Computational Argumentation Synthesis as a Language Modeling Task

Roxanne El Baff [1] **Henning Wachsmuth** [2] **Khalid Al-Khatib** [1]
Manfred Stede [3] **Benno Stein** [1]
[1] Bauhaus-Universität Weimar, Weimar, Germany, <first>{.<last>}+@uni-weimar.de
[2] Paderborn University, Paderborn, Germany, henningw@upb.de
[3] University of Potsdam, Potsdam, Germany, stede@uni-potsdam.de

Abstract

Synthesis approaches in computational argumentation so far are restricted to generating claim-like argument units or short summaries of debates. Ultimately, however, we expect computers to generate whole new arguments for a given stance towards some topic, backing up claims following argumentative and rhetorical considerations. In this paper, we approach such an argumentation synthesis as a language modeling task. In our language model, argumentative discourse units are the "words", and arguments represent the "sentences". Given a pool of units for any unseen topic-stance pair, the model *selects* a set of unit types according to a basic rhetorical strategy (logos vs. pathos), *arranges* the structure of the types based on the units' argumentative roles, and finally *"phrases"* an argument by instantiating the structure with semantically coherent units from the pool. Our evaluation suggests that the model can, to some extent, mimic the human synthesis of strategy-specific arguments.

1 Introduction

Existing research on computational argumentation largely focuses on the *analysis* side. Various analysis tasks are widely studied including identifying the claims along with their supporting premises (Stab and Gurevych, 2014), finding the relation between argumentative units (Cocarascu and Toni, 2017), and assessing the persuasiveness of arguments (Habernal and Gurevych, 2016).

Diverse downstream applications, however, necessitate the development of argumentation *synthesis* technologies. For example, synthesis is needed to produce a summary of arguments for a given topic (Wang and Ling, 2016) or to build a debating system where new arguments are exchanged between the users and the system (Le et al., 2018).

As a result, a number of recent studies addresses the argumentation synthesis task. These studies have proposed different approaches to generating claims or reasons for a given topic, partly with a particular stance towards the topic (Bilu and Slonim, 2016; Hua and Wang, 2018). However, the next important synthesis step is still missing in the literature, namely, to generate complete texts including both argumentative and rhetorical considerations. With the latter, we refer to Aristotle's three means of persuasion: logos (providing logical arguments), ethos (demonstrating credibility), and pathos (evoking emotions). As discussed by Wachsmuth et al. (2018), following a rhetorical strategy is key to achieving persuasion with argumentative texts.

This paper proposes a new computational approach that synthesizes argumentative texts following a rhetorical strategy. We do not tackle this task immediately "in the wild", i.e., generating an entirely new argumentative text for a freely-chosen topic and a possibly complex strategy. Rather, we consider a "controlled" synthesis setting, with the goal of successively creating models that are able to deal with more complex settings later on.

In particular, given a pool of argumentative discourse units (ADUs), our approach generates arguments for any unseen pair of topic and stance (e.g., "con abortion") as well as a basic rhetorical strategy (i.e., logos-oriented vs. pathos-oriented).[1] To abstract from the arguments' topics during training, we first identify different ADU types using clustering. Our approach then learns to *select* unit types matching the given strategy and to *arrange* them according to their argumentative roles. Both steps are realized as a language model where ADUs represent words and arguments are sentences. Finally, our approach *"phrases"* an argument by predicting the best set of semantically related ADUs for the arranged structure using supervised regression. Thereby, we ensure that the synthesized texts are

[1] We consider a single argument to be a sequence of ADUs where each ADU has a specific role: thesis, con, or pro.

Proceedings of The 12th International Conference on Natural Language Generation, pages 54–64,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

composed of meaningful units, a property that neural generation methods barely achieve so far.

In our evaluation, we utilize the dataset of Wachsmuth et al. (2018). This dataset contains 260 argumentative texts on 10 topic-stance pairs, where each text composes five ADUs in a logos-oriented or pathos-oriented manner. In our experiments, we train our approach on nine topic-stance pairs and then generate an argument for the tenth. The results demonstrate that our approach successfully manages to combine pairs of ADUs, but its performance on longer sequences of ADUs is limited.

Altogether, our contribution is three-fold:

1. A new view of argumentation synthesis that represents argumentative and rhetorical considerations with language modeling.
2. A novel approach that selects, arranges, and phrases ADUs to synthesize strategy-specific arguments for any topic and stance.
3. First experimental evidence that arguments with basic rhetorical strategies can be synthesized computationally.[2]

2 Related Work

Recently, some researchers have tackled argumentation synthesis statistically with neural networks. For instance, Wang and Ling (2016) employed a sequence-to-sequence model to generate summaries of argumentative texts, and Hua and Wang (2018) did similar to generate counterarguments. Using neural methods in text generation, it is possible to achieve output that is on topic and grammatically (more or less) correct. However, when the desired text is to span multiple sentences, the generated text regularly suffers from incoherence and repetitiveness, as for instance discussed by Holtzman et al. (2018) who examine texts that were produced by RNNs in various domains. While these problems may be tolerable to some extent in some applications, such as chatbots, bad text cannot be accepted in an argumentative or debating scenario, where the goal is to convince or persuade a reader (rather than to merely inform or entertain).

