Recognizing Citations in Public Comments

Jaime Arguello
Jamie Callan
Stuart Shulman

ABSTRACT. Notice and comment rulemaking is central to how U.S. federal agencies craft new regulation. E-rulemaking, the process of soliciting and considering public comments that are submitted electronically, poses a challenge for agencies. The large volume of comments received makes it difficult to distill and address the most substantive concerns of the public. This work attempts to alleviate this burden by applying existing machine learning techniques to the problem of recognizing *citation* sentences. A citation in this context is defined as a statement in which the author of the public comment references an external source of factual information that is associated with a specific person or organization. The problem is formulated as a binary classification problem: Is a specific person or organization mentioned in a sentence being referenced as an external source of information? We show that our definition of a citation is reproducible by human judges and that citations can be detected using machine learning techniques with some success. Casting this as a machine learning problem requires selecting an appropriate representation of the sentence. Several feature sets are evaluated individually and in combination. Superior results are obtained by combining feature sets. Syntactic features, which characterize the structure of the sentence rather than its content, significantly improve accuracy when combined with other features, but not when used in isolation. Although prediction

Jaime Arguello is a Ph.D. student at the Language Technologies Institute at Carnegie Mellon University. His work focuses on text data mining, information retrieval, and natural language processing.

Jamie Callan is a Professor at the Language Technologies Institute, a graduate department in Carnegie Mellon University's School of Computer Science. His research and teaching focus on text-based information retrieval. His recent research studies advanced search engine architectures, federated search across groups of search engines, adaptive information filtering, text analysis and organization, text mining, and automatic collection of instructional materials for an intelligent reading tutor. His earlier IR research included first generation Web-search systems, integration of text search with relational database systems, and information literacy in K-12 education.

Dr. Stuart W. Shulman is Director of the Sara Fine Institute in the School of Information Sciences at the University of Pittsburgh. He is also the founder and Director of the Qualitative Data Analysis Program (QDAP) at Pitt's University Center for Social and Urban Research, which is a fee-for-service coding lab working on projects funded by the National Science Foundation, the National Institutes of Health, DARPA, and other funding agencies. He has been Principal Investigator and Project Director on related National Science Foundation–funded research projects focusing on electronic rulemaking, human language technologies, coding across the disciplines, digital citizenship, and service-learning efforts in the United States. Dr. Shulman is the Editor-in-Chief of the *Journal of Information Technology & Politics*.

This work was supported in part by the eRulemaking project and NSF grants IIS-0240334 and IIS-0429293. We are grateful to the U.S. EPA and the U.S. Department of the Interior's FWS for providing the public comment data that made this research possible. We thank the coders at the Qualitative Data Analysis Program (QDAP) for their time and effort in producing our gold standard data and for providing the feedback that guided the evolution of our coding scheme. We also thank the anonymous reviewers for their comments. Any opinions, findings, conclusions, and recommendations expressed in this paper are the authors' and do not necessarily reflect those of the sponsors.

Address correspondence to: Jaime Arguello, Carnegie Mellon University, Pittsburgh, PA (E-mail: jaime@cs.cmu.edu).

Journal of Information Technology & Politics, Vol. 5(1) 2008
Available online at http://jitp.haworthpress.com

doi:10.1080/19331680802153683

error rate is adequate, coverage could be improved. An error analysis enumerates short-term and long-term challenges that must be overcome to improve recall.

KEYWORDS. Citation analysis, public comments, e-rulemaking, text mining, information extraction, machine learning

RECOGNIZING CITATIONS IN PUBLIC COMMENTS

U.S. law and standard regulatory practice requires U.S. regulatory agencies to give notice of a proposed rule and then to respond to *substantive* comments from lobbyists, companies, trade organizations, special interest groups, and the general public before issuing a final rule. Traditionally, public comments were submitted in paper form. However, during the last few years, the government has begun to allow comments to be submitted electronically in some cases. It is now easier for the public to examine and comment on proposed regulations. High-profile rules attract hundreds of thousands of e-mail comments. Finding the most substantive comments among these large corpora is a difficult task.

One criterion by which a policy maker could decide that a public comment is substantive is whether it references an external source of information. Policy makers might want to focus on evidence-bearing comments for several reasons. First, policy makers might want to examine the external information source cited or at least be aware of it. They might care to know where the public gets its facts. Second, a reference to an external source of information could be a proxy for the value of the author's comments. Possibly, an author that references an external source of information is better informed than one that does not. It may be more likely that his/her claims and opinions are based on factual evidence. Finally, citing external sources may be a mechanism for the author to establish credibility (Ziman, 1968). The presence of a citation may indicate the author's intent to write a constructive comment, rather than a substanceless rant. Thus, if comments need to be prioritized for the policy maker's consideration, which is likely the case when comment volume is high, focusing on comments containing references to external evidence might be useful.

In this work, we define a citation as a reference to an external source of information that is associated with a specific *person* or *organization*. The problem is framed as a binary classification problem, where the classification question is: "Given a mention of a specific person or organization, is it being referenced as (or associated with) an external source of information in the sentence?" The unit of classification is the person or organization mentioned. The evidence that informs the classification decision is limited to the sentence mentioning the person or organization. We show that our definition of a citation is reproducible by humans and detectable by automated methods. We also present a comparative evaluation of different machine learning settings for detecting citations.

From a technical perspective, we explored the following research questions:

- Can citations be detected automatically?
- If so, what features are most effective and under what assumptions?

Framing this as a text classification problem requires selecting an appropriate representation of the sentence. For example, under a traditional bag-of-words representation, the classification decision is informed by the words in the sentence (e.g., Does the word *reported* occur in the sentence?). Another representation may focus on the syntactic relations between the person or organization in question and its surrounding context (e.g., Is the entity the subject of the verb *reported*?).

In our experiments, we held the classification algorithm constant and evaluated different feature sets individually and in combination. In particular, we explored whether syntactically motivated features are required for detecting citations. This question is important because generating syntactic features requires more work than working under a bag-of-words representation. Furthermore, evidence can be borrowed from different parts of the sentence. Under a bag-of-words representation, features can originate from the person/organization potentially being cited or elsewhere. We explored the effects on classification accuracy of focusing on different parts of the sentence and discuss the implications of using different feature sets. Finally, we show that combining different representations (i.e., feature sets) improves classification accuracy. We explored two alternatives for feature set combination: (a) incorporating different feature sets into a single classifier and (b) combining the output of individual classifiers, each trained on a separate set of features. We show superior results under condition (b). Our best performing model catches about 66% of all true citations and gets about 80% of citation predictions correct. Our 66% coverage estimate is optimistic for reasons stated later.

The remainder of this article is organized as follows. The next section introduces the public comment corpora used in this work. This is done early on because the numerous examples presented throughout the article originate from these corpora, and knowing the context of each example may be informative to the reader. Then we define what is and is not a citation in this work. We next discuss the human annotation process and trace how our definition of a citation evolved until it produced adequate intercoder agreement. Data preprocessing is discussed in the following section. Our algorithms and experimental results are discussed next, followed by an in-depth error analysis, which suggests possible next steps. Related work and concluding remarks make up the final two sections.

CORPORA

The data used in this evaluation originates from public comments submitted to the U.S. Department of the Interior's Fish and Wildlife Service's (FWS) proposal to list the polar bear as "threatened" under the Endangered Species Act of 1973 (USDOI-FWS-2007-0008). This corpus (the *Polar Bears corpus*) was selected because preliminary analysis revealed sufficient references to external sources of evidence such as reports, research studies, media distributions, etc. The Polar Bears corpus contains more than 540,000 comments that were submitted to the FWS by e-mail. These comments tend to focus on the deterioration of the polar bear's habitat, primarily due to global warming and hunting. Examples originating from the Polar Bears corpus are marked with a PB.

Some of the examples presented in this article originate from comments submitted to the Environmental Protection Agency (EPA) in response to its 2004 proposal for new emission standards for hazardous air pollutants from coal- and oil-fired utility plants (USEPA-OAR-2002-0056). This corpus (the *Mercury corpus*) contains more than 530,000 comments that were submitted to the EPA by e-mail. These comments tend to focus on the negative effects of mercury contamination from coal-fired power plants. Examples originating from the Mercury Corpus are marked with a MR. Both corpora are available for research use.[1]

DEFINITIONS

What is a Citation?

We define a citation as a mention of a specific person or organization (also referred to as *named entity* or just *entity* from here forward) that is referenced as an external source of information. We impose the restriction that a citation must reference a specific person or organization to avoid ambiguous references, which were frequent in the corpora we examined. For instance, an ambiguous reference such as in *Studies show polar bears are resorting to cannibalism because they are starving*.^{PB} would not be considered a citation under this definition because the *studies* are not associated with a specific person/organization. On the other hand, the following example would be considered a citation. In all examples, mentions of a specific organization are marked with [·]$_{\text{org}}$ and mentions of a specific person are marked with [·]$_{\text{per}}$.

> *A study released in 2006 by the* [*US Geological Survey's Alaska Science Center*]$_{\text{org}}$ *suggests that bears in the southern Beaufort Sea may be turning to cannibalism as a result of the shrinking ice cover cutting off access to seal prey.*$^{\text{PB}}$

This would be a citation because the study is associated with a specific organization, the *U.S. Geological Survey's Alaska Science Center*. References to ambiguous information sources are not called citations based on the assumption that policy makers care primarily about specific data items. Note that it is possible for an ambiguous reference to be disambiguated somewhere else in the comment (e.g., imagine the first example being followed by the sentence *One such study was produced by the* [*USGS's Alaska Science Center*]$_{\text{org}}$). Since we focus on individual sentences in isolation, we would miss such cases.

