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

Citing previously published scientific papers is a necessary and common practice among researchers. It gives credit and acknowledgement to original ideas, and to researchers who did significant work in a particular field of research, and more importantly, upholds intellectual property. A reader of such research papers often encounters these citations made by the authors in various sentences throughout the paper. Often enough, if a reader wishes to gain a better understanding of the current context, it is necessary to follow through these citations and read up on these cited papers. Readers would also be interested to know where the information is in the cited paper.

However, as frequent readers might find, most citations are only *mentions*. They do not directly refer to some particular section of the cited paper, for example, to make reference to the evaluation results made by the authors of the cited paper. Instead, they are general citations. These citations are equally important. But since it is not immediately clear where the cited information is from, a reader has to invest additional time to read through the entire cited paper before being able to find out what or where the critical information is.

In general, we wish to improve the reading experience of scientific & research documents, from the various fields of research. Readers will be informed of where exactly the cited information is from in the cited paper. We aim to be able to identify which section or paragraph in the

referenced paper is the source of the information referred to in the citation, i.e. Citation Provenance.

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 (teufel2009annotation). 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. (teufel2006automatic) had already worked on the automatic classification of citation function, utilising features extracted from the *citing context*. (dongensemble) 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. (athar2011sentiment) used sentence structure based features extracted from the citing context and produced promising results.

In (citation-sensitive) and (csibs), 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. (nelken2006towards) 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 (li2010fast) 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, about 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 are 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 may refer to a paper as a whole. If the author cites for a key idea, e.g. Machine Learning, and Machine Learning makes up the entire or majority of the cited paper, it is a general citation.

2. Authors may 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 may refer to a term definition in the cited paper.

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

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

4. Authors may refer to digits or numerical figures in the cited paper. Usually for making reference to evaluation results in the cited paper. Authors may also complement the cited paper for its promising/excellent performance.

5. Authors may 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 Scope Of The Problem

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.

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 (parscit). Conveniently, ParsCit identifies the citing sentence, together with the citing context.

3.2.3 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 (parscit), the citing context can be extracted. The search query would be citing context which consists of the citing sentence.

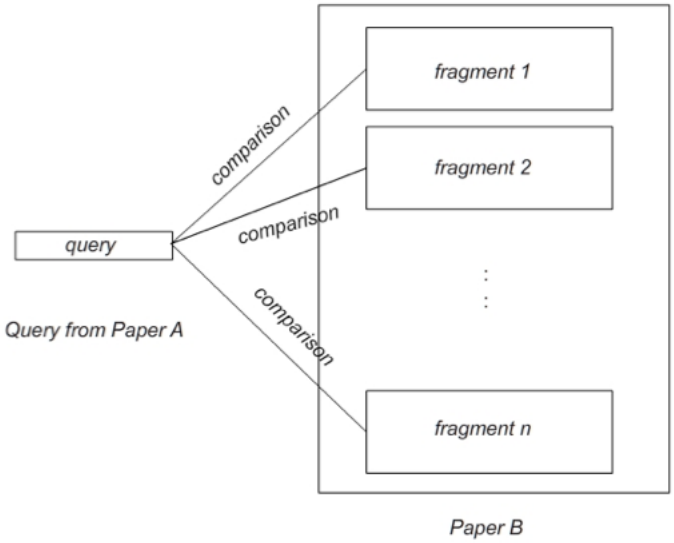

Figure 3.1: Modeling Our Problem

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

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

us of the cite links between the papers in the corpus. For instance, in the interlink data, a link 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 $(n \times m)$ data 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 MTurk workers to participate in our annotation task. After a trial 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. Participants must be able to understand the contents of the papers, thus, must be researchers or have some experience in reading scientific papers. While it is possible to target a selected category of MTurk workers for our tasks, the complexity of this annotation task requires participants with research experiences, which could be limited in numbers. Furthermore, most of the annotations collected from MTurk do not agree among the annotators and ourselves. To collect annotations that disagree among annotators most of the time, is not helpful for the problem we are trying to tackle. Thus we abandon collecting annotations via MTurk, and performed annotations manually on our own.

