Will it Blend? Blending Weak and Strong Labeled Data in a Neural Network for Argumentation Mining

Eyal Shnarch, Carlos Alzate, Lena Dankin, Martin Gleize,
Yufang Hou, Leshem Choshen, Ranit Aharonov, Noam Slonim
IBM Research
{eyals, lenad, leshem.choshen, ranita, noams}@il.ibm.com
{carlos.alzate, martin.gleize, yhou}@ie.ibm.com

Abstract

The process of obtaining high quality labeled data for natural language understanding tasks is often slow, error-prone, complicated and expensive. With the vast usage of neural networks, this issue becomes more notorious since these networks require a large amount of labeled data to produce satisfactory results. We propose a methodology to blend high quality but scarce labeled data with noisy but abundant weak labeled data during the training of neural networks. Experiments in the context of topic-dependent evidence detection with two forms of weak labeled data show the advantages of the blending scheme. In addition, we provide a manually annotated data set for the task of topic-dependent evidence detection.

1 Introduction

In recent years, neural networks have been widely used for natural language understanding tasks. Such networks demand a considerable amount of labeled data for each specific task. However, for many tasks, the process of obtaining high quality labeled data is slow, expensive, and complicated (Habernal et al., 2018). In this work, we propose a method for improving network training when a small amount of labeled data is available.

Several works have suggested methods for generating *weak labeled data* (WLD) whose quality for the task of interest is low, but that can be easily obtained. One approach for gathering WLD is to apply heuristics to a large corpus. For example, Hearst (1992) considered a noun to be the hypernym of another noun if they are connected by the *is a* pattern in a sentence.

Distant supervision is another form of WLD used in various tasks such as relation extraction (Mintz et al., 2009; Surdeanu et al., 2012) and sentiment analysis (Go et al., 2009). Other works use emojis or hashtags as weak labels describing the texts in which they appear (e.g., Davidov et al. (2010) in the context of sarcasm detection).

WLD can be freely obtained, however it comes with a price: it is often very noisy. Therefore, systems trained only on WLD are at a serious disadvantage compared to systems trained on high quality labeled data, which we term henceforth *strong labeled data* (SLD). However, we suggest that the easily accessible WLD is still useful when used alongside SLD, which is naturally limited in size.

In this work we propose a method for blending WLD and SLD in the training of neural networks. Focusing on the argumentation mining field, we create and release a data set for the task of topic-dependent evidence detection. Our evaluation shows that such blending improves the accuracy of the network compared to not using WLD or not blending it. This improvement is even more evident when SLD is not abundantly available.

We believe that blending WLD and SLD is a general notion that may be applicable to many language understanding tasks, and can especially assist researchers who wish to train a network but have a small amount of SLD for their task of interest.

2 Background

2.1 WLD and networks

In the field of neural networks, WLD has mainly been employed for pre-training networks. This was done in related fields such as information retrieval (Dehghani et al., 2017b) and sentiment analysis (Severyn and Moschitti, 2015; Deriu et al., 2017). Contrary to those works, we explore a way to utilize WLD *together* with SLD and *throughout* the training process.

Most similar to our work, Dehghani et al. (2017a) use WLD and SLD together, for sentiment classification. They train two separate networks, one with WLD only, and another with SLD only. They control the magnitude of the gradient updates to the network trained on WLD, using the scores provided by the network trained on SLD. Differently, we blend the two types of labeled data in a single network.

2.2 Argumentation mining

Argumentation mining is attracting a lot of attention (Lippi and Torroni, 2016). One line of research focuses on identifying arguments (claims and evidence/premises) within a text (Stab and Gurevych, 2014; Habernal and Gurevych, 2015; Persing and Ng, 2016; Eger et al., 2017). Another line of work seeks to mine arguments relevant for a given topic or claim, either from a pre-built argument repository where arguments are collected from online debate portals (Wachsmuth et al., 2017), or from unrestricted large scale corpora (Levy et al., 2014; Rinott et al., 2015; Levy et al., 2017). Our work falls into the latter category of *corpus wide topic-dependent* argumentation mining.

Previous work by Rinott et al. (2015) presented the task of detecting evidence texts that are relevant for claims of a given topic. They search in a preselected set of articles, in which the likelihood to find an evidence is considerably higher than in an arbitrary article from the corpus. In this work, we detect evidence directly supporting or contesting the topic (without an intermediate claim), and we search in the entire corpus, with no need for pre-selecting a small set of relevant articles.

