Exploiting Debate Portals for Semi-Supervised Argumentation Mining in User-Generated Web Discourse

Ivan Habernal[†] and **Iryna Gurevych**[†‡]

[†]Ubiquitous Knowledge Processing Lab (UKP-TUDA)
Department of Computer Science, Technische Universität Darmstadt
[‡]Ubiquitous Knowledge Processing Lab (UKP-DIPF)
German Institute for Educational Research
www.ukp.tu-darmstadt.de

Abstract

Analyzing arguments in user-generated Web discourse has recently gained attention in argumentation mining, an evolving field of NLP. Current approaches, which employ fully-supervised machine learning, are usually domain dependent and suffer from the lack of large and diverse annotated corpora. However, annotating arguments in discourse is costly, error-prone, and highly context-dependent. We asked whether leveraging unlabeled data in a semi-supervised manner can boost the performance of argument component identification and to which extent is the approach independent of domain and register. We propose novel features that exploit clustering of unlabeled data from debate portals based on a word embeddings representation. Using these features, we significantly outperform several baselines in the cross-validation, cross-domain, and cross-register evaluation scenarios.

1 Introduction

Argumentation mining, an evolving sub-field of NLP, deals with analyzing argumentation[1] in various genres, such as legal cases (Mochales and Moens, 2011), student essays (Stab and Gurevych, 2014a), and medical and scientific articles (Green, 2014; Teufel and Moens, 2002). Recently, the focus of argumentation mining has also shifted to the Web registers (such as comments to articles, forum posts, or blogs) which is motivated by the need of retrieving and understanding ordinary people's arguments to various contentious topics on the large scale. Applications include passenger rights and protection (Park and Cardie, 2014), hotel reviews (Wachsmuth et al., 2014), and controversies in education (Habernal et al., 2014).

Despite the plethora of existing argumentation theories (van Eemeren et al., 2014), the prevalent view in argumentation mining treats arguments as discourse structures consisting of several argument components, such as *claims* and *premises* (Peldszus and Stede, 2013). Current approaches to automatic analysis of argumentation usually follow the fully supervised machine-learning paradigm (Biran and Rambow, 2011; Stab and Gurevych, 2014b; Park and Cardie, 2014) and rely on manually annotated datasets. Only few publicly available argumentation corpora exist, as annotations are costly, error-prone, and require skilled human annotators (Stab and Gurevych, 2014a; Habernal et al., 2014).

To overcome the limited scope and size of the existing annotated corpora, semi-supervised methods can be adopted, as they gain performance by exploiting large unlabeled datasets (Settles, 2012). However, unlike in other NLP tasks where data can be cheaply labeled using for example distant supervision, employing such methods in argumentation mining is questionable. First, argumentation is an act of persuasion (Nettel and Roque, 2011; Mercier and Sperber, 2011) but not all user-generated texts can be treated as persuasive (Park and Cardie, 2014; Habernal et al., 2014), thus the selection of an appropriate unlabeled dataset represents a problem on its own. Second, argument components (e.g., *claims* or *premises*) are highly context-dependent and cannot be easily labeled in distant data using predefined patterns. So far, semi-supervised methods for argumentation mining remain unexplored.

In this article, we tackle argumentation min-

[1]Argumentation is *a verbal activity for which the goal consists of convincing the listener or reader of the acceptability of a standpoint by means of a constellation of propositions justifying or refuting the proposition expressed in the standpoint* (van Eemeren et al., 2002) or *the art of persuading others to think or act in a definite way, including all writing and speaking which is persuasive in form* (Ketcham, 1917).

ing of user-generated Web data by exploiting *debate portals*—semi-structured discussion websites where members pose contentious questions to the community and allow others to pick a side and provide their opinions and arguments in order to 'win' the debate.[2] Our first research question is whether debate portals (which contain noisy user-generated data) can be utilized in a semi-supervised manner for fine-grained identification of argument components. As a second research question, we investigate to what extent our methods are domain independent and evaluate their adaptation across several domains and registers.

Our contribution is three-fold. First, to the best of our knowledge, we present the first successful attempt to semi-supervised argumentation mining in Web data based on exploiting unlabeled external resources. We leverage these resources and derive features in an unsupervised manner by projecting data from debate portals into a latent argument space using unsupervised word embeddings and clustering. Second, our novel features significantly outperform state-of-the-art features in all scenarios, namely in cross-validation, cross-domain evaluation, and cross-register evaluation. Third, to ensure full reproducibility of our experiments, we provide all data and source codes under free licenses.[3]

2 Related work

Analysis of argumentation has been an active topic in numerous research areas, such as philosophy (van Eemeren et al., 2014), communication studies (Mercier and Sperber, 2011), and informal logic (Blair, 2004), among others. In this section, we will focus on the most related works on argumentation mining techniques in NLP in the first part, with an emphasis on Web data in the second part.

Mochales and Moens (2011) based their work on argumentation schemes (Walton et al., 2008) and experimented with Araucaria and ECHR datasets using supervised models to classify argumentative and non-argumentative sentences ($\approx 0.7F_1$) and their structure. Feng and Hirst (2011) classified argument schemes on the Araucaria dataset, reaching 0.6-0.9 accuracy. Experiments on this dataset were also conducted by Rooney et al. (2012), who classified sentences to four categories (*conclusion*, *premise*, *conclusion-premise*, and *none*) and achieved 0.65 accuracy. These approaches assume the text is already segmented into argument components. Stab and Gurevych (2014b) examined argumentation in persuasive essays and classified argument components into four categories (*premise*, *claim*, *major claim*, *non-argumentative*) using SVM and achieved 0.73 macro F_1 score. They further classified argument relations (support and attack) and reached 0.72 macro F_1 score. The best-performing features were structural features (such as the location or length ratios), as persuasive essays usually comply with a certain structure which can be seen as a potential drawback of this approach.

