Honours Year Project Report

Citation Provenance

By

Heng Low Wee
(U096901R)

Department of Computer Science

School of Computing

National University of Singapore

2011/12

Honours Year Project Report

Citation Provenance

By

Heng Low Wee
(U096901R)

Department of Computer Science

School of Computing

National University of Singapore

2011/12

Project No: H079820
Advisor: A/P Min-Yen Kan
Deliverables:
Report: 1 Volume
Source Code: 1 DVD

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Subject Descriptors:
TO BE COMPLETED

Keywords:
information retrieval, citation

Implementation Software and Hardware:

Software: Python, NLTK, scikit-learn
Hardware: MacBook Pro, Intel Core 2 Duo 2.4GHz, 4GB Memory.

Acknowledgement

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.1 Motivation

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

1.2 Goal

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat

non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Chapter 2

Related Work

2.1 Citation Analysis

Several authors had researched on works related to citation analysis. In these works, they could be categorised into several directions for development. One of them that has a major impact would be citation classification or similarly, the classification of citation function. This task is aimed at making sense of the rationale why the authors of a paper would cite the work of another, and thus better aid readers on understanding the key ideas presented in a paper. The reasons why authors would cite, are what was meant by the citation function. In an updated version of their paper, Teufel et al. presented an annotation scheme for annotating a citation's function (Teufel, Siddharthan, & Tidhar, 2009). In their scheme, citations are generalised into 4 main categories: Weak, Contrast, Postive & Neutral. Some of these categories are further broken down into more specific sub-categories, producing a total of 12 classes for annotating citations. (Teufel, Siddharthan, & Tidhar, 2006) had already worked on the automatic classification of citation function, utilising features extracted from the *citing context*. (Dong & Schäfer, 2011) presented an approach to citation classification that uses a combination of various supervised learning algorithms. Similarly, authors worked on analysing the sentiment of citations to determine the polarity of these citations. (Athar, 2011) used sentence structure based features extracted from the citing context and produced promising results.

In (Wan, Paris, Muthukrishna, & Dale, 2009) and (Wan, Paris, & Dale, 2010), Wan and his teams worked on building a research tool that acts a reading aid for readers when browsing through scientific papers. Wan et al. investigated the *literature browsing task* by conducting surveys on researchers who read scientific papers frequently to update themselves. In this initial study conducted by Wan et al., several key ideas were revealed. First, when researchers read scientific papers and see citations made by the author, their main concern, as time-constrained professionals, is whether the cited paper would be worth their time and effort, and money, to follow up on and at the same time, whether to believe in the citation. Second, readers faced the difficulty of finding the exact text that justify the citation. Third, the surveys revealed that readers found it useful if a reading tool could identify important sentences and key words in the cited paper. This study conducted by Wan et al. is based on the fundamental idea of improving the reading experience of practitioners and researchers. The goal is to save a reader's time by helping the reader make relevance judgements about the cited documents. As it is often that readers have to read up on the cited documents to gain a better insight on the current context, this task would be of relevance. The authors then developed the CSIBS based on their studies. The CSIBS tool helps reader determine whether to read on the cited papers by providing a contextual summary of the cited papers.

2.2 Sentence Alignment

Aligning sentences belonging to similar documents of the same language is an important research area for tasks related to summarisation and paraphrasing. (Nelken & Shieber, 2006) presented a novel algorithm for sentence alignment in monolingual corpora. They showed their approach, which is based on TF*IDF similarity score, produced great precision at aligning sentence, with precision score of 83.1%. A more recent work by (Li, Sun, & Xue, 2010) introduced a new sentence alignment algorithm called Fast-Champollion. Briefly, it splits the input text into alignment fragments and identifies the components of these fragments before aligning them using a Champollion-based algorithm.

Chapter 3

Our Approach

3.1 Terminology

To aid the reader, and to avoid misunderstanding and confusion, it is important that we first list some of the key terms we are using in our paper.

