Aspect-based Document Similarity for Research Papers

Malte Ostendorff[1,2] **Terry Ruas**[3] **Till Blume**[4] **Bela Gipp**[2,3] **Georg Rehm**[1]
[1]German Research Center for Artificial Intelligence, Berlin, Germany
[2]University of Konstanz, Konstanz, Germany
[3]University of Wuppertal, Wuppertal, Germany
[4]University of Kiel, Kiel, Germany
malte.ostendorff@dfki.de

Abstract

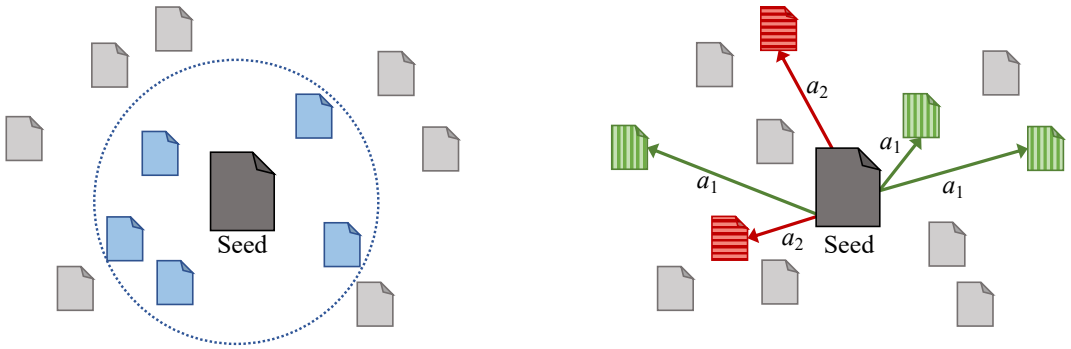
Traditional document similarity measures provide a coarse-grained distinction between similar and dissimilar documents. Typically, they do not consider in what aspects two documents are similar. This limits the granularity of applications like recommender systems that rely on document similarity. In this paper, we extend similarity with aspect information by performing a pairwise document classification task. We evaluate our aspect-based document similarity approach for research papers. Paper citations indicate the aspect-based similarity, i. e., the title of a section in which a citation occurs acts as a label for the pair of citing and cited paper. We apply a series of Transformer models such as RoBERTa, ELECTRA, XLNet, and BERT variations and compare them to an LSTM baseline. We perform our experiments on two newly constructed datasets of 172,073 research paper pairs from the ACL Anthology and CORD-19 corpus. According to our results, SciBERT is the best performing system with F1-scores of up to 0.83. A qualitative analysis validates our quantitative results and indicates that aspect-based document similarity indeed leads to more fine-grained recommendations.

1 Introduction

Recommender systems assist researchers in finding relevant papers for their work. When user feedback is sparse or unavailable, content-based approaches and corresponding document similarity measures are employed (Beel et al., 2016). Recommender systems present a candidate document depending on whether it is similar or dissimilar to the seed document. This coarse-grained similarity assessment (similar or not) neglects the many facets that can make two documents similar. Concerning the general concept of similarity, Goodman (1972), and Bär et al. (2011) even argue that similarity is an ill-defined notion unless one can say to what aspects the similarity relates. In recommender systems for scientific papers, the similarity is often concerned with multiple facets of the presented research, e. g., method, findings (Huang et al., 2020). Given that document similarity can differentiate research aspects, one could obtain tailored recommendations. For instance, a paper with similar *methods* but different *findings* could be recommended. Such a recommender system would facilitate the discovery of analogies in research literature (Chan et al., 2018). We describe the underlying multiple aspect similarity in research papers as *aspect-based document similarity*. Figure 1 contrasts aspect-based with aspect-free similarity (traditional). Following the research paper example, aspect a_1 concerns findings and aspect a_2 methods (red and green in Figure 1b).

In prior work (Ostendorff et al., 2020), we propose to infer an aspect for document similarity formulating the problem as a multi-class classification of document pairs. We extend our prior work to a multi-label scenario and focus on scientific literature instead of Wikipedia articles. Similar to the work of Jiang et al. (2019) and Cohan et al. (2020), we use citations as training signals. Instead of using citations for binary classification (i. e., similar and dissimilar), we include the title of the section in which a citation occurs, as a label for a document pair. The section titles of citations describe the aspect-based

Proceedings of the 28th International Conference on Computational Linguistics, pages 6194–6206
Barcelona, Spain (Online), December 8-13, 2020



(a) Aspect-free similarity.

(b) Aspect-based similarity.

Figure 1: Recommender systems rely on similarity measures between a seed and the k most similar target documents (a). This neglects the aspects which make two or more documents similar. In aspect-based document similarity (b), documents are related according to the inner aspects connecting them (a_1 or a_2).

similarity of citing and cited papers. Our datasets originate from the ACL Anthology (Bird et al., 2008) and CORD-19 (Wang et al., 2020).

