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

We investigate a new task in citation analysis, *citation provenance*, which is to determine the provenance of the claim supported in the paper referenced in a citation. We describe the challenges in collecting annotations for our training set, and present a two-tier approach in tackling this problem. From our evaluation results, we show that the features we introduce into our approach were able to differentiate citations that referred to the whole paper in general (*general*) versus ones the cited specific claims, evidence or parts of the paper (*specific*). We also show that the system is able to determine which is the cited fragment in the cited paper, given knowledge that a citation is *specific*.

Subject Descriptors:
TO BE COMPLETED

Keywords:
citation analysis, citation provenance, source of citation

Implementation Software and Hardware:

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

Acknowledgement

I would like to express my gratitude to all the volunteer participants from the NUS WING group for participating in my pilot annotation tests. I thank them for testing my annotation scheme, and appreciate the feedback that improved my project.

Million thanks to Jin Zhao, Tao Chen, and especially my supervisor to this project, A/P Min Yen Kan for providing their guidance during the duration of the project.

List of Figures

List of Tables

Table of Contents

Chapter 1

Introduction

Citing previously published scientific papers is an important practice among researchers. It gives credit and acknowledgement to original ideas, and to researchers who did significant work in enabling the current research. More importantly, it 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 these citations and read the cited papers to understand the basis for the current work. Often, when reading the claims of a sentence supported by a citation, readers wish to know where in the cited paper the information comes from.

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 what I term *general* citations. Other citations refer specifically to particular claims, parts or sections of a paper. These citations are equally important. However, since it may not be immediately clear where the cited information is from[1], a reader has to invest additional effort to locate the cited information. I'll refer to (Wan, Paris, Muthukrishna, & Dale, 2009) for their survey results to justify my claims. In the series of surveys they conducted, most of their participants found it difficult *finding the exact text to justify the citation.* I quote one of their participants' response

[1] page numbers or references to specific artifacts, such as sections or equation numbers sometimes help to localize such references, but are not often included.

directly: "*Citation usually does not include the position of the information* in the cited article... it might be necessary to read all of the article to find it in another reference and so on." (Wan et al., 2009)

Citation Provenance refers to the source of a citation. The task of determining citation provenance is to locate the information in the cited paper that justifies the citation. It improves the reading experience of scientific and research documents by showing where exactly the cited information is from in the cited paper. I aim to identify which section or paragraph in the referenced paper is the cited information.

In comparison with (Wan, Paris, & Dale, 2010), which only provided a summarisation solution, this paper describes the first attempt to provide a solution to the difficulty in locating the information that justifies a citation. I hope this would also encourage meaningful discussions to designing a new citation style that better captures the provenance of the cited information.

In the rest of this paper, we will first look at some past works that are related to what I am describing. In Chapter 4, I analyse the problem and describe my approach on tackling the problem. I present my experimental results in Chapter 5 followed by my conclusion.

Chapter 2

Related Work

Several authors had researched on works related to Citation Analysis. These works could be categorised into several directions for development. One of them that has a major impact is Citation Classification or similarly, the classification of Citation Function. It aims to determine why the authors of a paper would cite the work of another, and thus better aid readers understand 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 this 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) previously 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 et al., 2009) and (Wan et al., 2010), Wan and his teams built a research tool that acts as a reading aid for readers when browsing through scientific papers. Wan et al. investigated the *literature browsing task* through 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.

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, 2010a) 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.

Authors paraphrase the content they were referring to usually for greater clarity and to introduce variety. While (Shinyama, Sekine, & Sudo, 2002) presented an approach to acquire paraphrase automatically, in Citation Provenance we are, in a way, trying to achieve the opposite. By comparing the words and phrases used in a citation with paraphrases extracted from a cited work, one could possibly achieve better sentence alignment between the 2 documents.

Chapter 3

Problem Analysis

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.

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

Figure 3.1: Terminologies used in this 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.

3.0.1 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 (Councill, Giles, & Kan, 2008). Conveniently, ParsCit identifies the citing sentence, together with the citing context.

3.0.2 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 4.2 for a simple illustration). With the help of ParsCit (Councill et al., 2008), the citing context can be extracted. The search query would be citing context which consists of the citing sentence.

