Argument Facet Detection in Online Debates Based on Attention Weights and Clustering with Combined Similarity Matrices*

Sangah Lee**, Hyopil Shin***
(Seoul National University)

Lee, Sangah & Shin, Hyopil. 2021. **Argument Facet Detection in Online Debates Based on Attention Weights and Clustering with Combined Similarity Matrices.** *Korean Journal of Linguistics* 46-1, 107-134. The purpose of argument mining research is to analyze and understand the stances, content, and structures of large argumentative texts, such as online debates. For our research, we collected a list of identified arguments from online debates and attempted to use unsupervised methods to create a list of common justifications for each argument stance in each domain. We propose a model that clusters arguments by subtopics, or justifications, and then extracts the list of representative words for argument facets from each cluster. We were able to improve clustering performance by using a combination of three different similarity matrices (cosine similarity between BERT sentence embeddings, semantic textual similarity, and similarity between topic probability distributions) for the clustering algorithm. Our clustering produced 5%p and 7.5%p of ARI and V-measure values on average, which outperforms previous work in two of four domains. Additionally, we used a Transformer model to utilize the attention weights to discover argument facets, and we observed better performances compared to the method without attention weights. **(Seoul National University)**

Key words: argument mining, argument facets, argument clustering, semantic similarities, debate texts, attention weights

1. Introduction

Argument mining research aims to analyze the content and structure of argumentative texts. To do this, one must determine the ideological

* This work was supported by the 13th Overseas Faculty Training Program of College of Humanities at Seoul National University in 2018.

** First Author

*** Corresponding Author

posts' stances towards some given topic, detect the claims or arguments, and discover the evidence or justification for making such claims. For example, a government can check the public response to a policy and improve it based on its pros and cons by analyzing texts on forums or comments of news articles.

Such processes of understanding texts require various linguistic knowledge including lexical, syntactic, semantic, and pragmatic information. It is important to extract knowledge from the texts and convert it to appropriate features to be applied to models for text analysis. Several previous works of natural language processing constructed features including the existence and frequency of n-grams (Bag-of-words), part-of-speech tags, and dependency relations between words. Meanwhile, the latest general-purpose pretrained language models based on a large-scale text dataset, such as BERT (Devlin et al. 2019) and ELMo (Peters et al., 2018), learn such information themselves without hand-crafting linguistic features, using internal language modeling tasks utilizing contextual information within sentences. Proper usage of linguistic features and language models would help to process texts and to extract important information from them.

In this work we focus on analyzing the justifications used when making an argument, aiming to discover Argument Facets from online debates. Argument Facet refers to the important keywords which are commonly used in topic-relevant arguments. This corresponds to the concept of Aspects used in review texts. For example, a review about a restaurant may mention some common aspects for restaurants such as food, staff, and price. Similarly, we assume that argumentative texts would also have some common ideas to support their opinions. Based on this, we try to understand what kind of reasons or evidence people use when discussing ideological topics.

We deal with this problem as a twofold task: argument clustering and identifying facet words. First, we collected sentential arguments about ideological topics and grouped them into clusters, each consisting of arguments with similar justifications. This enables the following process to extract important words from sentences in each cluster representing

a type of reason or justification for argumentative stances of texts. Then for the second task follows as a pipeline, we applied a Transformer-based BERT model, one of the large-scale pretrained language models, to the arguments in each cluster to obtain attention weights assigned to each word in the argument sentence. We extracted the words which had higher attention weights per cluster and formed a list of representative words from that cluster. When lists of such words are given after clustering texts on a specific topic, we can understand how people think about the topic based on which reason or justification.

The Transformer architecture (Vaswani et al. 2017) is based on a multi-head self-attention mechanism, which calculates attention weights of words in a sentence via multiple heads and layers. Here, attention weights represent the calculated degree of attending of words to other words in a sentence based on context information, which we considered to be relevant to the importance of words in the sentence. The self-attention mechanism only considers words in a current sentence, while words in more than one sentences can be in regard in other cases. Each head parallelly performs the computation of attention weights repeated through several attention layers, where the output of a layer is used as input of the next layer, updating and refining the calculated attention weight values. BERT is one of the language models that slightly improved the internal structure of the Transformer.

While there is not a great deal of previous work related to this topic, most research in this area has used Hierarchical Agglomerative Clustering to cluster arguments according to similarity. In this paper, we used our own approach to clustering by defining three different similarities to combine and apply to a Spectral Clustering algorithm. The similarities between semantic vectors of sentences are constructed utilizing topic models and Transformer-based models based on sequences of words that compose sentences or documents. This resulted in clustering performances better than or close to the state-of-the-art.

After this, we used attention weights calculated for each sentence by the Transformer-based BERT model (the "BERT-Large" model (Devlin et al. 2019)) to select semantically important words in the sentence which

became the argument facets. Without being given any additional information about the subtopics or justifications of arguments, our experiments extracted the correct lists of facet words, which shows the positive effect of using attention weights. We regard this process of obtaining facet words as an unsupervised method to some degree since we do not have any gold label for those words.

The experiments and analyses within this paper focus on the points below.

- Constructing and combining different similarities between sentences for a clustering algorithm utilizing linguistic knowledge can improve the clustering of arguments.
- The attention weights assigned to each word in a sentence help selecting important words in the context of the sentence, resulting in the detection of facet words representing each type of justification well.