In the clearest case, a citation is an instance in which the author attributes concrete information (e.g., that *mercury increases the risk of kidney damage*) to a specific external source (e.g., the *World Health Organization*):

- *The* [*World Health Organization*]$_{\text{org}}$*, in 1991, concluded that urinary mercury increases the risk of kidney damage.*$^{\text{MR}}$
- *An* [*Atomic Energy Commission*]$_{\text{org}}$*'s document published in 1952 notes that fish concentrate 150,000 times the poisons they ingest.*$^{\text{MR}}$
- *Furthermore, news (* [*NY Times*]$_{\text{org}}$*, April 7, P A4) shows that text drawn from a 2000-page National Academy of Science report was edited in a way that minimizes the risk associated with mercury exposure.*$^{\text{MR}}$

Other types of sentences also qualify as citations. The author can state that he or she examined some external source of information without mentioning the claims made by the source (e.g., *After watching this week's edition of NOW on* [*PBS*]$_{\text{org}}$*, I am more educated on this issue.*$^{\text{MR}}$). The author can explicitly ask the policy maker to consider some external information source (e.g., *Please read the latest* [*Pentagon*]$_{\text{org}}$ *report before rolling back any pollution control.*$^{\text{MR}}$). Finally, the author can simply mention that the information source exists (e.g., [*Judith Bluestone*]$_{\text{per}}$*'s remarkable study, The Fabric of Autism, will be out in a few months and has an impressive array of research citations.*$^{\text{MR}}$). All of these qualify as citations.

What Is Not a Citation?

Not every sentence where the author mentions a person or organization while (possibly) supporting an argument is a citation. For example, mentions of someone's actions to support some claim are not citations (e.g., *The* [*Bush*]$_{\text{per}}$ *administration's refusal to curb global warning is a primary cause of the polar bear's melting habitat and their terrible plight.*$^{\text{PB}}$). Such cases are not citations because the claim being made is derived by the author from the named entity's action, not by the named entity itself. It is the author who is making the connection between the named entity and the argument. A related case, also not a citation, is a mention of a person or organization's reputation (e.g., *Given* [*EPA*]$_{\text{org}}$*'s tarnished record of improving air quality in our parks, any effort to further . . . will take us in exactly the wrong direction.*$^{\text{MR}}$). The EPA's *tarnished record* is not information that the EPA produced or distributed. Rather, it is the author's or the public's perception. Thus, it is not a reference to external evidence.

Another confusable case, not necessarily a citation, is when the author quotes (directly or indirectly) an external entity. One would think that when the author quotes, he/she is implying that the person or organization being quoted has produced information worth the policy maker's consideration. However, consider the following two quotes:

- *In the northern territories, where temperatures have risen an average of four degrees since 1950, wildlife experts such as Mr.* [*Mitch Taylor*]$_{\text{per}}$ *say that bears have never been healthier and more plentiful.*$^{\text{PB}}$
- [*Ghandi*]$_{\text{per}}$ *said that you can really tell about a people from the way they treat their animals.*$^{\text{PB}}$

The first statement is considered a citation, but not the second. The difference is subtle. With the second quote, rather than telling the policy maker about some external source of information, the author is requesting that the policy maker "do the right thing." The author is probably not seriously asking the policy maker to reference Gandhi in the final rule. This distinction is subjective. Therefore, perfect agreement among human annotators for this task is unrealistic.

A Note About Citations in Scientific Text

When one first thinks of a citation, academic literature comes to mind. In a scientific paper, a citation is a meaningful link between the work that is being presented and work done in the past. In a public comment, the notion of a citation is different. In academic literature authors cite for different reasons. In the annotation scheme of academic citation function presented in (Teufel, Siddharthan, & Tidhar, 2006a), citations fulfill four general functions: (a) to point out weaknesses in the cited approach, (b) to compare/contrast the work presented with the work cited, (c) to show compatibility with the cited work, and (d) other. This breakdown of why authors cite in academic work does not fit comfortably into the context of public comments. Authors of public comments are not presenting or defending work they have done. Rather, they are defending their stance on the proposed regulation. Thus, given this difference in authors' intent, our definition of a citation deviates from what is called a citation in academic text. The Related Work section surveys prior work on automatic citation analysis in academic literature.

HUMAN ANNOTATION

This section describes the human annotation process that produced the gold standard data used for training and testing. Human annotators were presented with a set of sentences, each containing one noun-phrase marked as either person ($[\cdot]_{per}$) or organization ($[\cdot]_{org}$). Human annotators were instructed to label a sentence as a citation if the marked person or organization is associated with the production or distribution of specific information; otherwise, to label it as non-citation. The unit of classification was the named entity marked. Annotators were shown sentences in isolation, outside the context of their public comment. Sentences mentioning more than one person or organization were duplicated and a different named entity was marked in each copy. For example, the sentence *The* [*Center for Disease Control and Prevention*]$_{org}$ *has reported that 1 out of every 6 women of childbearing age has mercury levels that are higher than the limits established by the* [*EPA*]$_{org}$.^{MR} was presented to the annotators (and our algorithms) as two separate instances:

- *The* [*Center for Disease Control and Prevention*]$_{org}$ *has reported that 1 out of every 6 women of childbearing age has mercury levels that are higher than the limits established by the EPA.*^{MR}
- *The Center for Disease Control and Prevention has reported that 1 out of every 6 women of childbearing age has mercury levels that are higher than the limits established by the* [*EPA*]$_{org}$.^{MR}

The first instance would be a citation, but not the second, since the *CDC* and not the *EPA* is the organization responsible for reporting this statistic.

During the final coding process, two annotators (called coder *A* and coder *B* from here forward) from the Qualitative Data Analysis Program at the University of Pittsburgh[2] annotated a set of 6,000 sentences. Agreement between coders was evaluated on 20 percent of the data (1,200 sentences). The remaining 5,800 sentences were divided evenly between the two coders. Intercoder agreement was evaluated in terms of Cohen's Kappa (κ), defined as

$$\kappa = \frac{Pr(a) - Pr(e)}{1 - Pr(e)},$$

where *Pr(e)* is the *expected* probability of agreement due to chance, and *Pr(a)* is the *actual* or observed relative agreement between the coders. κ ranges between +1 when agreement is perfect and −1 when agreement is perfectly negatively correlated. When $\kappa = 0$, agreement is at the level that would be reached if each coder annotated at random, consistent with the empirical class distribution. The 1,200-sentence set used to measure intercoder agreement was divided into two sets of 600 sentences each (S_{pre} and S_{post}). Agreement was evaluated on set S_{pre} prior to the full annotation task and on set S_{post} after the full annotation task. These two tests for intercoder agreement can be seen as pre-/posttests, where the post-test measures agreement after each coder annotated his or her half of the 5,800 sentence set. Intercoder agreement with respect to κ is shown in Table 1. What value of κ qualifies as high agreement depends on the task. However, Carletta (1996, p. 256) notes that “content analysis researchers generally think that $\kappa > .80$ is good reliability, with $.67 < \kappa < .80$ allowing tentative conclusions to be drawn.” Thus, intercoder agreement on both sets is considered good. The higher agreement on the post set, S_{post}, suggests this may be a task human coders can improve on with practice.

TABLE 1. Intercoder Agreement

Set	κ
S_{pre}	0.871
S_{post}	0.920

Coding Manual Evolution

The coding manual and our definition of a citation underwent two major modifications prior to the final coding manual used to produce the gold standard data. The next two subsections describe these modifications. This section discusses what we did wrong and is included for researchers interested in similar annotation tasks. Readers less interested in the human annotation aspect of this work can skip to the next section without loss of continuity.

Coding Manual Version 1.0

The main difference between our first and final attempt was that initially a distinction was drawn between sentences that attribute a fact to an external source and sentences that reference an external source without attributing a fact to it. Coders were instructed to label sentences into three classes: *citation, non-citation,* and *boundary*. A citation was defined as a sentence that attributes specific facts to the person or organization in question. For example, in the sentence *An* [*Atomic Energy Commission*]$_{org}$ *document published in 1952 notes that fish concentrate up to 150,000 times the poisons they ingest.*^{MR}, a specific fact (i.e., *fish concentrate up to 150,000 times the poisons they ingest*) is attributed to the *Atomic Energy Commission*. The boundary class covered the following cases, not covered by the citation class because a specific fact is not attributed to the source:

1. The author gives his or her opinion about information produced or distributed by some external source (e.g., *Or did nobody read the* [*Pentagon*]$_{org}$ *report on climate change, which will be devastating for us all.*^{MR}).
2. The author tells the policy maker that some external information source is worth considering (e.g., *Read the recent* [*Pentagon*]$_{org}$ *report and wake up*!^{MR}).
3. The author states that he or she examined some information produced or distributed by an external source, without mentioning the facts (e.g., *After watching this week’s edition of Now on* [*PBS*]$_{org}$, *I am more educated on this issue*.^{MR}).

The non-citation class covered the remaining cases.

Three hundred sentences from the EPA’s Mercury corpus were coded by six coders. For comparison with the final round of coding, we focus on agreement between coder *A* and *B*. Agreement between *A* and *B* was not high enough ($\kappa = .625$). Table 2 shows the contingency matrix between both coders for round one. As is shown in the table, citation and non-citation were confused only 4 times (2+2). The boundary class was confused with citation 9 times (3+6) and with non-citation 22 times (13+9), indicating that its definition was problematic for coders.

TABLE 2. Round-One Coding: Contingency Matrix (Total = 300)

		Coder *A*		
		citation	boundary	non-citation
Coder *B*	citation	17	3	2
	boundary	6	15	13
	non-citation	2	9	233

TABLE 3. Round-Two Coding: Contingency Matrix (Total = 600)

		Coder *A*		
		primary	secondary	non-citation
Coder *B*	primary	10	1	1
	secondary	2	0	0
	non-citation	27	0	559

An error analysis revealed that coders had difficulty deciding if a claim attributed to an external source is concrete enough to call the instance a citation rather than boundary. A “fact” attributed to a person or organization in a citation can range from a concrete fact (e.g., *The* [*CDC*]$_{\text{org}}$ *has cited scientific evidence that levels this high increase the risk of brain damage in newborns*.$^{\text{MR}}$) to claims that either lack credibility, are subjective, or belong to the author rather than the external source (e.g., *A recent* [*PBS*]$_{\text{org}}$ *report claimed that the Amazon forest will be completely wiped out*.$^{\text{MR}}$). For the second round of coding, the boundary class was merged with the citation class.

