Length, Interchangeability, and External Knowledge: Observations from Predicting Argument Convincingness

Peter Potash, Robin Bhattacharya, Anna Rumshisky
Department of Computer Science
University of Massachusetts Lowell
{ppotash,rbhattac,arum}@cs.uml.edu

Abstract

In this work, we provide insight into three key aspects related to predicting argument convincingness. First, we explicitly display the power that text length possesses for predicting convincingness in an unsupervised setting. Second, we show that a bag-of-words embedding model posts state-of-the-art on a dataset of arguments annotated for convincingness, outperforming an SVM with numerous hand-crafted features as well as recurrent neural network models that attempt to capture semantic composition. Finally, we assess the feasibility of integrating external knowledge when predicting convincingness, as arguments are often more convincing when they contain abundant information and facts. We finish by analyzing the correlations between the various models we propose.

1 Introduction

Predicting argument convincingness has mostly been studied in relation to the overall quality of a persuasive essay (Attali and Burstein, 2004; Landauer, 2003; Shermis et al., 2010), with a recent focus specifically on predicting argument strength (Persing and Ng, 2015; Wachsmuth et al., 2016). Zhang et al. (2016) have also attempted to predict argument convincingness, in the form of predicting debate winners. Unfortunately, these are very rare argumentative formats that are seldom encountered in everyday life. In practice, at least at the moment, we tend to digest a large quantity of our information from social media and engage in a tremendous amount of interpersonal communication using it. Since, in social media, communications are roughly a single paragraph, analyzing arguments in a persuasive essay or oxford-style debate is not applicable to our primary means of community engagement. Presenting an entire convincing argument within a single paragraph can be an invaluable skill in the modern world. This paper seeks to improve upon previous methodology for predicting argument convincingness.

Prompt: Is it better to have a lousy father or to be fatherless? **Stance:** It is better to have a lousy father.

Argument 1	Argument 2
It is better to have a lousy father because researchers at the McGill University have warned that growing up without a father can permanently change the structure of a child's brain and make him/her more aggressive and angry.	Having a lousy father is better because when a child does not have a father, it causes him/her to look for a father figure. During such searches, a child may end up getting sexual harassed or being emotionally exploited to various degrees.

Table 1: Example of an argument pair where Argument 1 is more convincing.

Habernal and Gurevych (2016b) have recently released a dataset of short, single-paragraph arguments annotated for convincingness, which we will refer to as UKPConvArg. For 16 issues, arguments with the *same* stance are compared with each other to determine, given a pair of arguments, which one is more convincing. Table 1 provides an example of an argument pair with arguments from the prompt 'Is it better to have a lousy father or to be fatherless'; and the stance: 'It is better to have a lousy father'. In this pair Argument 1 is chosen to be more convincing. Other such issues include: 'Does India have the potential to

Proceedings of the The 8th International Joint Conference on Natural Language Processing, pages 342–351,
Taipei, Taiwan, November 27 – December 1, 2017 ©2017 AFNLP

lead the world?', 'Which web browser is better, Internet Explorer or Mozilla Firefox?', and 'Should physical education be mandatory in schools'. In follow-up work, Habernal and Gurevych (2016a) examined the *reasoning* behind the annotations in their original corpus. That is, why arguments were selected as more convincing. Overwhelmingly, the reasons could be expressed by the following statement "Argument X has more details, information, facts or examples / more reasons / better reasoning / goes deeper / is more specific". Although Habernal and Gurevych (2016b) experimented with two promising models, the models were not intended to directly take into account the reasons why an argument could be more convincing, as expressed in the previous quotation. The primary task of the dataset is, given two arguments with the same stance toward a topic, determine which argument is more convincing – this corresponds to outputting a binary label. Most of our experiments focus on this task, as it was the annotation directive for annotating convincingness in Habernal and Gurevych (2016b). From the pairwise annotation, they also derived convincingness scores for individual arguments, which is posed as a regression task. We evaluate on this task in Section 3.1.

In our work, we improve upon the initial experiments of Habernal and Gurevych in 3 ways: (1) we offer heuristic-based methods that requiring no training or fitting of a model to data; (2) we explore modifications of the initial 'deep' model used by Habernal and Gurevych (2016a), which was a Bidirectional Long Short-Term Memory (BLSTM) network; (3) we test the feasibility of offering factually relevant knowledge in the form of Wikipedia articles related to the argument topics.