Regarding user-generated Web data, Biran and Rambow (2011) used naive Bayes for classifying justification of subjective claims from blogs and Wikipedia talk pages, relying on features from RST Treebank and manually-processed n-grams. In similar Web registers, Rosenthal and McKeown (2012) automatically determined whether a sentence is a *claim* using logistic regression and various lexical and sentiment-related features and achieved accuracy about 0.66-0.71. Park and Cardie (2014) classified propositions in user comments into three classes (*verifiable experiential*, *verifiable non-experiential*, and *unverifiable*) using SVM and reached 0.69 macro F_1 score. Goudas et al. (2014) identified *premises* in Greek social media texts using BIO encoding and achieved 0.42 F_1 score with Conditional Random Fields. The research gaps in the above-mentioned approaches are the following. First, the argumentation models are simplified to either *claims* or a few types of *premises/propositions*. Second, the segmentation of discourse into argument components is ignored (except the work of Goudas et al. (2014)). Recently, Boltužić and Šnajder (2015) employed hierarchical clustering to cluster arguments in online debates using embeddings projection, but in contrast to our work they performed only intrinsic evaluation of the clusters.

Debate portals have been used in a related body of research, such as classifying support and attack between posts by Cabrio and Villata (2012), or stance detection by Hasan and Ng (2013) or Gottipati et al. (2013). These approaches consider the complete documents (posts) but do not analyze the micro-level argumentation (e.g., *claims* or *premises*).

[2] For instance `createdebate.com` or `debate.org`

[3] `https://github.com/habernal/emnlp2015`

Doc #2823 (article comment, public-private-schools): [*claim:* I agree - Kids can do great in the public school system and parents DO need to be involved.] The more people leave, the worse its going to become. [*premise:* The public school system lets them deal with real life too, unfortunate that it may be but that is what's out there in college and the work force too.] [*premise:* There are still great teachers in the public schools - lets stand behind them.]

Doc #2224 (forumpost, single-sex-education): [*backing:* I went to an all boys school –] [*claim:* Can't say I particularly liked it, I would of much preferred gone to a co-ed.] [*premise:* It is closer to the 'real world' that way. Kids should grow up in the company of both sexes... They will be more at ease around the opposite sex when they are older and it just makes sense.] If it is purely education you are concerned about (and not so much behaviour), our year (at a private school) went shockingly bad in OP scores. We were the worst in 12 years and were beaten by LOTS of co-ed and public schools... So you can never tell. In saying that my sister really enjoyed going to an all girls school. Her year went really well too. Ask your daughters what they would prefer... [*backing:* Btw, I work at a co-ed school at the moment and the kids there get on just fine.]

Figure 1: Two examples of argument annotation of an article comment and a forum post.

3 Data

As data for training and evaluation of our methods, we use a corpus consisting of 340 English documents (approx. 90k tokens) annotated[4] with argumentation by Habernal et al. (2014). Compared to other corpora mentioned in the related work, this corpus is the largest one to date that covers different domains and spans several registers of user-generated Web content. In particular, the corpus comprises four registers (comments to articles, forum posts, blogs, and argumentative newswire articles) and covers six domains related to educational controversies (homeschooling, private vs. public schools, mainstreaming, single-sex education, prayer in schools, and redshirting).

The argumentation model used in this corpus is based on extended Toulmin's model (Toulmin, 1958). Each document contains usually one argument, where each argument consists of several argument components. There are five different components in this model, namely, the *claim* (the statement about to be established in the argument which conveys author's stance towards the topic), the *premise(s)* (propositions that are intended to give reasons of some kind for the claim), the *backing* (additional information used to back-up the argument), the *rebuttal* (attacks the claim), and the *refutation* (which attacks the rebuttal). Relations between the argument components are encoded implicitly in the function of the particular component type, for instance, premises are always attached to the claim. We made two observations in the data: the *claim* is often implicit (must be inferred by the reader), and some sentences have no argumentative function (thus are not labeled by any argument component).[5]

Figure 1 depicts two example annotations from the corpus. Argument components were annotated on the token level as non-overlapping annotation spans. We therefore represent the argument annotations using BIO encoding. Each token is labeled with one of the 11 categories (5 argument component types $\times$ B or I tag $+$ one O category for non-argumentative text).

4 Method

We cast the task of identifying argument components as a sequence tagging problem and employ SVM^{hmm} (Joachims et al., 2009).[6] For linguistic annotations and feature engineering, we rely on two UIMA-based frameworks – DKProCore (Eckart de Castilho and Gurevych, 2014) and DKProTC (Daxenberger et al., 2014).

Although the argument component annotations in the corpus are aligned to the token boundaries (token-level annotations), the minimal classification unit in our sequence tagging approach is set to the sentence level. First, this allows us to capture rich features that are available for entire sentences as opposed to the token level. Second, by modeling sequences on the token level we would lose the advantage of SVM^{hmm} to estimate dependencies between labels, as the label context is limited due to computational feasibility. On the token level, the label sequences are rather static (long sequences with the same label), as opposed to the sentence level. Before the classification step, we adjust all annotation boundaries (note that we use 11 BIO labels) so that they are aligned to the sentence boundaries and each sentence is then treated as a single classification unit with one label (for example, the first sentence from Figure 1 with token labels *Claim-B, Claim-I, Claim-I, ...* be-

[4] Available at `www.ukp.tu-darmstadt.de/data/argumentation-mining/`

[5] A publication containing a thorough analysis of the dataset is pending.