Coding Manual Version 2.0

In this round, the citation class was divided into *primary* and *secondary*. Primary covered cases where the person or organization produced the information, and secondary covered cases where the person or organization distributed the information produced by someone else (e.g., a newspaper article covering a scientist’s work). Both primary and secondary were defined to include cases where no facts are attributed to the source referenced. A sentence had to just make it clear that some information was produced or distributed by the named entity. Six hundred sentences from the FWS’s Polar Bears corpus were coded by five coders. Again, we focus on the two coders used in the final round of coding. Agreement was much worse than agreement for round one ($\kappa = 0.407$). Table 3 shows the contingency matrix between coder *A* and *B* for round two.

The first meaningful result was that although the coders expressed their preference for distinguishing between primary and secondary references to external data, secondary references were very uncommon. Coder *A* found one and coder *B* found two. For the final round of coding, the primary and secondary classes were again merged into a single citation class. The second meaningful result was that coder *A* coded 27 citations that *B* called non-citation. Inspecting the data revealed that coder *A* made many false positive mistakes and *B* many false negative ones with respect to the citation class.

Some comments mention the advocacy group that the author represents, for example, *Attached are the comments by the* [*Safari Club Foundation*]$_{\text{org}}$ *on the proposed listing of the polar bear as a threatened species* [. . .].$^{\text{PB}}$ Coder *A* considered such references as citations, while *B* considered them non-citations. Coders were instructed to ignore organizations that the author represents, unless they are mentioned as a source of information. Also, a number of statements cited Al Gore’s documentary *An Inconvenient Truth* as an external source (the PB corpus relates to global warming). *A* considered some references to this documentary as citations while *B* coded all references to it as non-citation. Coders were reminded that a media distribution can be an external source of evidence if it is relevant and presented by the author as a source of factual information.

DATA PREPROCESSING

As previously mentioned, the Polar Bears corpus was used to produce our evaluation data. The raw Polar Bears corpus contains 546,900 comments, which are a mixture of completely original comments as well as edited and unedited form letters. A form letter is a message written by an advocacy group that someone can either submit "as is" or after personalizing it. Duplicate text in either edited or unedited form letters was flagged using the Durian duplicate detection tool (Yang & Callan, 2006). Public comments containing no text flagged unique were completely ignored. The remaining text was sentence-segmented using OpenNLP,[3] yielding 131,839 unique sentences. To filter text originating from the e-mail's footer, all sentences with less than ten tokens (words+punctuation) were removed, resulting in 68,090 sentences. These were then named entity tagged using BBN Identifinder (Bikel, Schwartz, & Weischedel, 1999). A named entity tagger marks all proper nouns corresponding to a person or organization. Sentences without a mention of a specific person or organization were filtered out. Finally, each sentence mentioning more than one person or organization was duplicated and a different named entity was marked in each copy, as previously illustrated. From this final set, 6,000 sentences were randomly sampled. Each instance poses the question: Is the tagged person/organization being cited?

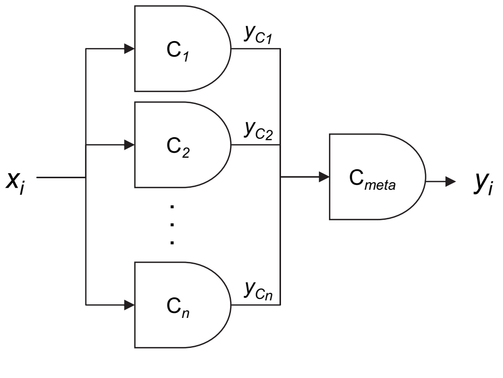
It should be noted that during the annotation process, coders were instructed to handle named entity tagging errors as follows. If the named entity tagged is off-center, such as in *As recently reported in the media, an official with the* [*Canadian Department of Fisheries*]$_{\mathrm{org}}$ *and Oceans said: We've noticed that the ice over the past 4 to 5 years has been deteriorating and it's giving us some concern.*$^{\mathrm{PB}}$, then the coder should consider the full named entity *Canadian Department of Fisheries and Oceans*. However, if the named entity is completely incorrect, such as in [*Agriculture*]$_{\mathrm{org}}$ *and forest studies have evidenced the importance of biodiversity in maintaining a healthy planet.*$^{\mathrm{PB}}$, then the coder should consider the instance a non-citation, regardless of the context.

ALGORITHMS

The goal of any text classification algorithm is to learn from annotated examples a model that can be applied to a previously unseen example to predict its class, in our case citation or non-citation. This is the typical supervised learning framework. A support vector machine (SVM) classifier (for details see Vapnik, 1995) was used in all experiments. SVM classifiers have produced good results in a wide range of text classification tasks. The SVM$^{\mathrm{light}}$ implementation (Joachims, 1998) was used with a linear kernel. This SVM classifier learns from training data a linear boundary that separates positive from negative examples (i.e., citations from non-citations) by the widest margin, while making as few misclassifications on the training data as possible (the training data may not be linearly separable).

Several feature sets were evaluated individually and in combination. Each feature set is a different way of representing the contents and/or syntactic structure of the sentence. A single feature in a feature set is an atomic unit of evidence, which a classifier leverages to predict the unknown class. One way to combine feature sets is simply to incorporate them into a single classifier. In the best case, the classifier learns how to leverage these potentially different sources of evidence to produce a reliable predictive model. A different way to combine features is via an ensemble method, depicted in Figure 1. In an ensemble method, each base classifier makes a prediction and then a meta-classifier takes as input the base classifiers' predictions and

FIGURE 1. Typical meta-classification set up. The example to be predicted, x_i, is input to all base classifiers, C_1, C_2, . . . , C_n. The base classifiers' individual predictions, y_{C1}, y_{C2}, . . . , y_{Cn}, are the input to the meta-classifier, C_{meta}, which makes the final prediction, y_i.



outputs a final prediction. The base classifiers, C_1, C_2, . . . , C_n, are said to form a *committee*. The meta-classifier, C_{meta}, somehow aggregates the predictions of its committee to produce a final prediction. A meta-classifier, for instance, might simply predict the majority class based on the predictions of its committee.

In our setup, the base classifiers and the meta-classifier are all SVM classifiers. Base classifiers differ from one another only in the features that each uses. Each base classifier is allowed to focus (and possibly become an expert) on a particular set of features. Their output was aggregated by training the meta-classifier to predict the right output based on the base classifiers' collective predictions. This ensemble method is known as *stacking* (Wolpert, 1990). Separating features into complementary views or feature sets to improve accuracy and reduce the burden of producing training data is also used in a technique called *co-training* (Blum & Mitchell, 1998).

The next two subsections describe our features sets. In the next subsection, we focus on our bag-of-words feature sets. Then, we focus on our semantically motivated syntactic features. Bag-of-words features focus on the words in the sentence. Our semantically motivated syntactic features focus on the sentence's structure. More specifically, they characterize how the person or organization potentially being cited fits syntactically into the sentence.

Bag-of-Words Features

As mentioned previously, one major research question explored in this work is whether syntactic information is at all required to determine that a person/organization in a sentence is being cited. To explore this question, four bag-of-words feature sets were evaluated. A bag-of-words model is one that ignores all ordering and relation information between features (the words in the sentence, in our case). These four bag-of-words feature sets differ exclusively in the parts of the sentence from where their features originate and are referred to as ALL, ONLY_NE, ALL_BUT_NE, and ALL_BUT_NE + ONLY_NE:

- ALL: All the words in the sentence are used as features.
- ONLY_NE: Only the words within the tagged named entity are used as features.
- ALL_BUT_NE: All the words in the sentence are used as features, except those within the tagged named entity.
- ALL_BUT_NE + ONLY_NE: All the words in the sentence are used as features.

However, a word occurring within the tagged named entity is treated as a different feature than the same word occurring outside the tagged named entity.

Individually, each bag-of-words classifier sheds light at the nature of the problem. If ALL performs well enough, then the extra machinery required to characterize how the named entity in question fits syntactically into the sentence may not be justified. The performance of ONLY_NE depends partly on the extent to which authors reference the same people or organizations in citations. If authors consistently cite the same sources, the ONLY_NE should perform well. Conversely, if authors cite a wide range of sources, but do so in a consistent manner, then ALL_BUT_NE should perform well. ALL and ALL_BUT_NE + ONLY_NE are related. The difference is that ALL_BUT_NE + ONLY_NE distinguishes whether a word occurs inside or outside the named entity.

For the implementation of these classifiers, all text was downcased and tokenized by splitting on white space and punctuation. Stopwords (i.e., topic-general function words) were removed and each term was stemmed (i.e., suffix-stripped) using the Porter stemmer (Porter, 1980). Term stems occurring only once, thus having no predictive power, were removed, resulting in a vocabulary (i.e., feature space) of 3,651 term stems (double this number for ALL_BUT_NE + ONLY_NE).

Semantically Motivated Syntactic Features

Generating meaningful syntactic features required two component technologies: (a) frame semantics and (b) dependency-tree parsing. Each component technology is described and motivated in the next two subsections. The third subsection describes how these two technologies were used together to generate useful features for classification.