In terms of heuristics, we examine the effectiveness of Metric Entropy (ME) of text to predict convincingness, which is inspired by the notion that written English language is well-formed, as opposed to random. Specifically, high ME corresponds to high randomness. The second heuristic uses a similarity to Wikipedia articles, with the hypothesis that the Wikipedia articles can act as a factual support reference for the arguments. We also hypothesize that Wikipedia articles have the potential to grade the quality of the writing in the arguments, on the assumption that arguments that better match the writing in Wikipedia articles are more likely to exhibit the qualities that make an argument convincing. For all methods that use the presence of Wikipedia articles, we use several variations of a corpus to determine how well the methods leverage topic-specific articles, as opposed to randomly selected articles.

In terms of supervised techniques, we first follow previous approaches to classifying paired data that create separate learned representations of elements in a pair that are then concatenated for the final predictive model (Bowman et al., 2015; Mueller and Thyagarajan, 2016; Potash et al., 2016b). Specifically, we experiment with creating separate representations using either a BLSTM or summing individual token embeddings. We then propose modifications of the supervised models to leverage external data. The models grow with increasing complexity, approaching a form of Memory Network (Sukhbaatar et al., 2015) that computes a weighted sum of representations of Wikipedia articles.

Our experimental results reveal several important insights into how to approach predicting convincingness. We summarize our findings as follows: 1) Unsupervised text length is an extremely competitive baseline that performs on par with highly-engineered classifiers and deep learning models; 2) The current state-of-the-art approach treats tokens as interchangeable, bypassing the need to model compositionality; 3) Wikipedia articles can provide meaningful external knowledge, though, naive models have trouble dealing with the noise in a large corpus of documents, whereas a model that attends to the Wikipedia corpus is better equipped to handle the noise.

2 Related Work

Habernal and Gurevych (2016b) present two methods in their dataset paper: (1) an SVM with numerous hand-crafted features; (2) a BLSTM that only uses word embeddings as input. Aside from the original corpus authors, only one other work has tested on the UKPConvArg dataset. Chalaguine and Schulz (2017) use a feature-selection method to determine the raw feature representation that serves as input into a feed-forward neural network. The authors conduct a thorough ablation study of the performance of individual feature types. The authors' best model records an accuracy of .766, compared to .781 and .757 of Habernal and Gurevych's SVM and BLSTM, respectively. Although the authors make an effort to determine the influence of individual feature type, their work continues to use supervised methods, which obscures the pure predictive power of individual features/metrics.

There are few datasets annotated for the convincingness of arguments. Zhang et al. (2016) published a dataset of debate transcripts, annotated with audience polling that occurs before and after the debate. In terms of argumentation, the key distinction between this dataset and that of Habernal and Gurevych (2016b) is that in the debate dataset, the debate teams have *opposing* stances on a topic, whereas Habernal and Gurevych's dataset has labels for arguments with the same stance towards a topic. Persing and Ng (2015) constructed a corpus of persuasive essays annotated for the essays' argument strength, which is slightly different to other annotated persuasive essay corpora, which have more of a focus on overall writing quality.

NLP datasets involving the processing of text pairs have become more prevalent. Examples include predicting textual entailment (Marelli et al., 2014; Bowman et al., 2015), predicting semantic relatedness/similarity (Marelli et al., 2014; Agirre et al., 2016), and predicting humor (Potash et al., 2016b; Shahaf et al., 2015). These tasks present interesting challenges from a modeling perspective, as methods must allow for semantic comparison between the texts.

Although relatively rare in the argument mining community, leveraging external knowledge sources is ubiquitous for the task of question-answering (Kolomiyets and Moens, 2011), using information retrieval techniques to mine the available documents for answers. Work such as Berant et al. (2013) forms a knowledge base from external documents, and maps queries to knowledge-base entries. Weston et al. (2014) have proposed a neural network-based approach for large-scale question-answering. In the argument mining community, Rinott et al. (2015) created a dataset for predicting potential support clauses for argumentative topics, while Braunstain et al. (2016) rank Wikipedia sentences for supporting answers made by online user answers. Conversely, Wachsmuth et al. (2017) approach the problem of measuring relevance amongst arguments themselves, proposing a methodology based on PageRank (Page et al., 1999).

