Detecting Forum Authority Claims in Online Discussions

Alex Marin, Bin Zhang, Mari Ostendorf
Department of Electrical Engineering
University of Washington
{amarin, binz}@uw.edu, mo@ee.washington.edu

Abstract

This paper explores the problem of detecting sentence-level forum authority claims in online discussions. Using a maximum entropy model, we explore a variety of strategies for extracting lexical features in a sparse training scenario, comparing knowledge- and data-driven methods (and combinations). The augmentation of lexical features with parse context is also investigated. We find that certain markup features perform remarkably well alone, but are outperformed by data-driven selection of lexical features augmented with parse context.

1 Introduction

In multi-party discussions, language is used to establish identity, status, authority and connections with others in addition to communicating information and opinions. Automatically extracting this type of social information in language from discussions is useful for understanding group interactions and relationships.

The aspect of social communication most explored so far is the detection of participant role, particularly in spoken genres such as broadcast news, broadcast conversations, and meetings. Several studies have explored different types of features (lexical, prosodic, and turn-taking) in a variety of statistical modeling frameworks (Barzilay et al., 2000; Maskey and Hirschberg, 2006; Liu, 2006; Liu and Liu, 2007; Vinciarelli, 2007; Laskowski et al., 2008; Hutchinson et al., 2010). Typically, these studies assume that a speaker inhabits a role for the duration of the discussion, so multiple turns contribute to the decision. Participant status is similar although the language of others is often more relevant than that of the participant in question.

Communication of other types of social information can be more localized. For example, an attempt to establish authority frequently occurs within a single sentence or turn when entering a discussion, though authority bids may involve multiple turns when the participant is challenged. Similarly, discussion participants may align with or distance themselves from other participants with a single statement, or someone could agree with one person at a particular point in the conversation and disagree with them at a different point. Such localized phenomena are also important for understanding the broader context of that participant's influence or role in the conversation (Bunderson, 2003).

In this paper, we focus on a particular type of authority claim, namely forum claims, as defined in a companion paper (Bender et al., 2011). Forum claims are based on policy, norms, or contextual rules of behavior in the interaction. In our experiments, we explore the phenomenon using Wikipedia discussion ("talk") pages, which are discussions associated with a Wikipedia article in which changes to the article are debated by the editors in a series of discussion threads. Examples of such forum claims are:

- *I do think my understanding of Wikipedia and policy is better than yours.*
- *So it has all those things going for it, and I do think it complies with [[WP:V]] and*

Proceedings of the Workshop on Language in Social Media (LSM 2011), pages 39–47,
Portland, Oregon, 23 June 2011.

[[WP:WTA]].

- *Folks, please be specific and accurate when you [[WP:CITE—cite your sources]].*

We treat each discussion thread as a unique "conversation". Each contiguous change to a conversation is treated as a unique "post" or turn. The dataset and annotation scheme are described in more detail in the companion paper.

Related previous work on a similar task focused on detecting attempts to establish topic expertise in Wikipedia discussions (Marin et al., 2010). Their work used a different annotation process than that which we build on here. In particular, the annotation was performed at the discussion participant level, with evidence marked at the turn level without distinguishing the different types of claims as in (Bender et al., 2011).

Treating the problem of detecting forum claims as a sentence-level classification problem is similar to other natural language processing tasks, such as sentiment classification. Early work in sentiment analysis used unigram features (Pang and Lee, 2004; Pang and Lee, 2005). However, error analyses suggested that highly accurate sentiment classification requires deeper understanding of the text, or at least higher order n-gram features. Kim and Hovy (2006) used unigrams, bigrams, and trigrams for extracting the polarity of online reviews. Gilbert et al. (2009) employed weighted n-grams together with additional features to classify blog comments based on agreement polarity. We conjecture that authority claim detection will also benefit from moving beyond unigram features.

The focus of the paper is on two questions in feature extraction:

- Can we exploit domain knowledge to address overtraining issues in sparse data conditions?
- Is parse context more effective than n-gram context?