The rest of this paper is organized as follows. Section 2 shows existing works on argument mining, especially detection and clustering of arguments. Section 3 illustrates the Hasan and Ng (2014) dataset we utilized to analyze the arguments included in texts. Section 4 proposes a clustering method using a novel combination of similarity matrices between argumentative sentences and provides its evaluation results. Section 5 shows our attention-based method of detecting major argument facets from each argument cluster. Section 6 concludes with the summarization of this paper and the future direction of improving our methods.

2. Related Work

There has been quite a bit of research relating to approaches for detecting topic-relevant arguments from a large corpus. Rinott et al. (2015) obtained a ranked list of candidate context-dependent evidence using logistic regression. Hua and Wang (2017) used a log-linear model

with linguistic, sentiment, discourse, and style features to perform argument detection as a ranking problem. Stab et al. (2018) constructed a system called "ArgumenText" to retrieve sentential arguments about given controversial topics. Finally, Reimers et al. (2019) used state-of-the-art contextualized vectors such as BERT and ELMo with models based on LSTM (Hochreiter and Schmidhuber 1997) for argument detection.

Based on the identified arguments, some research tries to cluster arguments by their contents. Argument clustering seems to be divided into two trends: predicting the similarity between argument pairs and grouping the arguments into separate clusters. Misra et al. (2016) provided data, called Argument Quality (AQ) and Argument Facet Similarity (AFS), which includes human-annotated similarities (6 scales) between argument sentences. They predicted these similarities using several features such as word embeddings and LIWC-based information (Pennebaker et al., 2001). Misra et al. (2017) extracted central propositions (CP) from online dialogs using Mechanical Turk summarization and clustered those CPs with the Agglomerative Clustering algorithm. Reimers et al. (2019) expanded the AFS dataset to construct the UKP Aspect Corpus, which includes pairwise similarities of arguments. The similarities here were simplified into 2 labels (similar or dissimilar). They predicted pairwise similarities by two methods: with and without using Agglomerative Hierarchical Clustering (Day and Edelsbrunner 1984). They also used contextualized embedding models (ELMo and BERT) with TF-IDF (Salton and Buckley 1988), InferSent (Conneau et al., 2017), and GloVe (Pennington et al. 2014) to calculate similarities.

On the other hand, Boltužić and Šnajder (2015) aimed to identify arguments using the online debate data from Hasan and Ng (2014). They used the Hierarchical Agglomerative Clustering algorithm (Xu and Wunsch 2005) with vector-space similarities and Semantic Textual Similarities (Agirre et al. 2012). Trabelsi and Zaiane (2019) provided a contrasting overview of the main viewpoints of online debates and their reasonings. Using models based on Latent Dirichlet Allocation (LDA) (Blei et al. 2003), they constructed phrases as the basic unit of arguments and labeled them

with (topic, viewpoint) information. Based on the information labels the arguments were agglomeratively clustered by the number of word overlaps between phrase pairs. Reimers et al. (2019) labeled clusters using topic models, focusing on verb expressions to select a phrase out of each cluster as a representative.

Following the latter of the two trends of argument clustering, we grouped the sentential arguments representing similar justifications to each other from Hasan and Ng (2014) by applying the Spectral Clustering algorithm. While we utilized topic models and a contextualized embedding model as aforementioned works, we defined different similarities between sentences by using each model separately and combined them for clustering. We used LDA for the topic model to construct a vector of topic probability distribution per sentence, and the Semantic Textual Similarity (STS) API (Han et al. 2013) we used for the second similarity method in this paper also used a topic modeling algorithm, Latent Semantic Analysis (LSA) (Hofmann 1999). For the embedding model, we used a large-scale BERT model (the "BERT-Large" model (Devlin et al. 2019)).

Some works defined the concept of argument facets as aspects or aspect terms, introducing frame which is a collection of arguments related to the same aspect. Naderi and Hirst (2017) classified frames of news articles using neural networks such as BiLSTM (Hochreiter and Schmidhuber 1997) and GRU (Cho et al. 2014). Ajjour et al. (2019) identified frames by two-level of clustering (topic-level and frame-level) using the K-means algorithm (Lloyd 1957; MacQueen 1967) based on Euclidean distance.

Trautmann (2020) released an aspect-based argument mining (ABAM) benchmark with an annotated corpus. In the study, aspect terms and the stances included in the arguments are labeled and classified per token, as a sequence labeling problem. The study also provided a list of the top 5 most common aspects for each topic, which is the intended output of our second subtask.

Unlike most previous research, but with a slight overlap with Trautmann (2020), we extracted important words from several arguments

in each cluster to construct a list of argument facets. To do this, we referenced the methods of He et al. (2017) and selected words for which the attention model calculated high attention weights. He et al. (2017) reconstructed the embedding vectors of sentences in review texts using neural word embeddings and an attention model and then extracted an inferred aspect from each document. In this work, we used the BERT-Large model, which includes a multi-head attention model with enough numbers of heads and layers.

3. Data

For this research, we used Hasan and Ng (2014) online debate data. This data includes four different domains ("abortion," "gay rights," "marijuana," "obama") consisting of user posts organized by two different opposing stances ("pro" and "con"). We extracted only the data from these posts which contained reason labels. In the paper by Hasan and Ng (2014), they made a predefined list of reason labels for each domain and labeled each sentence according to the justification it contained.