3 Heuristic Methods

As Habernal and Gurevych (2016b) note in their paper, comparing the SVM and BLSTM systems, it is desirable for methodologies to require minimal preprocessing of text. Along those lines, methods that use heuristics can circumvent the need for supervised training. We refer to the models in this section as heuristic models, as opposed to unsupervised models, because they do not fit themselves to data – they merely compare various metric values to determine convincingness. We experiment with two types of heuristics: ME and Wikipedia similarity. The motivation of these heuristics is as follows: Metric Entropy has previously been applied to the task of predicting tweet deletion (Potash et al., 2016a), with the idea that tweets with high ME are likely to be spam. Moreover, ME conveys how well-formed the language is in a piece of text, since higher ME means a higher randomness in the language. Conversely, Wikipedia similarity attempts to use external knowledge to measure the factual validity of the arguments, but also potentially measuring the writing quality of the arguments.

3.1 Metric Entropy

The Shannon Entropy of a text T containing a set of characters C is defined as:

$$H(T) = -\sum_{c \in C} P(c) \log_2 P(c) \tag{1}$$

where

$$P(c) = \frac{freq(c)}{len(T)} \tag{2}$$

and $freq(c)$ is the number of times c appears in T. Consequently, ME is the Shannon entropy divided by the text length, $len(T)$. Since ME produces a continuous output, it is sensible to evaluate it using the regression task from Habernal and Gurevych (2016b). Because ME is a combination of Shannon Entropy and text length, we also evaluate their effectiveness separately as well. We admit, however, that our initial experiments only included ME and Shannon Entropy, but given the vastly different performance of the two metrics, we decided to test length on its own as well.

3.2 Wikipedia Similarity

Suppose we have vector representations of an argument a and a Wikipedia article w. The similarity score, $sim(a, w)$ is simply the dot product of the two representations, aw^{T}. Therefore, given a corpus of Wikipedia articles W, we define the Wikipedia Similarity Score, WSS of an argument a as:

$$WSS(a) = \sum_{w \in W} aw^{\mathsf{T}} \tag{3}$$

For pairwise prediction, we predict the argument with the higher score as the more convincing argument.

We consider two possible representations for texts: 1) term-frequency (TF) count, and 2) Summing the embeddings of all the tokens in the text. For the TF representation, we use the `CountVectorizer` class from Scikit-learn (Pedregosa et al., 2011) to process the text and create the appropriate representation. For the embedding representation, we use GloVe (Pennington et al., 2014) 300 dimensions learned from the Common Crawl corpus with 840 billion tokens.

Our Wikipedia data is from the May 20th, 2017 dump[1]. We clean the raw Wikipedia data using *gensim* (Řehůřek and Sojka, 2010). We experiment with three different Wikipedia corpora. The first corpus has a set of 30 hand-picked Wikipedia articles, chosen to be of the same subject matter of the various topics in the argument convincingness corpora. We refer to this corpus as Wiki hand-picked (hp). The second corpus contains 38k random Wikipedia articles, chosen to be approximately the length of the hand-picked articles. The motivation behind the second corpus is to determine how valuable the topic-specific information is for assessing the validity of the arguments. The second corpus also simulates a situation where a model accesses an arbitrary knowledge base, as opposed to one that is hand-selected. We refer this corpus as Wiki random (ran). The third corpus combines the first two corpora, with the goal of determining how well the heuristic method can deal with the potential 'noise' of randomly chosen Wikipedia articles. We refer to this corpus as Wiki hp+ran.

4 Supervised Methods

Habernal and Gurevych (2016b) propose two supervised experiments for predicting argument convincingness: an SVM with numerous hand-crafted features, and a BLSTM that only uses word embeddings as input. While our heuristic methods show promising results, they do not yet achieve state-of-the-art on the argument convincingness dataset. In this section, we motivate our supervised experiments with a combination of results from Section 3.2 and Habernal and Gurevych. All models have the same cost function, which is the binary cross-entropy of the training set, based on the sigmoid activation of a continuous value from a 1-dimensional dense layer.

Model	Pearson	Spearman
SVM	0.351	0.402
BLSTM	0.270	0.354
SE	0.097	0.227
LEN	0.353	**0.425**
ME	**0.358**	0.422

Table 2: Results of the Metric Entropy experiments on the regression task. SE = Shannon Entropy, LEN = 1/text length, ME = Metric Entropy.