Our experiments compare the performance obtained using multiple methods for incorporating linguistic or data-driven knowledge and context into the feature space, relative to the baseline n-gram features. Section 2 describes the general classification architecture. Section 3 describes the various features implemented. Experimental results are presented in section 4. We conclude with some analysis in section 5 and remarks on future work in section 6.

2 System Description

We implement a classification system that assigns a binary label to each sentence in a conversation, indicating whether or not a forum authority claim is being made in that sentence. To obtain higher-level decisions, we apply a simple rule that any post which contains at least one sentence-level forum authority claim should be labeled positive. We use the sentence-level system to obtain turn-level (post-level) decisions instead of training directly on the higher-level data units because the forum claims are relatively infrequent events. Thus, we believe that the classification using localized features will yield better results; when using higher-level classification units, the positive phenomena would be overwhelmed by the negative features in the rest of the sample, leading to poorer performance.

Given a potentially large class imbalance due to the sparsity of the positive-labeled samples, tuning on accuracy scores would lead to very low recall. Thus, we tune and evaluate on F-score, defined as the harmonic mean of precision (the percent of detected claims that are correct) and recall (the percent of true claims that are detected).

The classifier used is a maximum entropy classifier (MaxEnt), implemented using the MALLET package (McCallum, 2002), an open-source java implementation. MaxEnt models the conditional probability distribution $p(c|\mathbf{x})$ of a forum claim c given the feature vector $\mathbf{x}$ in a log-linear form. Model parameters $\lambda_i^{(c)}$ are estimated using gradient descent on the training data log likelihood with L2 regularization.

Since our task is a two-class problem, and the objective is the F-score, we use a classification decision with decision threshold θ, i.e.

$$c^* = \begin{cases} \text{true} & \text{if } p(\text{true}|\mathbf{x}) > \theta, \\ \text{false} & \text{otherwise.} \end{cases}$$

where θ is tuned on the development set, and the optimal value is usually found to be much smaller than 0.5.

3 Features

Past work on various NLP tasks has shown that lexical features can be quite effective in categorizing linguistic phenomena. However, using a large number of features when the number of labeled training samples is small often leads to overtraining, due to the *curse of dimensionality* when dealing with high-dimensional feature spaces (Hastie et al., 2009). Thus, we investigate two task-dependent methods for generating lexical feature lists: a combined data- and knowledge-driven method using related Wikipedia content, and a knowledge-driven method requiring manual feature list generation.

We conjecture that using unigram features alone is often insufficient to capture the more complex phenomena associated with the forum claim detection task. Empirically, we find that even the word features most strongly correlated with the class variable are frequent in both classes. In particular, due to the class imbalance, such features are often more prevalent in the negative class samples than the positive class samples. We believe that additional information about the context in which such words appear in the data could be relevant for further increasing their discriminative power.

One method often used in the literature to capture the context in which a particular word appears is to define the context as its neighboring words, e.g. by using higher-order n-grams (such as bigrams or trigrams) or phrase patterns. However, this method also suffers from the *curse of dimensionality* problem, as seen from the feature set size increase for our training set when moving beyond unigrams (listed in table 1.)

Features	Counts
Unigrams	13,899
Bigrams	109,449
Trigrams	211,580

Table 1: N-gram feature statistics

To understand the meaning of a sentence, features based only on surface word forms may not be sufficient. We propose an alternate method that augments each word with information from the structure of a parse tree for each sentence in which that word appears.

Additionally, we use a small set of other (non-lexical) features, motivated by anecdotal examples from Wikipedia discussions.

3.1 Generating Word Feature Lists

We propose two knowledge-assisted methods for selecting lexical features, as described below, both of which are combined with data-driven selection of the most discriminative features based on mutual information.

