In-text citation analysis based on NLP and ML techniques

Xidan Kou

In-text Citation Analysis

- Citation Content Analysis – semantic content
- Citation Context Analysis

What is Citation Analysis?

Relationships between cited and citing publications

Consider both qualitative and quantitative factors

Applications of Citation Analysis

- Measure articles' impact qualitative and quantitative
- Acquire in-depth understanding for scientific literature context
- Improve searching algor
- Automatic summarization in scientific publications

Citation Analysis Methods

Traditional Method: bibliometric metadata

Advanced Method: NLP, ML, deep learning,...

Advanced Citation Analysis Methods

- N-grams
- Bag-of words
- word2vector

NLP

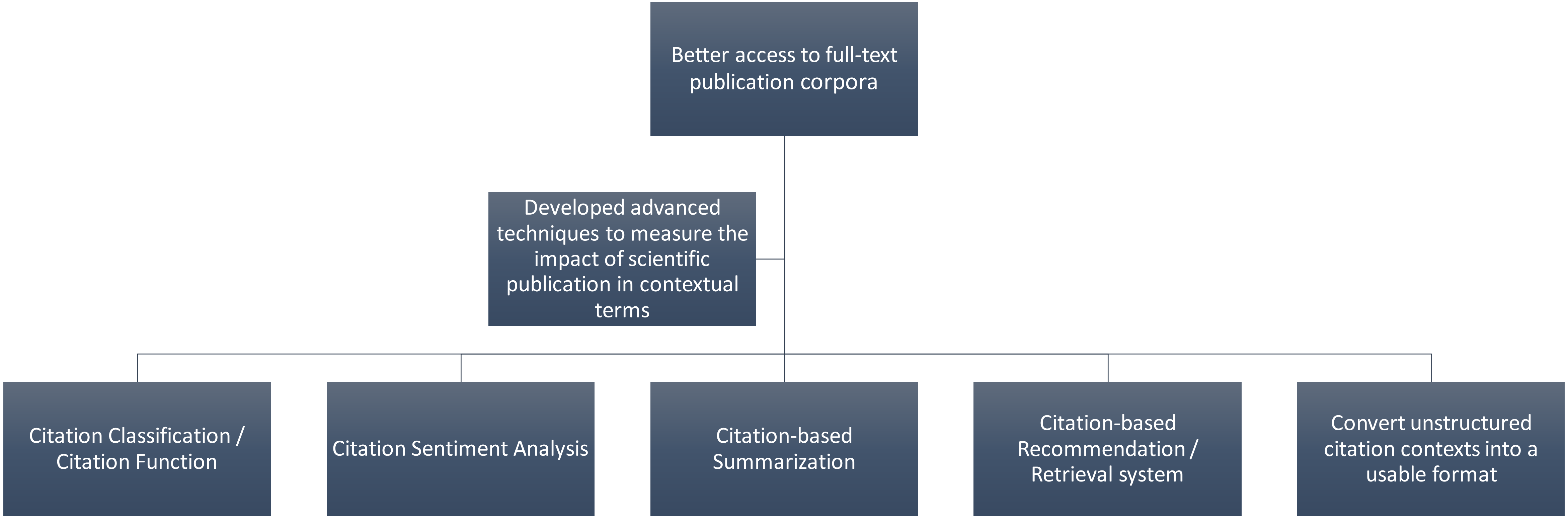
- SVM
- NB
- MaxEnt
- DT
- KNN
- LR

ML

- ANN
- CNN
- RNN
- LSTM

Deep Learning

How is Advanced Methods used in Citation Analysis?



Part 1 - In-text Citation Analysis

- Citation context window size
- Feature extraction for context identification
- In-text citation distribution
- Citations’ role according to position

Citation Context Window size

- Ritchie used ML techniques defined 9 categories of citation context

None	No citation context
1sent	Contains only the citing sentence
3sent	The citing sentence + one sentence before + one sentence after
1sentupto	Contains one sentence context, truncated at the next citation
3sentupto	Contains three sentence context, truncated at the next citation
Win50	Contains 50 words on the left and right of a citation
Win75	Contains 75 words on the left and right of a citation
win100	Contains 100 words on the left and right of a citation
full	Contains the full citing paper