4.1 Siamese BLSTM

The BSLTM model that Habernal and Gurevych (2016b) propose concatenates the text of the argument pairs, separated by a special delimiter. This single sequence is then run over by forward and backward LSTMs to produce the BLSTM embedding that is then used for logistic regression. We propose to model each argument in the argument pair separately, creating a representation for each argument pair that is then concatenated together for logistic regression output. The term 'Siamese' refers to the fact that the representations are created separately (we adopt the terminology from Mueller and Thyagarajan (2016)). Each argument goes through a BLSTM to produce its individual representation, using GloVe vectors as input to the BLSTM.

4.2 Siamese BOW Embedding

While a BLSTM model is very logical for most language tasks, given its sequential nature, work such as Joulin et al. (2016) shows that simply summing individual token embeddings can be extremely competitive for the task of text classification. Furthermore, in the current climate of increasingly complex deep learning models, it is important to continue to compare to simpler models. For this method, we represent an argument in an argument pair as the sum of its tokens' embeddings. Given the TF representation of a set of texts

[1] `https://dumps.wikimedia.org/enwiki/20170520/`

	Topic (Wiki corpus)	WS-TF hp	WS-TF ran	WS-TF hp+ran	WS-E hp	WS-E ran	WS-E hp+ran
Should physical edu.	No	0.792	**0.825***	**0.825***	0.783	0.783	0.783
be mandatory?	Yes	0.711	0.736	0.736	0.778	**0.784**	**0.784**
Ban Plastic	No	0.826	0.840	0.840	**0.851**	0.847	0.847
Water Bottles?	Yes	**0.905**	0.838	0.838	0.833	0.835	0.835
Christianity or Atheism	Atheism	0.713	0.777	0.777	**0.801**	**0.801**	**0.801**
	Christianity	**0.736**	0.716	0.716	0.697	0.705	0.705
Evolution vs. Creation	Creation	0.772	0.817	0.817	0.848	**0.846**	**0.846**
	Evolution	**0.678**	0.634	0.634	0.596	0.603	0.603
Firefox vs. Internet Exp	IE	0.785	0.668	0.668	**0.796**	0.792	0.792
	Firefox	0.774	0.768	0.768	**0.797**	0.793	0.793
Gay marriage -	Right	**0.802**	0.703	0.703	0.762	0.765	0.765
right or wrong?	Wrong	0.774	**0.841**	**0.841**	0.828	0.830	0.830
Should parents	No	0.766	0.796	0.796	**0.829**	0.821	0.821
use spanking?	Yes	0.648	0.672	0.672	0.808	**0.814***	**0.814***
If your spouse	No	**0.689**	0.601	0.604	0.683	0.677	0.677
committed murder [...]	Yes	0.682	0.673	0.673	0.795	**0.798***	**0.798***
India has the potential	No	0.784	0.776	0.776	**0.792**	**0.792**	**0.792**
to lead the world	Yes	**0.749**	0.714	0.714	0.685	0.687	0.687
Lousy father	Fatherless	0.707	0.711	0.711	**0.760**	**0.760**	**0.760**
or fatherless?	Lousy father	**0.675**	0.663	0.663	0.666	0.663	0.663
Is porn wrong?	No	**0.761**	0.703	0.703	0.746	0.749	0.749
	Yes	0.789	**0.838**	**0.838**	0.820	0.829	0.829
Is the school uniform	Bad	**0.706**	0.702	0.702	0.699	0.695	0.695
a good or bad idea?	Good	0.722	0.711	0.711	0.825	**0.827**	**0.827**
Pro choice vs. Pro life	Choice	0.681	0.678	0.678	**0.728**	**0.728**	**0.728**
	Life	0.807	0.726	0.726	0.807	**0.809**	**0.809**
TV is better than books	No	**0.747**	0.736	0.736	0.721	0.721	0.721
	Yes	0.774	0.770	0.770	**0.789**	0.780	0.780
Personal pursuit or	Common	0.728	**0.768**	**0.768**	0.720	0.718	0.718
common good?	Personal	**0.653**	0.610	0.610	0.641	0.650	0.650
Farquhar as the	No	**0.743**	0.682	0.682	0.714	0.723	0.723
founder of Singapore	Yes	0.660	0.702	0.702	**0.828**	0.820	0.820
AVERAGE		0.742	0.731	0.731	0.763	**0.764**	**0.764**

Table 3: Results of Wikipedia similarity experiments, using either a term-frequency representation (TF) or a sum of word embeddings (E). We experiment with three types of Wikipedia corpora: 30 hand-picked articles chosen to been highly relevant to the argument topics (hp); roughly 38k randomly chosen articles (ran); a combination of the first two corpora (hp+ran).

