News Aggregation with Diverse Viewpoint Identification Using Neural Embeddings and Semantic Understanding Models

Mark Carlebach[1*], **Ria Cheruvu**[1*], **Brandon Walker**[1*],
Cesar Ilharco[2], **Sylvain Jaume**[1†]
[1] Harvard University, [2] Google Research

Abstract

Today's news volume makes it impractical for readers to get a diverse and comprehensive view of published articles written from opposing viewpoints. We introduce a transformer-based news aggregation system, composed of topic modeling, semantic clustering, claim extraction, and textual entailment that identifies viewpoints presented in articles within a semantic cluster and classifies them into positive, neutral and negative entailments. Our novel embedded topic model using BERT-based embeddings outperforms baseline topic modeling algorithms by an 11% relative improvement. We compare recent semantic similarity models in the context of news aggregation, evaluate transformer-based models for claim extraction on news data, and demonstrate the use of textual entailment models for diverse viewpoint identification.

1 Introduction

The advent of news aggregators has ushered in a new age of information, exposing readers to continuous streams of articles from diverse outlets. However, the proliferation of data makes finding viewpoints presented in different articles challenging. We introduce a novel transformer-based news aggregation system that identifies diverse viewpoints, depicted in Figure 1. Rather than use preset criteria or learned behavior, our system provides a list of viewpoints covered in news articles and allows users to decide which viewpoints to explore. Our system consists of the following components, illustrated in Figure 2: (i) *Topic Modeling* organizes articles from multiple news sources into clusters, (ii) *Hypothesis Extraction* extracts an opinionated summary sentence (i.e., hypothesis) from each article, (iii) *Semantic Similarity* identifies differing viewpoints (i.e., sub-clusters) within each topic based on the hypotheses, (iv) *Premise Extraction* extracts a summary sentence (i.e., premise) from a group of articles associated with a viewpoint, (v) *Textual Entailment* evaluates the entailment between the hypothesis of each article and the premise of its subcluster. As part of this work, we define hypothesis extraction and premise extraction as subsets of claim extraction, where a claim is defined as a sentence expressing viewpoints associated with a news article. We define a hypothesis as a single summary sentence that represents an article's viewpoint, and use the terms hypothesis extraction and single-document subjectivity analysis interchangeably. We define a premise as a single summary sentence that represents viewpoints shared by multiple articles, and use the terms premise extraction and multi-document subjectivity analysis interchangeably.

2 Related Work

News Aggregation: Thorne et al. propose a system involving Document Retrieval, Sentence Selection, and Recognizing Textual Entailment (RTE) for fact extraction and verification (Thorne et al., 2018). Their system expects a claim as input to identify relevant documents, select sentences as evidence from the document, and finally classify the claim. Other authors evaluate the performance of transformer-based models against baseline models for debate data (Chen et al., 2019a; Chen et al., 2019b; Gretz et

[*]The first three authors contributed equally. Their listing order is random.
[†]Corresponding author. Email: sylvain@csail.mit.edu

Proceedings of the 7th Workshop on Argument Mining, pages 59–66
Barcelona, Spain (Online), December 13, 2020.

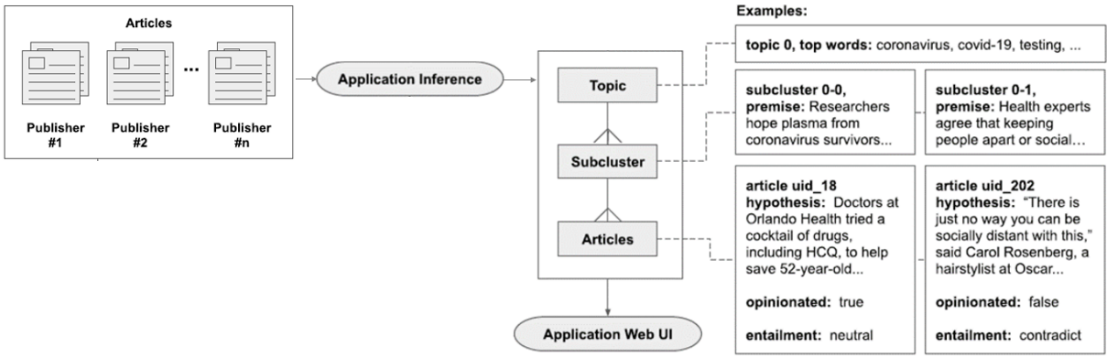

Figure 1: Architecture of our transformer-based system showing inference examples