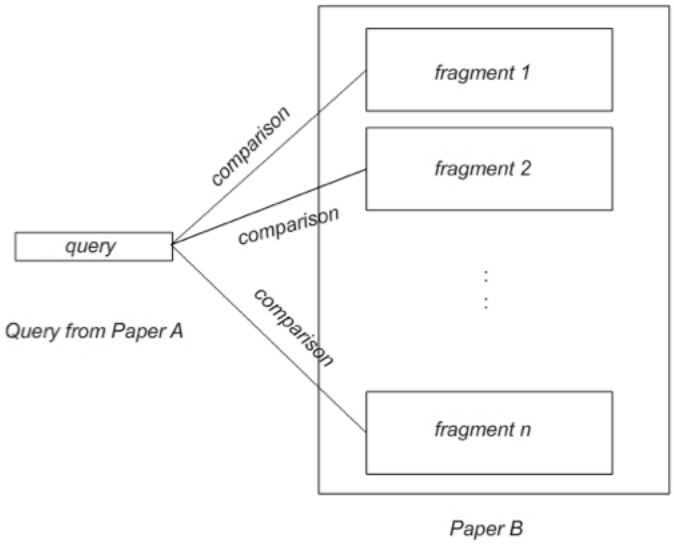

Figure 3.2: Modeling Our Problem

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

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

3.1.1 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

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

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.

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.

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

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

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 4.2 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 3.1: 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.

3.2 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 4.3 illustrates the flow of our approach.