T in matrix format A and a corresponding embedding matrix E, the BOW Embedding, $BOWE$, representation is equivalent to:

$$BOWE(T) = AE \tag{4}$$

For our application, our input will have two matrices, T_l and T_r, representing the left and right arguments in the pair. Once the individual representations are created, as with the Siamese BLSTM, we concatenate them together as the input for logistic regression. Lastly, instead of continuing to train the initialized embedding matrix E, we fix E, calling it E_{fixed}, and pass it through a fully-connected layer, W_{emb},

$$E_{learned} = E_{fixed} W_{emb} \tag{5}$$

Thus, $E_{learned}$ replaces E in Equation 4. Because we are summing embedding vectors to create the representation, the values of representations' dimensions could become large, causing a dramatically increased loss. While such methods as gradient clipping and gradient normalization could be used, we found it simple enough to divide the representation by 100.

4.3 Supervised Wikipedia Similarity

We now begin to modify the methodology described in Section 3.2 to add an increasing amount of complexity to better integrate the Wikipedia articles. The first model we propose uses the representations from Equation 4 to represent the arguments and Wikipedia articles, however, it is computed slightly differently for the arguments and wikipedia articles. While the argument representations use $E_{learned}$, the Wikipedia articles use E_{fixed}, and then the result of $BOWE(T)$ passes through a fully-connected layer, W_{wiki}. Just as we artificially normalized the argument representations, we divide the Wikipedia representations by 10,000, due to their greatly increased length compared to the argument text. Once we have the individual representations, we compute a similarity score as done in Equation 3. The one difference, though, is that we apply tanh to the result of the dot product to keep the summation in a manageable range, which aids training. The resulting similarity scores, one for each argument in the pair, become the features for a 2-dimensional logistic regression model. This model does not use dropout at the fully-connected layer.

4.4 Memory Network with Wikipedia

The model from Section 4.3 gives equal importance to the similarity scores from all Wikipedia articles. However, it's more intuitive for more relevant articles to have more importance. Therefore, we construct a model similar to the end-to-end Memory Network from Sukhbaatar et al. (2015). We create a weight for each score (also interpretable as a probability score P^j) for each Wikipedia article, w_i, and argument, a_j, as[2]:

$$P^j(w_i) = \text{softmax}(a_j w_i^\mathsf{T}) \tag{6}$$

which is used to create a weighted sum of the Wikipedia articles, s_j, for each argument j:

$$s_j = \sum_i^{|W|} P^j(w_i) w_i \tag{7}$$

[2]We note that we also experimented with an attention mechanism more akin that of Bahdanau et al. (2014), which uses a latent vector v to dot product with the sum $a_j + w_i$. However, this yielded the same results as the currently presented model.

We create the final representation, o_j, for argument j as follows:

$$o_j = a_j + s_j \tag{8}$$

which is the representation that is the input to the logistic regression layer (one for each argument in the pair).

5 Results

In each table that presents results, bold face indicates that a given system performed highest on a given topic within that table. An asterisk indicates that a given system performed highest on a given topic across *all* tables.

5.1 Heuristic Methods

Results of our ME experiments are shown in Table 2. We present the results on the regression task. The results of the Wikipedia similarity experiments are shown in Table 3.

5.2 Supervised Methods

Results of our supervised experiments are shown in Tables 4 and 5. We present the results of the Siamese BLSTM (SBLSTM), Siamese BOW Embeddings (SBOWE), Supervised Wikipedia similarity (SWS), and Memory Network with Wikipedia (MNW). Each model that uses Wikipedia articles is run with Wiki hp, Wiki ran, and Wiki hp+ran, as described in Section 3.2. All reported results are the average of three different runs. We report the accuracy on each topic, as well as the macro average across all topics. We compare our results with the SVM and BLSTM models from Habernal and Gurevych (2016b) in Table 4.