- Performance: 3sent > 1sent > 1sentupto , 3sentupto > win75, win 100 > win 50
- 4sentence/quasi norm : The citing sentence + one sentence before + two sentence after

Feature extraction for context identification

- Citation feature is important in automatic context identification
- Abu-Jbara, Ezra, and Radev computed 7 features by CRF; achieved the highest accuracy; should be used for future studies

Feature	Example
Demonstrative determiners	This, that, these
Conjunctive adverbs	However, accordingly, furthermore
Position	Position of current sentence with respect to the citation
2-3 g	The fist bi-gram and tri-gram in the sentence contains references other than the target.
Contains mention of the target reference	Contains a mention of the target reference
Multiple references	The citing sentence contains multiple references
contains closest noun phrase	Contains none phrase(method, corpus, tool)

In-text Citation Distribution

- Problem: Research are conducted with different datasets and sample size.

Distribution of citations [IMRaD]	Recurring Citations	Multiple in-text reference (MIR) and their location
• 41.8% in Introduction • 25.2% in Methods • 25.9% in Results • 7% in Discussion	• 74.3% of citations were cited only 1 time • 25.7% of citations were cited >= 2 times • Citation location analysis: Most cited reference was cited in a similar section • Citation Centext Analysis: first-time citations are perfunctory. Succeeding citations were more purposeful	• MIR frequently appear in all section • MIR are mostly found near verbs

- Finding: Large proportion of citation in Methods indicates Methodology Paper
  Even citation across sections indicates Review paper

Citation Role according to Position

- Terms' or verbs' frequencies appearing in the citation contexts in IMRaD structure
- Following shows the researches' contribution

Aljaber

- Citation context is a rich source of topically related terms
- Many terms in citing paper are semantically related to terms in cited paper
- The section/location of the citation term is related to its quality.

Bertin and Atanassova

- 50% of the verbs alone in 'Introduction' section
- 'show' is the most common word in both the 'Introduction' and 'discussion'

Fujiwara and Yamamoto

- Developed web-based search system / Coli system – extracting citation context

Small in biomedical domain

- Presented the top 10 words in citance that associated with scientific discoveries paper

Part 2 - Citation Classification

Automatic classification of citations

Feature extraction for citation classification

Role of linguistic features for classification

Important VS. non-important citation

Feature extraction for citation classification

- Following shows the accuracy of Automatic citation classification
- More classes achieves better accuracy

Researcher	Number of annotation categories used	Model Used	Accuracy achieved	Findings
Teufel et al.	12 classes	IBK	77%	• Presented the 12 classes annotation scheme • 'PMot' appears near to the beginning of the publications • Comparative classes (CoCoR-, CoCoR0) appear near the end of publications
Abu-Jbara et al.	6 classes	SVM	70.5%	• Combined 12 class to 6 classes • Pointed out the importance of structural and lexical features for the citation classification
Jha et al.	6 classes	SVM NB,LR	70.5%	• Used the same classes as Abu-Jbara • Had same result as Abu-Hbara

- See Appendix 1 for 12 classes annotation classes

Feature extraction for citation classification

- Attributes/ features in determining Citation classification / function
- Accuracy in feature extraction

Researcher	Features used	Macro F1 Value	Findings
Siddharthan and Teufel	• Scientific attribution • Lexical • Linguistic • Position-based	51%	• Adding scientific attribution features to the model only increase 2% accuracy
Dong and Schafer	• Texual (cue words) • Physical (citation location and density) • Syntactic features	64%	• Citing sentence that describes background of current work is in active voice • Citing sentence introduces the tools and methods is in passive voice
Jochim and Schutze	• Lexical • Word-level linguistic • linguistic structure • Location • Frequency • Sentiment • Self- reference • Named-entity-recognition	65%	• Lexical feature alone achieves 61% accuracy

Role of linguistic features for classification

Researcher	Features used	Model Used	Accuracy
Agarwal et al.	• Uni-grams • Bi-grams	SVM and MNB	92.2%
Sugiyama et al.	• Proper nouns • Previous and next sentence • Uni-gram • Bi-gram	SVM and MaxEnt	88.2%
Wang et al.	• 48 groups of Cue phrase	• High number of cue phrase identifies better than low number of cue phrase	
Small	• Hedging words ("using")	• Word "using" has higher predictability for method and non-method section	