[6] Keerthi and Sundararajan (2007) conclude that performance of SVM^{hmm} is comparable to another widely used method, Conditional Random Fields (Lafferty et al., 2001)

comes *Claim-B*). After classification, the labels are mapped back to tokens (so that, for example, *Claim-B* sentence label is transformed to *Claim-B, Claim-I, ...* token labels). However, all evaluations are performed on the token level and the performance is always measured against the original token labels. Using this approximation, we lose only about 10% of F_1 performance.[7]

4.1 Baseline features

Lexical baseline (FS0) We encode the presence of unigrams, bigrams, and trigrams in the sentence as 'one-hot' (binary) features.

Structural and syntactic features (FS1) Since the presence of discourse markers has been shown to be helpful in argument component analysis (e.g, "therefore" and "since" for *premises* or "think" and "believe" for *claims*), we encode the first and last three words as binary features. Furthermore, we capture the relative position of the sentence in the paragraph and the document, the number of part of speech 1-3 grams, maximum dependency tree depth, constituency tree production rules, and number of sub-clauses (Stab and Gurevych, 2014b). We used Stanford POS Tagger (Toutanova et al., 2003), Berkeley parser (Petrov et al., 2006), and Malt parser (Nivre, 2009).

Sentiment and topic features (FS2) We assume that claims express sentiment, thus we compute five sentiment categories (from very negative to very positive) using Stanford sentiment analyzer (Socher et al., 2013) and use these values directly as features. Furthermore, in order to help detecting off-topic and non-argument sentences, we employ topic model features. In particular, we use features taken from a vector representation of the sentence obtained by using Gibbs sampling on LDA model (Blei et al., 2003; McCallum, 2002) with topics trained on unlabeled data provided as a part of the corpus.[8]

Semantic and discourse features (FS3) Features based on semantic frames has been introduced in relevant works on stance recognition (Hasan and Ng, 2013). Our features, based on PropBank semantic role labels and obtained from NLP Semantic Role Labeler (Choi, 2012), extract various semantic information (agent, predicate + agent, predicate + agent + patient + (optional) negation, argument type + argument value) and discourse markers. Discourse relations also play an important role in argumentation analysis (Cabrio et al., 2013). We thus employ binary features (such as the presence of the sentence in a chain, the transition type, the distance to previous/next sentences in the chain, or the number of inter-sentence coreference links) obtained from Stanford Coreference Chain Resolver (Lee et al., 2013). Furthermore, we include features resulting from a PTDB-style discourse parser (Li et al., 2012), such as the type of discourse relation (explicit, implicit), the presence of discourse connectives, and attributions.

4.2 Unsupervised features

We enrich the above-mentioned features by utilizing external large unlabeled resources – *debate portals*. They fulfill several criteria, namely (a) they are 'argumentative' (meant as opposed to, for example, prose or encyclopedic genres), (b) they are comprised of user-generated content and (c) and there is at least some overlap with topics from our experimental corpus. On the other hand, they contain noisy texts of questionable quality and they do not provide any specific argumentative structure (in fact, these debates are simple discussions to a topic, where each post is only labeled with a *pro* or *contra* stance). Nevertheless, we assume that the posts from (unlabeled) debate portals contain valuable information that will help us with classifying argument components in labeled data. In order to do so, we employ clustering based on latent semantics, which we now formalize as *argument space* features.

We assume that phrases (sentences or documents) can be projected into a latent vector space, using, typically, a sum or a weighted average of all the word embeddings vectors in the phrase; see for example (Le and Mikolov, 2014). Neighboring vectors in the latent vector space exhibit some interesting properties, such as semantic similarity (thoroughly studied within the distributional semantics area). If the latent vector space is clustered, each n-dimensional vector gets reduced to a single cluster number; such clusters have been used directly as features in many tasks, such as NER (Turian et al., 2010), POS tagging (Owoputi

[7] In only 1% of the sentences there are two or more argument components in it; we arbitrarily choose the largest one.

[8] The number of topics was empirically set to 30, therefore for each sentence the topic distribution results into 30 real-valued features.

et al., 2013), or sentiment analysis (Habernal and Brychcín, 2013).

We build upon the above-mentioned approach (described by Søgaard (2013) as 'clusters-as-features' semi-supervised paradigm) and extend it further. We take both sentences and posts from the unlabeled debate portals, project them into a latent space using word embeddings and cluster them. The motivation is that these clusters will contain similar phrases or (similar 'arguments'). Centroids of these clusters would then represent a 'prototypical argument' (note that the centroids exist only in the latent vector space and thus do not correspond to any existing sentence or post). Then we project each sentence (classification unit) in the labeled data to the latent vector space, compute its distance vector to all the cluster centroids, and encode this distance vector directly as real-valued features. By contrast to the above-mentioned works using a single cluster label as a feature, the distance vector to cluster centroids resembles a soft labeling where each sentence belongs to several clusters with a certain 'weight'. We also use the latent vector space representation of the sentence directly as a feature vector.

As unlabeled data, we use data from two largest debate portals.[9] As a pre-processing step we removed all posts with less than one 'point' earned.[10] The data were then indexed using the Lucene framework and the top 100 debates for each of the 6 domains were retrieved which resulted into 5,759 posts ($\approx$ 35k sentences) in the unlabeled data in total. Our approach is formalized in the following paragraph.

Argument space features (FS4) Let $\vec{e}(w)$ be the embedding vector of word w and $\mathrm{tfidf}(w)$ be the TD-IDF value of w. Sentence $\vec{s} = (w_1, \ldots, w_n)$ is then projected into the embedding space $\mathbb{E}$ as $\vec{s}_e = \sum_{i=1}^{n} \mathrm{tfidf}(w_i)\vec{e}(w_i)n^{-1}$ so $\dim(\vec{s}_e) = \dim(\mathbb{E})$. Analogically to $\vec{s}$, we project the entire post $\vec{a} = (w_1, \ldots, w_m)$ to the same embedding space $\mathbb{E}$ such that $\vec{a}_e = \sum_{i=1}^{m} \mathrm{tfidf}(w_i)\vec{e}(w_i)m^{-1}$.