All models have dropout (Srivastava et al., 2014) of 0.5 at the dense layer (except for the model described in Section 4.3) and use a batch size of 32, as done by Habernal and Gurevych (2016b) in their BLSTM model. All models are implemented in TensorFLow (Abadi et al., 2016) and train for four epochs. The entire dataset has 11,650 argument pairs across all 32 topics. Since one topic is held-out for testing at a time, there is on average an 11,286/364 train/test split.

6 Discussion

6.1 Heuristic Methods

First, it is rather remarkable that text length alone, as a stand-alone metric, is able to record state-of-

	Topic	SVM	BLSTM	SBOWE	SBLSTM
Should physical edu. be mandatory?	No	0.79	**0.8**	0.788	0.750
	Yes	0.79	0.78	**0.879***	0.801
Ban Plastic Water Bottles?	No	0.85	0.76	**0.861**	0.760
	Yes	0.9	0.83	**0.910***	0.798
Christianity or Atheism	Atheism	0.81	0.8	**0.832**	0.771
	Christianity	0.68	0.75	0.747	**0.770**
Evolution vs. Creation	Creation	0.84	0.88	**0.893**	0.809
	Evolution	0.66	0.77	**0.809**	0.796
Firefox vs. Internet Explorer	IE	0.84	0.81	**0.931***	0.774
	Firefox	0.82	0.78	**0.893***	0.814
Gay marriage - right or wrong?	Right	0.76	0.74	**0.797**	0.735
	Wrong	0.82	0.87	**0.902**	0.799
Should parents use spanking?	No	0.84	0.78	**0.861***	0.745
	Yes	**0.79**	0.68	0.765	0.648
If your spouse committed murder [...]	No	0.71	0.64	**0.757**	0.633
	Yes	0.79	0.72	**0.795**	0.720
India has the potential to lead the world	No	0.82	0.77	**0.843**	0.747
	Yes	0.69	0.79	**0.874**	0.817
Is it better to have a lousy father or to be fatherless?	Fatherless	**0.77**	0.69	0.765	0.638
	Lousy father	0.67	0.6	**0.731**	0.584
Is porn wrong?	No	0.82	0.79	**0.835**	0.790
	Yes	0.85	0.85	**0.886**	0.785
Is the school uniform a good or bad idea?	Bad	0.75	0.78	**0.839**	0.829
	Good	**0.83***	0.74	0.795	0.681
Pro choice vs. Pro life	Choice	0.71	0.68	**0.741**	0.730
	Life	0.79	0.8	**0.862**	0.709
TV is better than books	No	0.78	0.73	**0.857**	0.740
	Yes	0.78	0.75	**0.860***	0.799
Personal pursuit or common good?	Common	0.72	**0.78***	0.773	0.712
	Personal	0.67	0.68	**0.696***	0.661
Farquhar as the founder of Singapore	No	0.79	0.63	**0.824**	0.736
	Yes	**0.85***	0.76	0.806	0.651
AVERAGE		0.781	0.757	**0.825***	0.742

Table 4: Results of supervised models that do not use Wikipedia. SVM and BLSTM results are reported from Habernal and Gurevych (2016b).

the-art results on the regression task. Although Chalaguine and Schulz (2017) directly showed the power of text length in a supervised setting, our results show an even simpler method for producing predictions on par with the previous state-of-the-art. There is intuitive reasoning for this result, since, as mentioned in Section 1, arguments are predominantly more convincing when they provide *more*; more facts, more information, more depth, etc. When evaluated on the pairwise binary prediction task, Metric Entropy and text length record 77.2% and 77.3% accuracy, respectively.

Reviewing the Wikipedia similarity results, it is evident that the BOW embedding representation does offer greater predictive power when compared to the term-frequency representation. This unsupervised method even outperforms the supervised methods BLSTM and SBLSTM. Furthermore, compared to other methods that use Wikipedia articles, this method is more insensitive to the content of the articles, as it actually shows a very slight improvement when the hand-picked articles are not present, which is the opposite of all the other Wikipedia-based methods.