- SVM model outperform other classifier in determining citations accurately

Important VS. non-important citation

Researchers focuses on two functions of citation(Important and non-important citation) instead of various citation functions

- Important Citation: Citations that extend or use the cited work in meaningful way
- Non-Important Citation: Citations that used in the literature review section

Researchers	Number of Features used	Model Used	Accuracy	Findings
Valenzuela at al.	12 features	RF	80%	• 85.4% of the citations were incidental • 14.6% were important
Hassan et al. (2017)	6 features	NB, KNN, SVM, RF, DT	84%	• RF outperforms
Pride and Knoth	52 features	• Combination of total # of direct citations, author overlap, and abstract similarity led to better classification results		
Hassan,Imran et al. (2018)		SVM,RF, LSTM	92.5%	• LSTM outperforms, achieved 92.5% accuracy

Part 3 - Citation-based sentiment analysis

Classification of citation into positive, negative, neutral classes

- Context window selection for sentiment classification
- Role of linguistic features for sentiment classification
- Influential features for sentiment classification
- Class unbalancing in sentiment classification

Context window selection for sentiment classification

- Two studies conducted by Athar and Teufel

Four-sentences context as classification features

- Should be favored in sentiment classification
- Tried two methods: using merged text VS. separate text
- separate text as feature outperforms
- Annotation of 4 classes annotation scheme- positive, negative, neutral, exclude

N-grams and dependency as classification features

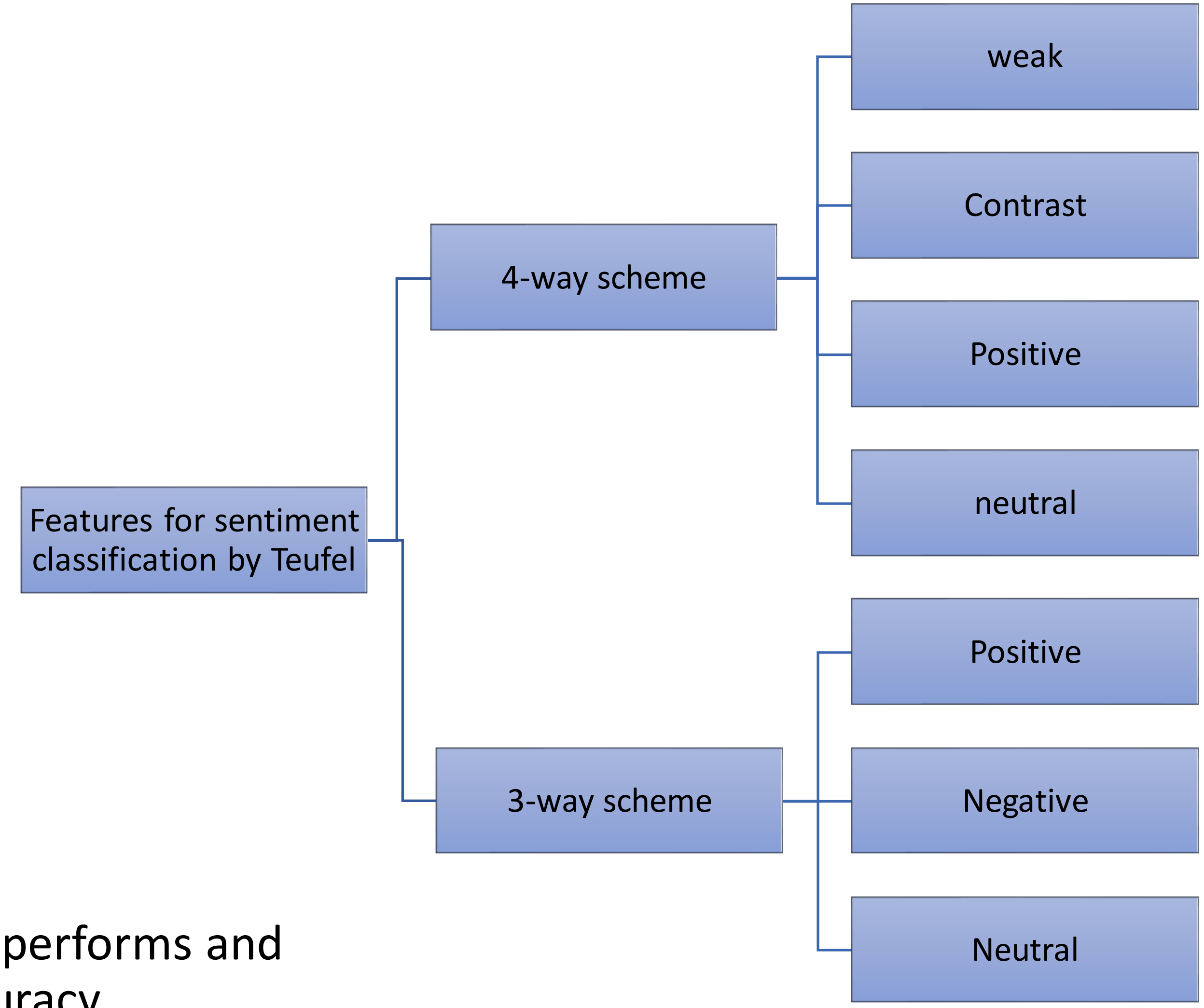
- Single sentence feature results in a loos of sentiment, due to lesser information