Holtzman et al. (2018) propose to alleviate incoherence and repetitiveness by training a set of discriminators, which aim to ensure that a text respects the Gricean maxims of quantity, quality, relation, and manner (Grice, 1975). To this end, they employ specific datasets, such as one that opposes authentic text continuation to randomly-sampled text. The discriminators learn optimal weightings for the various models and their combination, such that overall text quality is maximized. For argumentation, we hypothesize that one needs to go even further and eventually account for the author, implementing her underlying *intention* in the different parts of an argumentative text as well as in the relations between the parts.

In the past times of rule-based text generation, argumentation synthesis was a popular task (Zukerman et al., 2000). Approaches involved much handcrafted (linguistic and domain) knowledge and user modeling. For example, the system of Carenini and Moore (2006) compares attributes of houses (from a database) to desired target attributes (from a user model), to then recommend a house to the reader in a convincing text following the Gricean maxims. To this end, it selected house attributes potentially interesting to the user, arranged, and finally phrased them. The resulting texts resembled the arguments we work with here, which have been manually composed by experts (Wachsmuth et al., 2018) from the claims, evidence, and objections in the arg-microtext corpus (Peldszus and Stede, 2016). To achieve a similar level of output control, today's text-to-text generation models need to account for the various interdependencies between the text units to be combined.

Most related to our approach is the system of Sato et al. (2015), where a user can enter a claim-like topic along with a stance. The system then generates argumentative paragraphs on specific aspects of the topic by selecting sentences from 10 million news texts of the Gigaword corpus. Potentially relevant aspects are those that trigger evaluative judgment in the reader. The sentences are arranged so that the text starts with a claim sentence and is followed by support sentences, employing the approach of Yanase et al. (2015). The support sentences are ordered by maximizing the semantic connectivity between sentences. Finally, some rephrasing is done in terms of certain aspects of surface realization. In a manual evaluation, however, no text was seen as sounding natural, underlining the difficulty of the task. In contrast to Sato et al. (2015), we learn directly from input data what argumentative discourse units to combine and how to arrange them. We leave surface realization aside to keep the focus on the argument composition.

[2]The code for running the experiments is available here: https://github.com/webis-de/inlg19-argumentation-synthesis

Role	ID	Argumentative Discourse Unit
Thesis	t_1	German universities should on no account charge tuition fees
	t_2	the universities in Germany should not under any circumstances charge tuition fees
	t_3	tuition fees should not generally be charged by universities
	t_4	universities should not charge tuition fees in Germany
Con	c_1	one could argue that an increase in tuition fees would allow institutions to be better equipped
	c_2	those who study later decide this early on, anyway
	c_3	to oblige non-academics to finance others' degrees through taxes is not just
	c_4	unfortunately sponsoring can lead to disagreeable dependencies in some cases
Pro	p_1	education and training are fundamental rights which the state, the society must provide
	p_2	education must not be a question of money in a wealthy society such as Germany
	p_3	fees result in longer durations of studies
	p_4	funding-wise it ought to be considered how costs incurred by students from other (federal) states can be reimbursed
	p_5	if a university lacks the funds, sponsors must be found
	p_6	longer durations of studies are costly
	p_7	studying and taking higher degrees must remain a basic right for everyone
	p_8	there are other instruments to motivate tighter discipline while studying
	p_9	this would impede or prevent access to those who are financially weaker
	p_{10}	this would mean that only those people with wealthy parents or a previous education and a part-time job while studying would be able to apply for a degree programme in the first place
	p_{11}	universities are for all citizens, independent of their finances
	p_{12}	what is the good of a wonderfully outfitted university if it doesn't actually allow the majority of clever people to broaden their horizons with all that great equipment
Topic		Should all universities in Germany charge tuition fees? **Stance** Con

Table 1: The candidate thesis, con, and pro units for one topic-stance pair in the dataset of Wachsmuth et al. (2018).

Some other approaches have been proposed that recompose existing text segments in new arguments. In particular, Bilu and Slonim (2016) generated new claims by "recycling" topics and predicates that were found in a database of claims. Claim selection involves preferring predicates that are generally amenable to claim units and that are relevant for the target topic. Egan et al. (2016) created summaries of the main points in a debate, and Reisert et al. (2015) synthesized complete arguments from a set of manually curated topic-stance relations based on the fine-grained argument model of Toulmin (1958). However, we are not aware of any approach that synthesizes arguments fully automatically, let alone that follows rhetorical considerations in the synthesis process.

3 Data

To develop our model for argumentation synthesis, we exploit the dataset recently developed by Wachsmuth et al. (2018). The dataset comprises 260 *manually* generated argumentative texts. The generation of each text, for one topic-stance pair, has been conducted in a systematic fashion following the three canons of rhetoric (Aristotle, 2007):

1. *Inventio* $\sim$ *Selecting* a subset of argumentative discourse units (ADUs) from a pool of given ADUs for a topic-stance pair.
2. *Dispositio* $\sim$ *Arranging* the selected ADUs in a sequential order.
3. *Elocutio* $\sim$ *Phrasing* the arranged ADUs by adding connectives at unit-initial or unit-final positions.