Let K be the number of *sentence* clusters in $\mathbb{E}$ and $\vec{c}_k$ a centroid vector of cluster $k \in K$. Then $\vec{s}_c$ denotes the distance of sentence $\vec{s}_e$ to the sentence cluster centroids such that $\vec{s}_c = (\cos(\vec{s}_e, \vec{c}_1), \ldots, \cos(\vec{s}_e, \vec{c}_k))$ where $\dim(\vec{s}_c) = K$ and $\cos(\bullet, \bullet)$ denotes cosine similarity. Analogically, let L be the number of *post* clusters in $\mathbb{E}$ and $\vec{a}_l$ a centroid vector of cluster $l \in L$. Then $\vec{s}_a$ denotes the distance of sentence $\vec{s}_e$ to the post cluster centroids such that $\vec{s}_a = (\cos(\vec{s}_e, \vec{a}_1), \ldots, \cos(\vec{s}_e, \vec{a}_l))$. We construct the feature vector by concatenating $\vec{s}_e$, $\vec{s}_c$ and $\vec{s}_a$.

For word embeddings, we use pre-trained skip-gram word vectors[11] produced by Mikolov et al. (2013) ($\dim(\mathbb{E}) = 300$). To create clusters for the argument space features, we used CLUTO software package[12] with Repeated Bisection clustering method (Zhao and Karypis, 2002). We clustered the data using different hyper-parameters K and L (we experimented with $K = \{50, 100, 500, 1000\}$ and $L = \{50, 100, 500, 1000\}$).

5 Results

We investigate three evaluation scenarios. First, we report 10-fold cross validation over all 340 documents, where the data are randomly distributed across the folds regardless of the domain or register. In this scenario, the model can benefit from domain-dependent features for the testing data, such as lexical knowledge (FS0) or domain-relevant argument space features (FS4). Second, we evaluate the cross-domain performance; the model is always trained on five domains and tested on the sixth one. In this settings, we also remove all features that exploit distant data relevant to the test set. For instance, if the test domain is *mainstreaming*, we exclude all debates relevant to this domain before constructing the argument space features (FS4). This evaluates the model's cross-domain performance without any target domain data available. Finally, we test cross-register performance in two set-ups: we train the models using comments and forum posts and test on blogs and newswire articles, and then the other way round. We divided the data into these two parts based on similar properties of blogs/articles and comments/forums, such as the length, or the distribution of argumentative and non-argumentative text.

In the evaluation, we focus on F_1 scores achieved on *claims*, *premises*, *backing*, and non-argumentative text (the 'O' class). Although the

[9] `createdebate.com` and `convinceme.net`, licensed under Creative Commons (CC-BY and CC0, resp.)

[10] 'Points' is the sum of up-votes/down-votes by other users to the particular post. Zero-point posts were usually noisy and spam-like.

[11] `https://code.google.com/p/word2vec/`

[12] `http://www.cs.umn.edu/~karypis/cluto`

FS	B-B	B-I	C-B	C-I	O	P-B	P-I	Avg
Human	.664	.579	.739	.728	.833	.673	.736	.707
0	.154	.211	.118	.159	**.718**	.202	.272	.262
01⋆	.237	.254	.167	.129	.671	.280	.356	.299
4⋆	.194	.283	.225	.197	.715	.230	.292	.305
012⋆	.258	.282	.189	.172	.685	.276	.359	.317
1234†	.235	.315	.181	.145	.690	.290	.394	.321
0123⋆	.313	.333	.152	.140	.691	.287	.372	.327
01234†	**.265**	.332	.183	.167	.690	**.314**	**.405**	.337
34⋆	.232	**.344**	**.256**	**.235**	.704	.269	.372	.345
234†	.238	.339	.253	.227	.703	.291	.388	**.348**

Table 1: F_1 results for the 10-fold cross-validation scenario. Feature set combination (the **FS** column) naming is explained in Section 4.1. Class labels: **B-B/I** = *Backing-B/I*, **C-B/I** = *Claim-B/I*, **O** = *non-argumentative*, **P-B/I** = *Premise-B/I*. Star ($\star$) denotes that the row is significantly better than the previous row; dagger (†) means the row is not significantly better than the previous row, but is significantly better than the previous row minus one; $p < 0.001$ using exact Liddell's test (Liddell, 1983).

classifier is trained and tested on all 11 classes including *rebuttal* and *refutation*, we do not report performance of these two argument components—the results are very poor regardless of the parameters for two reasons. First, these classes are underrepresented in the data (*Rebuttal-B*, *Rebuttal-I*, *Refutation-B* and *Refutation-I* are present in only about 4% of sentences). Second, the inter-annotator agreement reached on these classes were reported to be very low (Habernal et al., 2014).

Cross validation results Table 1 shows results for the cross-validation scenario. The human baseline in the first row is an average score between three original annotators of the dataset. The baseline features (FS0) perform poorly, yet they beat the random assignment and majority vote (< 0.12 F_1). The argument space features (FS4) increase the performance in every combination. The best results for *claims* are achieved when only discourse, sentiment, and argument space features are involved (FS3 and FS4), whereas *premises* and *backing* benefit from the presence of lexical, syntactic, and semantic features (the richest feature set). The overall average best results are obtained from a feature combination with higher level of abstraction, in particular without low-level lexical features from FS0.