Role of linguistic features for sentiment classification

Researcher	Features used	Accuracy	Findings
Athar	Novel features(n-grams, dependency relation, scientific lexicon, sentence splitting)	89%	• Tri-grams and dependency features are the best; achieved the highest accuracy • Found by Ikram et al.: higher value (n = 5) of n-grams yields 2% better
Abu-Jbara	Features in Appendix 2	74%	Features associated with subjectivity outperforms

Citation polarity tools	F-1 Score
SEMANTRIA	96%
THEYSAY	85.91%

Influential features for sentiment classification



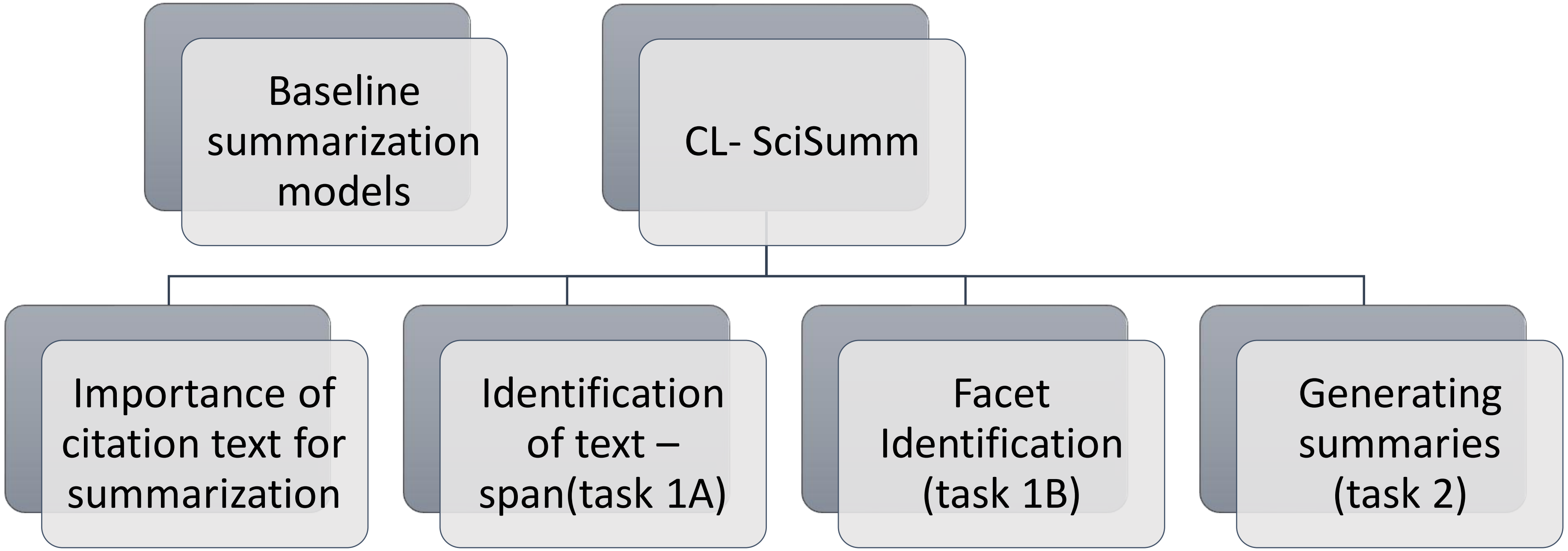
3-way scheme outperforms and achieved 83% accuracy