Specifically, Wachsmuth et al. (2018) selected a pool of 200 ADUs for 10 pairs of controversial topic and stance from the English version of the arg-microtexts corpus (Peldszus and Stede, 2016). As a preprocessing step, they "decontextualized" these ADUs manually by removing connectives, resolving pronouns, and similar. Each topic-stance pair comes with 20 such ADUs: four theses, four con units, and 12 pro units. Table 1 shows the ADU list for one topic-stance pair.

26 participants were asked by Wachsmuth et al. (2018) to create short argumentative texts for each topic-stance pair following one of two basic rhetorical strategies: (1) *logos-oriented*, i.e., arguing logically, and (2) *pathos-oriented*, i.e., arguing based on emotional appeals. For each topic-stance pair they created an argument by selecting one thesis, one con and three pro units that they thought could best form a persuasive argument following the given strategies. Table 2 shows two samples of generated arguments in the dataset.

The dataset contains 130 logos-oriented and 130

Strategy	ID	Text Manually Synthesized From Five Argumentative Discourse Units
Logos-oriented	c_1	one could argue that an increase in tuition fees would allow institutions to be better equipped,
	t_1	*however* German universities should on no account charge tuition fees.
	p_1	education and training are fundamental rights which the state, the society must provide,
	p_{12}	*because* what is the good of a wonderfully outfitted university if it doesn't actually allow the majority of clever people to broaden their horizons with all that great equipment.
	p_4	*Besides,* funding-wise it ought to be considered how costs incurred by students from other (federal) states can be reimbursed.
Pathos-oriented	p_1	education and training are fundamental rights which the state, the society must provide.
	t_2	*This is why* the universities in Germany should not under any circumstances charge tuition fees.
	c_1	one could argue that an increase in tuition fees would allow institutions to be better equipped,
	p_3	*however* fees result in longer durations of studies
	p_6	*and* longer durations of studies are costly.

Table 2: two sample arguments manually synthesized from the ADUs in Table 1, which are included in the dataset of Wachsmuth et al. (2018). The italiced connectives were added by the participants; they are *not* part of the ADUs.

pathos-oriented argumentative texts. We use these 260 texts to develop and evaluate our computational model for argumentation synthesis.

4 Approach

This section presents our computational approach to synthesize arguments for any pair of topic and stance, following one of two basic rhetorical strategies: arguing logically (*logos-oriented*) or arguing emotionally (*pathos-oriented*). A black-box view of the approach is shown in Figure 1.

As input, our approach takes a strategy as well as a pool of argumentative discourse units (ADUs) for any specific topic-stance pair x. Each ADU has the role of a *thesis* (in terms of claim with a stance on the topic), a *con* point (objecting the thesis), or a *pro* point (supporting the thesis). The approach then imitates the human selection, arrangement, and "phrasing" of a sequence of n ADUs, in order to synthesize an argument. Phrasing is done only in terms of picking semantically coherent ADUs for the arranged sequence; the addition of connectives between ADUs is left to future work.

Below, we detail how we realize each step (selection, arrangement, and phrasing) with a topic-independent model. For each step, we explain how it is trained (illustrated in Figure 2) and how it is applied to an unseen topic-stance pair (Figure 3).

4.1 Selection Language Model

This model handles the selection of a set of n ADUs for a topic-stance pair x and a rhetorical strategy. We approach the selection as a language modeling task where each ADU is a "word" of our language model and each argument a "sentence". To abstract from topic, the model actually selects ADU *types*, as explained in the following.

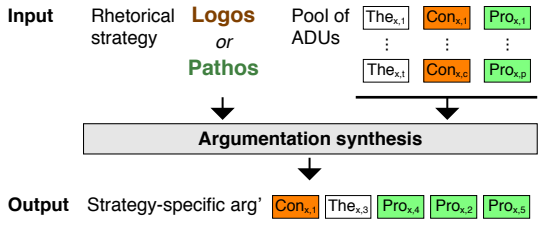

Figure 1: Black-box view of our argumentation synthesis approach. The input is a rhetorical strategy as well as a pool of thesis, con, and pro ADUs for some topic-stance pair x. The approach outputs a strategy-specific sequence of n ADUs as an argument for x (here, $n=5$).

4.1.1 Training of the Model

We start from a training set of ADUs for a set of m topic-stance pairs. To generalize the language model beyond the covered topics, each ADU is represented using features that aim to capture general emotion-related and logic-related characteristics, accounting for the two given strategies.

In particular, we first cluster the pool of all training ADUs based on their feature representation. As a result, each ADU is represented by a cluster label (A–F in Figure 2), where each label represents one ADU type. Now, for each of the strategies, we map each manually-generated sequence of ADUs to a sequence of cluster labels. Using these sequences of labels, we train one separated selection language model for each strategy.