al., 2020; Ein-Dor et al., 2020), which has applications for news data but is structured slightly differently. A single news article can be clustered under multiple topics and report multiple opinions within the same article. In contrast, debate data often directly align with one particular pre-defined topic and involve separate opinions. Our system differs from an argument search engine with indexing and retrieval (Stab et al., 2018; Wachsmuth et al., 2017). To adjust to the dynamic incoming stream and multiple sources of news data, we explore the generalization capability of language models to automate the news aggregation and viewpoint discovery problem. We have chosen to develop our own labeled article data set targeted specifically for news applications.

Topic modeling: A common approach to clustering documents based on textual content is Multinomial Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Many alternatives have been investigated to improve semantic coherence by representing words with word2vec embeddings (Mikolov et al., 2013a). One such approach is the Embedded Topic Model (ETM) (Dieng et al., 2020), which uniquely represents each document as latent topics, where each topic is an embedding in the semantic space of the words. In this paper, we use ETM for our topic modeling and investigate using transformer-based embeddings in lieu of word2vec embeddings to improve quality of clustering.

Semantic Similarity: Semantic Textual Similarity (STS) refers to the goal of quantifying the degree of similarity between two bodies of text by capturing the degree to which the meanings of the two inputs overlap (Cer et al., 2017). Until recently, state-of-the-art STS systems have relied heavily on word embedding approaches (Mikolov et al., 2013b), which lack the capability to fully capture semantic context. Methods such as InferSent were developed as a solution to embed multiples words, phrases, or sentences into a single representation (Conneau et al., 2017). The Universal Sentence Encoder (USE) models (Cer et al., 2018), BERT (Devlin et al., 2019), and other models such as RoBERTa (Liu et al., 2019) and GPT-3 (Brown et al., 2020) have since made significant improvements on InferSent.

Claim Extraction: A key component of opinion-oriented information extraction from articles is identifying sentence(s) expressing viewpoints associated with articles (Wilson et al., 2005b; Chen et al., 2019b). Early attempts towards solving the problem of single-document subjectivity analysis involved the use of Naïve Bayes classifiers, AdaBoost, and rule-based classifiers trained on the Multi-Perspective Question Answering (MPQA) Opinion Corpus (Wilson et al., 2005b) for identifying subjective expressions and similar tasks (Wilson et al., 2005a; Somasundaran and Wiebe, 2010). Recent work (Xu et al., 2019; Hoang et al., 2019; Han and Kando, 2019) has shown fine-tuned BERT models and BERT-based models (Cer et al., 2018) perform well against baseline models for sentiment analysis and opinion mining tasks. BERT has been applied to multiple passages/documents for question and answering tasks (Wang et al., 2019). However, few transformer-based models were applied for multi-document subjectivity analysis (Liu and Lapata, 2019). In this work, we implement hypothesis extraction as a sentence-classification task and consider BERT-based models against a Naïve Bayes classifier to determine if transformer-based models perform well for hypothesis extraction. We propose abstractive summarization models, such as BART (Lewis et al., 2019) and T5 (Raffel et al., 2019), for premise extraction.