2.3 SLD and WLD in argumentation mining

Publicly available strong labeled data (SLD) for argument mining is usually only a couple of thousand instances in size (e.g., Stab and Gurevych (2017) present one of the largest, with around 6,000 annotated positive instances). Recently, Habernal et al. (2018) have commented about the difficulty to collect valuable SLD from crowd sourcing for such tasks.

Several works utilize WLD for argumentation mining; Webis-Debate-16 (Al-Khatib et al., 2016) use the structure of online debates as distant supervision for the task of argument classification. Sentences from the first paragraph are considered as non-argumentative and the rest of the sentences are considered as argumentative.

For the topic-dependent claim detection task, Levy et al. (2017) showed that retrieving sentences with the word *that* followed by the concept representing the topic, yields candidates that are more likely to contain a claim for that topic than arbitrary sentences which contain the topic concept.

3 BlendNet

We present *BlendNet*, a neural network that is trained on a blend of WLD and SLD.

3.1 Network description

Our network is a bi-directional LSTM (Graves and Schmidhuber, 2005) with an additional attention layer (Yang et al., 2016).

The models are all trained with a dropout of 0.85, using a single dropout mask across all timesteps as proposed by Gal and Ghahramani (2016). The cell size in the LSTM layers is 128, and the attention layer is of size 100. We use the Adam method as an optimizer (Kingma and Ba, 2015) with a learning rate of 0.001, and apply gradient clipping with a maximum global norm of 1.0. Words are represented using the 300 dimensional GloVe embeddings learned on 840B Common Crawl tokens and are left untouched during training (Pennington et al., 2014).

We note that even though we chose this network architecture, there is nothing in the blending method we propose which is restricted to it, and blending can be easily applied to other networks.

3.2 WLD blending

WLD is a pair of disjoint sets, WLD_{pos} and WLD_{neg}. The two sets are constructed such that the probability of finding positive instances in WLD_{pos} is significantly higher than that of finding them in WLD_{neg}. This difference in probabilities is the source of the signal WLD provides. Importantly, the probability in WLD_{pos} can still be rather low.

As mentioned in Section 2.1, using WLD to pretrain neural networks has been proven to be effective. We extend this idea by allowing the use of WLD alongside SLD during the entire training process of the network. Our intuition is that even though WLD signal is noisy, there is potential in its additional massive amount, and integrating it can improve training when SLD is limited in size.

In every epoch (a pass through the entire SLD), the training data is enriched with WLD. However, since WLD is noisy, an exponentially decreasing fraction of it is blended into the network at each epoch.

Formally, we have m *initialization epochs* using the entire WLD with no SLD. After this pre-training phase, we continue with n *blending epochs*, in each using all the available SLD, and a fraction of the WLD which is determined by a *blend factor* $\alpha \in [0..1]$. In the k^{th} blending epoch ($k \in [0..n-1]$) we blend α^k of the WLD with the SLD, and feed the data in a random order to the network. Consequently, the first blending epoch uses full SLD and full WLD, and in every subsequent epoch the amount of WLD decays by a factor of α. The stopping point n will typically be empirically determined. We set it to a number that will guarantee that the last couple of epochs will be composed of mainly SLD, since eventually, this is the better signal for training.

One can come up with different methods for blending WLD and SLD. For instance, start training with all available SLD and gradually blend more and more WLD, or use all available WLD and SLD during the entire training. In Section 5 we refer to some alternatives and show that they do not achieve better results than the one presented above. However, we do not claim that our blending method is the only option or even the best one. The goal of this work is to suggest one method which works.

4 Data sets

We created a data set of 5,785 sentences with manual annotations for the task of topic-dependent evidence detection (this will serve as our SLD). It is available on the IBM Debater Datasets webpage.[1] We use it for training and for evaluation and describe it next. In Section 4.2 we describe two methods for freely obtaining weak labeled data for our task.

4.1 SLD annotation

Our strong labeled data (SLD) consists of pairs of a topic and a sentence. Topics were extracted from several sources, such as Debatepedia, an online encyclopedia dedicated to debates and argumentation. The data set includes 118 diverse topics, from domains such as politics, science and education. The topics generally deal with one clearly identifiable concept.

The sentences were extracted from Wikipedia and were annotated by crowd-sourcing. We used 10 annotators for each pair of topic and sentence; each annotator either confirms or rejects the sentence as evidence for the topic. We combine the annotators' votes into a binary label by majority. Ties are resolved as non-evidence.