For example, in the "marijuana" domain, there is a reason label "p-no_damage" which is an argument used in the "pro" stance towards marijuana legalization. A statement such as, "Does not cause any damage to our bodies," would be labeled with "p-no_damage."

Table 1 shows the number of reason labels and extracted sentences for each domain. Sentences which resulted in encoding errors were not included in the data.

Table 1. The size of the Hasan and Ng (2014) corpus.

domain	labels	sentences
"abortion"	15	732
"gay rights"	11	766
"marijuana"	12	689
"obama"	18	644

The exhaustive list of reason labels are shown in Table 2.

Table 2. The list of reason labels of Hasan and Ng (2014).

domain	pro labels	con labels
"abortion"	p-right, p-rape, p-not-human, p-mother_danger, p-baby_ill_treatment, p-birth_ctrl, p-not_murder, p-sick_mom, p-other	c-adopt, c-kill, c-baby_right, c-sex, c-bad_4_mom, c-other
"gay rights"	p-normal, p-right_denied, p-no_threat_for_child, p-born, p-religion, p-Other	c-religion, c-abnormal, c-threat_to_child, c-gay_problems, c-Other
"marijuana"	p-not_addictive, p-medicine, p-legal, p-right, p-no_damage, p-Other	c-health, c-mind, c-illegal, c-crime, c-addiction, c-Other
"obama"	p-economy, p-War, p-republicans, p-decision_policies, p-quality, p-health, p-foreign_policies, p-job, p-Other	c-economy, c-War, c-job, c-health, c-decision_policies, c-republicans, c-quality, c-foreign_policies, c-Other

4. Argument Clustering

4.1. Clustering Method

4.1.1. Clustering Algorithm

In this work, we used the Spectral Clustering algorithm, which is one

of the graph cut methods that uses the similarity between entities on a similarity graph and removes the edges with low similarity in order to split the graph into several subgraphs which results in similar data being clustered together. Here, the entities represent argumentative sentences including reason labels and the clustering process groups arguments that are semantically similar to each other.

The K-means and Hierarchical Agglomerative Clustering algorithms, the clustering algorithms used in previous research, were also tested, but spectral clustering resulted in the best performance in our experiments. The Scikit-learn library was used for implementation and the option "k-means" was used to assign labels in the embedding space. The number of clusters for each domain was determined according to the number of predefined reason labels in the Hasan and Ng (2014) data. Since using the number of predefined labels is not a perfect unsupervised method, determining a proper number of clusters without prior information would be important for future works.

Since clustering algorithms require similarity or distance between entities, we first had to implement an affinity matrix for our Spectral Clustering process. "Affinity" here refers to the similarity between arguments. For our spectral clustering, we defined three different similarity matrices and combined them.

4.1.2. Argument Similarity

We used three measures of similarity in order to create our matrices: semantic similarity between sentence embeddings (EMB), semantic textual similarity (STS), and similarity between topic probability distributions trained by a topic modeling algorithm, Latent Dirichlet Allocation (LDA).

- **EMB**: We transformed each sentential argument into a sentence embedding using a pretrained large-scale BERT (Devlin et al. 2019) model. BERT is a pretrained contextualized embedding model that learns generic linguistic knowledge via language modeling on a

large-scale text dataset. It is expected to perform well in most natural language processing tasks. We obtained sentence embeddings for arguments with a dimension of 1024 by averaging the embeddings of words in the sentence. Then we calculated the cosine distances of argument embeddings and converted them to similarities to form a similarity matrix.

- **STS**: We used the UMBC Semantic Textual Similarity API (Han et al. 2013) to get the semantic similarities between our arguments. The UMBC STS system used a topic modeling algorithm, Latent Semantic Analysis (LSA), and an English lexical database, WordNet, for the construction of semantic similarity features.
- **LDA**: We first obtained the topic distribution vector of each argument based on an LDA model using the gensim library (Rehurek and Sojka, 2010), and then calculated the Hellinger distances in order to get the similarities between those topic distributions. The topic modeling algorithm identifies topics used in documents by analyzing the distribution of word frequency in the documents.

Because EMB and LDA are expressed in distances, we converted them to their corresponding similarities by subtracting the distance between two arguments from one: $1-distance(s_1, s_2)$ for pairs of sentences. We first used EMB and STS separately to cluster the arguments and establish a baseline and then combined EMB, STS, and LDA similarities. This combination was calculated as the average of the three different similarities.

Table 3. Examples of clustered arguments in the "marijuana" domain. The numbers listed under the predicted cluster column are arbitrary; arguments with the same number were assigned to the same cluster.

argument	predicted cluster	gold cluster
marijuana has evidence of pharmaceutical uses, such as treatment of glaucoma.	6	p-medicine

it is proven that it helps with different things medically such as when going through chemo it gives you appetite, it helps with pain control	6	p-medicine
Cannabis has even been shown to promote new brain cell development, and its users report wide ranging benefits.	11	p-medicine
And don't forget so many lives ruined with excessive mandatory sentencing of jail times.	4	p-right
And even if it was harmful, how can anyone tell me what I can and can't put into my own body?	9	p-right
Also you must realize that a prohibition simply does not work. Pot smoking is still a very common phenomena despite it being illegal.	9	p-right
if we legalize pot there wil be a sharp increase in demand and consumption over a period of time	2	c-illegal
And if we want this to be legalize, then we will try to legalize too some harmful substances again soon after.	8	c-illegal
People abuse the drug as it is. If they legalize it, the abuse will grow exponentially. This applies for any illegal substance.	8	c-illegal
if you look up Cannabis and mental disorders, then you'll find that cannabis does more harm to you than good	9	c-mind
Marijuana causes some parts of the brain – such as those governing the emotions, memory and judgment – to spin out of control	11	c-mind
it depresses them, and marijuana fits in all of those positions	0	c-mind