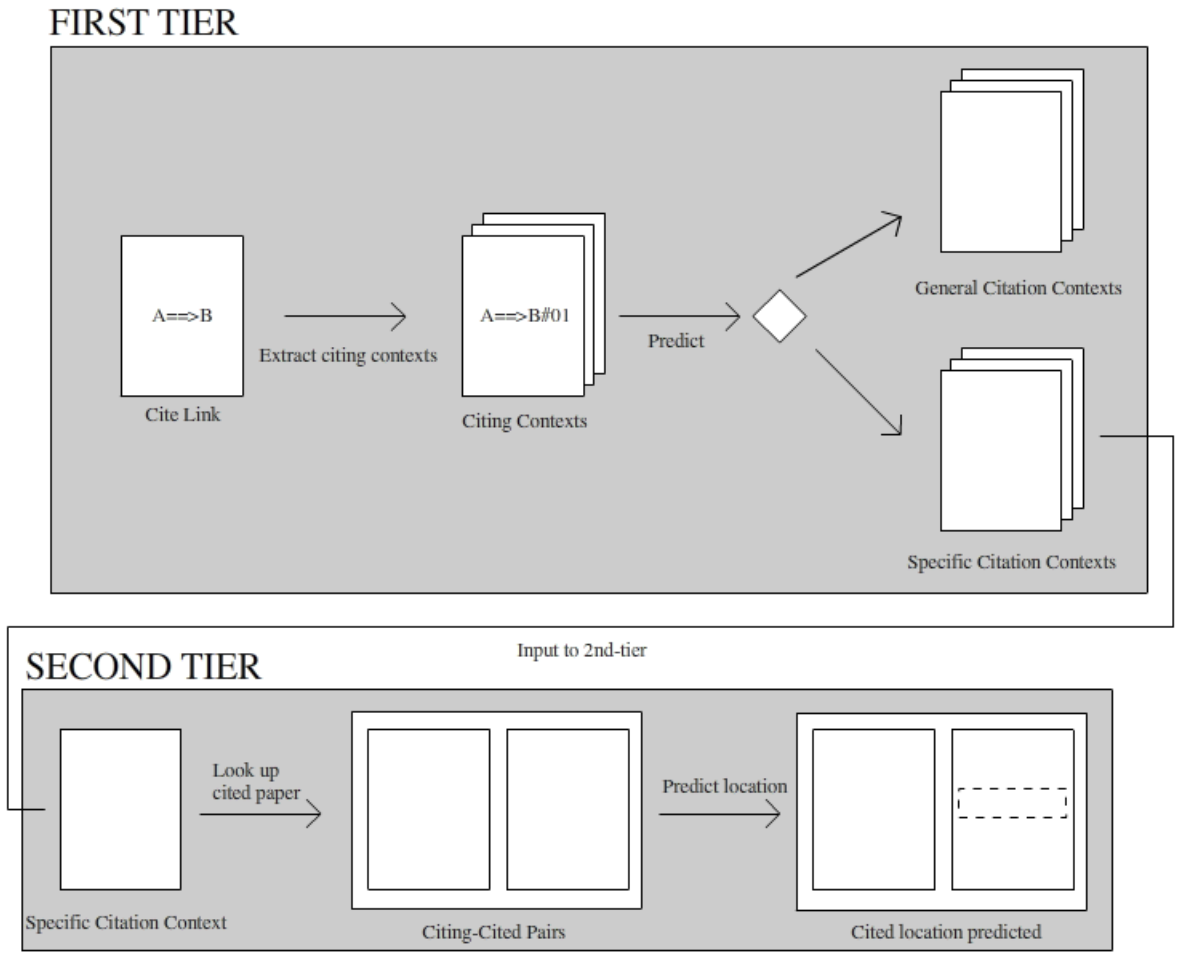

Figure 3.3: A Two-Tier Approach

3.2.1 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 (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 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 (Dong & Schäfer, 2011). 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` (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 (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.

3.2.2 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

Our Approach

4.1 Terminology

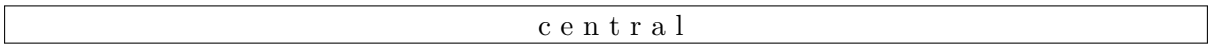

Figure 4.1: Testing

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.

4.2 Problem Analysis

4.2.1 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

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 4.1: Terminology

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.

4.2.2 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 (Councill et al., 2008). Conveniently, ParsCit identifies the citing sentence, together with the citing context.

4.2.3 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 4.2 for a simple illustration). With the help of ParsCit (Councill et al., 2008), the citing context can be extracted. The search query would be citing context which consists of the citing sentence.

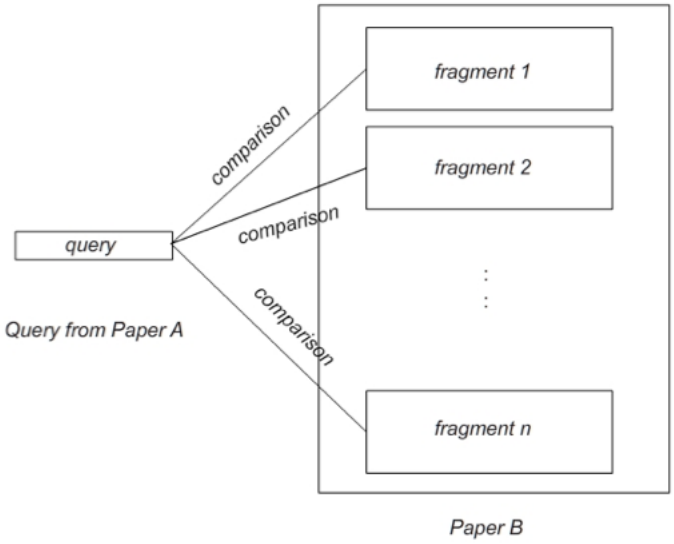

Figure 4.2: Modeling Our Problem

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

4.3 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,

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

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

4.3.1 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

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

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

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.

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 4.2 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 4.2: 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.

4.4 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 4.3 illustrates the flow of our approach.