Term	Description
Citing Paper	The paper that makes the citation
Cited Paper	The paper that is being cited by the citing paper
Cite Link	E.g. `E06-1034==>J93-2004`. A citation relation between a citing paper (`E06-1034`) and a cited paper (`J93-2004`)
Cite String	The citation mark. E.g. Nivre and Scholz (2004), [1], (23)
Citing Sentence	A sentence in the citing paper that contains the in-line citation. E.g. *That algorithm, in turn, is similar to the dependency parsing algorithm of* ***Nivre and Scholz (2004)****, but it builds a constituent tree and a dependency tree simultaneously.*
Citing Context	The block of text surrounding the citing sentence, usually 2 sentences before and after the citing sentence, for providing contextual information
Cited Fragment	A fragment, from a few lines to paragraphs, in the cited paper

Table 3.1: Terminology

3.2 Problem Analysis

3.2.1 Types of Citation

In the scope of our project, all citations could be classified into 2 types: **General**, and **Specific**. We define citations as such to be inline with our goal. That is, to be able to tell, if specific, where the cited information is in the cited document. Otherwise, the citation would be deemed general. To rid of ambiguity in our definition of a general/specific citation, we have the following guidelines:

General Citations

1. Authors refer to a paper as a whole. If the author cites for a key idea, e.g. Machine Learning, and the entire or majority of the cited paper is about Machine Learning, it is a general citation.

2. Authors refer to a paper as a form of mentioning. The authors merely mentions the cited paper out of acknowledgement of its contributions.

Specific Citations

1. Authors refer to a term definition in the cited paper.

2. Authors refer to a key idea/implementation in the cited paper. This key idea/implementation does not make up the entire cited paper.

3. Authors refer to an algorithm or a theorem in the cited paper. This algorithm/theorem does not make up the entire cited paper.

4. Authors refer to digits or numerical figures in the cited paper. Usually for making reference to evaluation results in the cited paper.

5. Authors quote a line/segment in the cited paper.

In general, for **Specific** citations, we would be able to specifically extract a fragment in the cited paper that represents the source of the information mentioned in the citation itself i.e. Citation Provenance.

3.2.2 Locating The Cited Information

Our problem is now reduced to determining whether a citation is General or Specific. If a citation is general, the reader can be directed, for example, to the Abstract section of the cited paper, but this is not the main focus of our task. If a citation is specific, the reader can be directed to that specific paragraph or lines respectively. Therefore during computation, the cited document can be broken down into fragments. Hence if given that a citation is specific, then there must exists a fragment that the citation refers to. For this we need to implement some ranking system that determines the location of this fragment.

3.2.3 Scope Of The Problem

In this project, we abstract away the problem of locating the in-line citations in a paper, and reduce our problem to only determining the type of a citation and its location. To solve the problem of locating the in-line citations, we utilize the open-source ParsCit system developed by (Councill, Giles, & Kan, 2008). Conveniently, ParsCit identifies the citing sentence, together with the citing context.

3.2.4 Modelling The Problem As Search

In web search engines, an user enters a search query, and a search engine would use this query to search within its search domain – millions of web pages – and then display the best matching web pages as compared to the search query. That would be equivalent to having a search query for an entire corpus of research papers. Our problem can also be modelled as a searching problem, but a reduced version as compared to web search engines.

Consider reading a paper, `A`. We know the citations made by `A`, and these cited papers are listed in the References section of `A`. From this our search domain for any query from `A` would be the contents of the list of cited papers. We reduce this search domain further when we are investigating a particular citation in `A`, say now paper `A` cites the paper `B`. Now, for this citation, the scope of search would be the sub-domain – contents of paper `B`. So instead of searching

for the best matching document in the corpus, we are now searching within `B`. Our problem analysis tells that we have to break down `B` into fragments, and the search query would be for these fragments (Refer to Figure 3.1 for a simple illustration). With the help of ParsCit (Councill et al., 2008), the citing sentence can be extracted. The search query would be citing context which consists of the citing sentence.