Table 3 shows the examples of clustered arguments from the "marijuana" domain. Each argument has its gold-standard cluster label (predefined reason label by Hasan and Ng (2014)) according to its justification, and the clustering process assigns the predicted cluster label to the argument. The predictions in Table 3 are based on the Spectral Clustering algorithm and the combination of the three different similarity matrices defined above.

4.2. Evaluation

Following the example of Boltužić and Šnajder (2015), we evaluated the performance of our clustering by following two metrics: the Adjusted Rand Index (ARI) (Hubert and Arabie 1985) and the information-theoretic V-measure (Rosenberg and Hirschberg 2007).

4.2.1. ARI

Rand Index is a measure which shows whether two sentences are grouped into the same cluster according to their class. It is probabilistically adjusted to be the Adjusted Rand Index (ARI). The values range from 0 to 1, with a value close to 1 representing high clustering performance.

4.2.2. V-measure

V-measure (V) is the metric for homogeneity (h) and completeness (c) of clustering. Homogeneity shows how many different classes are included in one predicted cluster. The fewer different classes included in a single cluster, the better the clustering performance. Completeness refers to how many members of the class are elements of the same predicted cluster. If all data points that are members of a given class are also elements of the same cluster, the clustered results are complete. The value of V-measure is the harmonic mean of homogeneity and

completeness on a scale of 0 to 1, with 1 being the best clustering performance.

4.3. Results and Analysis

We set two baselines for clustering performance: one based on the cosine similarity between sentential arguments' embeddings (EMB) and the second being the semantic textual similarity between arguments (STS). Previous works also use EMB and STS similarities, although the specific embedding model or API used may be different. Our performance results are based on the average of 10 repeated experiments. Tables 4 and 5 shows the comparison of our results with the performance of Boltužić and Šnajder (2015), which utilized the same dataset.

Our results showed higher clustering performance in the domains for "abortion" and "gay rights" than previous works. Overall, combining two similarity matrices tended to result in better clustering performance than observed both in previous works and our own baselines. The combination of all three similarity matrices yielded by far the best results.

Table 4. Clustering performances on the topics "abortion" and "gay rights" compared to previous work. The highest performances from previous work and our approach are provided in bold.

model (linkage)	"abortion"				"gay rights"			
metrics	*h*	*c*	*V*	ARI	*h*	*c*	*V*	ARI
BoW (complete)	.05	.04	.04	.01	.04	.04	.04	.01
Bow (Ward's)	.22	**.27**	**.24**	.07	.13	**.17**	**.15**	.04
Skip-gram (Complete)	.17	.24	.20	.03	.09	.10	.10	.04
Skip-gram (Ward's)	**.24**	.22	.23	**.08**	**.16**	.15	**.15**	**.07**
STS (Complete)	.06	.06	.06	.02	.05	.05	.05	.01

EMB	.28	.27	.27	.16	.16	.16	.16	.07
STS	.27	.27	.27	.11	.20	**.22**	.21	.08
LDA	.20	.18	.19	.06	.10	.10	.10	.03
EMB+STS	.32	.31	.32	.15	.20	.20	.20	.08
EMB+LDA	.32	.30	.31	.15	.18	.17	.17	.08
STS+LDA	.28	.25	.27	.11	.21	.20	.20	**.09**
EMB+STS+LDA	**.34**	**.32**	**.33**	**.16**	**.22**	.21	**.21**	.09

While the “marijuana” and “obama” domain didn't obtain better performance results than previous works, they still showed that the combined similarity matrix out-performs the single use of each similarity. Especially in the “marijuana” domain, we obtained results that were very close to state-of-the-art performances.

One factor which may harm the clustering performance is the “other” cluster. Not every argument in the data fits neatly in one of the predefined reason labels and so they are assigned to the “p-other” or “c-other” cluster. Because of the nature of this cluster, it may be difficult to label the arguments contained into a coherent group.

Another difficulty with this task relates to the problems of overgeneralization. While arguments in a domain may have different subtopics and justifications, they all relate to the same main topic (for example, “abortion” instead of “p-not_human”). Because of this, the clustering algorithm may not be able to catch the detailed differences in meaning between the subtopics and justifications.

Additionally, and arguably more importantly, implicit statements contained in the data would also affect the overall clustering performance. This includes stylistic differences of sentential arguments, such as anecdote, supposition, refutation, etc. Although there may be relevant content in such sentences, the lack of explicit words makes the process more difficult.