Textual Entailment: Recognizing textual entailment (RTE) involves identifying whether a hypothesis statement supports, contradicts, or is indifferent to a premise statement, regardless of whether the

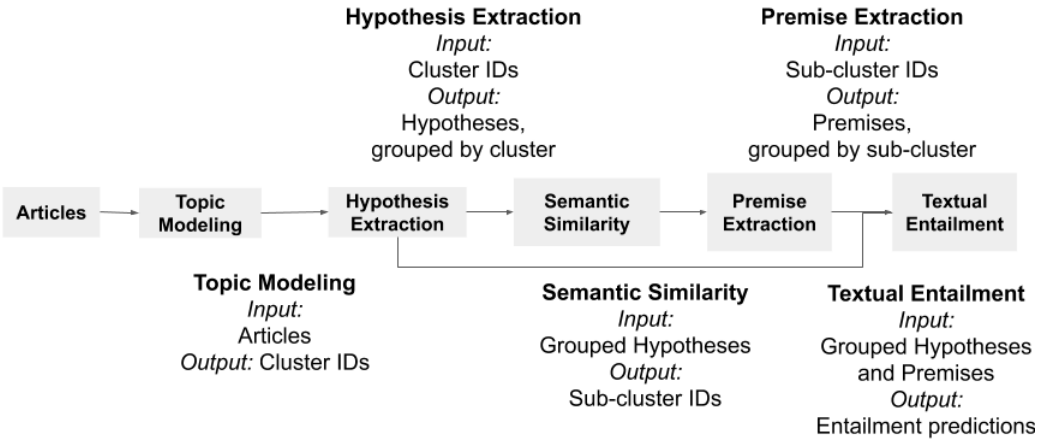

Figure 2: Our system processes news articles in five steps and generates entailment predictions.

premise and hypothesis lexically match (Sammons et al., 2012). Chen et al. leverage textual entailment to find evidence paragraphs in support of viewpoints (Chen et al., 2019b). Attempts towards RTE include Named Entity Recognition (NER) (Sammons et al., 2012), LSTMs with word embeddings, and transformer-based models, such as BART and RoBERTa, which deliver high performance for these tasks.

3 Methods

A demo of our system is available at https://harvard-almit.github.io/newsaggregator/. We tested this system on a dataset of over 1,000 scraped articles from various news outlets, such as Yahoo News and The Post and Courier. Figure 2 illustrates how news articles are processed across the components in our system: (i) *Topic Modeling* generates cluster IDs for the articles, (ii) *Hypothesis Extraction* generates hypotheses for the clustered articles, (iii) *Semantic Similarity* generates subcluster IDs using hypotheses, (iv) *Premise Extraction* generates premises for articles in each subcluster, and (v) *Textual Entailment*, consuming outputs of Hypothesis Extraction and Premise Extraction, generates entailment predictions and provides the results of our news aggregation system.

We perform topic modeling with three approaches using *20 Newsgroups* data (Lang, 1995) and Adjusted Rand Index Scoring (Hubert and Arabie, 1985): (i) Latent Dirichlet Allocation (LDA) using the *gensim* package (Řehůřek and Sojka, 2011), (ii) ETM with word2vec embeddings, and (iii) ETM with centroids of BERT based embeddings (Blei et al., 2003; Mikolov et al., 2013a; Dieng et al., 2020), where we select a value for *num_bert_centroids* as the number of embeddings ETM will use, and use k-means clustering (k=*num_bert_centroids*) from *FAISS* package (Johnson et al., 2019). For training hypothesis extraction models on the clustered articles, we use a modified version of the MPQA Opinion Corpus v3.0 consisting of expressive subjective elements (Deng and Wiebe, 2015). We train a multinomial Naïve Bayes classifier and fine-tune BERT, XLNet (Yang et al., 2019), and ALBERT (Lan et al., 2020) models using HuggingFace's *transformers* library (Wolf et al., 2020) on pre-processed MPQA data for sentence-level subjectivity analysis (i.e., binary opinion classification of sentences).