For clustering, we rely on topic-independent features that we expect to implicitly encode logical and emotional strategies: (1) psychological meaningfulness (Pennebaker et al., 2015), (2) eight basic emotions (Plutchik, 1980; Mohammad and Turney, 2013), and (3) argumentativeness (Somasundaran et al., 2007). In the following, we elaborate on the concrete features that we extract:

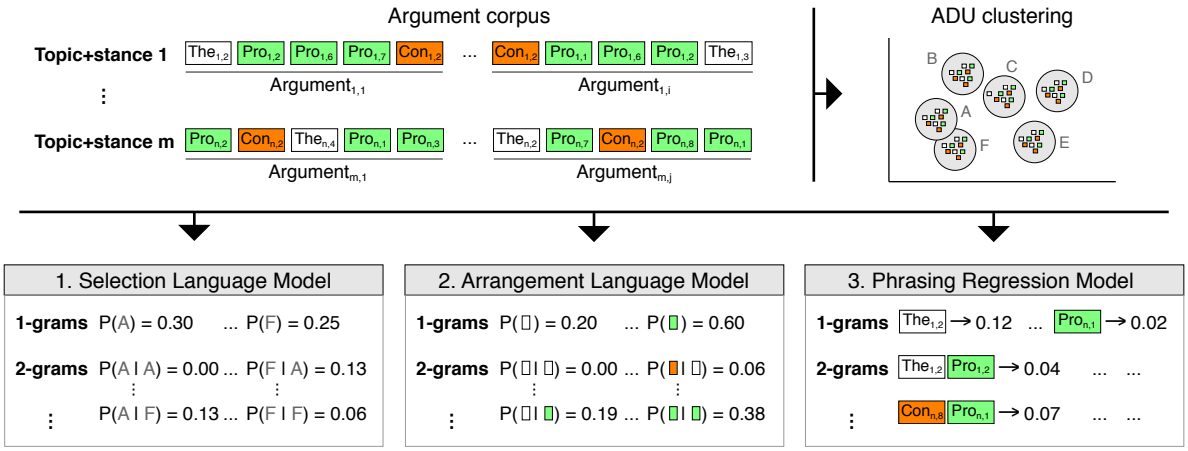

Figure 2: Illustration of training the three models of our argumentation synthesis approach. The input is a corpus of argumentative texts for m topic-stance pairs, each decomposed into a sequence of theses, con units, and pro units. Initially, the set of all these ADUs is clustered to obtain a set topic-independent ADU types, called A–F here. (1) *Selection language model*: Each argument is converted from a sequence of ADUs to a sequence of ADU types, where a language model is trained on these type sequences. (2) *Arrangement language model:* Each argument is converted from a sequence of ADUs to a sequence of ADU roles (thesis, pro, and con) where a language model is trained on these ADU role sequences. (3) *Phrasing regression model:* A linear regression model is trained which scores each ADU sequence with respect to its semantic coherence.

Linguistic Inquiry and Word Count (LIWC) LIWC is a lexicon-based text analysis that counts words in psychologically meaningful categories (Tausczik and Pennebaker, 2010). We use the version by Pennebaker et al. (2015), which contains the following 15 dimensions:

1. *Language metrics*, e.g., words per sentence.
2. *Function words*, e.g., pronouns and auxiliary verbs.
3. *Other grammar*, e.g., common verbs and comparisons.
4. *Affect words*, e.g., positive emotion words.
5. *Social words*, e.g., "family" and "friends".
6. *Cognitive processes*, e.g., "discrepancies" and "certainty".
7. *Perceptual processes*, e.g., "feeling".
8. *Biological processes*, e.g., "health".
9. *Core drives and needs*, e.g., "power" and "reward focused".
10. *Time orientation*, e.g., past-focused.
11. *Relativity*, e.g., "time" and "space".
12. *Personal concerns*, e.g., "work" and "leisure".
13. *Informal speech*, e.g., fillers and nonfluencies.
14. *Punctuation*, e.g., periods and commas.
15. *Summary variables*, as detailed below.

There are four summary variables, each of which is derived from various LIWC dimensions: (1) *analytical thinking* (Pennebaker et al., 2014), i.e., the degree to which people use narrative language (low value), or more logical and formal language (high); (2) *clout* (Kacewicz et al., 2014), i.e., the relative social status, confidence, and leadership displayed in a text; (3) *authenticity* (Newman et al., 2003), i.e., the degree to which people reveal themselves in an authentic way; and (4) *emotional tone* (Cohn et al., 2004), i.e., negative for values lower than 50 and positive otherwise.

NRC Emotional and Sentiment Lexicons We use the NRC lexicon of Mohammad and Turney (2013). The lexicon has been compiled manually using crowdsourcing and contains a set of English words and their associations with (1) *sentiment*, i.e., negative and positive polarities, and (2) *emotions*, i.e., the eight basic emotions defined by Plutchik (1980): anger, anticipation, disgust, fear, joy, surprise, sadness, and trust. These features are represented as the count of words associated with each category (e.g., the count of *sad words* in an ADU).