Frame Semantics

The goal of frame semantics is to manually enumerate the full range of semantic and syntactic combinatory possibilities of the words in a language (Ruppenhofer, Ellsworth, Petruck, Johnson, & Scheffczyk, 2006). A frame can be thought of as some meaningful event (e.g., death), action (e.g., forgive), or condition (e.g., certainty). FrameNet (Ruppenhofer et al., 2006) is a lexical database that maps certain words to frames and specifies for each frame its constituent semantic roles. A frame has two main ingredients: (a) a fixed set of *lexical units* (LUs) that individually mark the presence of the frame in discourse and (b) a set of core or optional *frame elements* (FEs) (a.k.a. semantic roles). For example, the apply_heat frame is associated with the lexical units *bake, boil, brown, cook*, and *simmer* and frame elements cook, food, container, heating_instrument, and duration. A frame's lexical unit can be thought of as the central word or phrase that triggers the frame in text. Frame elements are the semantic roles associated with the frame. For example, the apply_heat frame requires the element that is applying the heat (the cook) and the element that heat is being applied to (the food). FrameNet release 1.3[4] was used in this work. It contains 795 frames, 10,195 lexical units, and 7,124 frame elements. This release also contains a large set of annotated sentences with frames, LU, and FE each marked and labeled. For example, the sentence

$$\overbrace{\underbrace{\textit{Jim}}_{\text{cook}}\ \underbrace{\textit{boiled}}_{\text{LU}}\ \underbrace{\textit{the egg}}_{\text{food}}\ \underbrace{\textit{for 3 minutes.}}_{\text{duration}}}^{\text{apply\_ heat}}$$

exhibits the apply_heat frame, where the lexical unit *boiled* triggers the frame. *Jim* is the cook, *the egg* is the food, and *for 3 minutes* is filling the optional duration role. A lexical unit need not be a verb. The statement frame is associated with verb lexical units such as *acknowledge, address,* and *say,* as well as noun lexical units such as *report, denial,* and *explanation.*

Unfortunately, though not surprisingly, no frame in FrameNet maps directly to what we call a citation, though several frames seem relevant. For example, consider the following snippets from sentences known to be citations:

1. . . . *with the* $\overbrace{\underbrace{\textit{UN}}_{\text{speaker}}\ \underbrace{\textit{report}}_{\text{LU}}}^{\text{statement}}$ *underscoring* . . .

2. . . . *what the* $\overbrace{\underbrace{\textit{U.S. Humane Society}}_{\text{cognizer}}\ \underbrace{\textit{estimates}}_{\text{LU}}}^{\text{estimation}}$ *to be* . . .

3. . . . *and* $\overbrace{\underbrace{\textit{models by the IPCC}}_{\text{speaker}}\ \underbrace{\textit{predict}}_{\text{LU}}}^{\text{expectation}}$ *that*. . .

4. . . . $\overbrace{\begin{array}{c}\underbrace{\textit{accroding to}}_{\text{LU}}\\ \underbrace{\textit{a book by the Museum of Natural History}}_{\text{text}}\end{array}}^{\text{attribute\_information}}$ . . .

Sentence (1) exhibits the statement frame in which the *UN* is acting as the speaker. Sentence (2) exhibits the estimation frame in which the *U.S. Humane Society* is acting as the cognizer. Sentence (3) exhibits the expectation frame in which the *models by the IPCC* is acting as the speaker. Sentence (4) exhibits the attribute_information frame in which *a book by the Museum of Natural History* is acting as the text. It seems reasonable that, given a new sentence, if the entity in question is filling a similar frame-specific semantic role, then it is also being cited. A classifier could leverage from rules such as "If a sentence exhibits the estimation frame and the entity in question is acting as the speaker then the sentence is a citation." Thus, what is needed is a mechanism to determine that an entity (e.g., the *UN*) is filling a certain role (e.g., the speaker role) in a particular instance of a frame (e.g., the statement frame in Sentence 1, above).

The annotated sentences distributed with FrameNet exemplify how a frame-specific frame element may fit syntactically into a sentence that exhibits that frame. Specifically, annotated sentences exemplify possible syntactic relations between the frame element and the lexical unit. For example, annotations may show that in a statement frame, the speaker can occur as the subject of the verb *reported*, such as in

$$\underbrace{\textit{different studies}}_{\text{speaker}}\ \textit{reported that . . .}$$

or as the object of the prepositional phrase *reported by*, such as in

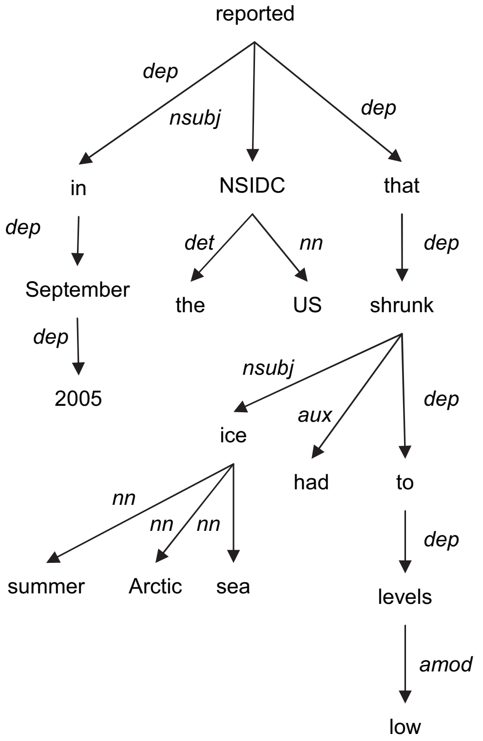
$$\textit{. . . reported by}\ \underbrace{\textit{different studies}}_{\text{speaker}}.$$

These syntactic relations can be captured by the dependency tree path between the FE and the LE. The next section describes dependency trees and how they are used to produce features for training a classifier.

Dependency Parse Trees

A dependency tree is a representation of a sentence that encodes the syntactic relations between the words in the sentence. Figure 2 shows the dependency parse tree for the sentence *In September 2005, the U.S. NSIDC reported that summer Arctic sea ice had shrunk to low levels.*PB Each word in the sentence is related to one parent, except the main verb, *reported*. Dependencies between child and parent are labeled according to a fixed set of syntactic relationships, such as *nsubj* (subject), *dobj* (direct object), and *amod* (adverb modifier). A dependency tree is a connected graph, so there is always a path from any word to any other word in the sentence. A word pair can share more than one connecting path, in which case choosing the shortest path has been shown to be a sensible heuristic for selecting the one that encodes the most meaningful syntactic relationship (Bunescu & Mooney, 2005; Erkan, Ozgur, & Radev, 2007).

FIGURE 2. Dependency tree representation of "*In September 2005, the U.S. NSIDC reported that summer Arctic sea ice had shrunk to low levels*."

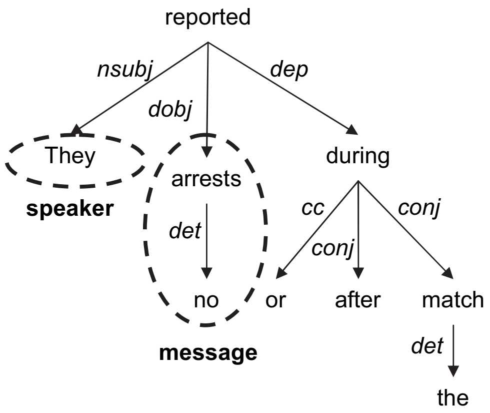


Suppose we are given the sentence in Figure 2 and we wish to know if there is a statement being made and, if so, (a) who is making the statement and (b) what the statement is. Suppose we have the dependency tree representation of the sentence. One annotated sentence in FrameNet exhibiting the statement frame is *They reported no arrests during or after the match.*, labeled as

$$\overbrace{\underbrace{\textit{They}}_{\text{speaker}}\ \underbrace{\textit{reported}}_{\text{LU}}\ \underbrace{\textit{no arrests}}_{\text{message}}}^{\text{statement}}\ \textit{during or after the match.}$$

The dependency tree representation of this sentence is shown in Figure 3, along with the spans of text filling the speaker and message roles. This tree shows *they*, the speaker, being

FIGURE 3. Dependency tree of FrameNet sentence "*They reported no arrests during or after the match*"., where *they* is marked as the speaker and *no arrests* is marked as the message.



the subject of the verb *reported* and *no arrests*, the message, being the direct object of the verb *reported*. We more concisely denote these syntactic rules as *reported* $\underrightarrow{\text{(nsubj)}}$ speaker and *reported* $\underrightarrow{\text{(dobj)}}$ message. These syntactic rules can then be applied to the dependency tree in Figure 2 to determine that *the U.S. NSIDC* is acting as the speaker and that *summer arctic sea ice had shrunk to the low levels* is the statement.

The next subsection describes how the pieces were put together to generate our FrameNet-based features.

Generating Frame-Target-Role Triples

All FrameNet-annotated sentences were dependency parsed using the Stanford parser.[5] Each FrameNet sentence is typically associated with one semantic frame, but may exhibit more than one. Each frame annotation shows the span of text that is the lexical unit and each span of text that corresponds to a frame element or semantic role. For each marked frame element (i.e., semantic role), we extracted the dependency tree path from the LU to the FE.[6] In total, we were able to process 134,471 example sentences exhibiting 592 frames (127 sentences per frame on average). Some annotated sentences were excluded due to parsing failure. A total of 85,093 syntactic patterns were found. Each syntactic pattern [e.g., *reported* $\underrightarrow{\text{(nsubj)}}$ speaker] describes how a frame-specific semantic role (e.g., the speaker in a statement frame) may fit syntactically in a sentence exhibiting that particular frame (e.g., the speaker is the subject of the verb *reported*). It describes the syntactic relationship between the semantic role and the lexical unit.

Using syntactic patterns directly as features posed a challenge. As mentioned, the number of syntactic-pattern features was 85,095, and our entire evaluation set was 6,000 sentences. In text classification, when the number of distinct features greatly outnumbers the number of instances in the training data, this is a problem. Features from the training set are less likely to reoccur in the test set, which reduces the predictive power of a learned model. Our solution was to resolve each syntactic pattern to its corresponding frame-LU-role triple (called frame-target-role triple from here forward). This reduced the feature set size from 85,095 patterns to 5,027 frame-target-role triples.

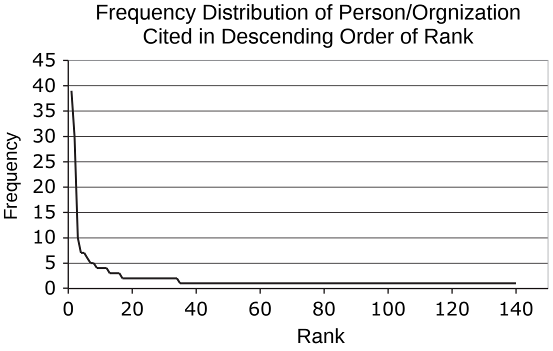
For each instance in the training and test set, its associated frame-target-role triples were generated by applying each of the 85,095 patterns to the sentence. A pattern is said to positively fit the sentence if the sentence's named entity (i.e., the marked person or organization—recall there is only one per sentence) falls into the role slot of the syntactic pattern. If a syntactic pattern fit the sentence, then the pattern's corresponding frame-target-role was included as a feature of the sentence. For example, in the sentence *The* [*IUCN*]$_{\text{org}}$ *is predicting a 30 percent reduction in the polar bear numbers in the next 45 years*.[PB], the pattern *predicting* $\underrightarrow{\text{(nsubj)}}$ speaker fits this sentence because the marked organization, the *IUCN*, is the noun subject of the verb *predicting*. This pattern corresponds to the frame-target-role triple statement-predicting-speaker. Thus, this triple would be added as one feature of this instance. Table 4 shows a few frame-target-role triples with some of their associated syntactic patterns.