Table 5. Clustering performances on the topics "marijuana" and "obama" compared to previous work. The highest performances from previous work and our approach are provided in bold.

model (linkage)	"marijuana"				"obama"			
metrics	*h*	*c*	*V*	ARI	*h*	*c*	*V*	ARI
BoW (complete)	.04	.04	.04	.00	.15	.15	.15	.03
Bow (Ward's)	.15	.20	.17	.02	.22	**.34**	.27	.04
Skip-gram (Complete)	.09	.22	.13	.02	.18	.26	.21	.04
Skip-gram (Ward's)	**.25**	**.24**	**.25**	**.19**	**.30**	.29	**.30**	**.10**
STS (Complete)	.05	.05	.05	.03	.11	.11	.11	.02
EMB	.21	.20	.20	.10	.20	.23	.22	.07
STS	.14	.15	.14	.05	.15	.19	.16	.03
LDA	.14	.15	.14	.05	.14	.13	.14	.02
EMB+STS	.22	.22	.22	.11	.21	.23	.22	.07
EMB+LDA	**.24**	.22	**.23**	.13	**.24**	**.24**	**.24**	.07
STS+LDA	.17	.15	.16	.08	.22	.22	.22	.06
EMB+STS+LDA	.24	**.23**	**.23**	**.14**	**.24**	**.24**	**.24**	**.08**

5. Discovering Argument Facets

In this phase, we attempted to obtain our argument facets, the list of keywords per cluster. We used a multi-head attention mechanism from a Transformer model, BERT, to get attention weights. The attention weights are expected to reflect the importance of each word in a sentence.

5.1. Method

First, we encoded the sentential arguments using the pretrained BERT

model (Devlin et al. 2019). BERT uses the Transformer model which includes the multi-head attention model. We selected the pretrained BERT-Large model, which consists of 16 heads, 12 layers, and 1024 dimensions for each word. Each head calculates attention weights for word tokens in a sentence separately. We averaged the attention weights for all heads and layers and assigned them to each word as the measure of that word's importance in the sentence.

For each sentence, we removed stopwords, punctuation, and topic words. Topic words are the specific words used for the domain names in the Hasan and Ng (2014) dataset: "abortion," "gay," "right," "marijuana," and "obama." These topic words were removed not to be the candidate facet words by simple matching. This is because such words could hamper the ability to distinguish detailed subtopics and justifications from the main topic. After these were removed, we then selected the words with the highest attention weights to be important words in each sentence as argument facets. The threshold for the number of words selected for each sentence was determined to be half the length of the sentence.

After this, we collected the selected words by cluster and sorted each list of words according to their frequencies. The words with the highest frequencies were regarded as the representative words of the clusters. Therefore, when we later came across these representative words while clustering, we could expect that the sentence would be clustered with the same justification as its corresponding representative word. Because we used words and not phrases as the basic unit of argument facets, the same word may appear in several different clusters.

Table 6 shows some examples of the words extracted from each sentence by attention weights and the sorted list of those words per cluster. We used the gold-standard (predefined) cluster labels in this section to better observe the pure effect of facet word extraction using the attention weights.

For example, for the cluster "p-not_murder" in the domain "abortion" in Table 6, the facet words extracted with the high attention weights include "murder," "kill," "destroys," and "cells," etc. The cluster includes arguments that support abortion with the reason which says "it is not

murder," and the extracted words can be used to infer the justification when the gold reason labels are not provided in an unsupervised manner. And the adequacy of the words to help the inference is evaluated by metrics in 5.2.

Table 6. The examples of extracted argument facets.

cluster	argument	words with high attention weights	collected words by cluster
p-not_murder	Abortion is not killing a human child.	killing	murder, kill, destroys, cells, mother, survive, week, baby, woman, inside, court, ...
p-not_murder	In America, the supreme court does not consider abortion to be murder.	court, consider, murder	
p-not_murder	It is well - established that it is not murder by the scientific community.	murder, established, scientific, community	
c-adopt	Giving up your child for adoption is an act of love rather than hate.	adoption, hate, love, child, rather	adoption, child, baby, give, want, families, couples, love, children, people, adopt, option, adopted, ...
c-adopt	If your not ready to raise a kid then put it up for adoption so it can be with a good family.	adoption, kid, family, ready, raise	
c-adopt	In any other situation, there are plenty of other options such as adoption.	options, adoption, situation, plenty	

5.2. Evaluation

We extracted highly-ranked words (by frequency) from each cluster to be the list of argument facets. In order to evaluate how well those words were detected, we used the same metrics as He et al. (2017): Coherence Score, User Evaluation, and Aspect Identification. Our baselines were constructed based on facet word extraction without attention weights.

5.2.1. Coherence Score

In order to evaluate whether the clusters had suitable argument facet words for the corresponding reason labels, and whether the collected words were coherent enough, we calculated the coherence score C as shown below. Given an aspect z, and a set of top N words of z, $S^z = w_1^z, \cdots, w_N^z$:

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} \log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_{l^z})}$$

where D_1 is the document frequency of word w, and D_2 is the co-document frequency of words w_1, w_2. The higher the coherence score is, the more coherent the arguments of a cluster are.

We averaged the coherence score of each cluster to get a scalar score per domain. For a baseline of this score, we identified the words with the highest frequency from the list of all words (regardless of the importance of the words) and calculated the coherence score for those words.

5.2.2. User Evaluation

This measure shows whether the argument facets assigned to each cluster are agreeable to humans. A human annotator assessed the top 20 words of the gold-standard clusters, and determined whether each

word was relevant to the justification, or subtopic, of the cluster. The labels and descriptions of the gold-standard, predefined reason clusters for each domain were provided for the annotator.