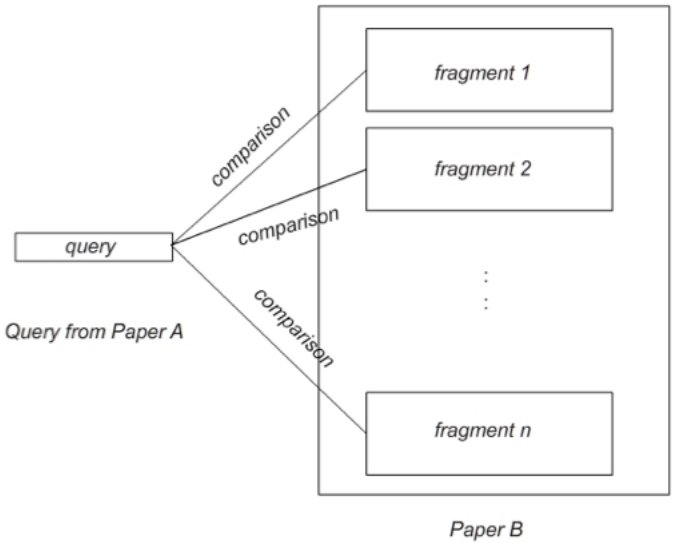

Figure 3.1: Modeling Our Problem

Our problem is now a *binary classification problem*, where we attempt to determine whether a fragment is either General or Specific.

3.3 Training Corpus

At this initial stage, we picked the ACL Anthology Reference Corpus[1] (ACL-ARC). The ACL-ARC consists of publications of the Computational Linguistics field. Note that in general, we wish to perform this citation provenance task on all publications from all fields of research. This corpus is chosen as a start, because it provides the *interlink data* that conveniently informs us of the cite links between the papers in the corpus. For instance, in the interlink data, a link

[1] http://acl-arc.comp.nus.edu.sg/

like `X98-103==>X96-1049` says that the paper `X98-103` cites `X96-1049`.

3.3.1 Collecting Annotations

Now that we have modelled our problem, we are able to specify the required data format for our task. For each cite link, there can be multiple in-line citations i.e. multiple citing contexts. For each citing context, we are comparing with each fragment in the cited paper. In other words, if a cite link has n citing contexts and the cited paper can be divided into m fragments, immediately we have $(n \times m)$ instances.

Our first attempt at collecting annotations was to require an annotator to specify the line numbers of the cited information that the citing context was referring to. The annotator would be provided the citing and cited paper in plain text format, and he/she will need to annotate on a separate file, specifying the line number range, e.g. line range `L12-55` of the cited paper. For this annotation task, we designed an annotation framework[2] where an annotator is presented with an user-friendly interface to select the lines in the cited paper that he/she deem Specific. We posted this task onto the Amazon Mechanical Turk (MTurk[3]) for a few Turk workers to participate in our annotation task. After a pilot round of collection, we reviewed this annotation scheme together with feedbacks from our small group of participants. First, this annotation task is a non-trivial one. The annotations collected from MTurk do not agree among the annotators and ourselves. Participants must be able to understand the contents of the papers, thus, must be researchers or have some experience in reading scientific papers. Second, this annotation scheme is too tricky, and would also cause us much problem when it comes to evaluation. Consider our implemented system that outputs a prediction for citation provenance in the form of a line number range. It is difficult to judge the correctness of this prediction, say `L30-67`, when compared against the annotated `L12-55`.

Our second attempt is more straightforward. Recall that we use ParsCit for extracting the citing context. ParsCit also divides a paper into logically adequate fragments according to

[2]http://citprov.heroku.com

[3]https://www.mturk.com

sections, sub-ssections, figures and tables etc. So instead of annotating by line number ranges, we annotated each fragment with 3 classes: General (g), Specific-Yes (y) and Specific-No (n). To be precise, we annotate g if a cite link is deemed General, and y only for the fragment(s) that is deemed Specific. For the other fragments that are not Specific, we annotate n. Table 3.2 summarises the statistics for annotation. Note that we only display percentage values for Specific instances.

Item	Statistics
No. of Cite Links	275 (7.6% Specific)
No. of Fragments	30943 (0.09% Specific-Yes, 12.9% Specific-No)

Table 3.2: Annotation Statistics

As one can see, Specific citations are very rare. From a machine learning point of view, immediately one can observe that the training data is skewed towards General citations. From this we may conclude that even if we attempt to gather more positive instances, the ratio between General and Specific should remain about the same. This situation we have with our annotations also contributes to our approach to the problem, as we explain in the following section.

3.4 A Two-Tier Approach

We propose a two-tier approach to our problem. In the first tier, it plays the role of a *filter*, and attempts to filter out the General citations, leaving behind the Specific citations to be passed to the second tier. Figure 3.2 illustrates the flow of our approach.