MPQA Arguing Lexicon Somasundaran et al. (2007) constructed a lexicon that includes the following arguing patterns: *assessments*, *doubt*, *authority*, *emphasis*, *necessity*, *causation*, *generalization*, *structure*, *conditionals*, *inconsistency*, *possibility*, *wants*, *contrast*, *priority*, *difficulty*, *inyour-*

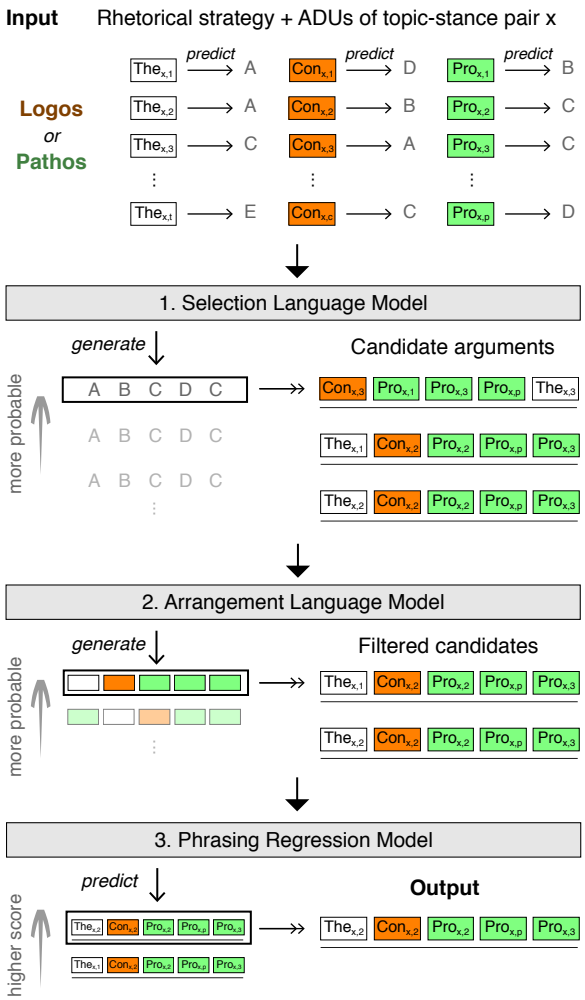

Figure 3: Illustration of applying our synthesis approach. Given the predicted type of each input ADU of the given topic-stance pair *x*, (1) the *selection* generates the most probable type sequence, (A, B, C, D, C). From the type sequence, a set of candidate arguments is decoded. (2) The *arrangement* filters out candidates not matching the most probable ADU role sequence, $(Thesis, Con, Pro, Pro, Pro)$. (3) *Phrasing* scores each remaining argument and outputs the top argument.

shoes, *rhetorical question*. We use the count of each arguing pattern in text as one feature (e.g., number of *assessments* patterns in an ADU).

4.1.2 Application of the Model

As shown in Figure 3, the selection language model takes the ADUs of an unseen topic-stance x as input. It then outputs a set of candidate arguments, in terms of sequences of ADUs. Each ADU is encoded into a cluster label (representing an ADU type). For example, one might have the following mappings, given the six labels A–F from Figure 2:

$$A \leftarrow \{The_{x,1}, The_{x,2}, Con_{x,3}\}$$
$$B \leftarrow \{Con_{x,2}, Pro_{x,1}\}$$
$$C \leftarrow \{The_{x,3}, Con_{x,c}, Pro_{x,2}, Pro_{x,3}\}$$
$$D \leftarrow \{Pro_{x,p}, Con_{x,1}\}$$
$$E \leftarrow \{The_{x,t}\}$$
$$F \leftarrow \{The_{x,4}, Con_{x,4}, Pro_{x,4}\}$$

The language model for either of the two rhetorical strategies generates a set of arguments where each argument is composed of n cluster labels, e.g., (A, B, C, D, C) for $n = 5$ in Figure 3. This set is ranked by probability of the associated sequence. For example, assume that (A, B, C, D, C) is most probable. Then we decode all possible ADU sequences for topic-stance x from (A, B, C, D, C) to a set of candidate arguments:

$$\begin{aligned}(A, B, C, D, C) \rightarrow \\ &\{The_{x,1}, The_{x,2}, Con_{x,3}\} \\ &\times \{Con_{x,2}, Pro_{x,1}\} \\ &\times \{The_{x,3}, Con_{x,c}, Pro_{x,2}, Pro_{x,3}\} \\ &\times \{Pro_{x,p}, Con_{x,1}\} \\ &\times \{The_{x,3}, Con_{x,c}, Pro_{x,2}, Pro_{x,3}\}\end{aligned}$$

The output of the model is a set of candidate arguments, which becomes the input of the arrangement language model.

4.2 Arrangement Language Model

In the arrangement process, we aim to imitate the human behavior of arranging ADUs for a specific topic-stance following a rhetorical strategy (here, *logos* or *pathos*). Again, we approach this problem as a language modeling task. Each ADU role (thesis, pro, or con) is a word of the language model and each argument a sentence.

4.2.1 Training of the Model

As sketched in Figure 2, we first convert the human-generated arguments from a sequence of ADUs to a sequence of ADU roles. Then, we use these sequences to train a language model for each strategy.

4.2.2 Application of the Model

As shown in Figure 3, the arrangement language model takes as input the candidate arguments that we get from the selection language model and outputs a set of filtered candidate arguments.

The language model for a specific strategy generates a set of argument structures where each such structure is a sequence of n ADU roles, e.g., $(Thesis, Con, Pro, Pro, Pro)$ for $n = 5$ in Figure 3. This set is ranked by the probability of the sequences. For example, assume that the most frequent sequence is $(Thesis, Con, Pro, Pro, Pro)$.