The semantic similarity module then clusters generated hypotheses of documents within a topic into clusters of semantically related articles, in which each article is associated with a single, more specific topic. We considered BERT, RoBERTa, and DistilBERT (Sanh et al., 2019) in a siamese network structure (Reimers and Gurevych, 2019), in addition to USE using the Pearson correlation coefficient, for the semantic similarity module. The models were fine-tuned on the Argument Facet Similarity Corpus by (Misra et al., 2016) and the STS-Benchmark dataset provided by the *SentEval* (Conneau and Kiela, 2018) package. For premise extraction, we fine-tuned a large BART model (406M parameters) (Lewis et al., 2019) and a small T5 model (60M parameters) (Raffel et al., 2019) using the *transformers* library on data taken from IBM's Project Debater Claim Stance Dataset (Bar-Haim et al., 2017). We reformatted this dataset into a summarization dataset for fine-tuning. We chose not to utilize the claims provided in the dataset, since the hypothesis extraction module already accomplishes this purpose, and used topics

(statements that represent a group of articles) from the dataset instead. The final output of the system is provided by the textual entailment module. For each article in a semantic similarity sub-cluster, premises and hypotheses generated from the claim extraction module are input to the textual entailment module to predict whether the hypothesis contradicts, entails, or is unrelated to the premise. We evaluated Fairseq's pre-trained RoBERTa and BART models fine-tuned for MNLI (Ott et al., 2019) using the transformers library for textual entailment on the claim stance dataset presented by (Bar-Haim et al., 2017), given that BART reportedly performs similar to RoBERTa on the MNLI task (Lewis et al., 2019).

4 Results

Our results for topic modeling show ETM with word2vec embeddings outperforming LDA by 32% on unseen data, and ETM with BERT based embeddings outperforms ETM with word2vec embeddings based on the number of BERT centroids. When predicting on unseen data, ETM trained with 100K BERT centroids outperforms ETM with 25,535 word2vec embeddings by 11%, suggesting the benefits of long sequence contextualized embeddings. We found when predicting on unseen data, the improvements of ETM trained with BERT embeddings do not continue beyond a certain number of centroids due to overfitting. However, when predicting on seen data, ETM trained with BERT embeddings's outperformance increases as the number of BERT centroids increases to 1 million centroids. When predicting on seen data, ETM trained with 1 million BERT centroids outperforms ETM with word2vec embeddings by 17%. For hypothesis extraction, on a held-out dataset of the MPQA data, we found XLNet outperforms the Naïve Bayes classifier baseline by 23%, and provides better performance compared to BERT and ALBERT on the F1 score while ALBERT achieved a higher Matthews correlation coefficient score (see Table 1). We found the BERT-based model is capable of extracting distinct hypotheses from different entities for a particular article.

MODEL	F1	MCC
NAÏVE-BAYES BASELINE	0.740	0.302
BERT	0.878	0.722
XLNET	**0.911**	0.736
ALBERT	0.893	**0.757**

Table 1: XLNet outperforms other models on F1 using held-out MPQA data.

MODEL	STS-B	AFS
USE	0.78413	0.44501
SBERT	0.84195	0.75800
SROBERTA	0.84266	0.75502
SDISTILBERT	0.84135	0.73400

Table 2: Siamese BERT-based models outperform USE on STS-B and AFS by 6%.