After the cross validation experiments, we also fixed the hyperparameters (using grid search) to $K = 1000, L = 100$ for the cluster sizes and $t = 1$ and $e = 0$ for the hyperparameters of SVM^{hmm} .

Cross-domain results For each domain, the cross-domain results are shown in Table 2. On average, the best results are about 0.10 F_1 points worse than in the cross-validation settings (Table 1). In all domains, the best average performance was achieved using only the argument space features (FS4); in four cases this system significantly outperforms all other systems ($p < 0.001$). Moreover, more high-level feature set combinations that also contain argument space features (such as FS2+FS3+F4 or FS3+FS4) yield usually better results for particular argument components in contrast to features based on lexical or syntactic information (FS0 and FS1). For identifying non-argumentative texts, there is no clear winner with respect to feature set abstraction (in three domains the best results are achieved using FS4 but in other three domains the baseline FS0 performs best).

Cross-register results The argument space features (FS4) performs best in average also in the cross-register evaluation (see Table 3). In recognizing *premises*, better results were achieved by a system trained on blogs and articles and tested on comments and forum posts. Recognizing *claims* exhibits similar behavior. On the other hand, recognizing non-argumentative text performs better in the opposite direction. On average, the cross-register results are much worse than cross-validation and slightly worse than cross-domain results.

5.1 Error analysis

First, we quantitatively investigate errors in the cross-validation scenario. The confusion matrix in Table 4 shows that about 50-60% of errors for each argument component were caused by misclassifying it as non-argumentative (the 'O' class). The system tends to prefer the 'O' predictions because of the high presence of non-argumentative sentences in the corpus (about 57%). *Backing* is often confused with *premises*; in particular, *Backing-B* with *Premise-B* in 14%, *Backing-I* with *Premise-I* in 17%. These two argument components have a similar function–to support the claim–so the differences in the discourse (which are sometimes very subtle) confuse the system. Note that despite the confusion between these classes, the *-I* and *-B* tags mostly remain the same (the system correctly predicts whether the argument component begins or not).[13]

We also analyzed the errors of the best-

[13]To provide the complete picture, we also show the previously unreported classes (*rebuttal* and *refutation*). *Rebuttal* is usually misclassified as non-argumentative or *premise*, *refutation* as either non-argumentative, *backing*, or *premise*.

FS	B-B	B-I	C-B	C-I	O	P-B	P-I	Avg
Target domain: Homeschooling								
01234	.039	.249	.000	.000	.145	.000	.000	.062
34	.000	.000	.000	.027	.005	.184	.386	.086
1234	.063	**.263**	.000	.000	.289	.000	.000	.088
234	.026	.030	.000	.000	.000	**.197**	**.387**	.091
01	.000	.000	.000	.000	.689	.000	.000	.098
012	.000	.000	.000	.000	.690	.000	.017	.101
0123	.000	.020	.000	.000	.689	.000	.018	.104
0	.000	.000	.063	.032	.683	.079	.098	.136
4◇	**.182**	.258	**.069**	**.069**	**.700**	.143	.224	**.235**
Target domain: Mainstreaming								
01234	.065	.262	.000	.000	.086	.000	.054	.067
34	.000	.000	.000	.000	.000	.184	**.352**	.077
234	.000	.287	.000	.000	.000	**.222**	.241	.107
0	.000	.000	.000	.000	**.689**	.054	.046	.113
1234	.126	.279	.000	.000	.060	.158	.221	.121
01	.000	.000	.000	.000	.666	.103	.079	.121
012	.000	.000	.000	.000	.663	.054	.141	.123
0123	.000	.000	.000	.000	.630	.261	.307	.171
4⋆	**.222**	**.448**	.000	.000	.674	.145	.247	**.248**
Target domain: Prayer in schools								
1234	.040	**.150**	.000	.000	.163	.000	.014	.052
0123	.000	.000	.000	.000	.080	.061	.292	.062
01234	.000	.115	.000	.000	.080	.149	.175	.074
234	**.058**	.042	.000	.000	.012	**.215**	**.303**	.090
34	.000	.000	**.098**	.105	.034	.203	.297	.105
0	.000	.111	.000	.000	.745	.000	.000	.122
01	.000	.115	.000	.000	**.810**	.000	.000	.132
012	.000	.000	.027	.045	.689	.120	.187	.153
4	.000	.146	.083	.048	.695	.168	.156	**.185**

FS	B-B	B-I	C-B	C-I	O	P-B	P-I	Avg
Target domain: Public vs. private schools								
01	.000	.000	.026	.004	.645	.000	.000	.096
012	.000	.000	.026	.005	.647	.000	.000	.097
01234	.000	.000	.000	.000	.591	.069	.093	.108
0	.026	.038	.000	.000	.638	.019	.054	.111
0123	.058	.064	.019	.023	.622	.019	.025	.119
234	.000	.089	.026	.042	.013	**.240**	.424	.119
1234	.000	.022	.000	.000	.496	.133	.239	.127
34	.051	.166	.037	.045	.011	.203	**.386**	.128
4◇	**.228**	**.251**	**.275**	**.270**	**.653**	.232	.220	**.304**
Target domain: Redshirting								
34	.073	.047	.000	.000	.144	.132	**.265**	.094
01234	.000	.000	**.076**	**.070**	.251	.024	.264	.098
234	.079	.162	.000	.000	.195	.101	.179	.102
0	.000	.000	.000	.000	**.740**	.000	.000	.106
01	.000	.000	.000	.000	.733	.000	.029	.109
012	.000	.000	.000	.000	.738	.049	.045	.119
1234	.102	.321	.000	.000	.118	.022	.277	.120
0123	**.304**	.356	.000	.000	.603	.082	.108	.208
4◇	.226	**.390**	.000	.000	.736	**.161**	.227	**.249**
Target domain: Single-sex education								
0123	**.137**	.178	.000	.000	.107	.000	.000	.060
012	.000	.033	.000	.000	.712	.000	.000	.106
01234	.138	.194	.024	.036	.247	.056	.148	.120
34	.065	.124	.000	.000	.073	**.208**	.379	.121
01	.000	.000	.000	.000	.708	.092	.209	.144
0	.000	.000	.000	.000	.728	.154	.130	.145
234	.061	.125	.000	.000	.395	.180	.269	.147
1234	.067	**.187**	**.078**	**.073**	.522	.067	.117	.159
4◇	.104	.185	.000	.000	**.689**	.204	**.397**	**.226**

