Experiments in Detection of Implicit Citations

Ahmed Abura'ed, Luis Chiruzzo, Horacio Saggion
Universitat Pompeu Fabra Universidad de la República Universitat Pompeu Fabra
Barcelona, Spain Montevideo, Uruguay Barcelona, Spain
ahmed.aburaed@upf.edu, luischir@fing.edu.uy, horacio.saggion@upf.edu

Abstract

The identification of explicit and implicit citations to a given reference paper is important for numerous scientific text mining activities such as citation purpose identification, scientific opinion mining, and scientific summarization. This paper presents experiments on the identification of implicit citations in scientific papers. As in previous work, and relying on an annotated dataset of explicit and implicit citation sentences, we cast the problem as classification, evaluating several machine learning algorithms trained on a set of task-motivated features. We compare our work with the state of the art on the annotated dataset obtaining improved performance. We also annotate a new dataset which we make publicly available to validate our approach. The results on the new dataset confirm our set of features outperforms previously published research.

Keywords: scientific text mining; citation context identification; sentence classification

1. Introduction

Current estimates indicate that the number of scientific publications grows at unprecedented rates: between 0.7 and 1.5 million new papers are published every year (Bornmann and Mutz, 2015; Jinha, 2010). In this context of scientific information overload, text mining tools are of utmost importance for researchers who must read scientific papers in order to assess their value and make use of their contents. Research papers are not isolated units, but they are inter-connected by means of co-citation relations or citation networks which are useful to quantitatively understand the value of a piece of scientific work. However, citation networks are limited in that they do not provide information about why a paper is being cited or what part of the reference paper the citing paper is referring to. This qualitative information is very important in order to allow fine-grained automatic analysis of scientific works. Identifying which sentences of a reference paper contain the information being referred to by a set of citing papers is a difficult task in part due to the short context provided by the explicit citation, so it becomes necessary to look beyond this explicit citation for other information in the citing paper that might be relevant.

Although the detection of explicit or *formal citations* (for example in the form of author name and paper year, or using a bracketed notation) is a problem that can be resolved with high precision, papers usually contain more information about their references that is not necessarily present in a sentence containing a formal citation. We call these sentences *implicit citations*. This paper describes some experiments on the detection of implicit citations in a paper, i.e. sentences that refer to the work done in another paper but do not contain an explicit citation marker.

Consider the fragment shown in Figure 1, which is an extract from (He et al., 2008) where they cite, amongst other papers, the Pyramid method defined in (Nenkova et al., 2007).

(217) The official evaluation comprises three methods under different assumption: ROUGE[4], PYRAMID[5], and BE[3].
...
(247) In essence, Pyramid evaluation method adopts the voting idea to give the different weight for different importance Summary Content Unit(SUC).
(248) For our approach, we essentially find the stationary distribution of random walk in evolutionary manifold-ranking (...).
(249) This idea is similar to the evaluation idea of Pyramid method and more importance is that we catched the evolutionary characteristic (...).
(250) Whereas we don't do any processing of coherence and got the less linguistic quality.
...
(256) We think that ROUGE and BE are suitable to evaluate the content selection of generative summary, (...) and PYRAMID is suitable to evaluate the content selection of extractive summary (...).

Figure 1: Extract from He et al., 2008, indicating the number of sentence in the document between parenthesis.

The explicit citation is in sentence 217: The marker "[5]" refers to (Nenkova et al., 2007). Several paragraphs later, from sentence 247 onwards, they describe properties of the Pyramid method and compare it to what they did and to other methods. In this example, we could consider sentence 217 as a formal citation, sentences 247, 249 and 256 as implicit citations, and sentences 248 and 250 are not considered citations. Notice that in these sentences, they use the name of the method defined in the reference paper instead of the author, but nonetheless they are talking about the same paper.

Authors can use several techniques for implicitly referring to a paper (Athar and Teufel, 2012), for example: using only the name of the main author, using pronouns that could be resolved as the mentioned work, or using keywords that refer to a distinguishing topic in the paper (in the previous example, the Pyramid method). Also, the implicit citations can be present far from the explicit citations.

The problem can be framed as a sentence classification task: considering one sentence of the citing paper at a time, try to identify if the sentence is talking about the work done in the target reference paper, but does not contain an explicit citation to it.