In Table 2, for the semantic similarity task, we found the BERT-based models in a siamese network outperformed USE, making them well-suited for our use-case. The results presented are the Pearson correlation of the cosine distance between the embedding vectors and the human-labeled similarity score. The results indicate that we have fairly high correlation ($r \approx 0.84$) recognizing semantically similar sentences and moderate correlation ($r \approx 0.75$) recognizing argument facets. This gives us an average r value across the two tasks of approximately 0.80. Additionally, we note that the smaller DistilBERT yields results similar to its larger counterparts. For premise extraction, we achieved a loss of 1.260 with T5 and a loss of 6.192 with BART on the validation dataset. The T5 model typically outputs 3 sentences. The output is further processed to include the longer sentence to prevent run-off sentences from occurring in the predicted premises. We found BART's predictions were limited to topics in the training data contra T5's predictions that were directly related to article content. Sample predictions from T5 include "The house would be a great place to promote the liberal arts movement" and "The study believes that warm climates would limit the spread of the virus if people are immune from it". For textual entailment, BART has slightly higher accuracy (67%) compared to RoBERTa (65%) on the claim stance dataset. However, for our datasets, RoBERTa outputted predictions with higher probability compared to BART. Given these results, we implemented XLNet for hypothesis extraction, SBERT for semantic similarity, T5 for premise extraction, and RoBERTA for textual entailment. As for ETM, since it proves to be highly resource-intensive, we opted for LDA instead. The following examples show two groups of premise, hypothesis and predicted entailment generated by our system.

Premise	Health experts agree that keeping people apart, or "social distancing," during the coronavirus pandemic is essential for bringing the outbreak under control.
Hypothesis	"There is just no way you can be socially distant with this," said Carol Rosenberg
Entailment	Contradiction

Premise	Word that money would soon land in bank accounts across the country has led to a surge of scam phone calls, with fraudsters falsely claiming people had to provide personal information to collect government money.
Hypothesis	Clicking a link takes them to what looks like an official website asking for personal information with instructions that the step is "necessary" to process their check
Entailment	Neutral

5 Discussion

We found topics generated by ETM for our dataset were more coherent compared to LDA for topic modeling. ETM with BERT based embeddings outperforms ETM with word2vec embeddings when ETM is trained with a number of BERT centroids greater than the number of word2vec embeddings associated with the corpus. Training ETM with BERT centroids involves significantly more computational work than training ETM with word2vec embeddings. On the other hand, in settings where both model creation and predictions are based on the same, full dataset, ETM with BERT-based embeddings does perform better and could be incorporated. We demonstrate that hypothesis extraction can be phrased as subjectivity analysis, and we found XLNET and ALBERT can deliver high performance for this task. A novel aspect of our methodology is that we employ the generated hypotheses as input to the semantic similarity module. Current state-of-the-art models treat semantic similarity as a pair-wise regression problem, making them computationally inefficient for clustering for news aggregation. We found that transformer-based models increased the quality of the clustering compared to USE. The results here show that we can efficiently find semantic clusters with standard clustering methods, e.g., k-Means++ (Arthur and Vassilvitskii, 2007), or density based clustering, e.g., DBSCAN (Ester et al., 1996), to present users with a diverse set of articles on a specific topic.

We show that premise extraction is closely related to the task of abstractive summarization, as premises must be constructed to enable a group of articles to agree or disagree with statements. We observed the small T5 model was able to significantly outperform the larger BART model. We demonstrate that multi-document and single-document claim extractions can be informative premises and hypotheses that are inputs to a textual entailment module. From the second premise-hypothesis pair in Section 4, we see there is a small contradiction that the model does not detect, but could potentially predict if a different premise-hypothesis pair was chosen, or if predictions were validated using phrases from the hypothesis (e.g., the model based its prediction on the phrase "what looks like an official website asking for personal information"). Consequentially, we found that different premise-hypothesis pairs within an article can lead to different predictions from the textual entailment module for the same article, due to opposing viewpoints described in the article and distinct word phrasing between sentences.

6 Conclusion

We have introduced a transformer-based news aggregation system, consisting of topic modeling, hypothesis extraction, semantic clustering, premise extraction, and textual entailment that allows readers to view articles from diverse viewpoints. Our results show relative improvements over baseline models in the range of 10-23% using Embedded Topic Modeling, semantic similarity through fine-tuned BERT models with a siamese network structure, and hypothesis extraction using large pre-trained language models for sentence-level subjectivity analysis. Our results also show a five-fold loss decrease when using a small T5 model, compared to a large BART model, for premise extraction. The system we have developed demonstrates that pre-trained BERT-based models of textual entailment can be used to identify diverse viewpoints.

