What Sentence are you Referring to and Why? Identifying Cited Sentences in Scientific Literature

Ahmed Abura'ed
Universitat Pompeu Fabra
Large Scale Text Understanding Systems Lab
TALN / DTIC
Barcelona, Spain
ahmed.aburaed@upf.edu

Luis Chiruzzo
Universidad de la República
Facultad de Ingeniería
Instituto de Computación
Montevideo, Uruguay
luischir@fing.edu.uy

Horacio Saggion
Universitat Pompeu Fabra
Large Scale Text Understanding Systems Lab
TALN / DTIC
Barcelona, Spain
horacio.saggion@upf.edu

Abstract

In the current context of scientific information overload, text mining tools are of paramount importance for researchers who have to read scientific papers and assess their value. Current citation networks, which link papers by citation relationships (reference and citing paper), are useful to quantitatively understand the value of a piece of scientific work, however they are limited in that they do not provide information about what specific part of the reference paper the citing paper is referring to. This qualitative information is very important, for example, in the context of current community-based scientific summarization activities. In this paper, and relying on an annotated dataset of co-citation sentences, we carry out a number of experiments aimed at, given a citation sentence, automatically identify a part of a reference paper being cited. Additionally our algorithm predicts the specific reason why such reference sentence has been cited out of five possible reasons.

1 Introduction

Scientific publications such as scientific papers are some of the most valuable sources of human knowledge, containing information as relevant as how to cure diseases, create medicaments or construct useful life-saving machines; they are permanent records of what has been discovered so far (Kuhn, 1970). The amount of scientific publications is growing at unprecedented rates (Bornmann and Mutz, 2015; Saggion and Ronzano, 2016) with recent estimates indicating that a new research paper is published every 13 seconds. Scientific research needs awareness of what has been discovered or published before us to progress: as Newton has put it, "If I have seen further it is by standing on the shoulders of Giants."

The value of a scientific article is sometimes assessed by the number of publications citing the article, although this way of measuring research is well established, it is only quantitative, leaving qualitative considerations out of the picture. Although over the years several works have been interested in qualitative aspects related to paper citations (Moravcsik and Murugesan, 1975; Spiegel-Rosing, 1977; Teufel et al., 2006; Abu-Jbara et al., 2013; Athar, 2011; Shotton, 2010; Fisas et al., 2016; Valenzuela et al., 2015), recently, in the context of the Computational Linguistics Scientific Summarization Challenge (Jaidka et al., 2016) (CL-SciSumm hereafter), one task was identified as relevant to citation analysis: given a scientific paper – the reference, and another paper citing it, the task is to determine what part of the reference the citing paper is referring to, and also what particular aspect is being cited (e.g. the aim of the paper? the method?). In this paper we investigate how to address the above task and present a system which uses unsupervised sentence similarity metrics and supervised machine learning to address the challenge.

The contributions of this paper are as follows:

- A thorough comparison of several sentence matching algorithms relying on discrete and continuous word representations;
- An state-of-the-art method for matching citing sentences to cited sentences in scientific literature; and
- A trainable method for identifying citing facets which, in terms of precision, achieves competitive performance.

Proceedings of Recent Advances in Natural Language Processing, pages 9–17,
Varna, Bulgaria, Sep 4–6 2017.
https://doi.org/10.26615/978-954-452-049-6_002