Using the output from the selection language model, we filter out all candidate arguments that do not match $(Thesis, Con, Pro, Pro, Pro)$, ending up with the following filtered arguments:

$$\{The_{x,1}, The_{x,2}\} \times \{Con_{x,2}\} \times \{Pro_{x,2}, Pro_{x,3}\} \times \{Pro_{x,p}\} \times \{Pro_{x,2}, Pro_{x,3}\}$$

The output of the model is a filtered set of candidate arguments, which becomes the input of the phrasing regression model.

4.3 Phrasing Regression Model

The set of arguments resulting from the selection and arrangement language models are based on topic-independent features. The missing step is to entail the topical relationship between the ADUs in each generated argument. We approach this task with supervised regression. As indicated above, our model does not really *phrase* an argument. Rather, it aims to choose the best among the given set of candidates in terms of semantic coherence.

4.3.1 Training of the Model

For each argument, we opt for a feature representation that embeds the content properties of ADUs in order to capture their content relationship. Concretely, we represent each argument by calculating the semantic similarities of each adjacent bigram in a human-generated argument. We train a linear regression model where each instance represents the features of one argument. To this end, we set a score to be the sum of the probabilities of ADU bigrams occurring in one argument.

The phrasing model scores each of the filtered arguments given as output by the arrangement model. The argument with the highest score is the final generated argument.

4.3.2 Application of the Model

At this point, the phrasing model is provided by the filtered arguments from the arrangement model. For each filtered argument, we extract the bigram features (semantic similarities). Next, using the phrasing model, we predict the score of each sequence. The sequence with the highest score is the generated argument. In Figure 3, this is:

$$(The_{x,2}, Con_{x,2}, Pro_{x,2}, Pro_{x,p}, Pro_{x,3})$$

5 Experiments

In this section, we report the results of evaluating the introduced approach to argumentation synthesis based on the dataset described in Section 3.

Strategy	2-grams	3-grams
Logos-oriented	9,110.6	9,466.3
Pathos-oriented	7,939.5	10,279.6

Table 3: *Selection*. Perplexity of the 2-gram and 3-gram language models for each strategy, averaged over 10 leave-one-topic-out runs using Laplace smoothing.

5.1 Experimental Set-up

Our experiments are designed in leave-one-topic-out cross-validation setting: From the 10 topic-stance pairs in the dataset, we use nine for training and the last as the test fold, and we repeat this once for each possible fold. This way, no topic-specific knowledge can be used in the synthesis process.

For each given basic rhetorical strategy (*logos-oriented* and *pathos-oriented*), we train one model each for the selection, the arrangement, and the "phrasing" of argumentative discourse units (ADUs) on the nine training folds. The arguments synthesized by their combination are then evaluated against the human-generated arguments in the test folds. The evaluation covers all three models as well as the final generated argument for each strategy. We report the average accuracy across all ten folds for each of the models.

5.2 Training: Selection Language Model

In each training/test experiment for one of the two strategies, we first abstract all ADUs across all strategy-specific topic-stance pairs by extracting the LIWC, NRC, and MPQA features, as described in Section 4.1. Then, we cluster the given training set using standard k-means (Ostrovsky et al., 2012). After some initial experiments, we decide to set k to 6, because this best balanced the distribution of arguments over clusters, and showed clear strategy-specific differences.[3] Using the resulting clustering model, we predicted the type A–F of each ADU in the test set (the tenth topic).

Given the ADU types, we next converted the human-generated training and test arguments from a sequence of ADUs to a sequence of ADU types. After that, we trained one 2-gram and one 3-gram selection language.[4] In Table 3, we report the mean perplexity of the models for both strategies.

[3] A more thorough evaluation of k is left to future work.

[4] We did not consider 1-grams, because arguments are inherently relational, hence requiring at least two ADUs.

Strategy	2-grams	3-grams
Logos	54.5	33.5
Pathos	45.9	23.9

Table 4: *Arrangement.* Perplexity of the 2-gram and 3-gram language models for each strategy, averaged over 10 leave-one-topic-out runs using Laplace smoothing.

As shown, the 2-gram perplexity is lower than the 3-gram perplexity in both cases. We assume that the reason lies in the limited size of the dataset and the narrow setting: Only 117 sentences (ADUs) are given per strategy for training, with a vocabulary size of 6 (number of ADU types). Based on the results, we decided to use the 2-gram selection language model to generate candidate arguments.

5.3 Training: Arrangement Language Model

To train arrangement as described in Section 4.2, we took all arguments of the nine training topics in each experiment. We converted each argument from a sequence of ADUs to a sequence of ADU roles (thesis, pro, and con). After that, we trained a 2-gram and 3-gram language model for each strategy. Table 4 lists the mean perplexity values over the 10 folds.

Here, the perplexity is lower for 3-grams than for 2-grams, which can be expected to yield better performance. Therefore, we used the 3-gram language model to filter the set of candidate arguments.

5.4 Training: Phrasing Regression Model

For phrasing (in terms of choosing the best ADU sequence), we first extracted features from each candidate, as described in Section 4.3. Then, we calculated the semantic similarities between each pair of adjacent ADUs as follows:

1. We obtained a 300-dimensional word embedding for each word in an ADU using the pre-trained GloVe common-crawl model (Pennington et al., 2014).[5]
2. We averaged the embeddings of all words in an ADU, resulted in one vector representing the ADU.
3. For each adjacent pair of ADUs, we computed the cosine similarity of their vectors.