The guidelines for the task present three criteria which all have to be met for a positive label. The sentence must clearly support or contest the topic, and not simply be neutral. It has to be coherent and stand mostly on its own. Finally it has to be convincing, something you could use to sway someone's stance on the topic: a claim is not enough, it has to be backed up.

The annotators agreement is 0.45 by Fleiss' kappa. This is a typical value in such challenging labeling tasks, comparable to previous reports in the literature, e.g., (Aharoni et al., 2014; Rinott et al., 2015). In addition, for 85% of the labeled instances, the majority vote included at least 70% of the annotators, further supporting the quality of the released data.

The 118 topics were randomly split into two sets: 83 topics for training (4,066 sentences), and 35 topics for testing (1,719 sentences). No sentences of the same topic appear in both sets. The prior for positive, i.e., an evidence instance, is about 40% for both sets. In addition, every occurrence of the topic concept in the candidate is replaced with a common token, to keep the training topic-independent. The topic concept is detected by an in-house wikification tool, similar to TagMe (Ferragina and Scaiella, 2010). The README, provided with this paper, includes additional information about the data set and the pre-processing.

4.2 WLD generation

Next we describe two sources of WLD we use in our experiments. For the first source, we use the method described by Levy et al. (2017) for unsupervised topic dependent claim detection. Following them, we construct the set of WLD$_{\text{pos}}$ by retrieving sentences from Wikipedia which match the query "that + topic concept", i.e. sentences which contain the word "that" followed by the

[1] See http://www.research.ibm.com/haifa/dept/vst/debating_data.shtml

concept of the topic (not necessarily adjacent). The WLD$_{\text{neg}}$ set is constructed by retrieving sentences that contain the topic concept and are not part of WLD$_{\text{pos}}$. Levy et al. (2017) showed that the likelihood of claims in WLD$_{\text{pos}}$ is double the likelihood in WLD$_{\text{neg}}$.

We believe that the query "that + topic concept" is indicative of argumentative content in general, and not just of claims. It is therefore a good fit for constructing WLD for the topic-dependent evidence detection task. Indeed, in the data set, described in Section 4.1, the prior for positive in the entire training set is close to 40%, but among the candidates that match the query, it is much higher – 52%. Applying this WLD method we were able to extract $253,352$ sentences from Wikipedia which contain the topic concept, 25% of them also contain "that" before the topic concept, and they are our WLD$_{\text{pos}}$.

For the second source of WLD, we use the Webis-Debate-16 corpus (Al-Khatib et al., 2016), using their argumentative vs. non-argumentative division. This division was automatically created by mapping the specific structure of idebate.org pages – introduction, points for/against, point/counterpoint – to the two classes. The sentences of the *introduction* are labeled by them as non-argumentative, under the assumption that they neutrally present the topic. We use them as our WLD$_{\text{neg}}$. The other sentences are labeled in Webis-Debate-16 as argumentative, thus we use them as our WLD$_{\text{pos}}$. Out of $16,402$ total instances, 66% are in WLD$_{\text{pos}}$. This data set doubly deserves the status of WLD in our task because the labels do not exactly match the evidence/non-evidence classification, and in addition it is produced automatically based on a coarse-grained mapping that is bound to introduce noise.

5 Experimental setup and results

We use the data set described in Section 4, training the network on the train set and evaluating its accuracy on the test set. We empirically explore several blending configurations and evaluate their impact on the accuracy of the network. To validate our assumption that WLD contribution would be more prominent when SLD is limited, we test each configuration with varying sizes of SLD between 500 and 4,000.

Following some preliminary exploration, on a different data set, we noticed that the parameter m, the number of initialization epochs, does not make a significant difference, and we set it to be 1 (trying $m > 1$ resulted in slightly worse accuracy).

As mentioned in Section 3.2, our stopping criterion was set to ensure that in any configuration, we have four blending epochs in which the input for the network is mostly SLD, i.e. it is at least 95% of the data seen by the network.

For the blending factor we tried $\alpha \in \{0, 0.05, 0.2\}$, and quickly learned that choosing a blending factor value larger than 0.05 is typically ineffective. Since the blending factor determines the numbers of epochs in which the WLD is significant, and since it is reasonable to limit this number due to the noisy nature of the WLD, it is not surprising that a small value of α is preferable. We note that setting $\alpha = 0$ means WLD is only used in the initialization epochs.

Finally, to keep results reliable, as SLD size can get quite small, we repeat each configuration run five times with different SLD slices to reduce variance. For each run we record the best accuracy out of all its epochs and report the micro average of the best accuracies of the five runs.