2 Related Work

Recent studies proposed to take advantage of the scientific papers citation network mainly to approach scientific literature summarization. This section discusses some of the most related studies that tried to identify which reference paper sentences have been cited and also tried to identify the discourse facet of the reference sentence. Nomoto et al. (2016) aimed to detect a part of the reference paper that is most related to a given citation made by a citation paper through producing a hybrid model consisting of TFIDF and Neural Network (NN), he used the Neural Network to adapt the embedding model used in the question answering domain to provide a scoring function and built the TFIDF part based on the test data of the CL-SciSumm 2016 corpus. The training data, comprising of triples of citance, the true reference and the set of false references for the citance, was used to perform a Stochastic gradient descent search while sentence selection was based on a dissimilarity score (i.e. MMR). Li et al. (2016) used a Support Vector Machines (SVM) classifier to identify the spans of text in the reference paper matching each citance. They also used a combination of methods such as: idf similarity, Jaccard similarity and context similarity leading into the following methods: Sentence fusion, Jaccard Cascade method, Jaccard Focused method, SVM method, etc. Cao et al. (2016) considered the problem of identifying the reference text sentences that most precisely reflect the citance as a ranking problem which they modeled using SVM Rank, they also used a decision tree classifier to identify the facet that a citance belongs to. Moraes et al. (2016) used cosine similarity based on TFIDF weights for sentences with multiple incremental modifications and SVMs with a tree kernel. Moreover, the best results were obtained by cosine similarity. Di Iorio et al. (2013) presented a tool named CiTalO to automatically deduce the nature of citations by combining techniques of ontology learning from natural language, sentiment-analysis, word-sense disambiguation, and ontology mapping. CiTalO extracts information about the nature, the motivations and the goals of each citation. Many configurations were applied with CiTalO including: filtered citations, in which all synsets of which WordNet's gloss is not aligned with the natural language description of the property in consideration were filtered out; sentiment, including sentiment polarity emerging from the text in which the citation is included. CiTalO was tested by comparing its results with a human classification of the citations. Finally, Saggion et al. (2016) presented two supervised approaches for identifying RP's text spans and facet classification. They trained the sentence matching approach using only reference sentences in the gold standard which proved to be uneffective in testing data.

3 CL-SciSumm Challenge and Corpus

In the CL-SciSumm 2016 evaluation (Jaidka et al., 2016) participants were given a set of clusters, each one composed of n documents where one is a reference paper (RP) and the n-1 remaining documents are referred to as citing papers (CPs) since they cite the reference paper. Given this set-up, researchers were asked to address the following two tasks:

- **Task A**: For each citance in the CP (i.e., a reference to the RP), identify the spans of text (cited text spans) in the RP that most accurately reflect the citance.
- **Task B**: For each cited text span, identify what **facet** of the paper it belongs to, from a predefined set of facets, namely: *Aim*, *Hypothesis*, *Implication*, *Results* or *Method*.

There was an additional summarization task which we do not address here due to space constraints.

3.1 Dataset

In the experiments to be presented in this paper, we have relied on the CL-SciSumm 2016 corpus which provides training, development and testing data arranged in clusters of reference paper and the papers citing it. The corpus also contains text files that represent the gold manual annotations indicating the facet and the text span(s) in the reference paper that best represent each citance. Figure 1 shows an example of one line of the gold manual annotations provided by the CL-SciSumm organizers.

4 Text Processing

Each document in the clusters was annotated using processing resources from the following freely available tools: GATE system (Maynard et al., 2002), the SUMMA library (Saggion, 2008) and

Citance Number: 60 | Reference Article: J96-3004.xml | Citing Article: W12-1011.xml | Citation Marker Offset: ['41'] | Citation Marker: Sproat et al., 1996 | Citation Offset: ['41'] | Citation Text: <S sid ="41" ssid = "5">Indeed, even native speakers can agree on word boundaries in modern Chinese only about 76% of the time (Sproat et al., 1996).</S> | Reference Offset: ['325'] | Reference Text: <S sid ="325" ssid = "34">The average agreement among the human judges is .76, and the average agreement between ST and the humans is .75, or about 99% of the interhuman agreement</S> | Discourse Facet: Results_Citation | Annotator: Ankita Patel |

Figure 1: One line example of the gold manual annotation provided by the CL-SciSumm 2016 corpus.

the Dr Inventor library (Ronzano and Saggion, 2015).