3.1.1 Leveraging "Parallel" Data

The Wikipedia data naturally has "parallel" data in that each talk page is associated with an article, and there are additional pages that describe forum policies and norms of behavior. By comparing article and talk pages, one can extract words that tend to be associated with editor discussions (words which have high TF-IDF in a discussion but low TF-IDF in the associated article). By comparing to the policies pages, one can identify words that are likely to be used in policy-related forum claims (words with high average TF-IDF in the corpus of policy and norms of behavior pages.) To select a single reduced set of words, we pick only the words with sufficiently high TF-IDF in the discussion pages. In practice, to avoid tuning additional parameters, we selected the settings which yielded the largest list (with approximately 520 words) and let the feature selection process trim down the list. Some words identified by the feature selection process include:

- words shared with the knowledge-driven list (discussed below): *wikipedia, policy, sources, guidelines, reliable, rules, please*
- relevant words not appearing in the knowledge-driven list: *categories, pages, article, wiki, editing*
- other words: *was, not, who, is, see*

3.1.2 Knowledge-Driven Word List

The knowledge-driven method uses lists of words picked by trained linguists who developed the guidelines for the process of annotating our dataset. Six lists were developed, containing keywords and short phrases related to:

- behavior in discussion forums (*reliable, respectful, balanced, unacceptable*)
- politeness (*please, would you, could you, would you mind*)
- positioning and expressing neutrality (*point of view, neutral, opinion, bias, good faith*)
- accepted practices in discussion forums (*practice, custom, conflict, consensus*)
- sourcing information (*source, citing, rules, policy, original research*)
- Wikipedia-specific keywords (*wikipedia, administrator, registered, unregistered*)

In all our experiments, the various word lists were concatenated and used as a single set of 75 words. Phrases were treated as single keywords for purposes of feature extraction, i.e. a single feature was extracted for each phrase. If another word on the list were a substring of a given phrase, and the phrase were found to appear in the text of a given sample, both the single word and the phrase were kept in that sample.

3.2 Adding Higher-Level Linguistic Context

As an alternative to using n-grams as lexical context, we propose using syntactic context, represented by information about the parse tree of each sentence in the data. Given the low amount of available training data, learning n-gram features we believe is likely to overtrain, due to the combinatorial explosion in the feature space. On the other hand, adding parse tree context information to each feature results in a much smaller increase in feature space, due to the smaller number of non-terminal tokens as compared to the vocabulary size. To extract such features, the data was run through a version of the Berkeley parser (Petrov et al., 2006) trained on the Wall Street Journal portion of the Penn Treebank.

For each sentence, the one-best parse was used to extract the list of non-terminals above each word in the sequence. The list was then filtered to a shorter subset of non-terminal tags. The words augmented with non-terminal parse tree tags were treated as individual features and used in the usual way. We used a context of at most three non-terminal tags (i.e. the POS tag and two additional levels if present.)

For simplicity, multi-word phrases from the knowledge-driven word list were either removed entirely, or split with each word augmented independently. Using this method resulted in the feature counts shown in table 2. In particular, we see that splitting phrases instead of removing them results in almost twice as many parse-augmented word features, in great part due to function words appearing in a variety of unrelated contexts.

Features	Counts
All unigrams	38,384
Data-driven list	5,935
Knowledge-driven list, no phrases	504
Knowledge-driven list, split phrases	908

Table 2: Parse feature statistics

3.3 Other Features

We use a number of additional features not directly related to lexical cues. We extract the following sentence complexity features:

- the length of the sentence
- the average length of the 20% longest words in the sentence

Additionally, we use a number of other features motivated by our analysis of the data. These features are:

- the number of words containing only uppercase letters in that sentence
- the number of (external) URLs in the sentence
- the number of links to Wikipedia pages containing norms of forum behavior or policies
- the number of other Wikipedia-internal links

4 Experiments

4.1 Dataset and Procedure

We use data from the Authority and Alignment in Wikipedia Discussions (AAWD) corpus described in our companion paper (Bender et al., 2011). The dataset contains English Wikipedia discussions annotated with authority claims by four annotators. Not all the discussions are annotated by multiple annotators. Thereby in the train/dev/eval split, we select most of the discussions that are multiply annotated for the dev and eval sets. The statistics of each set are shown in table 3.