TABLE 4. Example Frame-Target-Role Triples and Some of their Associated Syntactic Patterns

Frame-Target-Role Triple	Syntactic Pattern
statement-reported-speaker	*reported* $\underrightarrow{\text{(nsubj)}}$ speaker
	reported $\underrightarrow{\text{(prep)}}$ *by* $\underrightarrow{\text{(dep)}}$ speaker
	reported $\underrightarrow{\text{(prep)}}$ *from* $\underrightarrow{\text{(dep)}}$ speaker
evidence-revealed-support	*revealed* $\underrightarrow{\text{(nsubj)}}$ support
	revealed $\underrightarrow{\text{(prep)}}$ *by* $\underrightarrow{\text{(dep)}}$ support
	revealed $\underrightarrow{\text{(prep)}}$ *in* $\underrightarrow{\text{(dep)}}$ support
attributed_information-according-speaker	*according* $\underrightarrow{\text{(dep)}}$ *to* $\underrightarrow{\text{(dep)}}$ speaker
emotion_directed-pleased-experiencer	*pleased* $\underleftarrow{\text{(acomp)}}$ *feel* $\underrightarrow{\text{(nsubj)}}$ experiencer
	pleased $\underleftarrow{\text{(acomp)}}$ *was* $\underrightarrow{\text{(nsubj)}}$ experiencer
	pleased $\underleftarrow{\text{(acomp)}}$ *seemed* $\underrightarrow{\text{(nsubj)}}$ experiencer
emotion_directed-pleased-topic	*pleased* $\underleftarrow{\text{(acomp)}}$ *sounded* $\underrightarrow{\text{(dep)}}$ *about* $\underrightarrow{\text{(dep)}}$ topic

It is possible for a syntactic pattern to resolve to multiple frame-target-role triples. For example, the pattern *report* $\underrightarrow{\text{(nn)}}$ speaker and the pattern *report* $\underrightarrow{\text{(nn)}}$ topic describe the same syntactic pattern. The ambiguity stems from the fact that *X* in a noun phrase *the X report* can be the speaker of the statement (e.g., *the IPCC report*) or the topic of the statement (e.g., *the polar bear report*). In cases where a syntactic pattern resolves to multiple frame-target-role triples, each triple was added as a feature, weighted proportional to the level of ambiguity. In this case, both triples would be added with a weight of .5 because the ambiguity is between two frame-target-role triples.

FIGURE 4. Frequency distribution of persons/organizations cited, in descending order of frequency-based rank.



EXPERIMENTAL RESULTS

As previously mentioned, two humans coded a set of 6,000 sentences for training/testing. Two hundred and seventy nine (4.7%) sentences were labeled citation, and the remaining were labeled non-citation. The frequency distribution of entities referenced in citations was very heavy-tailed, as shown in Figure 4. Along the *x* axis, entities are ranked in descending order of times they were referenced in a citation. As the figure shows, a few persons and organizations were cited very frequently and many were cited only once or twice. The most frequently cited entity was the *Discovery Channel*, cited 39 times (14% of all citations). The second most frequently cited entity was *Al Gore*, cited 30 times (11% of all citations). The third most cited entity was the *IPCC*, cited ten times. Beyond the second-most cited entity, the number of citations per person/organization drops significantly. Eighteen entities were cited twice (36 references, accounting for 13% of citations). Most entities were cited only once (106 references, accounting for 38% of all citations).

Results are presented in terms of precision (P), recall (R), and F-measure (F1), calculated by micro- and macro-averaging, for reasons motivated below. Precision (P) measures prediction accuracy with respect to the target class, in this case the citation class: *Of the instances predicted citation, what percentage were correct*? Recall (R) measures coverage: *Of the instances that are true citations, what percentage were correctly predicted citation*? F-measure (F1) is the harmonic mean of precision and recall. Micro-averaged P, R, and F1 are defined as

$$\mathrm{P}_{\mathrm{micro}} = \frac{\#\ \text{correct citation predictions}}{\#\ \text{citation predictions}},$$

$$\mathrm{R}_{\mathrm{micro}} = \frac{\#\text{correct citation predictions}}{\#\ \text{true citations}},$$

$$\mathrm{F1}_{\mathrm{micro}} = \frac{(2 \times \mathrm{P}_{\mathrm{micro}} \times \mathrm{R}_{\mathrm{micro}})}{(\mathrm{P}_{\mathrm{micro}} + \mathrm{R}_{\mathrm{micro}})}.$$

Macro-averaged P, R, and F1 are defined as

$$\mathrm{P}_{\mathrm{micro}} = \frac{\sum_{x \in S_{\mathrm{pred}}} \frac{\#\ \text{correct citation predictions that reference } x}{\#\ \text{citation predictions that reference } x}}{|S_{\mathrm{pred}}|},$$

$$\mathrm{R}_{\mathrm{micro}} = \frac{\sum_{x \in S_{\mathrm{true}}} \frac{\#\ \text{correct citation predictions that reference } x}{\#\ \text{true citation that reference } x}}{|S_{\mathrm{true}}|},$$

$$\mathrm{F1}_{\mathrm{micro}} = \frac{(2 \times \mathrm{P}_{\mathrm{macro}} \times \mathrm{R}_{\mathrm{macro}})}{(\mathrm{P}_{\mathrm{macro}} + \mathrm{R}_{\mathrm{macro}})},$$

where S_{pred} defines the set of unique named entities in predicted citations and S_{true} defines the set of unique named entities in true citations.

Micro-averaging treats all citations as equally useful. Correctly predicting a citation that references a frequently cited entity affects micro-averaged P, R, and F1 the same as correctly predicting a citation that references a rarely cited entity. Micro-averaged metrics are dominated by what happens with the frequent cases (e.g., *Discovery Channel*, *Al Gore*, *IPCC*). In contrast, in computing $\mathrm{P}_{\mathrm{macro}}$, precision is calculated for each entity x in a predicted citation ($x \in S_{\mathrm{pred}}$) and averaged across entities. Similarly, in computing $\mathrm{R}_{\mathrm{macro}}$, recall is calculated for each entity x in a true citation ($x \in S_{\mathrm{true}}$) and averaged across entities. Correctly predicting a citation that references a rarely cited entity increases macro-averaged P, R, and F1 more than correctly predicting a citation that references a frequently cited entity. Macro-averaged metrics are dominated by what happens with the rare cases. Both micro- and macro-averaged P, R, and F1 are presented because the distribution of entities referenced in true citations was very heavy-tailed. A classifier that catches all references to the *Discovery Channel* and *Al Gore* catches 25% of all citations (i.e., $\mathrm{R}_{\mathrm{micro}} = 0.25$). However, a different classifier that also catches 25% of all citations but finds a wider variety of entities cited may be preferred in some cases.

Evaluation was performed via tenfold cross-validation. The data was divided into ten folds. During each cross-validation step, each classifier was trained on nine folds (90% of the data) and evaluated on the held-out fold (10% of the data). This process was repeated ten times. The P, R, and F1 values presented are the average of the ten cross-validation runs. Where specified, standard deviation numbers are also derived from these ten cross-validation results. The same folds were used in all experiments, allowing a per-fold comparison across all classifiers. For this reason, significance testing was done using a two-tailed paired t-test.

Eleven classifiers were evaluated:

1. ALL
2. ALL_BUT_NE
3. ONLY_NE
4. ALL_BUT_NE + ONLY_NE
5. FTR
6. FTR + ALL
7. FTR + ALL_BUT_NE
8. FTR + ONLY_NE
9. FTR + ALL_BUT_NE + ONLY_NE

10. META (ALL_BUT_NE + ONLY_NE)
11. META (FTR + ALL_BUT_NE + ONLY _NE)

The first four classifiers use only bag-of-words features. The fifth classifier, FTR, uses only our frame-target-role (FTR) triples as features. The next four classifiers (6, 7, 8, and 9) individually combine a bag-of-words feature set with our frame-target-role triples. The last two classifiers (10 and 11) are meta-classifiers. META (ALL_BUT_NE ONLY_NE) combines the output of ALL_BUT_NE and ONLY_NE, while META (FTR+ALL_BUT_NE + ONLY_ NE) combines the output of FTR, ALL_BUT_NE, and ONLY_NE. The two meta-classifiers were trained using two tiers of cross-validation. The second tier does ninefold cross-validation on the first tier's training folds. The main point is that these two tiers of cross-validation ensure that the meta-classifier is not trained indirectly with evidence derived from the test set.

Evaluation results in terms of micro- and macro-averaged P, R, and F1 are given in Table 5 and Table 6, respectively. Micro- and macro-averaged F1 is shown graphically in Figure 5.

The best f-measure (F1), both micro- and macro-averaged, was obtained by META (FTR+ALL_BUT_NE+ONLY_NE), which outperformed all other classifiers, including META (ALL_BUT_NE+ONLY_NE), by a statistically significant margin (p <.01). META (FTR + ALL_BUT_NE+ONLY_NE) found 66% of all true citations with an 18% prediction error rate (i.e., $1-P_{micro} = .18$). It should be noted that our estimate of recall is optimistic, as we only consider sentences in which at least one entity was tagged as a person or organization. Citations missed due to a miss by the named entity tagger are not factored into our recall estimate.

Several conclusions can be drawn from these results. First, in terms of macro- and micro-averaged F1, all three bag-of-words models that borrow evidence from the named entity potentially being cited (i.e., ALL, ONLY_NE, and ALL_BUT_NE + ONLY_NE) outperformed ALL_BUT_NE, which ignores the entity being cited (p < .05). In fact, ONLY_NE, which considers only features occurring within cited entities in the training set, was statistically indistinguishable from ALL and ALL_BUT_ NE + ONLY_NE, which consider the entire sentence.