The GATE system was used to tokenize, sentence split, part of speech tag, manage gazetteers and lemmatize each document. Teufel's (Teufel et al., 2000) action and concept Lexicons were used to create gazetteers lists to identify in text scientific concepts (e.g. *research*: analyze, check and gather; *problem*: violate, spoil and mistake, and *solution*: fix, cure and accomplish). The Dr Inventor's library for analysing scientific documents was additionally applied to each document to generate rich semantic information such as citation marker, BabelNet concepts (Navigli and Ponzetto, 2012), causality markers, co-reference chains, and rhetorical sentence classification. The library classifies each sentence of a paper based on a rhetorical category of scientific discourse among: Approach, Background, Challenge, Outcome and FutureWork. In other words, it predicts the probability of the sentence of belonging to one of the five discourses provided. See (Fisas et al., 2016) for more details about the corpus used for training the classifier. Finally, the SUMMA library was used to produce term vectors, normalized term vectors, BabelNet synset ID vectors, normalized babelnet synset ID vectors, terms n-grams (up to two) and part of speech n-grams (up to two) for each document.

5 Matching Citations to Reference Papers

In this section, we present methods to, given a citation sentence, identify the sentence or sentences in the reference paper being referred to. First, we present a method based on current continuous word representations. This method did not provide the expected performance. Then, we present a method based on more traditional bag-of-words representations which performed reasonably well when compared to results from the CL-SciSumm Challenge.

5.1 Sentence Embeddings

Word embeddings are continuous vector representations for words. This technique tries to map a set of words into a set of n-dimensional continuous vectors, where the dimension n is much lower than the number of words in the vocabulary. There are a number of ways for training such word representations, for example the word2vec algorithms Continuous Bag of Words and Skip-gram (Mikolov et al., 2013) are amongst the most popular ones. In our experiments we used a pre-trained word embeddings collection trained over 100 billion words from the Google News dataset[1]. It comprises 3 million words and phrases embedded in a 300 dimensional space. We also used two word embedding models (dimensions 100 and 300) trained with scientific papers text[2](Liu, 2017) from the ACL Anthology Reference Corpus (Bird, 2008). We performed experiments using only Google News vectors (GN), only ACL vectors of size 100 (ACL100) or 300 (ACL300), and also using the concatenation of GN and ACL100 or ACL300 vectors.

Having the embeddings for the words it is possible to generate embeddings for larger units like phrases or sentences. There are several ways of creating sentence embeddings, in our experiments we averaged the word embeddings for each word in the sentence. This is a very simple approach that has nonetheless given good results to problems such as extractive summarization (Kågebäck et al., 2014) and semantic classification (White et al., 2015).

Using these sentence embeddings, we tried to identify the sentences in the reference paper (RP) that more accurately reflect the citance by calculating the sentence embedding for each citance sentence (sentences from the citing paper) and the sentence embedding for each RP sentence. Then

[1]https://code.google.com/archive/p/word2vec/

[2]https://github.com/liuhaixiachina/Sentiment-Analysis-of-Citations-Using-Word2vec/tree/master/trainedmodels

return the most similar sentences from the RP to any of the citance sentences according to cosine similarity.

We used the `gensim` library (Řehůřek and Sojka, 2010) for working with word embeddings. Table 1 summarizes the performance of these experiments. The number following the experiment name indicates the number of retrieved sentences we considered for each experiment, we tried retrieving different number of sentences (2, 5, 8 and 10) and optimized against the development corpus for the best F1 score. The best result has 0.101 of F1 score averaged over the test corpus documents, and was attained considering the top 2 sentences from the RP according to cosine similarity.

5.2 Words, BabelNet Concepts and Sentence Similarity Measures

We relied on two different sentence representations produced by the SUMMA library and two different text similarity measures. Sentences were represented either as word-vectors (W) or BabelNet-vectors (B) where each vector component (a word or a BabelNet synset) was weighted using a tf*idf weighting schema. For calculating the BabelNet vectors, we used the BabelNet service to get the list of synsets used in each sentence, and we calculated tf*idf over that. Where the similarity measures are concerned we used cosine similarity (C) or the Jaccard coefficient (J). Given a citation sentence in the CP, and a sentence in the RP, we computed four similarity values: the cosine similarity using word vectors (WC), the cosine similarity using BabelNet vectors (BC), the jaccard similarity using word vectors (WJ), and the jaccard similarity using BabelNet Vectors (BJ). We then performed another experiment using a modified version of the jaccard similarity (MJ) that takes into consideration the inverted frequency of words as well. For this experiment we calculated the tf*idf weighting schema over both training and development sets of documents, performing stemming of words and using only the first characters of every word so that words like "structure" and "structural" are considered the same token. The modified jaccard similarity between two sentences s_1 and s_2 is defined in equation 1, and it gives more weight to matching words that are infrequent in the corpus.