Class unbalancing in sentiment classification

Problem: unbalancing data between of positive, negative, neutral classes

To Solve: eliminate neutral class; increase dataset; provide balancing data;

Part 4 - Citation- based summarization



Baseline summarization models

Summarization system

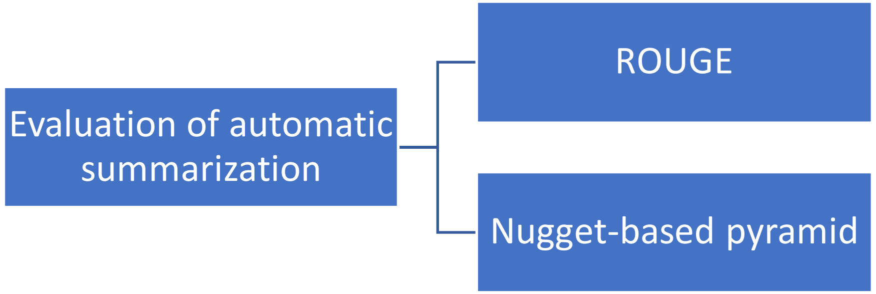
- MEAD
- LexRank

Steps of automatic summarization by Abu-Jbara and Radev

1. Reference tagging, context identification, sentence filtering
2. Extracted representative sentence were classified, similar sentence were added into classter, and the LaxRank value of each sentence was computed
3. The sentence was added into a summary based on the sentence ranking of cluster and LaxRank values.

Importance of citation text for summarization

- Teufel argues that citations include valuable subjective assessments of cited publications.
- Summaries generated from abstract perform better
- Without abstract, citance is a good substitute for automatic summarization
- Citation context have unique information can be used to improve the summarization result
- Use of citance can improve executive summaries of publications



Identification of text –span(task 1A)

- See Appendix 3 for task description

Researcher	Feature used	Method used	Accuracy/Findings
Kaplan rt al.	Cosine similarity	• Co-reference • Chain-based • Baseline 1: extracts only citance • Baseline 2: extract sentence before and after the citance	84% But small data, so less representative
Qazvinian and Radev	Lexical similarities	• Probabilistic inference	Uses of four sentences on each side of citance improved the pyramid score considerably
Nomoto		• TF-IDF • ANN	ANN > TF-IDF
Klampfl et al.		• TextSentenceRank • Tsr-sent-class • Sect-class-tsr	TextSentenceRank outperforms on extracting most relevant key tems and sentences

Facet Identification (task 1B)

- Issue: Class unbalancing(60% of text span in method section. 9% in hypothesis section)
- Studies considered facet identification problem as text-classification problem.
- Cao et al. stated facet identification problem is multi-label classification task
- NB and SVM tends to outperform
- Similarity-based features are more suitable than position-based features
- With class-inbalancing data, TF-IDF similarities and IDF similarities are robust features.

Generating summaries (task 2)

Researcher	Method used	Data	Findings
Mei and Zhai	LM: scoring matches between queries and documents	Small dataset: 14 articles from MEDLINE	Language-based model performs better than conventional summarization techniques
Tando and Jain	• LM • Used opinion vocabulary with uni-gram and bi-gram to describe cited publications' opinion	30 articles from MA search engine	Combination of adjectives, verbs, and bi-grams models beats the accuracy of the LM model.
Barrera and Verma		DUC 2002 and scientific magazine articles	Semantic linkage and topic-heading relevance produce useful summaries.
Conroy and Davis	• Vector-space model • Non-negative matrix factorization model		Non-negative matrix factorization model improves ROUGE scores.
Yasunaga et al.	Hybrid model (citation based & abstraction based)	100 sample article	Hybrid model performs better than single.

Part 5 - Citation Recommendation system

Match user queries with existing publications and recommend publications that could be cited.