Figure 1 depicts our results. Blending WLD throughout several epochs of training (the thick green curve with round dots), improves performance over using it only for initialization, as most previous works do (the dashed red curve), and over not using WLD at all (the blue curve with triangles). This effect is significantly more notable as we use less SLD. For example, in the left plot, which presents the usage of Webis-Debate-16 as WLD, we see that using 1,000 instances of SLD with WLD yields results comparable to using 2,500 SLD instances. Similarly, 2,000 SLD instances plus WLD, are comparable to using 3,000 SLD instances. The effect is smaller when the WLD is based on the "that + topic concept" query, but the trend is similar.

One may claim that the signal in WLD is stronger than we hypothesized and therefore the performance improves simply because we are adding labeled data for training. To test this claim we train the network with all available WLD and only it. The single triangles on the Y-axis of each plot show that the accuracy of the network with such training is much lower than using the entire SLD, reflecting the inferior quality of the WLD. In addition, we note that the accuracy on the test set of the "that + topic concept" query, which was

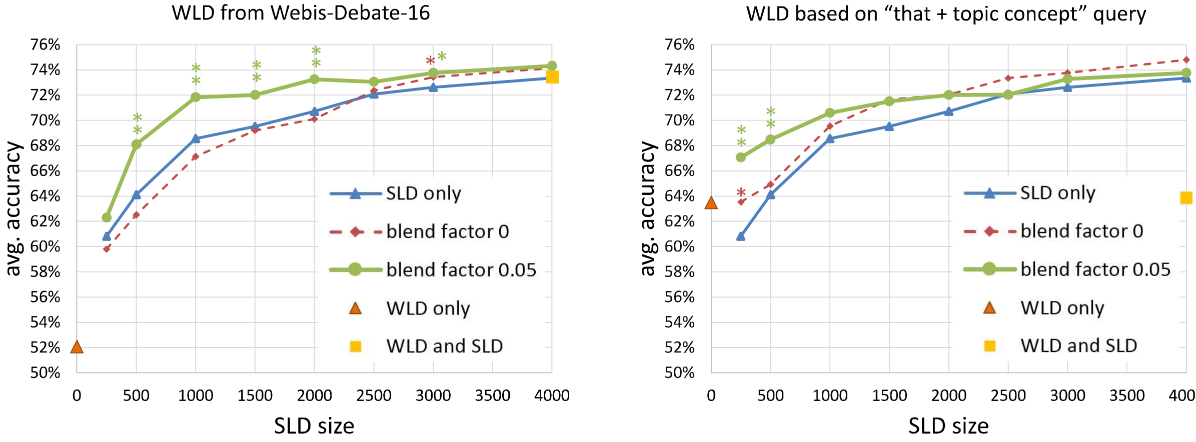

Figure 1: Micro-averaged accuracy on the SLD test set for the different sizes of SLD training data. A single asterisk (*) indicates significant results in comparison to *SLD only* and double asterisks indicate significant results also in comparison to *blend factor 0* (unpaired student t-test with $p < 0.05$).

used to collect one of our WLD types, is only 17%.

Another claim may be that just by utilizing WLD in addition to SLD the accuracy improves, and that there is no need for any blending method. To answer that, we unify the WLD and the SLD, without applying any blending method (single squares on the right border of each plot). For the WLD constructed by the “that + topic concept” query the accuracy is well below the accuracy achieved when using SLD alone, as can be seen in the right plot. On the left plot, we see that unifying the WLD with the SLD does not help nor harm compared to using the SLD alone.

We conclude that even though WLD is not nearly as accurate as SLD, it has the potential to improve performance, if blended correctly.

We also tried gradually *increasing* the amount of WLD in each blending epoch, instead of decreasing it. We tested several increasing factors on both types of WLD. Results were similar to the proposed blending method.

6 Conclusions

Neural networks have become widely useful in natural language understanding tasks. It is often the case that there is not enough high quality labeled data for the target task, leading to significant drops in network performance. On the other hand, for many tasks, weak labeled data can be easily obtained but is usually noisy.

In this work we explore a way to enable a network to take advantage of the large size of WLD without overriding the high quality of SLD.

In the method we present, training starts with initialization epochs in which only the WLD is used. It continues with blending epochs in which the data fed to the network is a dynamic mixture of WLD and SLD. The blending method we presented, assigns higher importance to the vast amount of WLD at the beginning of the training and decreases its impact as training progresses.