$$MJ(s_1, s_2) = \frac{\sum_{t \in s_1 \cap s_2} 2^{idf(t)}}{|s_1 \cup s_2|} \quad (1)$$

We optimize F-score on the training data searching for the best similarity threshold and the top number of sentences to retrieve for each citing sentence. Results for the test data are presented in Table 2. The table shows each configuration with the threshold used to retrieve sentences (i.e. similarity $>$ thr) and the number of top RP sentences retrieved for each CP citance. As it can be observed the best performance is achieved using the modified jaccard similarity. Our results improve the state of the art performance obtained in the CL-SciSumm 2016 evaluation for this task.

To better comprehend the results of our approach we did a comparison with the top four results obtained by participants at the CL-SciSumm 2016 challenge. See Table 3.

6 Identifying Citation Facets

In this section, we present experiments aiming at identifying the facet the cited text span belongs to. We modeled pairs of reference and citance sentences as a feature vector. Then, we used such pair representation to enable the training of classification algorithms tailored to determine whether a cited text span belongs to one out of five predefined facets: *Aim*, *Hypothesis*, *Implication*, *Results* or *Method*. We rely on the WEKA machine learning framework (Witten et al., 2016) as a tool to conduct our experiments. We first describe the set of features used to generate the feature vectors for instance representation to then describe the machine learning algorithms used.

6.1 Features

Sentence position: we use three features that are based on the location of the sentence in the document:

- Sentence position: the position of the sentence in the reference paper.
- Section Sentence position: the position of the sentence in the section of the paper.
- Facet position: five binary features indicating whether the sentence is in a section indicating one of the target facets. We designed a set of keywords to determine if a section title belongs to a given facets (e.g. the word "method" indicates a section dealing with the facet *Method*). In case any of the title's words belong to a given facet, the value of that facet feature will be 1 otherwise it will be 0.

Table 1: Performance of the sentence embedding experiments.

Method	Top	Avg. Precision	Avg. Recall	Avg. F-Measure
GN	2	0.079	0.13	0.096
ACL100	2	0.055	0.084	0.066
ACL300	2	0.074	0.117	0.089
GN+ACL100	2	**0.082**	**0.132**	**0.101**
GN+ACL300	2	0.081	0.129	0.099

Table 2: Performance of the word vectors and similarity metrics experiments on test data.

Method	Thr	Top	Avg. Precision	Avg. Recall	Avg. F-Measure
WC	0.0	8	0.025	0.179	0.044
BC	0.0	8	0.023	0.124	0.039
WJ	0.1	2	0.084	0.138	0.105
BJ	0.1	2	0.079	0.111	0.092
MJ	0.1	2	**0.123**	**0.201**	**0.151**

Text similarity: We rely on the cosine similarity between the pair of citing and reference sentence using word and BabelNet synsets vectors (see Section 5.2).

Rhetorical Category Probability Features: We mentioned in Section 4 that the Dr Inventor library predicts the probability of a sentence being in one of five possible categories (different from the CL-SciSumm task): Approach, Background, Challenge, Outcome and FutureWork. Even though Dr Inventor library has different mapping from our targeted discourse facets, we believe this information could be useful for classification. Therefore, We use such probabilities as features for citing and reference sentences. **Dr Inventor Sentence related features**: We utilize an additional set of features produced by Dr Inventor:

- Citation marker: three features to represent the number of citation markers in the reference sentence, citing sentence and the pair of sentences together.
- Cause and effect: two features to represent if the reference or citing sentence participates to the formulation of one or more causal relations by specifying the cause or the effect.
- Co-reference Chains: three features to represent the number of nominals and pronominals chained in the reference sentence, citing sentence and the pair of sentences together.