Table 2: F_1 results for the cross-domain evaluation scenario ranked by performance. Feature set combination naming (the **FS** column) is explained in Section 4.1. Class labels: **B-B/I** = *Backing-B/I*, **C-B/I** = *Claim-B/I*, **O** = *non-argumentative*, **P-B/I** = *Premise-B/I*. Diamond (◇) in the last (winning) row signals a significant difference between this row and all other rows while star (⋆) denotes that the row is significantly better than the previous row; $p < 0.001$ using exact Liddell's test (Liddell, 1983).

FS	B-B	B-I	C-B	C-I	O	P-B	P-I	Avg
Train: blogs, articles; Test: comments, forums								
01234	.063	**.259**	.027	.051	.147	.000	.064	.087
012	.000	.000	.000	.000	.643	.000	.000	.092
0	.010	.237	.000	.000	.352	.014	.036	.093
01	.000	.000	.000	.000	.643	.010	.013	.095
0123	.021	.032	.000	.000	**.645**	.005	.002	.101
1234	**.097**	.215	.052	.068	.369	.000	.013	.116
234	.042	.068	.065	.068	.534	.093	.168	.148
34	.030	.061	.098	.099	.221	**.211**	**.385**	.158
4◇	.076	.206	**.167**	**.158**	.611	.151	.209	**.225**

FS	B-B	B-I	C-B	C-I	O	P-B	P-I	Avg
Train: comments, forums; Test: blogs, articles								
34	.052	.130	.036	.037	.057	.000	.000	.045
01234	.000	.008	.000	.000	.003	.080	.301	.056
234	.055	**.182**	.033	.036	.121	.025	.015	.067
1234	.071	.176	.014	.021	.050	.061	.290	.098
0	.000	.000	.000	.000	**.773**	.012	.019	.115
01	.000	.000	.051	**.058**	.720	.025	.043	.128
012	.000	.000	.039	.037	.746	.063	.046	.133
0123	.000	.000	.000	.000	.679	.099	.227	.144
4◇	**.142**	.162	**.061**	.032	.693	**.161**	**.353**	**.229**

Table 3: F_1 results for the cross-register evaluation scenario ranked by performance. Feature set combination naming (the **FS** column) is explained in Section 4.1. Class labels: **B-B/I** = *Backing-B/I*, **C-B/I** = *Claim-B/I*, **O** = *non-argumentative*, **P-B/I** = *Premise-B/I*. Diamond (◇) in the last (winning) row signals a significant difference between this row and all other rows; $p < 0.001$ using exact Liddell's test (Liddell, 1983).

performing cross-domain system in detail.[14] We randomly sampled 40 documents and manually compared the predicted arguments with the gold data. We found that 11 predicted documents were simply wrong or no argument components were predicted at all (e.g., document #1640, #1658, #1021, #5258). Most of these errors occur in blogs, which seem to convey rather complex argumentation structure (#1666, #1197, #4586, #5258). In 8 documents, we identified that only some premises were (correctly) spotted by the system. This happened mostly in long comments (#452) and blogs (#400, #697, #4583). In 7 investigated documents, we identified errors caused by slightly different boundaries of recognized argument components (#4517, #2447, #2252, #4840) or when multiple segments were merged/split (#1604, #2180, #2310).

By analyzing the predicted output, we also found that in 12 documents the recognized argument components seemed to be valid to some extent, although this was our subjective judge. For instance, in #4285 (see Figure 2), the first *premise* was misclassified as a *claim*. The gold-data argument was annotated as an *enthymeme* (with implicit *claim* that advocates private schools), while in the prediction, the same proposition was identified as the an explicit *claim* supporting private

[14] Available also as PDF at `https://github.com/habernal/emnlp2015`; we use #ID to point to the particular documents.

↓ gold \ pred. →	Bac-B	Bac-I	Cla-B	Cla-I	O	Pre-B	Pre-I	Reb-B	Reb-I	Ref-B	Ref-I
Backing-B	54	12	12	1	106	31	8	0	0	0	0
Backing-I	12	3,238	1	353	5,089	17	1,777	1	18	0	45
Claim-B	7	3	41	5	107	19	9	1	1	0	2
Claim-I	0	160	0	713	2,095	1	456	0	25	0	25
O	97	3,170	53	1,135	36,061	156	5,459	4	178	1	38
Premise-B	35	17	17	2	290	142	28	6	0	0	1
Premise-I	18	1,680	2	544	10,779	51	7,015	2	234	2	41
Rebuttal-B	3	4	3	2	40	9	7	0	0	0	0
Rebuttal-I	1	199	0	47	1,063	10	859	0	0	0	0
Refutation-B	2	2	0	1	16	1	3	0	1	0	0
Refutation-I	0	86	0	7	592	2	148	0	6	0	0

Table 4: Confusion matrix for the best performing configuration in the cross-validation scenario.

schools with one *premise* why the education was not satisfying, which might be also another valid interpretation. The second example #2180 in Figure 2 shows that the boundaries of the predicted *premises* are mixed up (two recognized instead of three), but the longer *backing* is also meaningful. These examples demonstrate that argument analysis is in some cases ambiguous and allows for different valid interpretations.