This task has attracted considerable attention because of its applicability in several problems in scientific literature analysis. Our approach, which is based on training a classifier with task-motivated features improves over the state of the art in a publicly available dataset.

The contributions of this work are the following:

- a novel set of features for implicit citation identification;
- a set of experiments demonstrating the improved performance of the taken approach; and
- a novel data-set for the implicit citation identification task.

The software and data developed are being made available to the research community.

2. Related work

One of the early attempts at identifying sentences that were related to a citation but did not explicitly contain the citation marker was done by Nanba and Okumura (1999). In their work, they define a "reference area" which begins with the sentence that contains a citation marker and contains the following sentences that have a connection with the same subject. They use a set of cue words for identifying the sentences that belong to the reference area and use this information to build a multi-paper summarization system.

Kaplan et al. (2009) defines citation sites as the portions of text around a citation anchor in which the citation is discussed. These citation sites might be non-contiguous, but they limit the maximum distance from the anchor and they train a coreference resolution model to identify this non-contiguous fragments.

Qazvinian and Radev (2010) try to identify what they call "context sentences", which are sentences that contain an implicit citation. After analyzing some cases they report that those context sentences tend to occur in a small neighborhood of the explicit citation. They train a Markov Random Field model that tries to identify these context sentences, and use this information to build a summarization system by extracting keyphrases (Qazvinian et al., 2010).

Our work follows closely the research of Athar and Teufel (2012), which uses the implicit citations in order to enrich a citation sentiment analysis system. In order to do this, they build a corpus of papers annotated with formal (explicit) citations and informal (implicit) citations, all of them categorized as positive, negative or objective. They train a SVM model with a set of features that tries to capture relevant information for detecting implicit citations, even if they are non-contiguous, for example detecting other ways of referring to the work of an author inside a document that do not imply using the name of the author (see section 4.1.). Using this information they improve the performance of a citation sentiment classifier.

More recently, in Kaplan et al. (2016) they define a similar problem of citation block determination. They train SVM and CRF models including features such as location, topic modeling, discourse and coreference to determine if a sentence belongs to a citation block. The main difference is that they do not consider the implicit citations that might be non-contiguous to the citing sentence.

Another related problem is, instead of finding whole sentences, trying to identify which parts of a sentence with many citations is referring to a paper, i.e. the scope of the reference (Abu-Jbara and Radev, 2011), (Abu-Jbara and Radev, 2012).

Incorporating implicit citations together with formal citations can be applied to several tasks in the context of scientific literature analysis. For example, one problem that has been studied is the automatic creation of paper summaries. The importance of considering citations when creating a summary or a survey has been stressed by several authors (Nanba and Okumura, 1999),(Elkiss et al., 2008), (Mei and Zhai, 2008), (Qazvinian and Radev, 2008), and including implicit citations as well as explicit ones can improve the quality of the result (Qazvinian and Radev, 2010).

As mentioned above, another task that can be improved using implicit citations is the detection of function and polarity of a citation (Athar, 2011), (Athar and Teufel, 2012), (Li et al., 2013), (Abu-Jbara et al., 2013).

Our interest is related to the problem of citation-based scientific summarization. As defined in (Jaidka et al., 2016), given a citation in a paper, try to identify which sentences of the reference paper best reflect this citation. Given that our refernce-citation matching system has limited performance in the proposed tasks, see (AbuRa'ed et al., 2017) for example, it is natural to think that implicit citations could boost system performance (see (Moraes et al., 2017)).

3. Corpus

Athar's Citation Context Corpus consists of 203,803 sentences from the ACL Anthology Network (AAN) corpus. This data is presented as a set of HTML files where each file contains all papers in the AAN which cite a specific target paper. The file contains a table where each row corresponds to a citing paper, and each cell in that row represents one sentence in the citing paper. Each sentence is marked as a formal citation, an informal citation, or no citation at all, using a color code. Figure 2 shows an example of one target paper (HTML file). The colors represent if the citations are positive, negative or netural, and the shades represent formal or informal citations.