The results of this assessment are shown by *precision@n* (*p@n*). *P@n* is usually used for tasks that consist of obtaining a ranked list of sentences or documents and showing the proportion of the top-n documents that are relevant to a topic. If the system got r relevant document out of n documents, $p@n = r/n$. In this work, *p@n* would be the proportion of agreeable words out of proposed facet words.

To see the effect of attention weights, we set the same baseline as used with the coherence score: extracting the highest frequency words from the list of all the words (not filtered by attention weights) in the arguments.

5.2.3. Aspect Identification

For aspect identification, we evaluated whether the extracted argument facets – that is, the representative words of each sentence – are proper and whether the extraction method using the attention weights is effective. In order to do this, a human annotator was shown the most important word (word with the highest attention weight) for each sentence and asked to judge if that word was suitable for the gold cluster of the sentence.

He et al. (2017) evaluated the inferred aspect of each sentence related to the gold aspects by precision, recall, and F1 score. Instead of using the inferred aspect, we used the representative words to directly construct the list of argument facets. In this way, we slightly changed the metrics and used accuracy to measure the properness of the words, judging the properness as binary.

To establish a baseline, we used another way of extracting the most representative word of each sentence without using attention weights. He et al. (2017) did this by calculating the average between all the word embeddings in the sentence as an anchor and then selecting the word closest to that average to be the representative. We modified this

approach and replaced the average anchor with the sentence embeddings extracted from a large-scale BERT model (Devlin et al. 2019). Then we calculated the distance between the word embeddings and the BERT-Large embeddings using cosine distances.

5.3. Results and Analysis

5.3.1. Coherence Score

We calculated the coherence scores versus the top *n* terms for each domain, where *n* is the threshold of argument facet extraction from the list of words for each cluster. Here the data is clustered according to the predefined gold reason labels, and we compare the coherence scores for when attention weights were used versus unused.

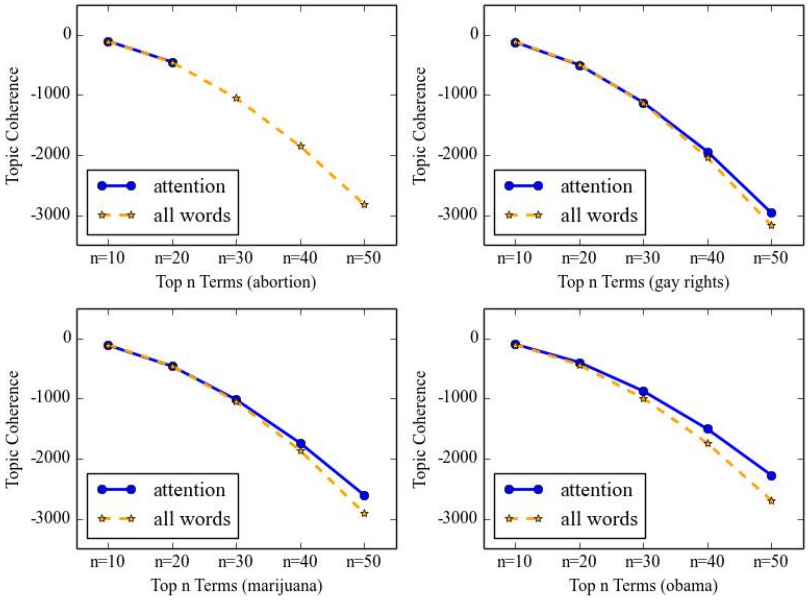

Figure 1. Coherence scores for the four domains.

As Figure 1 shows, smaller coherence values mean the extracted top n terms are less coherent and a larger n corresponds to lower coherence score. Three of the domains, not including "abortion," show that the extraction of argument facets using attention weights resulted in higher coherence score than extraction without attention weights. In other words, the facet words extracted based on attention weights are more coherent, and better show the effect of attention weights.

In the case of the "abortion" domain, the coherence score for the attention-based method is calculated only for n=10 and n=20. This is because some clusters in the domain had less than 30 facet words extracted, possibly resulting from shorter length of argument sentences or overlap of extracted words to each other. Larger n could disturb obtaining precise average coherence score of clusters.

Additionally, we excluded the "other" clusters when calculating the coherence scores for each of the four domains. We chose to do this because the extracted words from those "other" clusters would not be coherent due to their reference to a wider variety of less common justifications.

5.3.2. User Evaluation

We calculated the average of $p@n$ for the gold clusters, excluding the two "other" clusters, "p-other" and "c-other," for each domain. Here we set the n as 20, considering the minimum length of facet word lists of all clusters of all domains.

Table 7 shows the number of coherent clusters. The cluster is coherent when the majority, 10, of 20 facet words are judged as relevant to the cluster by a human annotator. Overall, the method of extracting argument facets based on the attention weights seems to have more coherent clusters than the method which did not use attention weights.