Figure 4 shows a histogram of the distribution of the cosine similarities of each adjacent pair of

[5]The used model can be found here: `http://nlp.stanford.edu/data/glove.42B.300d.zip`.

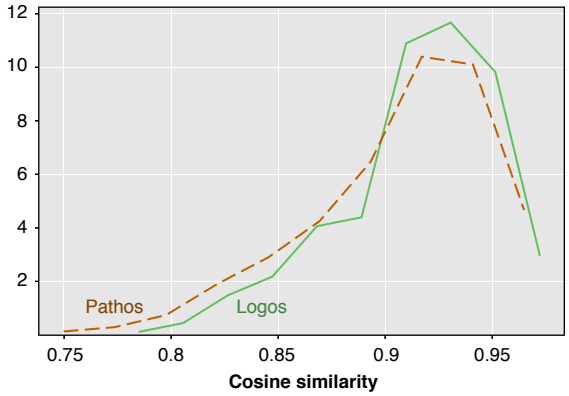

Figure 4: Histogram of the cosine similarity of the average word embeddings of adjacent pairs of ADUs in logos-oriented and in pathos-oriented arguments.

ADUs (i.e., each ADU 2-gram) in logos-oriented arguments and in pathos-oriented arguments. We observe a generally high similarity between neighboring ADUs for both strategies, with logos-oriented 2-grams being slightly more similar on average.

Given the ADU 2-grams, we train a linear regression model that predicts the sum of ADU 2-gram probabilities in each argument. In case of the logos strategy, the model has a mean squared error (MSE) of 0.05. In case of pathos the MSE is 0.03.

5.5 Results: Argumentation Synthesis

Up to this point, we trained all selection, arrangement, and phrasing models 10 times. Combining the three models for each strategy, we finally generated one argument per strategy for the topic-stance pair left out in each experiments. Hence, we ended up with 10 computationally synthesized arguments per strategy in total.

We evaluate each of these arguments by checking whether it matches any of the 13 human-generated ground-truth arguments given per topic-stance pair. The matching is quantified in terms of n-gram overlap with $n = \{1, \ldots, 5\}$.

For comparison, we consider a baseline that randomly generates arguments for each topic-stance pair as follows:

1. Select a random thesis unit from t_1 to t_4.
2. Select a random con unit from c_1 to c_4.
3. Select three random pro units from p_1 to p_{12}.
4. Randomly arrange the selected units.

Table 5 presents the accuracy of n-gram overlaps between each of the 13 human-generated arguments per topic-stance pair and the arguments computationally synthesized arguments by our model

Strategy	Approach	Sequential					Non-Sequential				
		1-gram	2-gram	3-gram	4-gram	5-gram	1-gram	2-gram	3-gram	4-gram	5-gram
Logos	Our model	**80.0%**	**15.0%**	0.0%	0.0%	0.0%	**80.0%**	**39.0%**	**9.0%**	0.0%	0.0%
	Baseline	76.0%	10.0%	**0.7%**	0.0%	0.0%	76.0%	3.1%	8.8%	**2.0%**	0.0%
Pathos	Our model	**88.0%**	**20.0%**	0.0%	0.0%	0.0%	**88.0%**	**48.0%**	**17.0%**	**4.0%**	0.0%
	Baseline	82.0%	11.5%	**0.7%**	0.0%	0.0%	82.0%	38.9%	10.7%	1.6%	0.0%

Table 5: Accuracy of n-gram overlaps between the human-generated arguments for each strategy and the arguments computationally synthesized by *our model* and the *baseline*. In the *sequential* case, the ordering is considered, in the *non-sequential* case, it is ignored. The better result in each experiment is marked bold, if any.

Strategy	ID	Argument Computationally Synthesized from Five Argumentative Discourse Units
Logos	t_4	**universities should not charge tuition fees in Germany.**
	c_3	to oblige non-academics to finance others' degrees through taxes is not just.
	p_9	this would impede or prevent access to those who are financially weaker.
	p_5	if a university lacks the funds, sponsors must be found.
	p_8	there are other instruments to motivate tighter discipline while studying.
Pathos	p_2	education must not be a question of money in a wealthy society such as Germany.
	c_1	one could argue that an increase in tuition fees would allow institutions to be better equipped.
	p_7	studying and taking higher degrees must remain a basic right for everyone.
	p_6	longer durations of studies are costly.
	t_2	**the universities in Germany should not under any circumstances charge tuition fees.**

Table 6: Comparison of two con arguments computationally synthesized with our model for the topic *Should all universities in Germany charge tuition fees?*, each being a sequence of five ADUs. A logos-oriented argument $(t_4, c_3, p_9, p_5, p_8)$ and a pathos-oriented argument $(p_2, c_1, p_7, p_6, t_2)$. The thesis of each argument is marked bold.

and by the baseline, with and without considering the ordering of ADUs. Our models outperform the baseline for 1-grams and 2-grams in all cases. For *sequential 3-grams*, however, it did not achieve any overlap with the human-generated arguments for either strategy. This may be explained by the fact that the employed selection and phrasing models are based on 2-grams only. For $n \geq 2$, the synthesis generally does not work well anymore. We believe that the small data size is a main cause behind this, although it may also point to the limitation of composing ADUs based on surface features. In the non-sequential case, though, our model performs comparably well for 3-grams, and it even manages to correctly synthesize some ADU 4-grams.