4. Experiments

In the experiments to be described we compare our approach to that of (Athar and Teufel, 2012) but since the software produced in that work is not available, we reimplemented all the features defined in that paper and attempted to replicate the results obtained by the authors using Support Vector Machines (SVM) classifier. Then, we extended the approach with newly task-motivated features, evaluating several classifiers using AUTO-WEKA (Thornton et al., 2013): which simultaneously select a learning algorithm and sets its hyper parameters, we configured AUTO-WEKA to run experiments for 7 hours in which 13 configurations were performed using 10-fold cross validation using several classifier including: Bayes net, Random Tree, SMO and Random Forest. Finally, we tested both (Athar and Teufel, 2012) and our approach against a new test set we annotated manually.

4.1. Original features

The original classifier described in (Athar and Teufel, 2012) is a SVM trained using the following set of binary features:

- **Formal Citation:** Two features indicating if the citation (for example as author name followed by year) appears in the previous or in the current sentence. The feature for the current sentence is meant to help the classifier discard sentences containing formal citations, as they are not the target of their work. The feature for the previous sentence, however, could help detect sentences immediately after a formal citation that might be still talking about the same subject.

- **Author name:** A feature indicating if the author name is in the sentence, but not together with the year as would happen in a formal citation. It has been shown that sometimes a paper that was already formally cited can be recalled by using only the author name.

- **Other citations:** This feature indicates if the sentence contains a citation different from the one the classifier is trying to the detect.

- **Determiner and *work noun*:** Work nouns are defined in (Siddharthan and Teufel, 2007) as nouns used to indicate other people's work. This feature would capture expressions like "the study" or "their result".

- **Third person pronoun:** This feature indicates a sentence that starts with a third person pronoun, in order to capture sentences like "They show that...".

- **Connector:** Used to mark if a sentence starts with a connector, from a list of 23 connectors like "however" or "moreover".

- **Subsection heading:** Three features indicating if the previous, current, or next sentence starts with a subsection heading. These features could help identify a topic shift in the analyzed sentence.

- **Acronyms:** Indicates if a sentence contains an acronym mentioned near a formal citation. In the example above, "PYRAMID" is an acronym used in place of a citation.

- **Lexical hooks:** For this feature, it is necessary to analyze all citing papers besides the one that is being classified. A lexical hook is defined as the most frequent capitalized phrase found around a formal citation to the reference paper. The intention is to capture other common ways of referring to a paper that do not imply the name of the author or an acronym, for example the phrase "Pyramid method" in the example above.

- **N-gram features:** They consider features for n-grams of length 1 to 3 in the sentence. Besides using these features in the final classifier, they train a classifier that only uses n-grams as a baseline for comparison. In our case we used the SUMMA library (Saggion, 2008) for calculating the n-gram features.

According to (Athar, 2014), the most relevant features were lexical hooks, acronyms, and if the previous sentence contains a formal citation.

4.2. Our features

After an empirical examination of the corpus and the task at hand, we defined content-based, contextual features that could incorporate novel information to the classifier:

- **Word Embeddings Cosine Similarity:** The more similar a text is to another, the more likely it is that it might be referring to it. We utilized a set of pre-trained word2vec models with 300 dimensions representing each vector in the vector space, for this set of features we calculated the centroid of each sentence and the centroid of the reference paper abstract, and then measured the cosine similarity of these two vectors. We generated a set of three features corresponding to using three different embeddings collections: Google News embeddings [1], the ACL Anthology Reference Corpus embeddings (Liu, 2017) (trained over a corpus of ACL papers (Bird, 2008)), and the BabelNet embeddings (Mancini et al., 2016) (trained over a corpus of documents disambiguated using BabelNet (Navigli and Ponzetto, 2012) synsets).

- **Context Vectors Cosine Similarity:** Another semantic similarity measure we used is context vectors. Using the SUMMA library (Saggion, 2008) we calculated the tf*idf vectors of each sentence and the ref-