6 Conclusion

In this article, we proposed a semi-supervised model for argumentation mining of user-generated Web content. We developed new unsupervised features for argument component identification that exploit clustering of unlabeled argumentative data from debate portals based on word embeddings representation. With the help of these features we significantly improved performance of the argumentation mining system and outperformed several baselines. While the improvement was decent in cross-validation scenario, we gained almost 100% improvement in cross-domain and cross-register settings.

We evaluated the methods on a publicly available corpus annotated with argumentation that origins from user-generated Web data. By a detailed analysis of the errors, we pointed out the strengths (such as domain adaptability) and weaknesses (such as unsatisfying results for *rebuttal* and *refutation* components), as well as the challenges for the argumentation mining task (such as boundary identification issues or ambiguous arguments). If we put our results into the context of existing works, the most relevant one by (Goudas et al., 2014) achieved 0.42 F_1 score on identifying only premises. We get comparable results in the cross-validation settings (F_1 0.31-0.40) yet with more complex argumentation model (five different components).

Although argumentation mining in user-generated Web discourse has a long way to go (our methods currently achieve only about 50% of human performance), we see a huge potential for various future tasks, such as information seeking for better-informed personal decision making or support for argument quality assessment. To foster the research within the community, we provide all source codes and data required for the experiments under free licenses.

Acknowledgements

This work has been supported by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant Nº I/82806 and by the German Institute for Educational Research (DIPF). Access to the CERIT-SC computing and storage facilities provided under the programme Center CERIT Scientific Cloud, part of the Operational Program Research and Development for Innovations, reg. no. CZ. 1.05/3.2.00/08.0144, is greatly appreciated. Lastly, we would like to thank the anonymous reviewers for their valuable feedback.

References

Or Biran and Owen Rambow. 2011. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, 5(4):363–381.

J. Anthony Blair. 2004. Argument and its uses. *Informal Logic*, 24:137151.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, March.

Filip Boltužić and Jan Šnajder. 2015. Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity. In *Proceedings of the 2nd*

Gold

[*premise:* I sent my kid to private school so that she could get a better education.] [*backing:* She was at a public school that was 90% hispanic.] [*premise:* The problem was not their race but the fact that they were way behind in reading langauge and math. This situation was holding my kid and preventing her from excelling in her studies.] Do you think I should of just left my kid in this class? Give me a break!

Predicted

[*claim:* I sent my kid to private school so that she could get a better education.] She was at a public school that was 90% hispanic. [*premise:* The problem was not their race but the fact that they were way behind in reading langauge and math.] This situation was holding my kid and preventing her from excelling in her studies. Do you think I should of just left my kid in this class? Give me a break!

(a) Doc #4285 (article comment, public vs. private schools)

Gold

[*claim:* Personally i'd go co-ed.]
[*backing:* As someone who went to a same sex school for 8 years, I found it lacked the diversity you get in a co-ed environment.] [*premise:* I found the attitude and behaviour of students in the co ed school to be better, and i attribute that to the influence of the opposite sex.] [*premise:* There's no doubt boys behave a little different when girls are watching, and i also found boys were quite good at limiting the bitchyness girls are renowned for. So both kept one another in line, and made for a more positive and dynamic environment.]
[*premise:* I also think there's a few extra life lessons and skills children can learn at co ed schools. Dating, relationships, interacting with the opposite sex, i think children at co ed schools tend to have a far better grasp of these skills then students who've only attended same sex schools.]

Predicted

[*backing:* Personally i'd go co-ed.
As someone who went to a same sex school for 8 years, I found it lacked the diversity you get in a co-ed environment. I found the attitude and behaviour of students in the co ed school to be better, and i attribute that to the influence of the opposite sex.] [*premise:* There's no doubt boys behave a little different when girls are watching, and i also found boys were quite good at limiting the bitchyness girls are renowned for.] [*premise:* So both kept one another in line, and made for a more positive and dynamic environment.
I also think there's a few extra life lessons and skills children can learn at co ed schools. Dating, relationships, interacting with the opposite sex, i think children at co ed schools tend to have a far better grasp of these skills then students who've only attended same sex schools.]

(b) Doc #2180 (forum post, single-sex education)

Figure 2: Examples of gold data annotations (on the left-hand side) and system predictions in the best-performing cross-domain evaluation scenario (on the right-hand side).

Workshop on Argumentation Mining, pages 110–115, Denver, Colorado. Association for Computational Linguistics.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 208–212, Jeju Island, Korea. Association for Computational Linguistics.

Elena Cabrio, Sara Tonelli, and Serena Villata. 2013. From discourse analysis to argumentation schemes and back: Relations and differences. In João Leite, Tran Cao Son, Paolo Torroni, Leon Torre, and Stefan Woltran, editors, *Proceedings of 14th International Workshop on Computational Logic in Multi-Agent Systems*, volume 8143 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

Jinho D. Choi. 2012. *Optimization of Natural Language Processing Components for Robustness and Scalability*. Ph.D. Thesis, University of Colorado Boulder, Computer Science and Cognitive Science.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. DKPro TC: a Java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66, Baltimore, Maryland, June. Association for Computational Linguistics.

Richard Eckart de Castilho and Iryna Gurevych. 2014. A broad-coverage collection of portable NLP components for building shareable analysis pipelines. In Nancy Ide and Jens Grivolla, editors, *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT (OIAF4HLT) at COLING 2014*, pages 1–11, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Vanessa Wei Feng and Graeme Hirst. 2011. Classifying arguments by scheme. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 987–996, Portland, Oregon. Association for Computational Linguistics.