We evaluate our blending method on the task of topic-dependent evidence detection, leveraging two WLD sources, and show that it improves performance for each source. The impact of blending increases as the amount of SLD decreases.

Additionally, we release a data set of 5,785 manually labeled sentences to encourage reproducibility and further work on evidence detection.

The impact of the two WLD we tried is evidently different: the Webis corpus seems to help more than the “that + topic concept” query. This calls for future work of understanding what makes a good fit between WLD and SLD. The amount of WLD does not seem to be an important factor, as we see that blending the smaller WLD of the two achieves better performance. It is probably highly related to the quality of the WLD. Sentences retrieved from Wikipedia are of many forms and domains, while the Webis corpus is composed of sentences from debates, which might explain why the network is able to leverage it better.

For future work we intend to examine ways to find better WLD and to make better use of it. For example, instead of choosing one type of WLD, we can combine several WLD types together.

References

Ehud Aharoni, Anatoly Polnarov, Tamar Lavee, Daniel Hershcovich, Ran Levy, Ruty Rinott, Dan Gutfreund, and Noam Slonim. 2014. A benchmark dataset for automatic detection of claims and evidence in the context of controversial topics. In *Proceedings of the first Workshop on Argumentation Mining*, pages 64–68.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, Jonas Köhler, and Benno Stein. 2016. Cross-domain mining of argumentative text through distant supervision. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, California, 12–17 June 2016, pages 1395–1404.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, CoNLL, pages 107–116.

Mostafa Dehghani, Aliaksei Severyn, Sascha Rothe, and Jaap Kamps. 2017a. Learning to learn from weak supervision by full supervision. In *Proceedings of the workshop on Meta-Learning at Advances in Neural Information Processing Systems 31(NIPS 2017)*, pages 65–74.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017b. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74. ACM.

Jan Deriu, Aurelien Lucchi, Valeria De Luca, Aliaksei Severyn, Simon Müller, Mark Cieliebak, Thomas Hofmann, and Martin Jaggi. 2017. Leveraging large amounts of weakly supervised data for multi-language sentiment classification. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1045–1052. International World Wide Web Conferences Steering Committee.

Steffen Eger, Johannes Daxenberger, and Iryna Gurevych. 2017. Neural end-to-end learning for computational argumentation mining. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics,* Vancouver, Canada, 30 July–4 August 2017, pages 11–22.

Paolo Ferragina and Ugo Scaiella. 2010. Tagme: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 1625–1628. ACM.

Yarin Gal and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. In *Advances in neural information processing systems*, pages 1019–1027.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural Networks*, 18(5-6):602–610.

Ivan Habernal and Iryna Gurevych. 2015. Exploiting debate portals for semi-supervised argumentation mining in user-generated web discourse. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 2127–2137.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. The argument reasoning comprehension task: Identification and reconstruction of implicit warrants. In *NAACL*, page to appear. Association for Computational Linguistics.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 15th International Conference on Computational Linguistics,* Nantes, France, 23-28 August 1992, pages 539–545.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations,* San Diego, 2015.

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Ehud Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of the 25th International Conference on Computational Linguistics,* Dublin, Ireland, 23–29 August 2014, pages 1489–1500.

Ran Levy, Shai Gretz, Benjamin Sznajder, Shay Hummel, Ranit Aharonov, and Noam Slonim. 2017. Unsupervised corpus–wide claim detection. In *Proceedings of the 4th Workshop on Argument Mining held at EMNLP 2017,* Copenhagen, Denmark, 8 September 2017, pages 79–84.

Marco Lippi and Paolo Torroni. 2016. Argumentation mining: State of the art and emerging trends. *ACM Transactions on Internet Technology (TOIT)*, 16(2):10.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing,* Singapore, 2–7 August 2009, pages 1003–1011.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Isaac Persing and Vincent Ng. 2016. End-to-end argumentation mining in student essays. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* San Diego, California, 12–17 June 2016, pages 1384–1394.

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 *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Lisbon, Portugal, 17–21 September 2015, pages 440–450.

Aliaksei Severyn and Alessandro Moschitti. 2015. Unitn: Training deep convolutional neural network for twitter sentiment classification. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pages 464–469.

Christian Stab and Iryna Gurevych. 2014. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Doha, Qatar, 25–29 October 2014, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43:619–660.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Natural Language Learning,* Jeju Island, Korea, 12–14 July 2012, pages 455–465.

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*, pages 1117–1127.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1480–1489.