	Topic (Wiki corpus)	SWS hp	SWS ran	SWS hp+ran	MNW hp	MNW ran	MNW hp+ran
Should physical edu.	No	0.797	**0.819**	0.794	0.802	0.792	0.775
be mandatory?	Yes	0.880	0.846	0.851	**0.877**	0.878	0.868
Ban Plastic Water Bottles?	No	0.821	0.844	0.811	0.829	0.852	**0.862***
	Yes	0.894	0.893	0.901	0.899	**0.906**	**0.906**
Christianity or Atheism	Atheism	0.822	0.804	0.821	0.800	0.838	**0.844***
	Christianity	**0.777***	0.727	0.747	0.765	0.756	0.743
Evolution vs. Creation	Creation	**0.904***	0.834	0.872	0.883	0.886	0.892
	Evolution	0.813	0.802	0.783	**0.832***	0.795	0.800
Firefox vs. Internet Exp	IE	0.901	0.888	0.889	**0.925**	0.903	0.906
	Firefox	0.876	0.884	0.876	**0.880**	0.840	0.856
Gay marriage -	Right	**0.815***	0.771	0.762	0.814	0.787	0.786
right or wrong?	Wrong	0.903	0.889	0.885	**0.908***	0.891	0.901
Should parents	No	0.813	0.816	0.840	0.835	**0.857**	0.853
use spanking?	Yes	0.773	0.748	0.735	0.773	0.782	**0.786**
If your spouse	No	**0.761***	0.733	0.728	0.760	0.732	0.748
committed murder [...]	Yes	0.779	0.780	0.761	0.789	**0.798***	0.750
India has the potential	No	0.833	0.824	0.820	0.842	0.847	**0.848***
to lead the world	Yes	0.861	0.869	**0.880***	0.867	0.870	0.856
Lousy father	Fatherless	**0.780***	0.760	0.751	**0.780**	0.746	0.753
or fatherless?	Lousy father	0.704	0.678	0.711	0.725	0.724	**0.732***
Is porn wrong?	No	0.791	0.836	0.834	0.824	**0.839***	0.816
	Yes	0.883	0.861	0.879	**0.892***	**0.892***	**0.892***
Is the school uniform	Bad	0.840	0.837	0.831	**0.851***	0.815	0.843
a good or bad idea?	Good	0.771	0.752	0.762	0.771	**0.792**	**0.792**
Pro choice vs. Pro life	Choice	**0.746***	0.721	0.723	0.733	0.716	0.722
	Life	0.856	0.834	**0.866***	0.852	0.854	0.850
TV is better than books	No	0.856	0.861	0.834	**0.864***	0.846	0.846
	Yes	0.837	0.849	**0.853**	0.835	0.847	0.849
Personal pursuit	Common	0.760	0.727	0.714	0.763	**0.766**	0.719
or common good?	Personal	0.682	0.669	0.686	0.680	0.687	**0.691**
Farquhar as the	No	0.794	0.783	0.799	0.820	**0.831***	0.823
founder of Singapore	Yes	0.820	0.776	0.794	0.806	0.814	**0.821**
AVERAGE		0.817	0.804	0.806	**0.821**	0.818	0.817

Table 5: We experiment with three types of Wikipedia corpora: 30 hand-picked articles chosen to been highly relevant to the argument topics (hp); roughly 38k randomly chosen articles (ran); a combining the first two corpora (hp+ran).

6.2 Supervised Methods

The first result to note is that the BOW Embedding model posts a new state-of-the-art on the dataset. This shows that the current best approach to predicting argument convincingness treats word order as interchangeable. Although, it is reasonable to surmise that facts and information are dependent on local compositionality, current methods to model such linguistic phenomena under-perform.

When comparing supervised models that integrate Wikipedia articles, we see that the MNW model is better equipped to handle the noise from a large corpus of documents, when compared to the SWS results, which shows roughly a 1% drop in accuracy when the *ran* corpus is added to the *hp* corpus.