In Table 6, we exemplify the top-scored arguments for one topic-stance pair, synthesized by our approach for logos and for pathos respectively. They indicate that our model was able to learn strategy-specific differences.[6] In particular, the logos argument starts with the thesis (t_2), as argumentation guidelines suggest. It then reasons based on consequences and alternatives. Matching intuition, the pathos argument appeals more to emotion, reflected in phrases such as "wealthy society" and "under any circumstances". Particularly the thesis (t_4) has a more intense tonality than t_2, and putting it at the end creates additional emphasis.

6 Conclusion

This paper has presented a topic-independent computational approach to imitate the process of selecting, arranging, and phrasing argumentative discourse units (ADUs) — so to speak, to synthesize arguments. We have proposed to operationalize the necessary synthesis knowledge in the form of a combined language and regression model that predicts ADU sequences. So far, we have evaluated our approach on a small dataset only that contains 260 argumentative texts following either of two rhetorical strategies. For a controlled experiment setting based on this data, we have reported preliminary results of medium effectiveness regarding the imitation of human-generated arguments.

A big challenge for the future is to move from such a controlled setting to a real-world scenario, where arguments have to be formed for a freely-chosen topic from material that is mined from the web. Still, our topic-independent approach defines a first substantial step in this direction.

[6] Notice that the coherence of the arguments may be optimized by inserting discourse markers, such as a "but" before p_7 in the pathos argument. As stated above, however, this is beyond the scope of the paper at hand.

References

Aristotle. 2007. *On Rhetoric: A Theory of Civic Discourse* (George A. Kennedy, Translator). Clarendon Aristotle series. Oxford University Press.

Yonatan Bilu and Noam Slonim. 2016. Claim synthesis via predicate recycling. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 525–530. Association for Computational Linguistics.

Giuseppe Carenini and Johanna D. Moore. 2006. Generating and evaluating evaluative arguments. *Artificial Intelligence*, 170(11):925–952.

Oana Cocarascu and Francesca Toni. 2017. Identifying attack and support argumentative relations using deep learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1374–1379. Association for Computational Linguistics.

Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science*, 15(10):687–693.

Charlie Egan, Advaith Siddharthan, and Adam Wyner. 2016. Summarising the points made in online political debates. In *Proceedings of the Third Workshop on Argument Mining (ArgMining2016)*, pages 134–143, Berlin, Germany. Association for Computational Linguistics.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics, Vol. 3*, pages 41–58. Academic Press, New York.

Ivan Habernal and Iryna Gurevych. 2016. Which argument is more convincing? Analyzing and predicting convincingness of web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1589–1599. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Maxwell Forbes, Antoine Bosselut, David Golub, and Yejin Choi. 2018. Learning to write with cooperative discriminators. Technical Report 1805.06087, arXiv.

Xinyu Hua and Lu Wang. 2018. Neural argument generation augmented with externally retrieved evidence. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 219–230. Association for Computational Linguistics.

Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2014. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143.

Dieu Thu Le, Cam-Tu Nguyen, and Kim Anh Nguyen. 2018. Dave the debater: a retrieval-based and generative argumentative dialogue agent. In *Proceedings of the 5th Workshop on Argument Mining*, pages 121–130, Brussels, Belgium. Association for Computational Linguistics.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin*, 29(5):665–675.

Rafail Ostrovsky, Yuval Rabani, Leonard J Schulman, and Chaitanya Swamy. 2012. The effectiveness of lloyd-type methods for the k-means problem. *Journal of the ACM (JACM)*, 59(6):28.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: 1st European Conference on Argumentation (ECA 16)*. College Publications.

James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The Development and Psychometric Properties of LIWC2015.

James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one*, 9(12):e115844.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. In *Theories of emotion*, pages 3–33. Elsevier.

Paul Reisert, Naoya Inoue, Naoaki Okazaki, and Kentaro Inui. 2015. A computational approach for generating Toulmin model argumentation. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 45–55. Association for Computational Linguistics.

Misa Sato, Kohsuke Yanai, Toshinori Miyoshi, Toshihiko Yanase, Makoto Iwayama, Qinghua Sun, and Yoshiki Niwa. 2015. End-to-end argument generation system in debating. In *Proc. ACL-IJCNLP 2015 System Demonstrations*.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2007. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6.

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

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

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

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of COLING 2018, the 27th International Conference on Computational Linguistics*. The COLING 2018 Organizing Committee. To appear.

Lu Wang and Wang Ling. 2016. Neural network-based abstract generation for opinions and arguments. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 47–57. Association for Computational Linguistics.

Toshihiko Yanase, Toshinori Miyoshi, Kohsuke Yanai, Misa Sato, Makoto Iwayama, Yoshiki Niwa, Paul Reisert, and Kentaro Inui. 2015. Learning sentence ordering for opinion generation of debate. In *Proceedings of the 2nd Workshop on Argumentation Mining*, pages 94–103. Association for Computational Linguistics.

Ingrid Zukerman, Richard McConachy, and Kevin B. Korb. 2000. Using argumentation strategies in automated argument generation. In *First International Conference on Natural Language Generation (INLG 00)*, pages 55–62.