3.4.1 First Tier

The First Tier is our attempt to filter out the General citations. In this tier, we are performing a 2-class citation classification task. We are not interested in determining whether the citation is one of the 12 class as defined by (Teufel et al., 2009), but only whether it is General or Specific. For each cite link we extract its citing contexts. Then for these contexts we extract

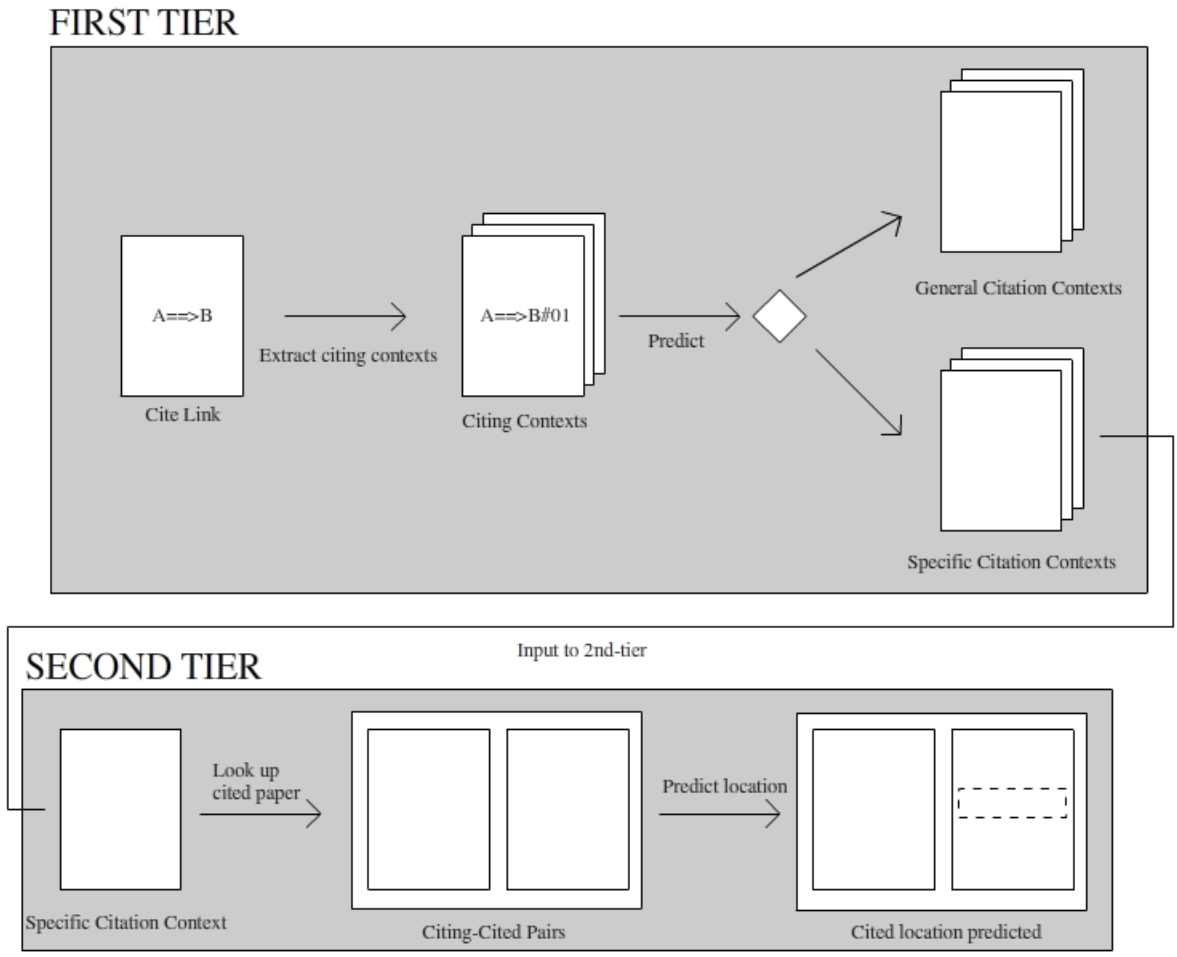

Figure 3.2: A Two-Tier Approach

feature vectors in order to pass it into our prediction model. We do adopt similar features that were presented in previous works on citation classification.