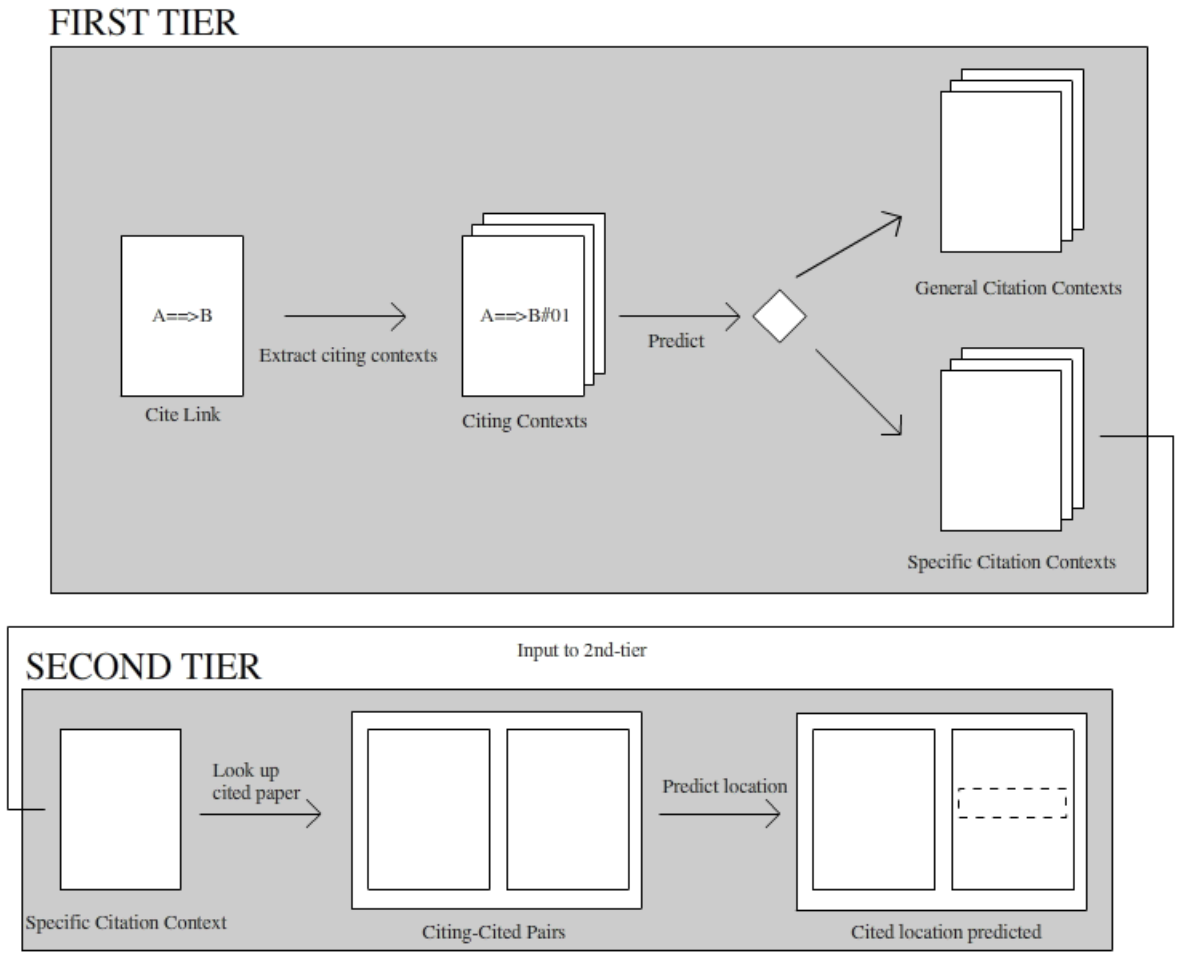

Figure 4.3: A Two-Tier Approach

4.4.1 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 (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 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 (Dong & Schäfer, 2011). 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` (Manning et al., 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 (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.

4.4.2 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 5

Evaluation

We performed 2 evaluations, one for each tier as described early in Chapter 4.4. We are able to do this because the tiers are independent of each other.

5.1 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 (Li, Liu, & Ng, 2010b), 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. Out of these 56 instances, we have 1 : 1 ratio of Specific versus General instances. The reason for choosing a 1 : 1 ratio is because we wish to measure our approach's ability to differentiate between General and Specific when given a balanced training set.

We trained our model using various classifiers, and then performed `Leave-One-Out` (`LOO`) and `n-Fold` ($n = 14$, each fold has 4 instances) evaluation using the 56 instances. The main reason for using `Leave-One-Out` is because we are working with limited instances and we wish to maximise them for training. To test our method's performance when given less training data, we use the `n-Fold` strategy. When trained on SVM, we observed that in terms of *feature weights*, feature 1(a) and 5 were assigned the highest weights. For instance, feature 1(a), *Density*, was assigned with a weight of magnitude 0.703. This suggests the amount of in-line citations and

the amount of cue words within the citing context play an important part in our prediction task.