	Train	Dev	Eval
# files	226	56	55
# sentences	17512	4990	4200

Table 3: Data statistics

A number of experiments were conducted to assess the performance of the various feature types proposed. We evaluate the effect of individual features when used in a MaxEnt classifier, as well as combined features.

We tune the number of features selected by the mutual information between a feature and the class labels, which is a common approach applied in text categorization (Yang and Pedersen, 1997). Feature selection and parameter tuning of the decision threshold θ are performed independently for each condition. We include the number of features selected in each case alongside the results. The performance of the various systems described in this paper is evaluated using F-score. The numbers corresponding to the overall best performance obtained on the dev and eval sets are highlighted in boldface in the appropriate table.

4.2 N-gram Features

First, we examine the performance of lexical features extracted at different n-gram lengths. We used maximum n-gram sizes 1, 2, and 3, and the counts of n-grams were used as features for MaxEnt. The results are summarized in table 4.

Maximum n-gram length	# selected features	Dev	Eval
1	50	0.321	0.270
2	50	0.331	0.300
3	20	0.333	0.290

Table 4: N-gram feature results

4.3 "Smart" Word Features

The second set of experiments compares the performance of various methods of selecting unigram lexical features. We compare using the full vocabulary with the two selection methods, outlined in section 3.1. The combination of the two simpler selection methods was also examined, under the assumption that the parallel-data-driven features may be more complete, but also more likely to overtrain, since they were derived directly from the data. The results are summarized in table 5.

Feature	# selected features	Dev	Eval
All words	50	0.321	0.270
Parallel corpus words	10	0.281	0.231
Hand-picked words	50	0.340	0.272
Parallel corpus + hand-picked words	100	0.303	0.259

Table 5: Smart word feature results

4.4 Parse-Augmented Features

A third set of experiments examines the effect of adding parsing-related context to the features. We use the same set of features as in section 3.2. For the knowledge-driven features, we present both versions of the parse features, the one in which phrases were split into their constituent words before augmentation with parse features, and the one from which phrases were removed altogether. The results are summarized in table 6.

Word list to derive features from	# selected features	Dev	Eval
All words	50	0.352	0.445
Parallel corpus words	20	0.336	0.433
Hand-picked words (no phrases)	50	0.314	0.306
Hand-picked words (split phrases)	50	0.328	0.310
Parallel corpus + hand-picked words (no phrases)	50	0.367	0.457
Parallel corpus + hand-picked words (split phrases)	50	0.359	0.450

Table 6: Parse-augmented feature results

We perform a small empirical analysis of features in the model with parse-augmented features for all words. Table 7 contains some of the most common features, their counts for each class, and model

weight (if selected.) As expected, the feature with the highest relative frequency in the positive class gets the highest model weight. Other features with high absolute frequency in the positive class also get some positive weight. All other features are discarded during model training.

Feature	# false	# true	Weight
Wikipedia_NNP_NP_PP	60	10	1.035
Wikipedia_NNP_NP_S	57	12	1.121
Wikipedia_NNP_NP_NP	26	16	1.209
Wikipedia_NNP_NP_VP	13	3	-
Wikipedia_JJ_NP_NP	6	0	-
Wikipedia_NNP_NP_FRAG	1	3	2.115

Table 7: Parse feature examples

4.5 Other Features

A fourth set of experiments shows the effect of Wikipedia-specific markup features described in Section 4.5. The results for the Wikipedia policy page feature are listed in table 8. The other features were found to not be useful, resulting in F-scores of less than 0.1.