First Tier Features

1. Physical Features

   We adopted the physical features as presented in (Dong & Schäfer, 2011). They are:

   (a) *Location*: in which section the citing sentence is from.

   (b) *Popularity*: no. of references in the citing sentence.

   (c) *Density*: no. of unique references in the citing sentence and its neighbour sentences.

   (d) *AvgDens*: the average of Density among the citing and neighbour sentences.

2. Number Density

   A numerical feature we formulated that computes the density of numerical figures in the citing context. The intuition is that Specific citations refer to evaluation results in the cited paper. E.g. "...Nivre and Scholz (2004) obtained a precision of 79.1%...".

3. Publishing Year Difference

   A numerical feature that represents difference in the publishing year between the citing and cited paper. The intuition is that higher difference suggests cited paper is older and presented a fundamental idea.

4. Citing Context's Average TF-IDF Weight

   A numerical feature that indicates the amount of *valuable* (as determined by TF-IDF (Manning, Raghavan, & Schütze, 2008)) words in the citing context. Higher values suggest important words and thus specific keywords.

5. Cue Words

   Another numerical feature adopted from (Dong & Schäfer, 2011), that computes the amount of cue words (pre-defined manually by us) that appear in the citing sentence and its neighbour sentences. We defined 2 classes of cue words: Cue-General and Cue-Specific. Cue-General = $\{proposed, introduced, sketched, discussed, suggested \ldots\}$ and Cue-Specific = $\{obtained, scored, precision, probabilities, experimental \ldots\}$. These cue words are selected based on the examples we observed in our training corpus.

From our training corpus we extracted these features to build our First Tier Model for prediction.

3.4.2 Second Tier

In our Second Tier, it is another abstraction of our problem. We assume all the inputs into the second tier are Specific citations, and then we attempt to predict which of the fragments in the cited paper is the cited fragment.

Second Tier Features

1. Surface Matching

   A numerical feature that measures the amount of word overlap between the citing sentence and a fragment in the cited paper.

2. Number Near-Miss

   A numerical feature that measures the amount of numerical figures overlap between the citing sentence and a fragment in the cited paper. This feature will preprocess each fragment, rounding numerical figures or converting to percentage values, when it try to match the numerical figures in the citing sentence. The intuition for this feature is from our observation that most Specific citations refer to evaluation results in the cited paper.

Chapter 4

Evaluation

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Chapter 5

Conclusion

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Chapter 6

Future Work

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

References

Athar, A. (2011). Sentiment analysis of citations using sentence structure-based features. *Proceedings of the ACL* (pp. 81–87), 2011.

Councill, I., Giles, C., & Kan, M. (2008). Parscit: An open-source crf reference string parsing package. *Proceedings of LREC*, Vol. 2008 (pp. 661–667), European Language Resources Association (ELRA), 2008.

Dong, C., & Schäfer, U. (2011). Ensemble-style self-training on citation classification. , 2011.

Li, P., Sun, M., & Xue, P. (2010). Fast-champollion: a fast and robust sentence alignment algorithm. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 710–718), Association for Computational Linguistics, 2010.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press.

Nelken, R., & Shieber, S. (2006). Towards robust context-sensitive sentence alignment for monolingual corpora. *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 161–168), 2006.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006). Automatic classification of citation function. *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing* (pp. 103–110), Association for Computational Linguistics, 2006.

Teufel, S., Siddharthan, A., & Tidhar, D. (2009). An annotation scheme for citation function. *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80–87), Association for Computational Linguistics, 2009.

Wan, S., Paris, C., & Dale, R. (2010). Supporting browsing-specific information needs: Introducing the citation-sensitive in-browser summariser. *Web Semantics: Science, Services and Agents on the World Wide Web*, *8*(2-3), 2010.

Wan, S., Paris, C., Muthukrishna, M., & Dale, R. (2009). Designing a citation-sensitive research tool: an initial study of browsing-specific information needs. *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, NLPIR4DL '09 (pp. 45–53), Stroudsburg, PA, USA, 2009: Association for Computational Linguistics.