References

David Arthur and Sergei Vassilvitskii. 2007. k-means++: the advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics.

Roy Bar-Haim, Indrajit Bhattacharya, Francesco Dinuzzo, Amrita Saha, and Noam Slonim. 2017. Stance classification of context-dependent claims. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, volume 1, pages 251–261, Valencia, Spain, April 3-7.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January.

T. Brown, B. Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, G. Krüger, Tom Henighan, R. Child, Aditya Ramesh, D. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, E. Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, J. Clark, Christopher Berner, Sam McCandlish, A. Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada, August. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Sihao Chen, Daniel Khashabi, Chris Callison-Burch, and Dan Roth. 2019a. Perspectroscope: A window to the world of diverse perspectives. *ACL system demonstration track*, abs/1906.04761.

Sihao Chen, Daniel Khashabi, Wenpeng Yin, Chris Callison-Burch, and Dan Roth. 2019b. Seeing things from a different angle: Discovering diverse perspectives about claims. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 542–557, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark, September. Association for Computational Linguistics.

Lingjia Deng and Janyce Wiebe. 2015. Mpqa 3.0: An entity/event-level sentiment corpus. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1323–1328, Denver, Colorado, May 31 – June 5.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of North American Association for Computational Linguistics: Human Language Translation (NAACL-HLT) 2019*, pages 4171–4186, Minneapolis, Minnesota, June 2-7.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. In *Transactions of the Association for Computational Linguistics*, volume 8, pages 439–453.

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - a working solution. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In Evangelos Simoudis, Jiawei Han, and Usama M. Fayyad, editors, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, pages 226–231. AAAI Press.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Wen-Bin Han and Noriko Kando. 2019. Opinion mining with deep contextualized embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 35–42.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland*, 167, pages 187–196. Linköping University Electronic Press.

Lawrence Hubert and Phipps Arabie. 1985. Comparing clusterings. *Journal of Classification*, 2:193–218.

J. Johnson, M. Douze, and H. Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, pages 1–1.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 331–339.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *ArXiv*, abs/1910.13461.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. *ArXiv*, abs/1905.13164.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Amita Misra, Brian Ecker, and Marilyn Walker. 2016. Measuring the similarity of sentential arguments in dialogue. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287, Los Angeles, September. Association for Computational Linguistics.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL-HLT*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Radim Řehůřek and Petr Sojka. 2011. Gensim-statistical semantics in python. In *Proceedings of EuroScipy*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Mark Sammons, Vinod Vydiswaran, and Dan Roth. 2012. Recognizing textual entailment. *Multilingual Natural Language Applications: From Theory to Practice*, pages 209–258.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *EMC2: 5th Edition co-located with NeurIPS*.

Swapna Somasundaran and Janyce Wiebe. 2010. Recognizing stances in ideological on-line debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

Christian Stab, Johannes Daxenberger, Chris Stahlhut, Tristan Miller, B. Schiller, Christopher Tauchmann, Steffen Eger, and Iryna Gurevych. 2018. Argumentext: Searching for arguments in heterogeneous sources. In *NAACL-HLT*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, June. Association for Computational Linguistics.

Henning Wachsmuth, Martin Potthast, Khalid Al Khatib, Yamen Ajjour, Jana Puschmann, Jiani Qu, Jonas Dorsch, Viorel Morari, Janek Bevendorff, and Benno Stein. 2017. Building an argument search engine for the web. In *ArgMining@EMNLP*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. In *EMNLP/IJCNLP*.

Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. 2005a. Opinionfinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pages 34–35.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005b. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv, abs/1910.03771*.

Hu Xu, Bing Liu, Lei Shu, and Philip S Yu. 2019. Bert post-training for review reading comprehension and aspect-based sentiment analysis. *Proceedings of North American Association for Computational Linguistic - Human Language Translation (NAACL-HLT) 2019*, pages 2324–2335, June.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada.