Identifying Prominent Arguments in Online Debates Using Semantic Textual Similarity

Filip Boltužić and **Jan Šnajder**
University of Zagreb, Faculty of Electrical Engineering and Computing
Text Analysis and Knowledge Engineering Lab
Unska 3, 10000 Zagreb, Croatia
{filip.boltuzic,jan.snajder}@fer.hr

Abstract

Online debates sparkle argumentative discussions from which generally accepted arguments often emerge. We consider the task of unsupervised identification of prominent argument in online debates. As a first step, in this paper we perform a cluster analysis using semantic textual similarity to detect similar arguments. We perform a preliminary cluster evaluation and error analysis based on cluster-class matching against a manually labeled dataset.

1 Introduction

Argumentation mining aims to detect the argumentative discourse structure in text. It is an emerging field in the intersection of natural language processing, logic-based reasoning, and argumentation theory; see (Moens, 2014) for a recent overview.

While most work on argumentation mining has focused on well-structured (e.g., legal) text, recently attention has also turned to user-generated content such as online debates and product reviews. The main motivation is to move beyond simple opinion mining and discover the reasons underlying opinions. As users' comments are generally less well-structured and noisy, argumentation mining proper (extraction of argumentative structures) is rather difficult. However, what seems to be a sensible first step is to identify the *arguments* (also referred to as *reasons* and *claims*) expressed by users to back up their opinions.

In this work we focus on online debates. Given a certain topic, a number of prominent arguments often emerge in the debate, and the majority of users will back up their stance by one or more of these arguments. The problem, however, is that linking users' statements to arguments is far from trivial. Besides language variability, due to which the same argument can be expressed in infinitely many ways, many other factors add to the variability, such as entailment, implicit premises, value judgments, etc. This is aggravated by the fact that most users express their arguments in rather confusing and poorly worded manner. Another principal problem is that, in general, the prominent arguments for a given topic are not known in advance. Thus, to identify the arguments expressed by the users, one first needs to come up with a set of prominent arguments. Manual analysis of the possible arguments does not generalize to unseen topic nor does it scale to large datasets.

In this paper, we are concerned with automatically identifying prominent arguments in online debates. This is a formidable task, but as a first step towards this goal, we present a cluster analysis of users' argumentative statements from online debates. The underlying assumption is that statements that express the same argument will be semantically more similar than statements that express different arguments, so that we can group together similar statements into clusters that correspond to arguments. We operationalize this by using hierarchical clustering based on semantic textual similarity (STS), defined as the degree of semantic equivalence between two texts (Agirre et al., 2012).

The purpose of our study is twofold. First, we wish to investigate the notion of prominent arguments, considering in particular the variability in expressing arguments, and how well it can be captured by semantic similarity. Secondly, from a more practical perspective, we investigate the possibility of automatically identifying prominent arguments, setting a baseline for the task of unsupervised argument identification.

2 Related Work

The pioneering work in argumentation mining is that of Moens et al. (2007), who addressed mining of argumentation from legal documents. Recently, the

Proceedings of the 2nd Workshop on Argumentation Mining, pages 110–115,
Denver, Colorado, June 4, 2015. ©2015 Association for Computational Linguistics

focus has also moved to mining from user-generated content, such as online debates (Cabrio and Villata, 2012), discussions on regulations (Park and Cardie, 2014), and product reviews (Ghosh et al., 2014).

Boltužić and Šnajder (2014) introduced *argument recognition* as the task of identifying what arguments, from a predefined set of arguments, have been used in users comments, and how. They frame the problem as multiclass classification and describe a model with similarity- and entailment-based features.

Essentially the same task of argument recognition, but at the level of sentences, is addressed by Hasan and Ng (2014). They use a probabilistic framework for argument recognition (reason classification) jointly with the related task of *stance classification*. Similarly, Conrad et al. (2012) detect spans of text containing *arguing subjectivity* and label them with *argument tags* using a model that relies on sentiment, discourse, and similarity features.

The above approaches are supervised and rely on datasets manually annotated with arguments from a predefined set of arguments. In contrast, in this work we explore unsupervised argument identification. A similar task is described by Trabelsi and Zaïane (2014), who use topic modeling to extract words and phrases describing *arguing expressions*, and also discuss how the arguing expressions could be clustered according to the arguments they express.

3 Data and Model

Dataset. We conduct our study on the dataset of users' posts compiled by Hasan and Ng (2014). The dataset is acquired from two-side online debate forums on four topics: "Obama", "Marijuana", "Gay rights", and "Abortion". Each post is assigned a stance label (*pro* or *con*), provided by the author of the post. Furthermore, each post is split up into sentences and each sentence is manually labeled with one argument from a predefined set of arguments (different for each topic). Note that all sentences in the dataset are argumentative; non-argumentative sentences were removed from the dataset (the ratio of argumentative sentences varies from 20.4% to 43.7%, depending on the topic). Hasan and Ng (2014) report high levels of inter-annotator agreement (between 0.61 and 0.67, depending on the topic).

For our analysis, we removed sentences labeled with rarely occurring arguments (<2%), allowing us to focus on prominent arguments. The dataset we work with contains 3104 sentences ("Abortion" 814, "Gay rights" 824, "Marijuana" 836, and "Obama" 630) and 47 different arguments (25 pro and 22 con, on average 12 arguments per topic). The majority of sentences (2028 sentences) is labeled with pro arguments. The average sentence length is 14 words.

Argument similarity. We experiment with two approaches for measuring the similarity of arguments.

Vector-space similarity: We represent statements as vectors in a semantic space. We use two representations: (1) a bag-of-word (BoW) vector, weighted by inverse sentence frequency, and (2) a distributed representation based on the recently proposed neural network skip-gram model of Mikolov et al. (2013a).

As noted by Ramage et al. (2009), BoW has shown to be a powerful baseline for semantic similarity. The rationale for weighting by inverse sentence frequency (akin to inverse document frequency) is that more frequently used words are less argument-specific and hence should contribute less to the similarity.

On the other hand, distributed representations have been shown to work exceptionally well (outperforming BoW) for representing the meaning of individual words. Furthermore, they have been shown to model quite well the semantic composition of short phrases via simple vector addition (Mikolov et al., 2013b). To build a vector for a sentence, we simply sum the distributed vectors of the individual words.[1]

For both representations, we remove the stopwords before building the vectors. To compute the similarity between two sentences, we compute the cosine similarity between their corresponding vectors.

Semantic textual similarity (STS): Following on the work of Boltužić and Šnajder (2014), we use an off-the-shelf STS system developed by Šarić et al. (2012). It is a supervised system trained on manually labeled STS dataset, utilizing a rich set of text comparison features (incl. vector-space comparisons). Given two sentences, the system outputs a real-valued similarity score, which we use directly as the similarity between two argument statements.

[1] We use the pre-trained vectors available at `https://code.google.com/p/word2vec/`

Clustering. For clustering, we use the hierarchical agglomerative clustering (HAC) algorithm (see (Xu et al., 2005) for an overview of clustering algorithms). This is motivated by three considerations. First, HAC allows us to work directly with similarities coming from the STS systems, instead of requiring explicit vector-space representations as some other algorithms. Secondly, it produces hierarchical structures, allowing us to investigate the granularity of arguments. Finally, HAC is a deterministic algorithm, therefore its results are more stable.

HAC works with a distance matrix computed for all pairs of instances. We compute this matrix for all pairs of sentences s_1 and s_2 from the corresponding similarities: $1 - cos(v_1, v_2)$ for vector-space similarity and $1/(1 + sim(s_1, s_2))$ for STS similarity. Linkage criterion has been shown to greatly affect clustering performance. We experiment with complete linkage (farthest neighbor clustering) and Ward's method (Ward Jr, 1963), which minimizes the within-cluster variance (the latter is applicable only to vector-space similarity). Note that we do not cluster separately the statements from the pro and con stances. This allows us to investigate to what extent stance can be captured by semantic similarity of the arguments, while it also corresponds to a more realistic setup.

4 Cluster Analysis

4.1 Analysis 1: Clustering Models

Evaluation metrics. A number of clustering evaluation metrics have been proposed in the literature. We adopt the external evaluation approach, which compares the hypothesized clusters against target clusters. We use argument labels of Hasan and Ng (2014) as target clusters. As noted by Amigó et al. (2009), external cluster evaluation is a non-trivial task and there is no consensus on the best approach. We therefore chose to use two established, but rather different measures: the Adjusted Rand Index (ARI) (Hubert and Arabie, 1985) and the information-theoretic V-measure (Rosenberg and Hirschberg, 2007). ARI of 0 indicates clustering expected by chance and 1 indicates perfect clustering. The V-measure trade-offs measures of homogeneity (h) and completeness (c). It ranges from 0 to 1, with 1 being perfect clustering.

Results. We cluster the sentences from the four topics separately, using the gold number of clusters for each topic. Results are shown in Table 1. Overall, the best model is skip-gram with Ward's linkage, generally outperforming the other models considered in terms of both ARI and V-measure. This model also results in the most consistent clusters in terms of balanced homogeneity and completeness. Ward's linkage seems to work better than complete linkage for both BoW and skip-gram. STS-based clustering performs comparable to the baseline BoW model. We attribute this to the fact that the STS model was trained on different domains, and therefore probably does not extend well to the kind of argument-specific similarity we are trying to capture here.

We observe quite some variance in performance across topics. Arguments from the "Gay rights" topic seems to be most difficult to cluster, while "Marijuana" seems to be the easiest. In absolute terms, the clustering performance of the skip-gram model is satisfactory given the simplicity of the model. In subsequent analysis, we focus on the skip-gram model with Ward's linkage and the "Marijuana" topic.

4.2 Analysis 2: Clustering Quality

Cluster-class matching. To examine the cluster quality and clustering errors, we do a manual cluster-class matching for the "Marijuana" topic against the target clusters, using again the gold number of clusters (10). Cluster-matching is done on a class majority basis, resulting in six gold classes matched. Table 2 shows the results. We list the top three gold classes (and the percentage of sentences from these classes) in each of our clusters, and the top three clusters (and the percentage of sentences from these clusters) in each of the gold classes. Some gold classes (#4, #9) are frequently co-occurring, indicating their high similarity. We characterize each cluster by its medoid (the sentence closest to cluster centroid).

Error analysis. Grouping statements into coherent clusters proved a challenging task. Our preliminary analysis indicates that the main problems are related to (a) need for background knowledge, (b) use of idiomatic language, (c) grammatical errors, (d) opposing arguments, and (e) too fine/coarse gold argument granularity. We show some sample errors in Table 3, but leave a detailed error analysis for future work.

Ex. *#knowledge* demonstrates the need for background knowledge (exports are government regu-

Model (linkage)	"Obama" *h*	*c*	*V*	ARI	"Marijuana" *h*	*c*	*V*	ARI	"Gay rights" *h*	*c*	*V*	ARI	"Abortion" *h*	*c*	*V*	ARI
BoW (Complete)	.15	.15	.15	.03	.04	.04	.04	.00	.04	.04	.04	.01	.05	.04	.04	.01
BoW (Ward's)	.22	**.34**	.27	.04	.15	.20	.17	.02	.13	**.17**	**.15**	.04	.22	**.27**	**.24**	.07
Skip-gram (Complete)	.18	.26	.21	.04	.09	.22	.13	.02	.09	.10	.10	.04	.17	.24	.20	.03
Skip-gram (Ward's)	**.30**	.29	**.30**	**.10**	**.25**	**.24**	**.25**	**.19**	**.16**	.15	**.15**	**.07**	**.24**	.22	.23	**.08**
STS (Complete)	.11	.11	.11	.02	.05	.05	.05	.03	.05	.05	.05	.01	.06	.06	.06	.02

Table 1: External evaluation of clustering models on the four topics

lated). A colloquial expression (*pot*) is used in Ex. *#colloquial*. In *#oppose*, the statement is assigned to a cluster of opposing argument. In Ex. *#general* our model predicts a more coarse argument.

Another observation concerns the level of argument granularity. In the previous analysis, we used the gold number of clusters. We note, however, that the level of granularity is to a certain extent arbitrary. To exemplify this, we look at the dendrogram (Fig. 1) of the last 15 HAC steps on the "Marijuana" topic. Medoids of clusters divided at point *CD* are (1) *the economy would get billions of dollars (...) no longer would this revenue go directly into the black market.* and (2) *If the tax on cigarettes can be $5.00/pack imagine what we could tax pot for!*. These could well be treated as separate arguments about *economy* and *taxes*, respectively. On the other hand, clusters merged at *CM* consists mostly of gold arguments (1) *Damages our bodies* and (2) *Responsible for brain damage*, which could be represented by a single argument *Damaging our entire bodies*. The dendrogram also suggests that the 10-cluster cut is perhaps not optimal for the similarity measure used.

5 Conclusion

In this preliminary study, we addressed unsupervised identification of prominent arguments in online debates, using hierarchical clustering based on textual similarity. Our best performing model, a simple distributed representation of argument sentence, performs in a 0.15 to 0.30 V-measure range. Our analysis of clustering quality and errors on manually matched cluster-classes revealed that there are difficult cases that textual similarity cannot capture. A number of errors can be traced down to the fact that it is sometimes difficult to draw clear-cut boundaries between arguments.

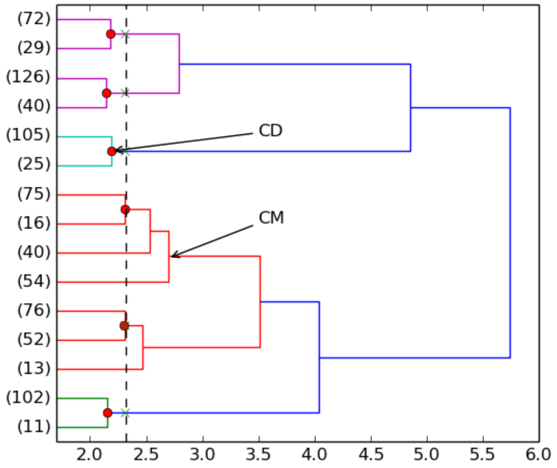

Figure 1: Dendrogram for the "Marijuana" topic (the dashed line shows the 10-clusters cut)

In this study we relied on simple text similarity models. One way to extend our work would be to experiment with models better tuned for argument similarity, based on a more detailed error analysis. Also of interest are the internal evaluation criteria for determining the optimal argument granularity.

A more fundamental issue, raised by one reviewer, are the potential long-term limitations of the clustering approach to argument recognition. While we believe that there is a lot of room for improvement, we think that identifying arguments fully automatically is hardly feasible. However, we are convinced that argument clustering will prove valuable in human-led argumentative analysis. Argument clustering may also prove useful for semi-supervised argument recognition, where it may be used as unsupervised pre-training followed by supervised fine-tuning.

Acknowledgments. We thank the anonymous reviewers for their many comments and suggestions.

Hypothesized clustering			Gold classes		
Id	Classes	Cluster medoid	Id	Clusters	Gold argument
1	**10 (54%)** 2 (12%) 6 (10%)	*Tobacco and alcohol are both legal and widely used in the US, (...) If the abuse of marijuana is harmful, isn't the abuse of tobacco or alcohol equally life threatening? (...)*	1	5 (23%) 9 (19%) 10 (18%)	*Used as a medicine for its positive effects*
2	**4 (92%)** 9 (8%)	*The biggest effect would be an end to brutal mandatory sentencing of long jail times that has ruined so many young peoples lives.*	2	1 (33%) 9 (28%) 3 (15%)	*Responsible for brain damage*
3	**9 (44%)** 4 (25%) 7 (8%)	*Legalizing pot alone would not end the war on drugs. It would help (...) my personal opinion would be the only way to completely end the war on drugs would be to legalize everything.*	3	9 (41%) 3 (23%) 10 (23%)	*Causes crime*
4	**8 (37%)** 1 (22%) 10 (17%)	*What all these effects have in common is that they result from changes in the brain's control centers (...) So, when marijuana disturbs functions centered in the deep control centers, disorienting changes in the mind occur (...)*	4	9 (40%) 3 (26%) 10 (12%)	*Prohibition violates human rights*
5	**1 (45%)** 6 (18%) 8 (10%)	*People with pre-existing mental disorders also tend to abuse alcohol and tobacco. (...) the link between marijuana use and mental illness may be an instance when correlation does not equal causation.*	5	6 (25%) 7 (25%) 4 (18%)	*Does not cause any damage to our bodies*
6	**5 (63%)** 10 (31%) 1 (6%)	*There are thousands of deaths every year from tobacco and alcohol, yet there has never been a recorded death due to marijuana.*	6	9 (29%) 1 (19%) 7 (16%)	*Damages our bodies*
7	**10 (48%)** 5 (13%) 6 (12%)	*as far as it goes for medicinal purposes, marijuana does not cure anything (...) It is for the sole purpose of numbing the pain in cancer patients (...) and also making patients hungry so they eat more and gain weight on their sick bodies*	7	9 (39%) 3 (30%) 1 (9%)	*Highly addictive*
8	**9 (92%)**	*the economy would get billions of dollars in a new industry if it were legalized (...) no longer would this revenue go directly into the black market.*	8	4 (44%) 7 (16%) 9 (16%)	*If legalized, people will use marijuana and other drugs more*
9	**4 (30%)** 9 (13%) 10 (11%)	*(...) I think it ridiculous that people want to legalise something that has four - seven times the amount of tar (the cancer causing agent) in one cone than in one cigarette (...)*	9	8 (53%) 3 (25%) 9 (10%)	*Legalized marijuana can be controlled and regulated by the government*
10	**10 (30%)** 9 (19%) 4 (15%)	*But I'm not gonna tell anyone they can't smoke pot or do meth because I don't like it.*	10	1 (36%) 7 (21%) 10 (18%)	*Not addictive*

Table 2: Manual cluster-class matching for the "Marijuana" topic and the gold number of clusters

Id	Statement	Hypothesized clustering argument	Gold argument
#knowledge	*Pot is also one of the most high priced exports of Central American Countries and the Carribean*	*Not addictive*	*Legalized marijuana can be controlled and regulated by the government*
#colloquial	*If I want to use pot, that is my business!*	*Legalized marijuana can be controlled and regulated by the government*	*Prohibition violates human rights*
#opposing	*(...) immediately following the legalization of the drug would cause widespread pandemonium. (...)*	*Legalized marijuana can be controlled and regulated by the government*	*If legalized, people will use marijuana and other drugs more*
#general	*The user's psychomotor coordination becomes impaired (...), narrow attention span, "depersonalization, euphoria or depression (...)*	*Damages our bodies*	*Responsible for brain damage*

Table 3: Error analysis examples for the "Marijuana" topic

References

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 task 6: A pilot on semantic textual similarity. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 385–393.

Enrique Amigó, Julio Gonzalo, Javier Artiles, and Felisa Verdejo. 2009. A comparison of extrinsic clustering evaluation metrics based on formal constraints. *Information retrieval*, 12(4):461–486.

Filip Boltužić and Jan Šnajder. 2014. Back up your stance: Recognizing arguments in online discussions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 49–58.

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*, pages 208–212.

Alexander Conrad, Janyce Wiebe, et al. 2012. Recognizing arguing subjectivity and argument tags. In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 80–88.

Debanjan Ghosh, Smaranda Muresan, Nina Wacholder, Mark Aakhus, and Matthew Mitsui. 2014. Analyzing argumentative discourse units in online interactions. In *Proceedings of the First Workshop on Argumentation Mining*, pages 39–48.

Kazi Saidul Hasan and Vincent Ng. 2014. Why are you taking this stance? Identifying and classifying reasons in ideological debates. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 751–762.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1):193–218.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. In *Proceedings of ICLR*, Scottsdale, AZ, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Marie-Francine Moens, Erik Boiy, Raquel Mochales Palau, and Chris Reed. 2007. Automatic detection of arguments in legal texts. In *Proceedings of the 11th International Conference on Artificial Intelligence and Law*, pages 225–230.

Marie-Francine Moens. 2014. Argumentation mining: Where are we now, where do we want to be and how do we get there? In *Post-proceedings of the forum for information retrieval evaluation (FIRE 2013)*.

Joonsuk Park and Claire Cardie. 2014. Identifying appropriate support for propositions in online user comments. *ACL 2014*, pages 29–38.

Daniel Ramage, Anna N Rafferty, and Christopher D Manning. 2009. Random walks for text semantic similarity. In *Proceedings of the 2009 workshop on graph-based methods for natural language processing*, pages 23–31.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *EMNLP-CoNLL*, volume 7, pages 410–420.

Frane Šarić, Goran Glavaš, Mladen Karan, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Takelab: Systems for measuring semantic text similarity. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 441–448, Montréal, Canada, 7-8 June.

Amine Trabelsi and Osmar R Zaïane. 2014. Finding arguing expressions of divergent viewpoints in online debates. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 35–43.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Rui Xu, Donald Wunsch, et al. 2005. Survey of clustering algorithms. *Neural Networks, IEEE Transactions on*, 16(3):645–678.