Feature	Dev	Eval
Wikipedia policy page	0.341	0.622

Table 8: Other feature results

4.6 Combined Features

The previous sets of experiments reveal that the feature of links to Wikipedia policy page is the most discriminative individual feature. Therefore, in the next set of experiments, we combine other features with the Wikipedia policy page feature to train MaxEnt models. We did not include any of the other features whose results were summarized in section 4.5, due to their very low individual performance. The results are shown in table 9.

4.7 Turn-level Classification

We propagate the sentence-level classification output to the turn-level if that turn has at least one sentence classified as forum claim. For simplicity, instead of running experiments on all the feature configurations, we use only the one that provides the highest dev set F-score, which is the MaxEnt classifier with Wikipedia policy page markup and hand-picked keyword features combined. The resulting F-score is 0.57 for the development set and 0.66 for the evaluation set.

Features other than Wikipedia policy page markup	# selected features	Dev	Eval
N-gram features			
unigram	20	0.448	0.550
unigram + bigram	50	0.447	0.551
unigram + bigram + trigram	100	0.446	0.596
Smart word features			
Parallel corpus words	20	0.427	0.483
Hand-picked words	50	**0.468**	0.596
Parallel corpus + hand-picked	100	0.451	0.569
Parse-augmented features			
All words	50	0.398	0.610
Parallel corpus words	100	0.381	0.623
Hand-picked words (no phrases)	20	0.392	**0.632**
Hand-picked words (split phrases)	100	0.392	0.558
Parallel corpus + hand-picked words (no phrases)	50	0.400	0.596
Parallel corpus + hand-picked words (split phrases)	50	0.398	0.607

Table 9: Combined feature results

5 Discussion

5.1 Data Variability

One of the most notable observations in the experiments above is the high degree of data variability. A simple rule-based classifier that uses only the Wikipedia policy page markup feature gives the best results on the evaluation set, but it is not nearly as effective on the development set. Simply put, the markup is a reliable cue when it is available, but it is not always present. Table 10 demonstrates this

through the precision and recall results of the dev and eval sets. The variability also extends to the utility of parse features.

	Dev	Eval
Precision	0.703	0.862
Recall	0.225	0.487

Table 10: Precision and recall of the rule-based system

To better understand this issue, we reran the best case configurations on the dev and eval sets with the role of the dev and eval sets reversed, i.e. using the eval set for feature selection. For the best case configuration on the dev set (Wikipedia policy page markup and hand-picked keywords), 50 and 20 features are selected when tuned on dev and eval sets, respectively, and the latter feature set is a subset of the former one. For the best case configuration on the eval set (Wikipedia policy page markup and parse-augmented features derived from hand-picked words without phrases), the same 20 features are selected when tuned on dev or eval sets. For each configuration, the combined feature set from the two different selection experiments was then used to train a new model, which was evaluated on the combined dev and eval test sets. The precision/recall trade-off is illustrated in figure 1, which can be compared to a precision of 0.78 and recall of 0.32 using the rule-based system on the two test sets combined. While this is a "cheating experiment" in that the test data was used in feature selection, it gives a better idea of the potential gain from parse-augmented lexical features for this task. From the figure, both best-case configurations outperform the rule-based system, and an operating point with more balanced precision and recall can be chosen. Furthermore, the system with parse-augmented features is able to operate at a high recall while still maintaining reasonable precision, which is desirable in some applications.

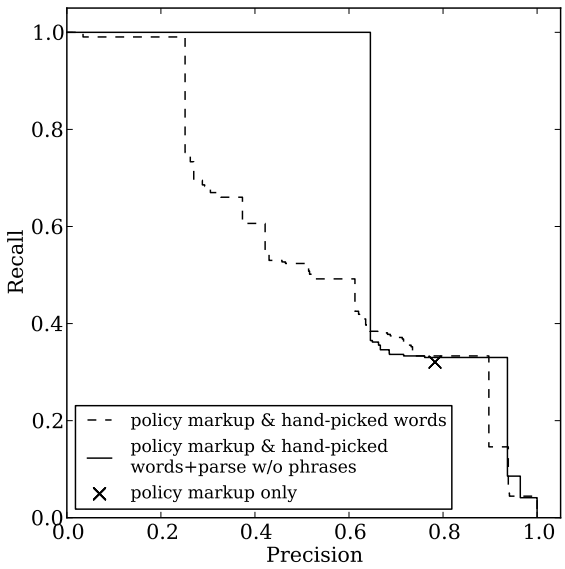

Figure 1: Precision-recall curve

5.2 Feature Analysis

The variability of data in this task poses challenges for learning features that improve over a simple knowledge-driven baseline. However, the results in section 4 provide some insights.