- The development of citation recommendation system used two techniques
  - Topic modeling
  - Neural Networks

Topic modeling

- Data Sparsity and noise issue should be solved

Researcher	Model used	Findings
Nallapati et al.	Link-PLSA-LDA (Combination of PLSA and LDA)	Link-PLSA-LDA performs better than PLSA and LDA
Tang and Zhang	RBM-CS	RBM-CS performs better than LM
He et al.	Developed the system to recommend for citing with similarity scores	The system outperforms the uni-gram, bi-grams, dependency model.
Wang and Blei	A collaborative topic regression model that combines the merits of probabilistic topic modeling and traditional collaborative filtering	Was able to make relatively useful recommendations

Neural Networks

Use LSTM to improve recommendation system

- Learns the embedding of citation contexts
- Measures their relevance
- Documents of the highest relevance

Neural networks techniques outperform topic modeling techniques in citation recommendation system

Part 6 - Challenges associated with in-text citation

Citation context contains multiple references, but only part may be relevant for focal publications

Most of the publications in this research area used data from the ACL Anthology

Some access but still limited access to full-text datasets due to copyright restrictions.

The End

Appendix 1

Table 7 Annotation scheme for citation function (Teufel et al., 2006: 105)

Category	Description
Weak	Weakness of the cited approach
CoCoGM	Contrast/Comparison in goals or methods (neutral)
CoCo-	Author's work is stated to be superior to cited work
CoCoR0	Contrast/Comparison in results (neutral)
CoCoXY	Contrast between two cited methods
PBas	Author uses cited work as the basis or starting point
PUse	Author uses tools/algorithms/data/definitions
PModi	Author adapts or modifies tools/algorithms/data
PMot	This citation is positive about the approach used or problem addressed (used to motivate work in the current paper)
PSim	Author's work and cited work are similar
PSup	Author's work and cited work are compatible/provide support for each other
Neut	Neutral description of cited work, or not enough textual evidence for above categories, or unlisted citation function

Appendix 2

Table 10 Features used for analysing citation purposes and polarity (Abu-Jbara et al., 2013: 601)

Feature	Description
Reference count	Number of references that appear in the citation context
Is separate	Whether the target reference appears within a group of references or separate (i.e., single reference)
Closest verb/ adjective/adverb	The lemmatized form of the closest verb/adjective/adverb to the target reference or its representative or any mention of it. Distance is measured based on the shortest path in the dependency tree
Self-citation	Whether the citation from the source paper to the target reference is a self-citation
Contains 1st/3rd person pronoun	Whether the citation context contains a first/third-person pronoun
Negation	Whether the citation context contains a negation cue
Speculation	Whether the citation context contains a speculation cue
Closest subjectivity cue	The closest subjectivity cue to the target reference or its representative or any anaphoric mention of it
Contrary expressions	Whether the citation context contains a contrary expression
Section	The heading of the section in which the citation appears
Dependency relations	All the dependency relations that appear in the citation context

Appendix 3

CL-SciSumm Task

The task is defined as follows:

- **Given:** A topic consisting of a Reference Paper (RP) and Citing Papers (CPs) that all contain citations to the RP. In each CP, the text spans (i.e., citances) have been identified that pertain to a particular citation to the RP.
- **Task 1A:** For each citance, identify the spans of text (cited text spans) in the RP that most accurately reflect the citance. These are of the granularity of a sentence fragment, a full sentence, or several consecutive sentences (no more than 5).
- **Task 1B:** For each cited text span, identify what facet of the paper it belongs to, from a predefined set of facets.
- **Task 2 (optional bonus task):** Finally, generate a structured summary of the RP from the cited text spans of the RP. The length of the summary should not exceed 250 words.

Wording Explanation

Phrase	Explanation	Phrase	Explanation
Citation Context	Text near citation in citing paper	TF-IDF	Weighting scheme scheme to evaluate the importance of a word for a document in a corpus
Citance	The citing sentence	TextSentenceRank	Graph-based ranking algorithm to extract key sentences or key terms
N-grams	A sequence of N words	MEDLINE	Medical literature database
Macro F1 score	assess the quality of problems with multiple binary labels or multiple classes	C-cite	Citation cite(block of text that contains both citation and its context)
Polarity Relation	Attitudes of authors' approval or disapproval for the work they cited		