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

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

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 `L50-78`, when compared against the annotated `L12-55` and that the prediction *overlaps* the annotation by 5 lines. This variable amount of overlap is not definitive at all, and is difficult for us to decide at what extent of overlap only do we consider the prediction correct. Thus we switched to the alternative.

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 sections, sub-sections, figures and tables etc. So instead of annotating by line number ranges, we annotate each of the fragments of the cited papers. We annotate them with 3 classes: General (g), Specific-Yes (y) and Specific-No (n). To be precise, we annotate g (for all its fragments) 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. After prolonged periods of searching for valid Specific citations in our training corpus, we argue that even if we attempt to gather more positive instances, the ratio between General and Specific should remain about the same. This challenging 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.

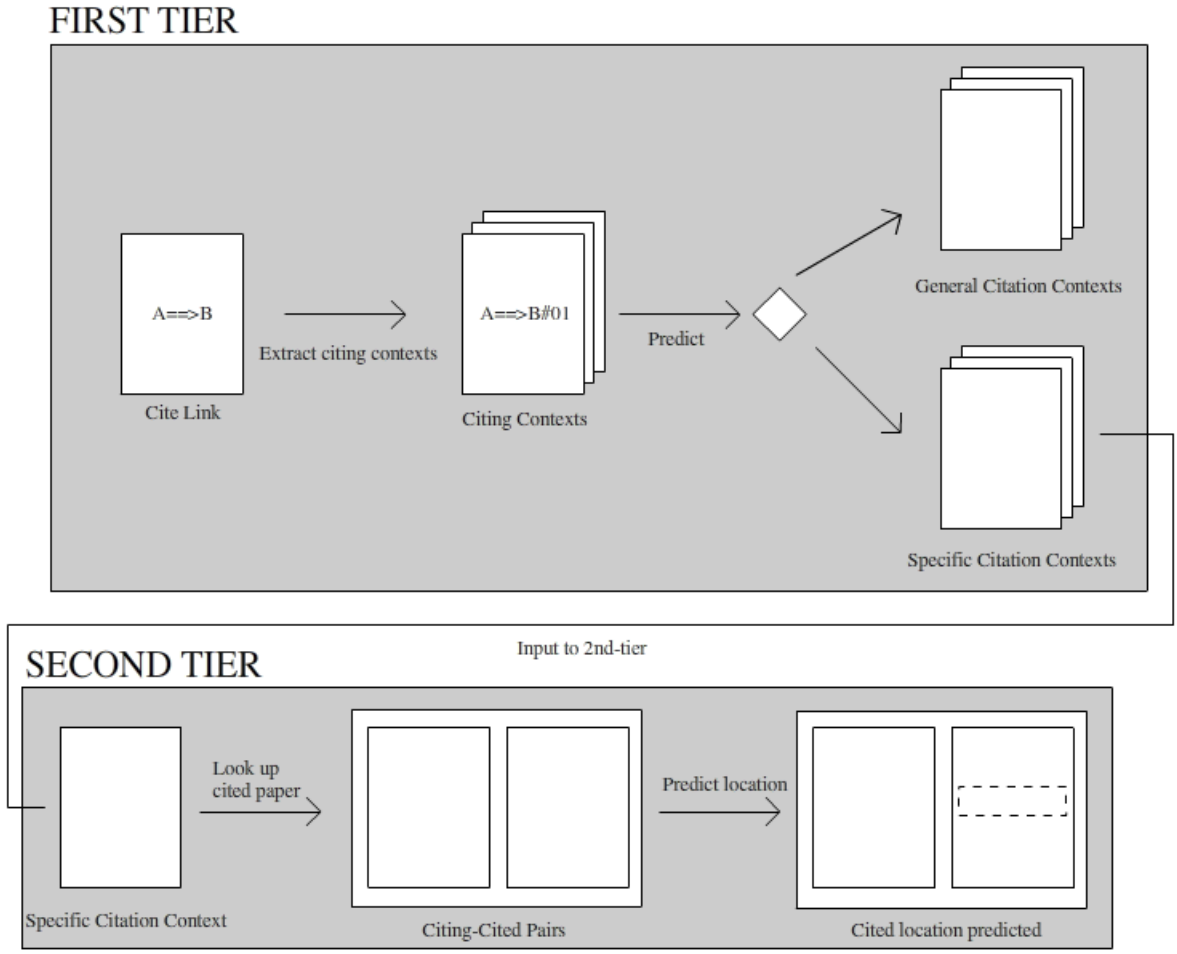

Figure 3.2: A Two-Tier Approach

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, which already is a challenging task in the research area of citation analysis. We are not interested in determining whether the citation is one of the 12 class as defined by (teufel2009annotation), but only whether it is General or Specific. For each

cite link we extract its citing contexts. Then for these contexts we extract feature vectors in order to pass it into our prediction model. We 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 (dongensemble). They are:

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

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

   (c) *Density*: no. of unique citation marks 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 that measures the density of numerical figures in the citing context. The intuition is that Specific citations tend 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, and thus cited for General purposes.

4. Citing Context's Average `tf-idf` Weight

   A numerical feature that indicates the amount of *valuable* (as determined by `tf-idf` (irtextbook)) words in the citing context. Higher values suggest important words and thus specific keywords.

5. Cue Words

   Another numerical feature adopted from (dongensemble), 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 (refer to Appendix A for list of cue words). 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. It is independent from the First tier. 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 tries to match the numerical figures in the citing sentence. The intuition for this feature is from our observations that most Specific citations refer to evaluation results in the cited paper.