First, unigram features alone provide poor performance. Adding bigrams improves the performance on both the development and the evaluation sets, while further adding trigrams degrades the eval set performance. This indicates that there are some discriminative high-order n-grams, but also too many noisy n-grams to extract the discriminative n-grams effectively with a small amount of training data.

The smarter word features do not perform as well as n-gram features when used alone (i.e. as unigrams), but they provide an improvement over n-grams when used with parse features. With parse features, the parallel corpus words are more effective than the hand-picked words, but the best performance is achieved with the combination. When combined with the Wikipedia policy page markup features, the hand-picked words are the most useful, with the best eval set results obained with the parse-augmented version.

Overall, the best performance seems to be obtained by using the combined feature set of Wikipedia policy page markup and hand-picked keyword features with parse augmentation. However, the test set variability discussed in section 5.1 suggests that it would be useful to assess the findings on additional data.

5.3 Further Challenges

By definition, a forum authority claim is composed of a mention of Wikipedia norms and policies to sup-

port a previously-mentioned opinion proposed by the participant. While the detection of mentions of Wikipedia norms is relatively easy, we conjecture that part of the difficulty of this task lies in identifying whether a mention of Wikipedia norms is for the purpose of supporting an opinion, or just a mention as part of the general conversation. For example, the Wikipedia policy *neutral point of view (NPOV)* is a frequently used term in talk pages. It can be used as support for the participant's suggested modification, or it can be just a mention of the policy without the purpose of supporting any opinion. For example, the sentence *This section should be deleted because it violates NPOV* is a forum claim, because the term *NPOV* is used to support the participant's request. However, the sentence *Thank you for removing the NPOV tag* is not a forum claim, as the participant is not presenting any opinion. For these reasons, the word *NPOV* alone does not provide enough information for reliable decisions; contextual information, such as n-grams and parse-augmented features, must be explored. On the other hand, a direct reference to a Wikipedia policy page is much less ambiguous, as it is almost always used in the context of strengthening an opinion or claim.

Another factor that makes the task challenging is the sparsity of the data. It is time-consuming to produce high quality annotations for forum claims, as many claims are subtle and therefore difficult to detect, even by human annotators. Given the limited amount of data, many features have low occurrences and cannot be learned properly. The data sparsity is an even bigger problem when the feature space is increased, for example by using contextual features such as n-grams and parse-augmented words. On the other hand, while it may be easier to capture the mention of Wikipedia policies using a limited set of keywords or phrases, it is difficult to model the behavior of presenting an opinion when the data is sparse, as the following forum claim examples show:

- *I think we can all agree that this issue bears mentioning, however the blurb as it stands is decidedly not NPOV, nor does it fit the formatting guidelines for a Wikipedia article.*
- *As a reminder, the threshold for inclusion in Wikipedia is whether material is attributable to a reliable published source, not whether it is true.*
- *If you think that some editor is violating NPOV, you can pursue dispute resolution, but it's no justification for moving or removing valid information.*
- *If you'd like to talk the position that quotes from people's opinions do not belong here, fine, but it is extremely POV to insist only on eliminating editorials that you disagree with, while not challenging quotes from your own POV.*

The examples above require deeper understanding of the sentences to identify the embedding of opinions. Modeling such phenomena using word-based contextual features when the training data is sparse is particularly hard. Even with parse-augmented features that do not increase the feature dimensionality as fast as n-grams, a certain amount of data is needed to obtain reliable statistics. Clustering of the features into a lower dimensional space would provide one possible solution to this issue, but how the clustering can be done robustly remains an open question.