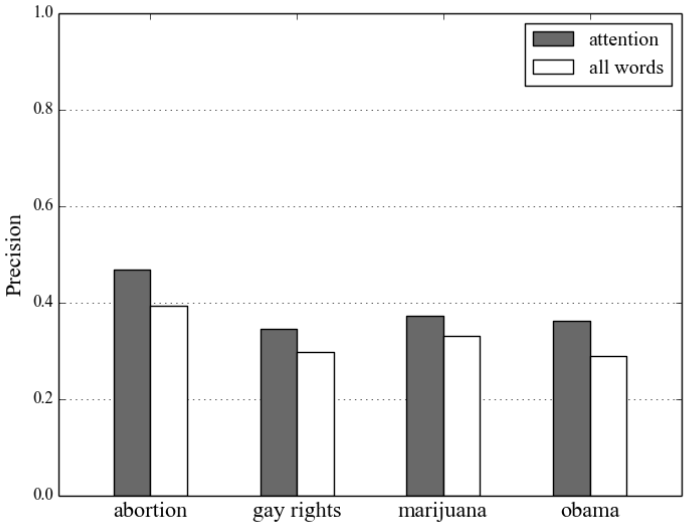

Figure 2. *P@n* values for the argument facet words.

Table 7. The number of coherent clusters.

domain	attention	all words	total number of clusters
"abortion"	6	2	13
"gay rights"	1	0	9
"marijuana"	2	0	10
"obama"	2	0	16

Figure 2 shows the *p@n* values for the facet words checked by the annotator. The figure represents the number of facet words that were judged to be relevant over the total number of facet words (20). As we can observe from the number of coherent clusters seen in Table 7, the precision values of clusters are higher when attention weights were utilized (darker part in the figure) compared to when only frequencies were used (brighter part in the figure).

Generally, the number (or proportion) of coherent clusters is very small. This may be because people use various words and phrases, including very implicit expressions, to justify their opinions. This may include words which are not directly relevant to the subtopic, or justification of the argument. Similarly, facet word lists for some clusters contained various words which only appeared with a frequency of 1.

Additionally, we observed that some words are not unrelated to the clusters but make distinguishing between clusters difficult. For example, in the "obama" domain, there are both pro and con clusters for the subtopics "foreign policies" and "health care" ("p-foreign_policies," "c-foreign_policies," "p-health," "c-health"). This is because people may hold different opinions on whether such subtopics are good or bad. To complicate this further, most of the arguments in such domains mention the policies of several specific fields, which results in the appearance of words such as "reform," "bill," "pass," and "regulation" commonly in many different clusters. We didn't consider such words as the gold-standard representative words for the clusters as they would dilute the coherent cluster.

5.3.3. Aspect Identification

Table 8. Accuracies of facet words from each cluster.

cluster	using attention weights	without attention weights
"p-not_murder"	0.765	0.471
"p-rape"	0.667	0.455
"p-birth_ctrl"	0.636	0.455
"c-adopt"	0.718	0.385
"c-kill"	0.683	0.443
"c-bad_4_mom"	0.667	0.333

Table 8 shows the accuracy of the top six clusters from the "abortion" domain for the purpose of comparing the effectiveness of attention

weights. We chose to focus on this domain because it showed the best clustering performance.

According to the table, the accuracy is much higher when attention weights from the transformer model were used to decide the most representative facet word for each sentence.

In contrast, using the word embeddings and the average of sentence embedding, the words that were extracted were not as relevant as the words chosen with attention weights. For example, the justification, "Put the baby up for adoption," from the "c-adopt" cluster, collects "babies," "child," etc. as its facet words. Similarly, the justification from the opposition, "Abortion is not murder," from the "p-not_murder" cluster selects "fetus" for their top ranked word. In some cases, words such as "also," "never," and "well" are extracted as important words from the clusters, despite that they are not helpful when distinguishing clusters and their justifications.

On the other hand, the method using attention weights collects different words than the examples above. The "c-adopt" cluster shows words such as "adopt," "adoption," and "give," while the "p-not_murder" cluster shows words such as "killing" and "murder" as the words with highest attention weight in the sentence. This evidence supports the method of using attention weights because we were able to obtain the expected words without being given any additional information about the reasons or justifications for each sentence.

Extracting facet words using attention weights performed better than using only frequencies or average of all sentence embeddings in a cluster instead of attention weights. The results above show that the extracted words based on the attention weights are proper to represent the justifications involved in sentential arguments. In other words, the usage of a large-scale pretrained Transformer model based on various linguistic knowledge could help to capture semantically important words as argument facet.

6. Conclusion

In this work, we proposed a model which clustered topic-relevant (pre-identified) arguments according to their subtopic or justification of specific stances for a domain, and extracted argument facets in the form of a word list representing each cluster.

We used the combination of customized similarity matrices for Spectral Clustering to improve performance. These similarity matrices consisted of cosine similarity of BERT embeddings, Semantic Textual Similarity, and similarity between LDA-based topic model probability distributions. Our experiments provided results better than or close to the results of previous work. Additionally, we used attention weights calculated by the Transformer model to discover argument facets, and observed better performance than when facet words were extracted without attention weights.

In the future, we can improve the performance of clustering by taking into account the various problems and difficulties associated with clustering methods, including the way of deciding the number of clusters in an unsupervised manner. We may also work to improve the method of discovering argument facets by exploring facets on the phrase-level as opposed to only the word-level. By constructing higher-level units of argument facets such as phrases, we would more effectively be able to represent their meanings.

References

Agirre, Eneko, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. "SemEval-2012 task 6: A pilot on semantic textual similarity," *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*, 385-393.