Swapna Gottipati, Minghui Qiu, Yanchuan Sim, Jing Jiang, and Noah A. Smith. 2013. Learning topics and positions from Debatepedia. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1858–1868, Seattle, Washington, USA, October. Association for Computational Linguistics.

Theodosis Goudas, Christos Louizos, Georgios Petasis, and Vangelis Karkaletsis. 2014. Argument extraction from news, blogs, and social media. In Aristidis Likas, Konstantinos Blekas, and Dimitris Kalles, editors, *Artificial Intelligence: Methods and Applications*, pages 287–299. Springer International Publishing.

Nancy L Green. 2014. Argumentation for scientific claims in a biomedical research article. In Elena Cabrio, Serena Villata, and Adam Wyner, editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 5–10, Bertinoro, Italy, July. CEUR-WS.

Ivan Habernal and Tomáš Brychcín. 2013. Semantic spaces for sentiment analysis. In *Text, Speech and Dialogue*, volume 8082 of *Lecture Notes in Computer Science*, pages 482–489, Berlin Heidelberg. Springer.

Ivan Habernal, Judith Eckle-Kohler, and Iryna Gurevych. 2014. Argumentation mining on the web from information seeking perspective. In Elena Cabrio, Serena Villata, and Adam Wyner, editors, *Proceedings of the Workshop on Frontiers and Connections between Argumentation Theory and Natural Language Processing*, pages 26–39, Bertinoro, Italy, July. CEUR-WS.

Kazi Saidul Hasan and Vincent Ng. 2013. Stance classification of ideological debates: Data, models, features, and constraints. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 1348–1356, Nagoya, Japan, October. Asian Federation of Natural Language Processing.

Thorsten Joachims, Thomas Finley, and Chun-Nam John Yu. 2009. Cutting-plane training of structural SVMs. *Machine Learning*, 77(1):27–59.

Sathiya Keerthi and Sellamanickam Sundararajan. 2007. CRF versus SVM-struct for sequence labeling. Technical report, Yahoo! Research.

Victor Alvin Ketcham. 1917. *The theory and practice of argumentation and debate*. Macmillan, New York.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA. Morgan Kaufmann Publishers Inc.

Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In Tony Jebara and Eric P. Xing, editors, *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, volume 32, pages 1188–1196, Beijing, China. JMLR Workshop and Conference Proceedings.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-centric, Precision-ranked Rules. *Computational Linguistics*, 39(4):885–916.

Lianghao Li, Xiaoming Jin, Sinno Jialin Pan, and Jian-Tao Sun. 2012. Multi-domain active learning for text classification. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1086–1094, Beijing, China. ACM.

Douglas Liddell. 1983. Simplified exact analysis of case-referent studies: matched pairs; dichotomous exposure. *Journal of Epidemiology & Community Health*, 37(1):82–84.

Andrew Kachites McCallum. 2002. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences*, 34(2):57–74; discussion 74–111.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22, April.

Ana Laura Nettel and Georges Roque. 2011. Persuasive argumentation versus manipulation. *Argumentation*, 26(1):55–69.

Joakim Nivre. 2009. Non-projective dependency parsing in expected linear time. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*, ACL '09, pages 351–359, Suntec, Singapore. Association for Computational Linguistics.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith. 2013. Improved Part-of-Speech Tagging for Online Conversational Text with Word Clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, Georgia. Association for Computational Linguistics.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. In *Proceedings of the First Workshop on Argumentation Mining*, pages 29–38, Baltimore, Maryland, June. Association for Computational Linguistics.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic*

Annotation Workshop and Interoperability with Discourse, pages 196–204, Sofia, Bulgaria, August. Association for Computational Linguistics.

Slav Petrov, Leon Barrett, Romain Thibaux, and Dan Klein. 2006. Learning accurate, compact, and interpretable tree annotation. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, ACL-44, pages 433–440, Sydney, Australia. Association for Computational Linguistics.

Niall Rooney, Hui Wang, and Fiona Browne. 2012. Applying kernel methods to argumentation mining. In *Proceedings of the Twenty-Fifth International Florida Artificial Intelligence Research Society Conference*, pages 272–275. Association for the Advancement of Artificial Intelligence.

Sara Rosenthal and Kathleen McKeown. 2012. Detecting opinionated claims in online discussions. In *2012 IEEE Sixth International Conference on Semantic Computing*, pages 30–37, Palermo, Italy. IEEE.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Anders Søgaard. 2013. *Semi-Supervised Learning and Domain Adaptation in Natural Language Processing*. Morgan & Claypool Publishers.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1501–1510, Dublin, Ireland, August. Dublin City University and Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October. Association for Computational Linguistics.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles: Experiments with relevance and rhetorical status. *Computational Linguistics*, 28:409–445.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Edmonton, Canada. Association for Computational Linguistics.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, number July, pages 384–394, Uppsala, Sweden. Association for Computational Linguistics.

Frans H. van Eemeren, R. Grootendorst, and A. F. Snoeck Henkemans. 2002. *Argumentation: Analysis, evaluation, presentation*. Lawrence Erlbaum, Mahwah, NJ, USA.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer, Berlin/Heidelberg.

Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A review corpus for argumentation analysis. In Alexander Gelbukh, editor, *15th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing 14)*, pages 115–127, Kathmandu, Nepal. Springer.

Douglas Walton, Christopher Reed, and Fabrizio Macagno. 2008. *Argumentation Schemes*. Cambridge University Press.

Y. Zhao and G. Karypis. 2002. Criterion functions for document clustering: Experiments and analysis. Technical report, Department of Computer Science, University of Minnesota, Minneapolis.