We evaluated using various classifiers and Table 5.1 summarises the performance for each classifier.

Classifier/Values	Avg. Precision	Avg. Recall	Avg. F_1-Score
	`LOO` / `n-Fold`	`LOO` / `n-Fold`	`LOO` / `n-Fold`
SVM	0.84 / 0.71	0.84 / 0.70	0.84 / 0.69
NaiveBayes	0.70 / 0.70	0.66 / 0.66	0.64 / 0.64
DecisionTree	0.72 / 0.66	0.71 / 0.66	0.71 / 0.66

Table 5.1: First Tier Results

We examine the confusion matrix for the best performing SVM classifier that we ran for the `Leave-One-Out` strategy. Recall that our First Tier's objective is to filter out the General citations. Our goal is to attain higher numbers in both the *g*-*g* and *s*-*s* cells in the confusion matrix. We achieved this in Table 5.2 and we can conclude that our First Tier performed well in differentiating General and Specific citations.

	actual g	actual s
predicted g	24	4
predicted s	5	23

Table 5.2: Confusion Matrix for SVM with Leave-One-Out

5.2 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 to get a $1:1$ ratio for Specific versus General instances.

Classifier/Values	Avg. Precision	Avg. Recall	Avg. F_1-Score
	`LOO` / `n-Fold`	`LOO` / `n-Fold`	`LOO` / `n-Fold`
SVM	0.85 / 0.84	0.84 / 0.82	0.84 / 0.82
NaiveBayes	0.80 / 0.78	0.79 / 0.77	0.78 / 0.77
DecisionTree	0.89 / 0.86	0.89 / 0.86	0.89 / 0.86

Table 5.3: Second Tier Results

In this case, the Decision Tree classifier performed slightly better than SVM. Similarly, our goal is to attain higher numbers in both the *g*-*g* and *s*-*s* cells in the confusion matrix and in Table 5.4 we achieved good results. This means, given a Specific citation, this approach would perform well in determining whether a fragment in the cited paper is the cited fragment or not.

	actual g	actual s
predicted g	23	5
predicted s	3	25

Table 5.4: Confusion Matrix for Naive Bayes

(Refer to Appendix B for more details of our experimental results.)

Chapter 6

Discussion

Citation Provenance is a task that has relatively little developments done on it. In our paper we attempt to define the nature of the problem, and presented a possible approach to tackle it. One of the main challenges we had with this task is the limited number of Specific citations in scientific papers. (Teufel et al., 2009) showed that the percentage of neutral citations was 62.7%. We can say that the percentage of General citations is at least as much, because our definition of a Specific citation is much more restricted compared to the 12 classes defined in (Teufel et al., 2009). We observed, and conclude that in general research & scientific papers, most citations are neutral and General.

We argue, that even though the percentage of Specific citations is low and that the value of applications that perform such task seems low, citation provenance would prove to be an important reading tool that helps readers understand and navigate between papers that are linked via citations. We support with evidence the validity of our claim, that our prototype application submitted as part of the `CodeForScience`[1] 2012 competition organised by Elsevier was well received among the judging panel that consisted of professionals from fields related to information technology and libraries.

[1] http://www.codeforscience.com/singapore

Sometimes, in-line citations to scientific papers in journals and books capture the chapter numbers and page numbers. The main reason is because the length of the cited document is very long compared to the citing document. An example of such citation is *(J. Doe, 2012, sec. 6.5, 174-85)*. In this citation it captures the section number, "*sec. 6.5*", and page numbers, "*174-85*", to a book or journal. Note that the granularity of such style is not specific enough for our problem as a section can be arbitrary lengthy. In our case, we consider computational linguistic papers that are usually less than 20 pages, which is much shorter than books and journals. For this we sketch a new citation style that better captures citation provenance.

Our sketched style is straightforward: To numerically label each segment or fragment in the cited paper. This applies to text bodies, figures and tables. An example for a Specific citation: *(B. White, 2011, **B23**)*. Notice we added another a **B** to **23**, which could be a better way to distinguish between text bodies (B), figures (F) and tables (T). **23** simply means the 23rd segment of the type B. Suppose the cited paper is already labelled, when a reader sees a citation a paper, the reader sees there is the additional information at the end of the citation and understands it is a Specific citation. To read up on the cited paper would be a breeze.