Scientific Gazetteer Features: As mentioned in Section 4 the documents were enriched with Teufel's action and concept Lexicon gazetteers lists producing the total of 58 lists. Each list is used to produce a feature which is the ratio of words in the sentence matching the list to the number of words in the sentence. The features are computed for the reference sentence, the citing sentence, and their combination, giving rise to 174 features.

Bag-of-word Features: eight string features are produced to represent the *bigram lemmas*, *POS-tags bigram*, *lemmas* and *POS-tags* for both the reference and the citing sentences.

7 Facets Experiments

After having the complete set of features based on the pair of sentences, we trained algorithms over the training dataset based on 432 training instances distributed as follows: Aim (72), Implication (26), Result (76), Hypothesis (1), Method (257). We evaluated the performance of several classification algorithms including: Support Vector Machines(SMO), Naive Bayes, IBK, Random Committee, Logistic and Random Forest. Then, we performed 10-fold cross validation experiments with the training data in order to decide which algorithm to use during testing. To achieve the best approach, we investigated feature selection on training data. We split the set of features by the top ten, twenty and thirty percent of features then we evaluated the classification algorithms on each set of features in addition to the whole training feature set. We choose the best three performing algorithms for each set of features to create our models. See Table 4 for a list of feature sets and the best three algorithm (10-fold cross validation).

8 Facets Classification Results

We performed nine system runs to identify the facets for each matched pair of reference and

Table 3: A comparison with the top four results by participants at CL-SciSumm 2016

Team	best Approach	Avg. Precision	Avg. Recall	Avg. F
Nomoto et al. (2016)	TFIDF + neural network,	0.091	0.111	0.100
Li et al. (2016)	Sentence Fusion + Jaccard Focused	0.082	**0.262**	0.125
Cao et al. (2016)	SVM Rank	0.088	0.131	0.103
Moraes et al. (2016)	TF-IDF+ST+SL	0.096	0.224	0.133
Our approach	MJ	**0.123**	0.201	**0.151**

Table 4: A list of the best three algorithm for each feature set based on 10-fold cross validation. Precision, Recall and F-measure results for each configuration are shown.

Features	Algorithm	Avg. Precision	Avg. Recall	Avg. F-Measure
top 10%	SMO	0.920	0.921	0.920
	Logistic	0.924	0.926	0.924
	Random Committee	0.882	0.880	0.874
top 20%	SMO	**0.936**	**0.938**	**0.936**
	Logistic	0.918	0.919	0.917
	Naive Bayes	0.876	0.877	0.875
top 30%	SMO	0.934	0.935	0.934
	Logistic	0.929	0.926	0.926
	IBK	0.891	0.889	0.889
all features	SMO	0.929	0.931	0.928
	Logistic	0.914	0.910	0.910
	IBK	0.908	0.907	0.907

citing sentences in the testing dataset; we used the gold annotations to identify the matched sentences, after identifying the facet, we produced the output annotations in a format consistent with the CL-SciSumm 2016 corpus. We evaluated each system by comparing its output with the gold annotations provided in the corpus. Table 5 presents the evaluation results of the nine system runs based on models generated from different features sets. What can be noticed from the evaluation result is that the best performing system is the one using the Support Vector Machines (SMO) using the whole feature set.

However, this evaluation considers an *ideal scenario* in which the facet classifier is invoked for gold standard matches. Therefore, we also evaluated the Support Vector Machines (SMO) classifier used to identify the facets over each pair of matched sentences according to the best matching system described in Section 5.

Finally, to improve our facet system, we retrained its model after adding information from the development dataset making the total of 897 training instances distributed as follows: Aim (145), Implication (71), Result (168), Hypothesis (19), Method (494) and combining each pair of reference and citing sentences togother before producing the Bag-of-word Features making the total number of string features to four instead of eight. The merging between the reference and citing sentences is motivated by representing each pair of reference and citing sentences as one entity instead of having two separate entities on each part of the pair. The evaluation can be seen in Table 6.

In comparison with CL-SciSumm 2016 challenge teams our approach scores directly after the best team's results. Please see Table 7 for such comparison.