In this dataset, 62% of citations referenced a named entity that was cited more than once. This helps explain ONLY_NE's performance in terms of micro-average F1. It is surprising, however, that ONLY_NE performed as well as it did in terms of macro-averaged F1. Again, macro-averaged F1 is dominated by precision and recall with respect to entities cited rarely (i.e., have few training examples). Of the 140 unique named entities mentioned in the 279 sentences labeled citation by the human coders, only 20 (14%) were referenced in more than one citation and not mentioned in a single

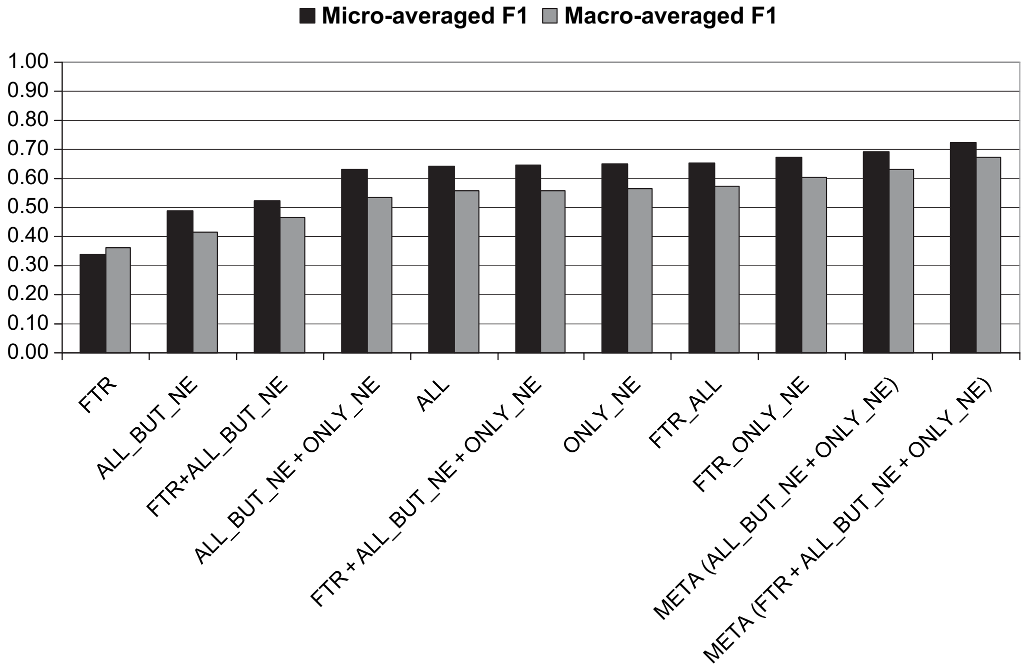
TABLE 5. Micro-Averaged Evaluation Results (± Standard Deviation)

		P_{micro}	R_{micro}	$F1_{micro}$
BOW	ALL	0.869 ± 0.096	0.517 ± 0.082	0.642 ± 0.069
	ALL_BUT_NE	0.818 ± 0.136	0.352 ± 0.084	0.487 ± 0.099
	ONLY_NE	0.806 ± 0.045	0.549 ± 0.087	0.648 ± 0.056
	ALL_BUT_NE+ONLY_NE	0.917 ± 0.090	0.492 ± 0.101	0.633 ± 0.089
FTR	FTR	0.643 ± 0.189	0.233 ± 0.090	0.339 ± 0.122
FTR+BOW	FTR + ALL	0.869 ± 0.100	0.53 ± 0.088	0.654 ± 0.080
	FTR + ALL_BUT_NE	0.829 ± 0.132	0.388 ± 0.095	0.523 ± 0.106
	FTR + ONLY_NE	0.828 ± 0.047	0.578 ± 0.109	0.675 ± 0.074
	FTR + ALL_BUT_NE + ONLY_NE	0.921 ± 0.084	0.506 ± 0.094	0.648 ± 0.086
Meta	META (ALL_BUT_NE + ONLY_NE)	0.813 ± 0.052	0.613 ± 0.109	0.694 ± 0.073
	META (FTR+ALL_BUT_NE + ONLY_NE)	0.813 ± 0.053	0.664 ± 0.107	0.725 ± 0.062

TABLE 6. Micro-Averaged Evaluation Results (± Standard Deviation)

		P_{macro}	R_{macro}	$F1_{macro}$
BOW	ALL	0.825 ± 0.124	0.427 ± 0.069	0.557 ± 0.063
	ALL_BUT_NE	0.788 ± 0.153	0.288 ± 0.093	0.416 ± 0.114
	ONLY_NE	0.806 ± 0.076	0.445 ± 0.093	0.564 ± 0.057
FTR	FTR	0.642 ± 0.190	0.256 ± 0.098	0.362 ± 0.129
FTR + BOW	FTR + ALL	0.826 ± 0.131	0.445 ± 0.080	0.573 ± 0.081
	FTR + ALL_BUT_NE	0.807 ± 0.143	0.332 ± 0.099	0.464 ± 0.115
	FTR + ONLY_NE	0.844 ± 0.074	0.482 ± 0.120	0.603 ± 0.079
	FTR + ALL_BUT_NE + ONLY_NE	0.900 ± 0.114	0.409 ± 0.079	0.557 ± 0.082
Meta	META (ALL_BUT_NE + ONLY_NE)	0.822 ± 0.081	0.528 ± 0.117	0.632 ± 0.073
	META (FTR + ALL_BUT_NE + ONLY_NE)	0.820 ± 0.086	0.587 ± 0.112	0.674 ± 0.057

FIGURE 5. Micro- and macro-averaged F1 results.



non-citation. These can be considered easy cases for ONLY_NE. So, how could ONLY_NE achieve a macro-averaged recall of .445? The answer is that there was significant vocabulary overlap between unique entities referenced in citations, for two reasons. First, synonymous or coreferential unique entities (e.g., *Discovery Channel*, *Discovery TV Channel*, and *The Discovery Channel* or *Al Gore* and *Gore*) were not manually mapped to a common canonical form. In other words, it was possible for two unique entities to refer to the same actual person or organization. Second, even across conceptually distinct entities, there was some vocabulary overlap with words such as *center, association, institute, magazine, academy, report, service*, and *federation*. Of the 106 unique entities that were referenced in only one citation (potentially difficult cases for ONLY_NE), 68 (64%) contained at least one non-stopword term stem that appeared in another named entity that was also cited.

Second, as is shown in Figure 5, FTR was the only classifier for which macro-averaged F1 was higher than micro-averaged F1. Therefore, of the classifiers evaluated, it was the least influenced by entities cited in the training data. Not being influenced by the entities cited in the training data is a desired property for model transfer, training a model on one corpus and applying it to another corpus. For instance, a classifier such as ONLY_NE might not transfer well across corpora in which different entities are referenced in citations. FTR's performance was low relative to the other classifiers. However, this result suggests that syntactic features, which ignore content and focus on structure, have potential value.

Third, both meta-classifiers outperformed their multiple-feature-set, single-classifier counterparts. In terms of micro- and macro-averaged F1, META (ALL_BUT_NE + ONLY_NE) outperformed ALL_BUT_NE + ONLY_NE and META (FTR + ALL_BUT_NE + ONLY_NE) outperformed FTR + ALL_BUT_NE + ONLY _NE (*p* <. 01). Combining these feature sets in a meta-classification framework worked better than merging the feature sets at the input of a single classifier.

Finally, although FTR performs worse than all other classifiers, frame-target-role triples added value in a meta-classification framework. META (FTR + ALL_BUT_NE + ONLY_NE outperformed META (ALL_BUT_NE + ONLY_NE) in terms of micro- and macro-averaged F1 (*p* < .01). The only difference between these two meta-classifiers is that the better performing one uses FTR's predictions as an additional input.

Learning Curves

Figures 6 and 7 show micro- and macro-averaged f-measure (F1) as a function of the amount of training data used to train each model, from 10% to 90% of the data in 10% increments. The total amount of training data was 6,000 sentences. We focus on FTR, ALL, ALL_BUT_NE, ONLY_NE, FTR + ALL_BUT_NE + ONLY_NE, META (ALL_BUT_NE + ONLY_NE), and META (FTR + ALL_BUT_NE + ONLY_NE).

These learning curves reveal several trends. First, in terms of micro- and macro-averaged F1, ONLY_NE and both meta-classifiers perform at comparable levels with little training data, but the meta-classifiers improve over ONLY_NE with more training data. The difference in performance is greater in terms of macro-averaged F1, reinforcing the claim that by combining feature sets both the meta-classifiers gravitate less toward citations referencing entities cited in the training data. Second, at low levels of training data (10%–30%) FTR outperformed ALL_BUT_NE. However, with more training data, ALL_BUT_NE outperformed FTR in terms of micro-averaged F1. This suggests that although ALL_BUT_NE ignores features from within the tagged named entity, it is still biased toward citations referencing entities in the training data.

ERROR ANALYSIS

Frame-Target-Role Triples

There are at least two reasons why the FTR classifier alone did not perform better: (a) low coverage and (b) ambiguous frame-target-role triples. In terms of coverage, of the 279 sentences labeled citation by the coders, 123 (44%) had no matching syntactic pattern, thereby no frame-target-role triple. For those sentences, the classifier had to make a blind prediction. There were two reasons for a sentence having no matching syntactic pattern. The first is that our inventory of frame-target-role triples is limited by what is covered in FrameNet. Not every word maps to a lexical unit in FrameNet. For example, the verb *publish* is not associated with a frame in FrameNet. There were eight citations in which the referenced person or organization appeared as the subject of the verb *publish*. Limited coverage is a known problem with FrameNet, and prior work has focused on extending it (Green, Dorr, & Resnik, 2004).

FIGURE 6. Learning curves: micro-averaged F1.