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

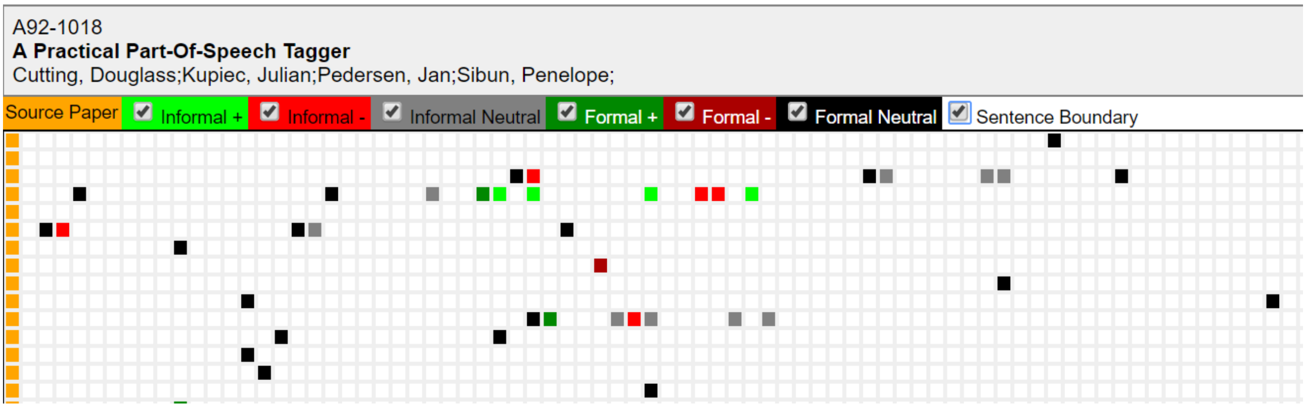

Figure 2: An example of a target paper's HTML file from Athar's Corpus

erence paper abstract, and we used the cosine similarity of these pairs of vectors as features. We have two features using the tf*idf measure for lemmas and for BabelNet synsets which are extracted using BabelFy (Moro et al., 2014). The idf tables for lemmas and synsets were computed using a subset of around four thousand ACL anthology papers.

- **Scientific Gazetteer:** Teufel's (Teufel, 2000) action and concept Lexicons were used to create gazetteers lists to identify scientific references (e.g. *research*: 'analyze', 'check' and 'gather'; *problem*: 'violate', 'spoil' and 'mistake', and *solution*: 'fix', 'cure' and 'accomplish'). We created two features to count how many words in the sentence belong to the "action" or the "concept" category of the gazeteers.

- **Co-reference Chains:** The Dr. Inventor (DRI) Text Mining Framework (Ronzano and Saggion, 2015) provides co-reference resolution over the scientific papers. We generated a feature that detects cases in which a sentence is referring to an explicit formal citation through a coreference chain.

- **Rhetorical Category:** DRI also predicts the probability of a sentence being in one of five possible rhetorical categories (i.e. Approach, Background, Challenge, Outcome and Future work). We added a feature which indicates the index of the highest probability facet the target sentence represents, we believe that indicating the sentence facet could be informative for our classification task.

- **Cause and Effect:** DRI annotates causal relations in scientific papers, we used a feature to detect the existence of any causality relation in the sentence.

- **Citations:** The more formal citations a sentence could have the less likely it will be an implicit citation. This feature will simply count the number of formal citations the sentence has for any related work.

- **Distance to closest formal citation** The closer a sentence to a formal citation to the reference paper the more likely it is an implicit citation. We generate one feature to calculate the distance between the sentence and the closest formal citation to the reference paper.

- **Title tokens:** Implicit citations could contain tokens from the reference paper title, one feature has been calculated in which the value represents the number of tokens in the reference paper title appearing in the sentence.

- **POS n-grams:** We also added Features for part of speech n-grams of length 1 to 3 in the sentence.

4.3. Features analysis

In order to understand the discriminative power of our features, we ran the information gain feature selection algorithm with the attribute ranking search method by utilizing WEKA on the set of features excluding n-grams over the training dataset. Such test provides a better insight of which features are more important than others. Table 1 shows the top 12 features selected by the algorithm.

Table 1: Top 12 features ranked by the information gain algorithm. The features marked with * are new features.

Distance to closest formal citation*	0.0062925
Title Tokens*	0.0054377
Context Vectors cosine similarity*	0.0049679
Lexical Hooks	0.0041353
Acronyms	0.0040977
Babelnet Context Vectors cosine similarity*	0.0040918
Previous Formal Citation	0.003768
ACL Word2Vec cosine similarity*	0.002034
Google News Word2Vec cosine similarity*	0.001385
Author name	0.000991
Co-reference Chains*	0.0007706
BabelNet Word2Vec cosine similarity*	0.000753

What can be noticed from table 1 is that the majority of the top features (8 - marked as * on the table - out of 12) of the training dataset are from the newly generated features used by our system.