What can be noticed at Table 7 is that our approach achieves higher precision than Li et al. approach but their approach has higher recall outperforming our system as a result. However, we believe that there are two reasons for such contrast: First, Li et al. used Jaccard Focused method for their matches system and then applied a Voting method which combines the results from three methods (Subtitle Rule, High Frequency Word and SVM classifier) to achieve the best results with the most votes for Facet Identification. On the other hand, our system did not use a voting mechanism instead we used a set of features including features which are equivalent to their methods (except their High Frequency Word feature). Second, our approach performed poorly over some of the testing clusters due to some noise in the gold annotations, such as a higher number of references to the title than expected (cluster P98-1046: 9 out of 31 annotations). We also show a set of results that exclude some of the most problematic clusters.

Table 5: Evaluation results of the nine system runs over the gold standard matched pair of sentences based on models generated from different features sets

Features	Algorithm	Avg. Precision	Avg. Recall	Avg. F-Measure
top 10%	SMO	0.7081	0.6243	0.6618
	Logistic	0.5263	0.4638	0.4916
	Random Committee	0.7294	0.6463	0.6835
top 20%	SMO	0.7458	0.6603	0.6985
	Logistic	0.5520	0.4889	0.5171
	Naive Bayes	0.5422	0.4766	0.5057
top 30%	SMO	0.7494	0.6634	0.7019
	Logistic	0.5263	0.4689	0.4946
	IBK	0.5131	0.4493	0.4777
all	SMO	**0.7519**	**0.6657**	**0.7043**
	Logistic	0.3223	0.2797	0.2986
	IBK	0.5889	0.5162	0.5486

Table 6: Evaluation results of the improved best SMO system over the matched pair of sentences based on the best performing match system

Features	Algorithm	Avg. Precision	Avg. Recall	Avg. F-Measure
all	SMO	0.635	0.155	**0.242**

Table 7: A comparison with the top team (best results) at the CL-SciSumm 2016 applied on the testing dataset

System	Testing Clusters	Algorithm	Avg.Precision	Avg.Recall	Avg.F-Measure
Li et al.	all	Jaccard Focused	0.581	0.230	0.314
Our approach	all	SMO	0.635	0.155	0.242
Our approach	excluding worst cluster	SMO	0.683	0.168	0.263
Our approach	ex. worst two clusters	SMO	0.738	0.185	0.288
Our approach	ex. worst three cluters	SMO	0.748	0.200	0.308

9 Conclusion

In this paper, we have presented several unsupervised sentence similarity metrics used to identify the sentences in a reference paper that most accurately reflect a citing sentence in a citing paper, and a supervised machine learning system to identify the facet that sentence belongs to, from a predefined set of facets. We used the CL-SciSumm 2016 corpus in which we utilized text processing and summarization tools to enrich the corpus with annotations which were used to compute the features for the supervised machine learning system as well as the vectors used by the unsupervised sentence similarity metrics. Our systems performed well when compared to participants of the CL-SciSumm 2016 challenge; for matches identification we obtained 0.151 F-score, an improvement over the best system so far which achieved 0.134 F-score. For facet identification we obtained 0.704 F-score in cross-validation experiments and 0.242 F-score when the facet identification system is combined with the sentence matching algorithm. Although our F-score result was lower than the best system which got 0.314, our system achieved better Precision than the best system. Our facet approach performed poorly over some of the testing clusters due to errors in the gold annotations. These errors could be attributed to noisy OCR output from processing the original PDF files or to the fact that a single annotator was responsible for each document cluster. We plan to extend this work by addressing the summarization task and also comparing how our models work on better curated datasets.

Acknowledgments

This work is (partly) supported by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502) and by the TUNER project (TIN2015-65308-C5-5-R, MINECO/FEDER, UE).

References

Amjad Abu-Jbara, Jefferson Ezra, and Dragomir R Radev. 2013. Purpose and polarity of citation: Towards NLP-based bibliometrics. In *HLT-NAACL*. pages 596–606.

Awais Athar. 2011. Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 Student Session*. Association for Computational Linguistics, Stroudsburg, PA, USA, pages 81–87.

Steven Bird. 2008. The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics .

Lutz Bornmann and Rüdiger Mutz. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *JASIST* 66(11):2215–2222.