Ajjour, Yamen, Milad Alshomary, Henning Wachsmuth, and Benno Stein. 2019. "Modeling frames in argumentation," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2915-2925.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent dirichlet allocation," *the Journal of machine Learning research* 3, 993-1022.

Boltužić, Filip and Jan Šnajder. 2015. "Identifying prominent arguments in online debates using semantic textual similarity," *Proceedings of the 2nd Workshop on Argumentation Mining*, 110-115.

Cho, Kyunghyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. "On the properties of neural machine translation: Encoder-decoder approaches," *arXiv preprint* arXiv:1409.1259.

Conneau, Alexics, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. "Supervised learning of universal sentence representations from natural language inference data," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 670-680.

Day, William HE, and Herbert Edelsbrunner. 1984. "Efficient algorithms for agglomerative hierarchical clustering methods," *Journal of Classification* 1(1), 7-24.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint* arXiv:1810.04805.

Han, Lushan, Abhay L Kashyap, Tim Finin, James Mayfield, and Jonathan Weese. 2013. "Umbc ebiquity-core: Semantic textual similarity systems," *Second Joint Conference on Lexical and Computational Semantics (* SEM) Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, 44-52.

Hasan, Kazi Saidul and Vincent Ng. 2014. "Why are you taking this stance? identifying and classifying 791 reasons in ideological debates," *Proceedings of the 2014 Conference on Empirical Methods in Natural 793 Language Processing (EMNLP)*, 751-762.

He, Ruidan, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2017. "An unsupervised neural attention model for aspect extraction," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 388-397.

Hochreiter, Sepp and Jürgen Schmidhuber. 1997. "Long short-term memory," *Neural Computation* 9(8), 1735-1780.

Hofmann, Thomas. 1999. "Probabilistic latent semantic analysis," *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI'99)*, 289-296.

Hua, Xinyu and Lu Wang. 2017. "Understanding and detecting supporting arguments of diverse types," *arXiv preprint* arXiv:1705.00045.

Hubert, Lawrence and Phipps Arabie. 1985. "Comparing partitions," *Journal of Classification* 2(1), 193-218.

Lloyd, Stuart P. 1957. "Least squares quantization in PCM," *Technical Report RR-5497, Bell Lab*.

MacQueen, James. "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1(14), 1967.

Misra, Amita, Pranav Anand, Jean E Fox Tree, and Marilyn Walker. 2017. "Using summarization to discover argument facets in online ideological dialog," *arXiv preprint* arXiv:1709.00662.

Misra, Amita, Brian Ecker, and Marilyn A Walker. 2016. "Measuring the similarity of sentential arguments in dialog," *17th Annual Meeting of the Special Interest Group on Discourse and Dialogue,* 276.

Naderi, Nona and Graeme Hirst. 2017. "Classifying Frames at the Sentence Level in News Articles," *Proceedings of the International Conference Recent Advances in Natural Language Processing, {RANLP} 2017*, 536-542.

Pennebaker, James W., Martha E. Francis, and Roger J. Booth. 2001. "Linguistic inquiry and word count: LIWC 2001," *Mahway: Lawrence Erlbaum Associates* 71(2001).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. 2014. "GloVe: Global Vectors for Word Representation," *Empirical Methods in Natural Language Processing (EMNLP)*, 1532-1543.

Rehurek, Radim and Petr Sojka. 2010. "Software framework for topic modelling with large corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.* Citeseer.

Reimers, Nils, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. "Classification and clustering of arguments with contextualized word embeddings," *arXiv preprint* arXiv:1906.09821.

Rinott, Ruty, Lena Dankin, Carlos Alzate Perez, Mitesh M Khapra, Ehud Aharoni, and Noam Slonim. 2015. "Show me your evidence-an automatic method for context dependent evidence detection," *Proceedings of the 2015 conference on empirical methods in natural language processing*, 440-450.

Rosenberg, Andrew and Julia Hirschberg. 2007. "V-measure: A conditional entropy-based external cluster evaluation measure," *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, 410-420.

Salton, Gerard, and Christopher Buckley. 1988. "Term-weighting approaches in automatic text retrieval," *Information Processing & Management* 24(5), 513-523.

Stab, Christian, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, Benjamin Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. "Argumentext: Searching for arguments in heterogeneous sources," *Proceedings of the 2018 conference of the North American chapter of the association for computational linguistics: demonstrations*, 21-25.

Trabelsi, Amine and Osmar R Zaiane. 2019. "Contrastive reasons detection and clustering from online polarized debate," *arXiv preprint* arXiv:1908.00648.

Trautmann, Dietrich. 2020. "Aspect-Based Argument Mining," *Proceedings of the 7th*

Workshop on Argument Mining, 41-52.
Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. "Attention is All you Need," *Advances in Neural Information Processing Systems* 30, 5998-6008.
Xu, Rui, and Donald Wunsch. 2005. "Survey of clustering algorithms," *IEEE Transactions on Neural Networks* 16(3), 645-678.

Sangah Lee, Postdoctoral Researcher
BK Post Doctoral Program in Graduate School of Data Science, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
E-mail: visualjan@snu.ac.kr

Hyopil Shin, Professor
Graduate School of Data Science and Dept. of Linguistics, Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea
E-mail: hpshin@snu.ac.kr

Received: February 10, 2021
Revised: March 18, 2021
Accepted: March 26, 2021