3. Bigrams Matching

   A numerical feature that measures the amount of bigrams overlap between the citing sentence and a fragment in the cited paper. This feature is to preserve word order when comparing the citing sentence and the fragment. This feature is also targeted at Specific citations that refer to the cited paper for term definitions and quoting.

4. Cosine Similarity

A common numerical feature used in information retrieval tasks to measure similarity between the query and a candidate document. In our case, citing sentence and the fragment.

Similarly we extracted these features from our training data to build our Second Tier Model for prediction.

Chapter 4

Evaluation

We performed 2 evaluations, one for each tier as described early in Chapter 3.4. We are able to do this because the tiers are independent of each other.

4.1 Results - First Tier

Recall that we have 275 annotated cite links, either General (g) or Specific (s), and that we have very limited instances of Specific cite links, a situation mentioned in (li2010negative), that we have a highly unbalanced ratio between General instances and Specific instances. So for our evaluation, we first gathered all Specific instances, and then randomly select General instances, twice the number of Specific instances. Out of these 84 instances, we have 1 : 2 ratio of Specific versus General instances.

We trained our model using various classifiers, and then performed `Leave-One-Out` evaluation using the 84 instances. In other words, in one round of evaluation we predict 84 times. For each classifier, we performed 10 rounds of evaluation and we compute the average score for each round. Table 4.1 summarises the accuracy for each classifier.

SVM	NaiveBayes	DecisionTree
0.702	**0.774**	0.679

Table 4.1: First Tier Results

4.2 Results - Second Tier

For Second Tier evaluation, we are predicting whether each fragment in the cited paper is a Specific one. We have over 30 thousand training instances for second tier, and so for the same reason, we had to select our training set manually similarly, except that we have a $1:1$ ratio for Specific versus General instances.

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.

Appendix A

Cue Words

The following is the list of cue words used in one of our feature. During feature extraction, all words are stemmed before we make any comparison.

A.1 Cue-General

proposed, propose, presented, present, suggested, suggests, described, describe, discuss, discussed, gave, introduction, introduced, shown, showed, sketched, sketch, talked, adopted, adopt, based, originated, originate, built, researchers, comparative, comparison, following, previously, previous

A.2 Cue-Specific

obtains, obtained, score, scored, high, F-score, Precision, precision, Recall, recall, estimated, estimates, reported, reports, probability, probabilities, peaked, experimental, experimented, rate, error