4.4. Test data

In order to further validate our approach and try to compare it to the previous one, we annotated a small set of test doc-

Cluster	Papers	Sents	Formal	Implicit
Nenkova2007	5	1550	9	21
Kaplan2004	6	1896	24	11
Blunsom2008	7	2750	19	18
BunescuPasca2006	12	5531	44	76
CardieWagstaff1999	9	3895	20	24
Total	39	15622	116	150

Table 2: Composition of the test corpus

uments. We collected five target papers from ACM Transactions on Computational Logic Journal, ACL and NAACL conferences, for each paper we collected the papers citing it and had identified the explicit and implicit citations following the same approach as (Athar and Teufel, 2012), but without considering the sentiment polarity of citations. The only annotations used are Formal Citation or Implicit Citation. Table 2 shows the composition of this small test corpus.

5. Results

Table 3 shows the results of our replication of Athar's experiments (the n-gram baseline and the features) as well as the results using our set of features, using 10-fold cross validation over the training data. For our method we have applied a set of machine learning algorithms to check which one yields the best results using 10-fold cross validation and unlike Athar's approach which used SVM, our best model was using Random Forest algorithm.

Experiment	Precision	Recall	F-Measure
Athar's baseline	0.643	0.293	0.403
Athar's features	0.609	0.362	0.454
Our novel features	0.684	0.370	0.480

Table 3: Cross validation results

One thing we can see from the results is that the experiments using our implementation of Athar's features did not yield the same performance as reported in (Athar and Teufel, 2012). They reported an F-measure of 0.513 for the implicit citation class, but in our case we got 0.454 for the same experiment. This is expected, as we did not use the same tools they used to compute features, so we could not replicate exactly the same experimental setting. For both experiments, however, the classifiers beat the n-gram baseline.

In order to better understand the kind of differences between results, we tried both models over the test data. The results are shown in table 4.

First note that the performance of both classifiers dramatically dropped on the test data. The worst performance was for the Blunsom2008 cluster, where both classifiers predicted no implicit citations. The best performance was for the Nenkova2007 cluster, where both classifiers had at least one hit. Although the classifier with our features performs

Experiment	Precision	Recall	F-Measure
Athar's system	0.0500	0.0067	0.0118
Our system	0.1613	0.0333	0.0552

Table 4: Results over test data

a little better than the previous one, both results were very poor. One possible explanation for this is that the features used for both classifiers were too specific and could not generalize to these new examples. Another possibility is that the test clusters themselves were very hard to classify. In either case, there is still a lot of room for improvement, and more data and experiments are needed in order to solve this task in a satisfactory way.

6. Conclusions

We presented the results of our experiments on the detection of implicit citations/references to a reseach paper, with the aim of using this method for improving the performance of a reference scope detection system. We calculated the features used in a previous work and created a new set of features that were found relevant for the classifier. We first trained an implicit citations classifier as specified in (Athar and Teufel, 2012), and then built a new classifier using all the features. The new classifier performs better than the previous published work when evaluated with a cross-validation methodology. In both cases the results were lower than the ones reported in (Athar and Teufel, 2012), but we consider this could happen because our experimental setting is different, so using our features with the same experimental setting as Athar might lead to even better results.

In order to further analyze the results of the classifiers, we annotated a small test set of scientific documents. On the newly created test set the performance of both classifiers drop, still our new classifier shows better results. We are in the process of annotating more test data and analyzing why the classifiers behave different with these new papers.

There are several avenues of possible research to improve this research. More data for training and evaluation might be necessary to create better classifiers. Also, it would be very interesting to test more advanced techniques, for example using deep learning methods.

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).

7. Bibliographical References

Abu-Jbara, A. and Radev, D. (2011). Coherent citation-based summarization of scientific papers. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language*

Technologies-Volume 1, pages 500–509. Association for Computational Linguistics.

Abu-Jbara, A. and Radev, D. (2012). Reference scope identification in citing sentences. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 80–90. Association for Computational Linguistics.