6.3 Model Correlations

Table 6 presents correlations between various models when comparing the accuracies of the individual topics. First, text length has a .96 correlation with the SVM model. This means that

	BLSTM	LEN	MNW	SBLSTM	SBOWE	SVM	SWS	WS-E	WS-TF
BLSTM	1.000	0.508	0.739	0.733	0.740	0.534	**0.785**	0.519	0.585
LEN	0.508	1.000	0.574	0.202	0.647	**0.964**	0.585	0.915	0.530
MNW	0.739	0.574	1.000	0.726	0.969	0.608	**0.975**	0.465	0.651
SBLSTM	**0.733**	0.202	0.726	1.000	0.722	0.277	0.723	0.173	0.528
SBOWE	0.740	0.647	**0.969**	0.722	1.000	0.681	0.948	0.552	0.683
SVM	0.534	**0.964**	0.608	0.277	0.681	1.000	0.615	0.904	0.584
SWS	0.785	0.585	**0.975**	0.723	0.948	0.615	1.000	0.528	0.630
WS-E	0.519	**0.915**	0.465	0.173	0.552	0.904	0.528	1.000	0.505
WS-TF	0.585	0.530	0.651	0.528	**0.683**	0.584	0.630	0.505	1.000

Table 6: Correlations between systems. Bold indicates the highest correlation for a given row.

the main predictive power of the SVM model can be distilled into using the text length to predict argument convincingness. What is perhaps more surprising is how high LEN correlates with WS-E. This could potentially be explained by the fact that articles with more words will sum together more embeddings, resulting in vectors with larger norms, which create larger dot-products when taken with the argument representations. However, the same argument can be made for the TF representation, so a more valid reason remains to be seen (note though that SBOWE and WS-TF have a low correlation with LEN). Secondly, we see that all models based on BOW embeddings have a very high correlation with each other, which is an intuitive finding.

7 Conclusion

In this work we have shown three key insights into the task of predicting argument convincingness: 1) Heuristic text length is an extremely competitive baseline that performs on par with highly-engineered classifiers and deep learning models; 2) The current state-of-the-art approach treats tokens as interchangeable, bypassing the need to model compositionality; 3) Wikipedia articles can provide meaningful external knowledge, though, naive models have trouble dealing with the noise in a large corpus of document, whereas a model that attends to the Wikipedia corpus is better equipped to handle the noise. Future work can focus on models that better handle compositionality, as well as integration of external knowledge, with an aim to surpass our new state-of-the-art on the corpus. One simple way to potentially enhance our MNW model is to perform multiple hops, a technique shown to greatly increase performance when using Memory Networks for other applications (Sukhbaatar et al., 2015).

Acknowledgments

This work was supported in part by the U.S. Army Research Office under Grant No. W911NF-16-1-0174.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Eneko Agirre, Carmen Banea, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval@ NAACL-HLT*, pages 497–511.

Yigal Attali and Jill Burstein. 2004. Automated essay scoring with e-rater® v. 2.0. *ETS Research Report Series*, 2004(2).

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, volume 2, page 6.

Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

Liora Braunstain, Oren Kurland, David Carmel, Idan Szpektor, and Anna Shtok. 2016. Supporting human answers for advice-seeking questions in cqa sites. In *European Conference on Information Retrieval*, pages 129–141. Springer, Cham.

Lisa Andreevna Chalaguine and Claudia Schulz. 2017. Assessing convincingness of arguments in online debates with limited number of features. In *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *EMNLP*, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *ACL (1)*.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Oleksandr Kolomiyets and Marie-Francine Moens. 2011. A survey on question answering technology from an information retrieval perspective. *Information Sciences*, 181(24):5412–5434.

Thomas K Landauer. 2003. Automated scoring and annotation of essays with the intelligent essay assessor. *Automated essay scoring: A crossdisciplinary perspective*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *AAAI*, pages 2786–2792.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays.

Peter Potash, Eric Bell, and Joshua Harrison. 2016a. Using topic modeling and text embeddings to predict deleted tweets. *Proceedings of AAAI WIT-EC*.

Peter Potash, Alexey Romanov, and Anna Rumshisky. 2016b. # hashtagwars: Learning a sense of humor. *arXiv preprint arXiv:1612.03216*.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Ruty Rinott, 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. In *EMNLP*, pages 440–450.

Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside jokes: Identifying humorous cartoon captions. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1065–1074. ACM.

Mark D Shermis, Jill Burstein, Derrick Higgins, and Klaus Zechner. 2010. Automated essay scoring: Writing assessment and instruction. *International encyclopedia of education*, 4(1):20–26.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*, pages 2440–2448.

Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *COLING*, pages 1680–1691.

Henning Wachsmuth, Benno Stein, and Yamen Ajjour. 2017. pagerank for argument relevance. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, volume 1, pages 1117–1127.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational flow in oxford-style debates. *arXiv preprint arXiv:1604.03114*.