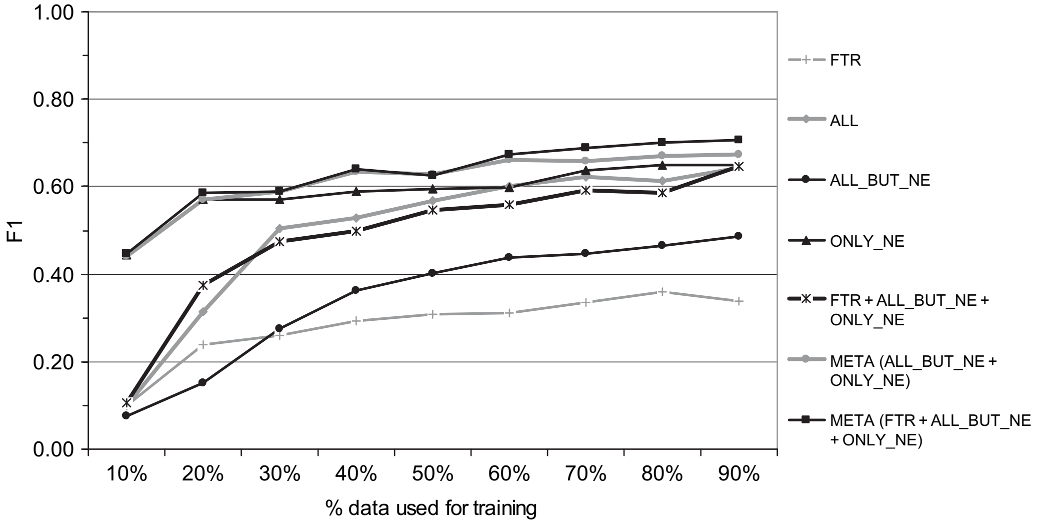
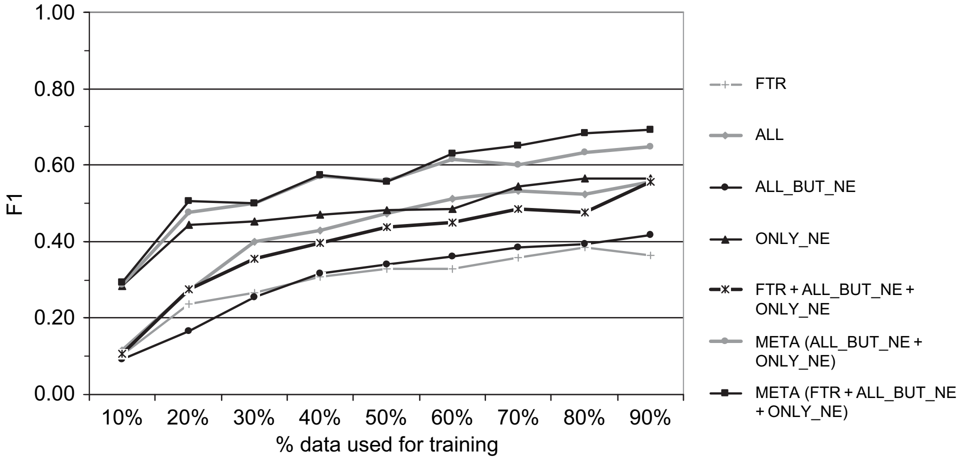


FIGURE 7. Learning curves: macro-averaged F1.



The second reason is that the annotated corpora distributed with FrameNet do not represent all the syntactic possibilities for a frame-target-role triple. This problem potentially could be alleviated by requiring only a soft matching between the syntactic context of a cited entity and a pattern in our inventory.

As expected, some frame-target-role triples provide more evidence of a citation than others. Consider the two frame-target-role triples statement-report-speaker and statement-said-speaker. The first triple provides stronger evidence of a citation because the lexical unit *report* inherently suggests that the claim being made is concrete and not obvious. Indeed, the first triple was associated with 15 citations and only two non-citations. However, speakers often make use of less reliable frame-target-role triples in citations. Consider the following two sentences. The first one is a citation, but not the second.

1. . . . *an official with the* [*Canadian Department of Fisheries and Oceans*]$_{\text{org}}$ said: *we've noticed that the ice over the past four or five years has been deteriorating*. . .$^{\text{PB}}$
2. *President* [*Bush*]$_{\text{per}}$ *in his presidential campaign said that he was campaigning on an environmental platform*.$^{\text{PB}}$

The only way to tell that the first statement is a citation is to consider the content of the information being attributed to the speaker. However, syntactically motivated features such as those used by FTR focus on structure rather than content.

Improving Overall Recall

Our best performing classifier, META (FTR+ALL_BUT_ENT+ONLY_NE), found approximately 66% of all citations (an optimistic estimate, as previously mentioned) with prediction accuracy of 81% micro-averaged. Ninety-four citations were missed. Seventy-seven out of these 94 misses (87%) were predicted non-citation by all three base classifiers that produced the three input features to this meta-classifier (FTR, ALL_BUT_NE, and ONLY_NE). Fifty-eight of these 94 misses (62%) were references to entities that were referenced only once. These were difficult cases for ONLY_NE, which focuses only on terms from the marked named entity. Forty-one of these 94 misses (44%) were sentences for which no syntactic pattern matched. These were featureless data points for FTR.

An improvement in recall would likely come from an improvement in named entity tagging accuracy. Remember that annotators were instructed to label named entity tagging errors as non-citations. Mistakes by the named entity tagger pose a challenge because they make non-citations look like citations. For example, in our evaluation, all base classifiers (and hence our best meta-classifier) missed the citation [*Nick Lunn*]$_{\text{per}}$ *shows the polar bears of Hudson Bay coming on land earlier as the sea ice melts earlier*.$^{\text{PB}}$ The reason was that four non-citations shared a similar local context:

1. [*Science*]$_{\text{org}}$ *shows that polar bear populations are declining as a result of global warming*.$^{\text{PB}}$
2. [*Science*]$_{\text{org}}$ *shows us a shrinking world for the polar bear*.$^{\text{PB}}$
3. [*Research*]$_{\text{org}}$ *shows that the sea ice in the Arctic has retreated farther*. . .$^{\text{PB}}$
4. [*Science*]$_{\text{org}}$ *shows that polar bears are threatened with extinction from*. . .$^{\text{PB}}$

These were labeled non-citation by the human coders because the first word in each was mistakenly labeled as an organization. Because the verb *shows* co-occurred more often with non-citations than citations, this feature and the syntactic pattern *shows* $\xrightarrow{\text{(nsubj)}}$ speaker were associated with the non-citation class, decreasing recall.

Difficult Cases

In some cases, the difference between a citation and a non-citation can be subtle. There is no reason to believe that such cases are uncommon across different corpora. Consider the following sentences:

1. . . . *an official with the* [*Canadian Department of Fisheries and Oceans*]$_{\text{org}}$ *said: we've noticed that the ice over the past four or five years has been deteriorating*. . .$^{\text{PB}}$
2. [*Carl Sagan*]$_{\text{per}}$ *said: there are no useless threads in the fabric of life*.$^{\text{PB}}$
3. *Like* [*Ben Franklin*]$_{\text{per}}$ *said: we all hang together or we hang separately*.$^{\text{PB}}$

The first reference is a citation. The second and third are not. However, the difference between the first quote and the last two is not directly accessible to an automated system that lacks domain (or even common) knowledge and uses evidence only from the sentence in question. A human might draw a distinction between the first statement and the last two based on three criteria:

1. Authority: It is easy for a human to guess that the *Canadian Department of Fisheries and Oceans* is more authoritative than *Carl Sagan* or *Ben Franklin* when it comes to the polar bear and its habitat.
2. Relevance: The information attributed to the *Canadian Department of Fisheries and Oceans* is more relevant than the claims attributed to *Carl Sagan* and *Ben Franklin*. It uses language that more closely matches the language central to the debate (e.g., words such as *ice* and *deteriorating* rather than *fabric* and *hang*).
3. Vagueness: The last two statements are vague, meaning that their interpretation is more open-ended than the claim made in the first sentence. The information attributed to the *Canadian Department of Fisheries and Oceans* is more concrete.

Relevance is something that has been widely studied in text classification and information retrieval. A possible improvement could come from adding a feature that quantifies the similarity between the language of the sentence in question and the language characterizing the proposal, which is exemplified in the text from the Federal Register, the publicly accessible document that describes the proposed regulation.

Measuring authority and vagueness are more difficult problems. Quantifying authority might require an external resource such as the Web or the Wikipedia.[7] Some named entities referenced in citations in the Polar Bears corpus (e.g., the IPCC, the United Stated Geological Survey, the Canadian Wildlife Service, and the Natural Resources Defense Council) have Wikipedia entries that are linked (directly or indirectly) to the Wikipedia article on polar bears. A direct or indirect link between two Wikipedia articles marks a relationship between the source and target entity. Entities more tightly linked to the page on polar bears may be more authoritative than those less tightly linked. However, this approach is at the mercy of what is covered in Wikipedia and may not lead to a corpus general solution.

Of the three dimensions listed above, vagueness seems to be the most difficult to quantify by automated means. In fact, we are unaware of any work that has attempted to automatically quantify degree of vagueness in text. Some prior work has focused on classifying statements into *subjective* and *objective* (Wiebe, Wilson, Bruce, Bell, & Martin, 2004). However, subjectivity and vagueness are not exactly the same phenomenon. A citation may contain a specific subjective component (i.e., the author cites hard evidence and then expresses his or her opinion). Another possibly relevant vein of research is that of automatically recognizing metaphors in literary text. Pasanek and Scully (in press) show that an SVM classifier using a bag-of-words representation can tell metaphors and non-metaphors apart.

RELATED WORK

To our knowledge, no prior work has focused specifically on detecting sentences that make reference to an external source of information in informal text such as public comments. Most prior work on automatic citation analysis focuses on formal text, such as academic literature. A citation in academic text is defined as a meaningful link between the work being presented and work done in the past and is usually presented in a stylized manner. In the next subsection, we review prior work on citation analysis in academic text. Then, we survey relevant syntax-based extraction technology that has been applied to similar problems.