Chapter 7

Conclusion

We touched on a new task for citation analysis, Citation Provenance. In this task, we are trying to first determine whether a in-line citation is General or Specific, and second, if Specific, to locate where in the cited paper is the referenced information for the citation itself.

One of the challenges in this task is the highly unbalanced ratio between General versus Specific citations, i.e. Specific citations are very rare. Also, the annotation task is very challenging and would require experienced researchers who understands the content of the papers to be annotated.

We presented a two-tier approach towards this problem. With the first acting as a filter to separate the General citations from the Specific ones and the second one to predict which of the fragments in the cited paper are referenced by the citation. We arranged for our training set to have a balanced ratio of General versus Specific instances, so that the results would reflect the extent our approach is able to differentiate the 2 types of citations we defined. We evaluated our approach on 3 classifiers and we were able to obtain promising results.

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. (2010a). 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.

Li, X., Liu, B., & Ng, S. (2010b). Negative training data can be harmful to text classification. *Proceedings of the 2010 conference on empirical methods in natural language processing* (pp. 218–228), 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.

Shinyama, Y., Sekine, S., & Sudo, K. (2002). Automatic paraphrase acquisition from news articles. *Proceedings of the second international conference on Human Language Technology Research* (pp. 313–318), Morgan Kaufmann Publishers Inc., 2002.

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.

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

Appendix B

Results Details (1st Tier)

B.1 Results: Leave-One-Out

	Precision	Recall	F_1-Score
g	0.83	0.86	0.84
s	0.85	0.82	0.84

	actual g	actual s
predicted g	24	4
predicted s	5	23

Table B.1: SVM $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.61	0.89	0.72
s	0.80	0.43	0.56

	actual g	actual s
predicted g	25	3
predicted s	16	12

Table B.2: Naive Bayes $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.70	0.75	0.72
s	0.73	0.68	0.70

	actual g	actual s
predicted g	21	7
predicted s	9	19

Table B.3: Decision Tree $P/R/F_1$ Scores and Confusion Matrix

B.2 Results: n-Fold

	Precision	Recall	F_1-Score
g	0.76	0.57	0.65
s	0.66	0.82	0.73

	actual g	actual s
predicted g	16	12
predicted s	5	23

Table B.4: SVM $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.61	0.89	0.72
s	0.80	0.43	0.56

	actual g	actual s
predicted g	25	3
predicted s	16	12

Table B.5: Naive Bayes $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.66	0.68	0.67
s	0.67	0.64	0.65

	actual g	actual s
predicted g	19	9
predicted s	10	18

Table B.6: Decision Tree $P/R/F_1$ Scores and Confusion Matrix

Appendix C

Results Details (2nd Tier)

C.1 Results: Leave-One-Out

	Precision	Recall	F_1-Score
g	0.91	0.75	0.82
s	0.79	0.93	0.85

	actual g	actual s
predicted g	21	7
predicted s	2	26

Table C.1: SVM $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.74	0.89	0.81
s	0.86	0.68	0.76

	actual g	actual s
predicted g	25	3
predicted s	9	19

Table C.2: Naive Bayes $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.92	0.86	0.89
s	0.87	0.93	0.90

	actual g	actual s
predicted g	24	4
predicted s	2	26

Table C.3: Decision Tree $P/R/F_1$ Scores and Confusion Matrix

C.2 Results: n-Fold

	Precision	Recall	F_1-Score
g	0.91	0.71	0.80
s	0.76	0.93	0.84

	actual g	actual s
predicted g	20	8
predicted s	2	26

Table C.4: SVM $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.73	0.86	0.79
s	0.83	0.68	0.75

	actual g	actual s
predicted g	24	4
predicted s	9	19

Table C.5: Naive Bayes $P/R/F_1$ Scores and Confusion Matrix

	Precision	Recall	F_1-Score
g	0.88	0.82	0.85
s	0.83	0.89	0.86

	actual g	actual s
predicted g	23	5
predicted s	3	25

Table C.6: Decision Tree $P/R/F_1$ Scores and Confusion Matrix