6 Conclusions

We have presented systems to detect forum authority claims, which are claims of credibility using forum norms, in Wikipedia talk pages. The Wikipedia policy page markup feature was found to be the most effective individual feature for this task. We have also developed approaches to further improve the performance by knowledge-driven selection of lexical features and adding context in the form of parse information.

Future work includes extending the contextual features, such as parse-augmented word features, to other types of linguistic information, and automatically learning the types of contexts that might be most useful for each word. Feature clustering methods will also be investigated, in order to reduce feature space dimensionality and deal with data sparsity. To improve the effectiveness of the parse features, domain adaptation of the parser or use of a parser trained on data closer matched to our target domain could be investigated. We will also plan to extend this work to other types of authority claims in

Wikipedia and to other multi-party discussion genres.

Acknowledgments

This research was funded by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA). All statements of fact, opinion or conclusions contained herein are those of the authors and should not be construed as representing the official views or policies of IARPA, the ODNI or the U.S. Government.

References

R. Barzilay, M. Collins, J. Hirschberg, and S. Whittaker. 2000. The rules behind roles: Identifying speaker role in radio broadcasts. In *Proceedings of AAAI*, pages 679–684.

E. M. Bender, J. Morgan, M. Oxley, M. Zachry, B. Hutchinson, A. Marin, B. Zhang, and M. Ostendorf. 2011. Annotating social acts: Authority claims and alignment moves in wikipedia talk pages. In *Proceedings of ACL – Workshop on Language in Social Media*.

J. S. Bunderson. 2003. Recognizing and utilizing expertise in work groups: A status characteristics perspective. *Administrative Science Quarterly*, 48(4):557–591.

E. Gilbert, T. Bergstrom, and K. Karahalios. 2009. Blogs Are Echo Chambers: Blogs Are Echo Chambers. In *Proceedings of HICSS*, pages 1–10.

T. Hastie, R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer, September.

B. Hutchinson, B. Zhang, and M. Ostendorf. 2010. Unsupervised broadcast conversation speaker role labeling. In *Proceedings of ICASSP*, pages 5322–5325.

S. M. Kim and E. Hovy. 2006. Automatic identification of pro and con reasons in online reviews. In *Proceedings of COLING-ACL*, pages 483–490.

K. Laskowski, M. Ostendorf, and T. Schultz. 2008. Modeling vocal interaction for text-independent participant characterization in multi-party conversation. In *ISCA/ACL SIGdial Workshop on Discourse and Dialogue*, pages 194–201.

F. Liu and Y. Liu. 2007. Soundbite identification using reference and automatic transcripts of broadcast news speech. In *Proceedings of ASRU*, pages 653–658.

Y. Liu. 2006. Initial study on automatic identification of speaker role in broadcast news speech. In *Proceedings of HLT*, pages 81–84.

A. Marin, M. Ostendorf, B. Zhang, J. T. Morgan, M. Oxley, M. Zachry, and E. M. Bender. 2010. Detecting authority bids in online discussions. In *Proceedings of SLT*, pages 49–54.

S. Maskey and J. Hirschberg. 2006. Soundbite detection in broadcast news domain. In *Proceedings of Interspeech*, pages 1543–1546.

A. K. McCallum. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

B. Pang and L. Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of ACL*, pages 271–278.

B. Pang and L. Lee. 2005. Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*, pages 115–124.

S. Petrov, L. Barrett, R. Thibaux, and D. Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of COLING-ACL*, pages 433–440.

A. Vinciarelli. 2007. Speakers role recognition in multiparty audio recordings using social network analysis and duration distribution modeling. *IEEE Transactions on Multimedia*, 9(6):1215–1226.

Y. Yang and J. O. Pedersen. 1997. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of ICML*, pages 412–420.