Ziqiang Cao, Wenjie Li, and Dapeng Wu. 2016. PolyU at CL-SciSumm 2016. In *BIRNDL@ JCDL*. pages 132–138.

Angelo Di Iorio, Andrea Giovanni Nuzzolese, and Silvio Peroni. 2013. Towards the automatic identification of the nature of citations. In *SePublica*. pages 63–74.

Beatríz Fisas, Francesco Ronzano, and Horacio Saggion. 2016. A multi-layered annotated corpus of scientific papers. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016.*.

Kokil Jaidka, Muthu Kumar Chandrasekaran, Sajal Rustagi, and Min-Yen Kan. 2016. Overview of the CL-SciSumm 2016 shared task. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016.*. pages 93–102.

Mikael Kågebäck, Olof Mogren, Nina Tahmasebi, and Devdatt Dubhashi. 2014. Extractive summarization using continuous vector space models. In *Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC)@ EACL*. Citeseer, pages 31–39.

Thomas S. Kuhn. 1970. *The structure of scientific revolutions*. University of Chicago Press, Chicago.

Lei Li, Liyuan Mao, Yazhao Zhang, Junqi Chi, Taiwen Huang, Xiaoyue Cong, and Heng Peng. 2016. CIST system for CL-SciSumm 2016 shared task. In *BIRNDL@ JCDL*. pages 156–167.

Haixia Liu. 2017. Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177* .

Diana Maynard, Valentin Tablan, Hamish Cunningham, Cristian Ursu, Horacio Saggion, Kalina Bontcheva, and Yorick Wilks. 2002. Architectural elements of language engineering robustness. *Natural Language Engineering* 8(2-3):257–274.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *ICLR Workshop* .

Luis Moraes, Shahryar Baki, Rakesh Verma, and Daniel Lee. 2016. University of Houston at CL-SciSumm 2016: SVMs with tree kernels and sentence similarity. In *BIRNDL@ JCDL*. pages 113–121.

Michael J Moravcsik and Poovanalingam Murugesan. 1975. Some results on the function and quality of citations. *Social studies of science* 5(1):86–92.

Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.* 193:217–250.

Tadashi Nomoto. 2016. Neal: A neurally enhanced approach to linking citation and reference. In *BIRNDL@ JCDL*. pages 168–174.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, pages 45–50. http://is.muni.cz/publication/884893/en.

Francesco Ronzano and Horacio Saggion. 2015. Dr. Inventor Framework: Extracting structured information from scientific publications. In *International Conference on Discovery Science*. Springer, pages 209–220.

Horacio Saggion. 2008. SUMMA: A robust and adaptable summarization tool. *Traitement Automatique des Langues* 49(2):103–125.

Horacio Saggion, Ahmed AbuRa'ed, and Francesco Ronzano. 2016. Trainable citation-enhanced summarization of scientific articles. In *Proceedings of the Joint Workshop on Bibliometric-enhanced Information Retrieval and Natural Language Processing for Digital Libraries (BIRNDL) co-located with the Joint Conference on Digital Libraries 2016 (JCDL 2016), Newark, NJ, USA, June 23, 2016.*. pages 175–186.

Horacio Saggion and Francesco Ronzano. 2016. Natural language processing for intelligent access to scientific information. In *COLING 2016, 26th International Conference on Computational Linguistics, Tutorial Abstracts, December 11-16, 2016, Osaka, Japan*. pages 9–13.

D. Shotton. 2010. CiTO, the citation typing ontology. *Journal of Biomedical Semantics* 1.

Ina Spiegel-Rosing. 1977. Science studies: Bibliometric and content analysis. *Social Studies of Science* 7(1):97–113.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, Stroudsburg, PA, USA, SigDIAL '06, pages 80–87.

Simone Teufel et al. 2000. *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Marco Valenzuela, Vu Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

Lyndon White, Roberto Togneri, Wei Liu, and Mohammed Bennamoun. 2015. How well sentence embeddings capture meaning. In *Proceedings of the 20th Australasian Document Computing Symposium*. ACM, page 9.

Ian H Witten, Eibe Frank, Mark A Hall, and Christopher J Pal. 2016. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.