Abu-Jbara, A., Ezra, J., and Radev, D. (2013). Purpose and polarity of citation: Towards nlp-based bibliometrics. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 596–606.

AbuRa'ed, A., Chiruzzo, L., and Saggion, H. (2017). What sentence are you referring to and why? identifying cited sentences in scientific literature. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 9–17, Varna, Bulgaria, September. INCOMA Ltd.

Athar, A. and Teufel, S. (2012). Detection of implicit citations for sentiment detection. In *Proceedings of the Workshop on Detecting Structure in Scholarly Discourse*, pages 18–26. Association for Computational Linguistics.

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. In *Proceedings of the ACL 2011 student session*, pages 81–87. Association for Computational Linguistics.

Athar, A. (2014). Sentiment analysis of scientific citations. Technical report, University of Cambridge, Computer Laboratory.

Bird, S. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

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

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D., and Radev, D. (2008). Blind men and elephants: What do citation summaries tell us about a research article? *Journal of the Association for Information Science and Technology*, 59(1):51–62.

He, R., Liu, Y., Qin, B., Liu, T., and Li, S. (2008). Hitir's update summary at tac 2008: Extractive content selection for language independence. In *TAC*.

Jaidka, K., Chandrasekaran, M. K., Rustagi, S., and Kan, M.-Y. (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)*, pages 93–102.

Jinha, A. E. (2010). Article 50 million: an estimate of the number of scholarly articles in existence. *Learned Publishing*, 23(3):258–263.

Kaplan, D., Iida, R., and Tokunaga, T. (2009). Automatic extraction of citation contexts for research paper summarization: A coreference-chain based approach. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 88–95. Association for Computational Linguistics.

Kaplan, D., Tokunaga, T., and Teufel, S. (2016). Citation block determination using textual coherence. *Journal of Information Processing*, 24(3):540–553.

Li, X., He, Y., Meyers, A., and Grishman, R. (2013). Towards fine-grained citation function classification. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 402–407.

Liu, H. (2017). Sentiment analysis of citations using word2vec. *arXiv preprint arXiv:1704.00177*.

Mancini, M., Camacho-Collados, J., Iacobacci, I., and Navigli, R. (2016). Embedding words and senses together via joint knowledge-enhanced training. *arXiv preprint arXiv:1612.02703*.

Mei, Q. and Zhai, C. (2008). Generating impact-based summaries for scientific literature. *Proceedings of ACL-08: HLT*, pages 816–824.

Moraes, L., Baki, S., Verma, R., and Lee, D. (2017). Identifying reference spans: topic modeling and word embeddings help ir. *International Journal on Digital Libraries*, pages 1–12.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Nanba, H. and Okumura, M. (1999). Towards multi-paper summarization using reference information. In *IJCAI*, volume 99, pages 926–931.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Nenkova, A., Passonneau, R., and McKeown, K. (2007). The pyramid method: Incorporating human content selection variation in summarization evaluation. *ACM Transactions on Speech and Language Processing (TSLP)*, 4(2):4.

Qazvinian, V. and Radev, D. R. (2008). Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics.

Qazvinian, V. and Radev, D. R. (2010). Identifying non-explicit citing sentences for citation-based summarization. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 555–564. Association for Computational Linguistics.

Qazvinian, V., Radev, D. R., and Özgür, A. (2010). Citation summarization through keyphrase extraction. In *Proceedings of the 23rd international conference on computational linguistics*, pages 895–903. Association for Computational Linguistics.

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

Saggion, H. (2008). SUMMA: A robust and adaptable summarization tool. *Traitement Automatique des Langues*, 49(2).

Siddharthan, A. and Teufel, S. (2007). Whose idea was this, and why does it matter? attributing scientific work to citations. In *Human language technologies 2007: The conference of the North American chapter of the Association for Computational Linguistics; proceedings of the main conference*, pages 316–323.

Teufel, S. (2000). *Argumentative zoning: Information extraction from scientific text*. Ph.D. thesis, Citeseer.

Thornton, C., Hutter, F., Hoos, H. H., and Leyton-Brown, K. (2013). Auto-WEKA: Combined selection and hyperparameter optimization of classification algorithms. In *Proc. of KDD-2013*, pages 847–855.