Citation Analysis in Academic Text

Citation analysis in academic text can be divided into four problems: (a) recognizing links between documents, (b) locating citations in the body of the text, (c) disambiguating which paper (from the references section) a citation in the body refers to, and (d) determining the purpose or function of a citation. See Teufel, Siddharthan, and Tidhar (2006b) for an example of work on predicting citation function. With respect to the first task, academic papers contain a references section that lists, usually in a consistent format, the papers cited in the body of the document. Thus, establishing a link between a paper and prior work comes down to correctly parsing the references section. Task (b), locating citations in the body of the text, sounds like it would be related to our work. However, the vast majority of citations in the body of academic text are *formal* citations, using the terminology from Powley and Dale (2007). Formal citations mostly follow a consistent, recognizable format (e.g., the Author-Year pair). Powley and Dale (2007) report high accuracy (i.e., greater than .99 precision and greater than .96 recall) in detecting formal citations.

Some citations in academic literature are known as *informal* citations, such as *This approach has many strong points, but does not provide a very satisfactory account of the adherence to discourse conventions in dialogue*., borrowed from Siddharthan & Teufel (2007). Informal citations look more like citations in public comments. Prior work on detecting informal citations treats it as a two step process. First, high-recall/low-precision heuristics are used to collect candidate informal citations. For example, Kim & Webber (2006) examine sentences with the pronoun *they*. Siddharthan and Teufel (2007) focus on a predefined set of referring expressions, including pronouns and work nouns such as *approach, study, investigation*, and *result*. In the second step, the problem is formulated as multiclass classification, where there is one class for every document in the references, one class allotted to the current document, and one class for "no-document." Indirectly, this two-step process is doing informal citation detection. If the referring expression refers to a paper in the references, then it is a citation. Detecting citations in public comments cannot be framed the same way as detecting informal citations in academic text. Public comments do not contain a references section, and, more generally, a finite set of citable sources does not exist for any corpus.

Syntax and Extraction

Motivated by the argument that syntax (the grammatical relation between words) is important for recognizing events in text, recent work has focused on dependency-tree paths for extraction and classification. The common approach, as in our FTR approach, is to use syntactic patterns as features in a machine learning setting. Approaches differ in how they decide which types of syntactic patterns are worth adding as features and which ones are not. For example, Yangarber (2003) and Stevenson and Greenwood (2005) focus on subject-verb-object relations, where each pattern is a verb and its subject and/or object. The motivation behind their approach is that the entities of interest favor certain predicates more than others (e.g., *organizations* tend to *acquire, merge*, and *sell* but not *eat, sleep*, and *marry*). This intuition also motivates Riloff's system, Autoslog (Riloff, 1993), which focuses on a fixed set of grammatical relations, including prepositional phrases and noun phrases.

One important hurdle in using syntactic information for extraction is that not every part of the syntactic pattern is important. Also, a syntactic pattern may include two or more words. Longer syntactic patterns are less frequent and have potentially less predictive power. Recent efforts have focused on soft matching between syntactic patterns to avoid this sparcity problem. Erkan etal. (2007) use dependency parsing to determine if two proteins mentioned in the same sentence are said to interact in some metabolic process. The path in a dependency parse from one protein to the other is used as evidence of their relation. Their soft-matching approach uses edit distance, which quantifies the similarity between two ordered sequences. Bunescu & Mooney (2005) use the dependency tree path between two entities in a sentence to classify their relation with respect to a fixed set of relations of interest (e.g., *person* is affiliated with *organization* or *person* has a family connection with *person*). Their dependency path is augmented with semantic information, to make a matching more plausible. In our approach, frame-target-role triples are resolved by applying our inventory of patterns in an exact match strategy. A possible improvement could come from implementing a soft-matching heuristic. However, combining FTR with simpler representations alleviates some of the shortcomings of hard matching.

CONCLUSION

We investigated the problem of manually and automatically recognizing citations in public comment corpora, addressed here for the first time. A significant portion of this work focused on refining our citation annotation manual up to the point where it produced acceptable inter-annotator agreement. We used existing machine learning techniques to learn to automatically detect citations from training data. Several feature sets were evaluated individually and in combination. Each bag-of-words feature set focused on a different part of the sentence in question (e.g., the entire sentence, only the potentially cited named entity, and the entire sentence, except the named entity). Our FrameNet-motivated features focused on how potentially cited named entities fit syntactically into their local context. We obtained our best results by combining feature sets in a meta-classification framework. This classifier found 66% of all citations (an optimistic estimate of recall) with prediction accuracy of 81%. Although this is an encouraging result, more work is needed to produce a solution that does not require extensive human annotation to produce training data.

One worrisome result was that macro-averaged recall was substantially lower than micro-averaged recall for most algorithms. A classifier that finds more references to unknown external sources (i.e., named entities not seen in training data citations) might be more valuable to a policy maker. Also, a classifier that favors references to entities already seen in the training data is less likely to transfer well across corpora. Ideally, we want a model that can be trained on one corpus and used to detect citations in a different corpus. Otherwise, human effort must be expended to produce training data for each new corpus. In our evaluation corpus, 62% of citations referenced an entity that was cited more than once. The Polar Bears corpus may not be an outlier in this respect. The challenge is one of representation. Under a bag-of-words representation, if authors consistently cite the same entities and do so in an inconsistent way, a model will favor features (i.e., terms) related to the entities cited in the training data. However, while these features are effective for the same corpus as the training data, they might not transfer well to a different corpus in which different entities are cited. Syntactic features, which ignore all content and focus on structure, might be more transferable across corpora (macro-averaged F1 was higher than micro-averaged F1 using our syntactic features). However, more work is needed to increase the coverage and accuracy of our syntactic-based model.

An attractive research direction to explore in future work is that of *co-training* (Blum & Mitchell, 1998). In co-training, different classifiers leverage distinct feature sets and are bootstrapped to make use of unlabeled data to improve performance. We obtained our best result by combining classifiers in a meta-classification framework, where each base classifier focused on a different feature set. This positive result indicates that our base classifiers made different types of mistakes, which is a desired property for co-training to work. A confident prediction made by one classifier could be used as training data for a different classifier. A bootstrapping framework could also be useful in model transfer. For instance, a model based on syntactic features might transfer better across corpora than a model based on content features (different corpora cover different topics), though it may suffer from low recall. If a good content-based model can be learned with little training data, then a few highly reliable predictions from a syntax-based model could provide enough traction for a higher recall, content-based model to be learned.

NOTES

[1] http://erulemaking.cs.cmu.edu/data.php

[2] http://www.qdap.pitt.edu.

[3] http://opennlp.sourceforge.net/

[4] http://framenet.icsi.berkeley.edu

[5] http://nlp.stanford.edu/downloads/lex-parser.shtml

[6] For consistency, each dependency-path pattern arbitrarily starts with the lexical unit and concludes with the frame element.

[7] http://en.wikipedia.org/wiki/polar_bear

REFERENCES

Bikel, D. M., Schwartz, R. L., & Weischedel, R. M. (1999). An algorithm that learns what's in a name. *Machine Learning, 34* (1–3), 211–231.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of COLT* (pp. 92–100). New York: Morgan Kaufmann Publishers.

Bunescu, R. C., & Mooney, R. J. (2005). A shortest path dependency kernel for relation extraction. In *Proceedings of HLT* (pp. 724–731). Association for Computational Linguistics.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics, 22*(2), 249–254.

Erkan, G., Ozgur, A., & Radev, D. R. (2007). Sem supervised classification for extracting protein interaction sentences using dependency parsing. In *Proceedings of EMNLP* (pp. 228–237). Association for Computational Linguistics.

Green, R., Dorr, B. J., & Resnik, P. (2004). Inducing frame semantic verb classes from WordNet and LDOCE. In *Proceedings of ACL* (pp. 375–382). Association for Computational Linguistics.

Joachims, T. (1998). Making large-scale support vector machine learning practical. In B. Scholkopf, C. Burges, & A. Smola (Eds.), *Advances in kernel methods: Support vector machines*. MIT Press.

Kim, Y., & Webber, B. (2006). Automatic reference resolution in astronomy aricles. In *Proceedings of CODATA*. The Committee on Data for Science and Technology.

Pasanek, B., & Scully, D. (in press). Mining millions of metaphors. In *Literary and linguistic computing.* Oxford Journals.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Powley, B., & Dale, R. (2007). Evidence-based information extraction for high accuracy citation and author name identification. In *Proceedings of RIAO*. C.I.D.

Riloff, E. (1993). Automatically constructing a dictionary for information extraction tasks. In *Proceedings of AAAI* (pp. 811–816). AAAI Press/The MIT Press.

Ruppenhofer, J., Ellsworth, M., Petruck, M. P. L., Johnson, C. R., & Scheffczyk, J. (2006). *Framenet II: Extended theory and practice, main project document*.

Siddharthan, A., & Teufel, S. (2007). Whose idea was this, and why does it matter? Attributing scientific work to citations. In *Proceedings of NAACL/HLT* (pp. 316–323). Association for Computational Linguistics.

Stevenson, M., & Greenwood, M. A. (2005). A semantic approach to IE pattern induction. In *Proceedings of ACL* (pp. 379–386). Association for Computational Linguistics.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006a). An annotation scheme for citation function. In *Proceedings of SIGDIAL* (pp. 80–87). Association for Computational Linguistics.

Teufel, S., Siddharthan, A., & Tidhar, D. (2006b, July). Automatic classification of citation function. In *Proceedings of EMNLP* (pp. 103–110). Association for Computational Linguistics.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics, 30*(3), 277–308.

Wolpert, D. H. (1990). *Stacked generalization* [Tech. Rep. No. LA-UR-90-3460]. Complex Systems Group, Los Alamos, NM.

Yang, H., & Callan, J. (2006). Near-duplicate detection by instance-level constrained clustering. In *Proceedings of SIGIR* (pp. 421–428). Association for Computational Machinery.

Yangarber, R. (2003). Counter-training in discovery of semantic patterns. In *Proceedings of ACL* (pp. 343–350). Association for Computational Linguistics.

Ziman, J. M. (1968). *Public knowledge: An essay concerning the social dimensions of science*. Cambridge, England: Cambridge University Press.