Dixit

The objective of the competition is to identify the mention of datasets within scientific publications. Your predictions will be short excerpts from the publications that appear to note a dataset.

Submissions are evaluated on a Jaccard-based FBeta score between predicted texts and ground truth texts, with Beta = 0.5 (a micro F0.5 score). Multiple predictions are delineated with a pipe (|) character in the submission file.

Kaggle Competition Jupyter Notebook review

The model used:

spaCy NLP

- This extracts language that indicates a dataset is cited in the text.
- The test data set is really small.
- I am not sure I understand the Jaccard FBeta scoring system but it looks like it measures the F score adjusted by a predetermined beta value.