In summary, our contributions are: (1) We extend traditional document similarity to aspect-based in a multi-label multi-class document classification task. (2) We demonstrate the aspect-based document similarity is well-suited for research papers. (3) We evaluate six Transformer-based models and a baseline for the pairwise document classification task. (4) We publish our source code, trained models, and two datasets from the computational linguistics and biomedical domain, to foster further research.

2 Related Work

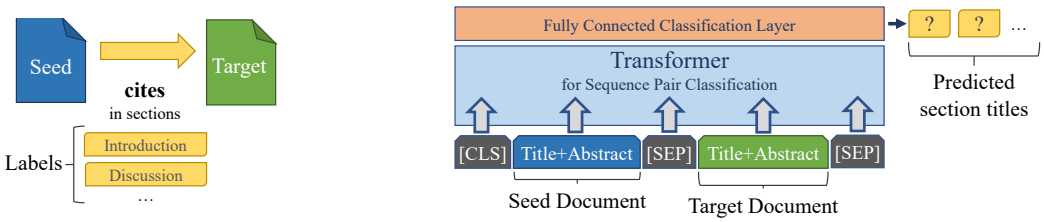
In the following, we discuss work on text similarity, recommendation, and applications of Transformers.

Bär et al. (2011) discuss the notion of similarity as often ill-defined in the literature and used as an "umbrella term covering quite different phenomena". Bär et al. (2011) also formalize what text similarity is and suggest content, structure, and style are the major dimensions inherent to text. For literature recommendation, the content and user information are the most predominant dimensions to consider (Beel et al., 2016).

Chan et al. (2018) explore aspect-based document similarity as a segmentation task instead of a classification task. They segment the abstracts of collaborative and social computing papers into four classes, depending on their research aspects: background, purpose, mechanism, and findings. Cosine similarity computed on segment representations allows the retrieval of similar papers for a specific aspect. Huang et al. (2020) apply the same segmentation approach on the CORD-19 corpus (Wang et al., 2020). Kobayashi et al. (2018) follow a related approach for citation recommendations. They classify sections into discourse facets and build document vectors for each facet. Nevertheless, segmentation is a suboptimal alternative as it breaks the coherence of documents. With pairwise document classification, the similarity is aspect-based without sacrificing the document coherence.

Our experiments investigate Transformer language models (Vaswani et al., 2017). BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), XLNet (Yang et al., 2019), and ELECTRA (Clark et al., 2020) improve many NLP tasks, e. g., natural language inference (Bowman et al., 2015; Williams et al., 2018) and semantic textual similarity (Cer et al., 2017). Reimers and Gurevych (2019) demonstrate how BERT models can be combined in a Siamese network (Bromley et al., 1993) to produce embeddings that can be compared using cosine similarity. Adhikari et al. (2019) and Ostendorff et al. (2019) explore BERT for the classification of single documents with respect to sentiment or topic. Beltagy et al. (2019) and Cohan et al. (2020) study domain-specific Transformers for NLP tasks on scientific documents.

Moreover, Cohan et al. (2020) are the first to use Transformers to encode titles and abstracts of papers to generate recommendations. Mohamed Hassan et al. (2019) also use BERT for recommendations, but only to encode paper titles. Other recent recommender systems rely on other techniques such as



(a) Dataset.

(b) Document Pair Classification.

Figure 2: We use the citations' section titles as labels for the pair of citing and cited paper (seed and target). The sections define the aspects of the similarity. A Transformer model with titles and abstracts as input is used for classification.

co-citation analysis, TF-IDF, or Paragraph Vectors (Kanakia et al., 2019; Collins and Beel, 2019).

In prior work (Ostendorff et al., 2020), we model aspect-based similarity as a pairwise multi-class document classification task. We use the edges of the Wikidata knowledge graph as aspect information for the similarity of Wikipedia articles. The used task definition allows only single-label classification. For research papers, this definition is not adequate. Two papers can be similar in multiple aspects. Accordingly, we incorporate the multi-class classification task and expand it to a multi-label one.

For our experiments, we utilize citations and the section titles in that the citations occur as classification labels. Nanni et al. (2018) demonstrate a related approach in the context of entity linking. They argue that in many situations a link to an entity offers only relatively coarse-grained semantic information. To account for different aspects in which an entity is mentioned, Nanni et al. (2018) link the entities not only to their respective Wikipedia articles but also to the sections that represent the different aspects.

With segment-level similarity and pairwise multi-class single-label classification, preliminary approaches addressing the aspect-based similarity are available. In particular, Transformer models seem promising with their success for similarity, classification, and other related tasks.

3 Experiments

We present our methodology (Figure 2) for classifying the aspect-based similarity of research papers.

3.1 Datasets

The generation of human-annotated data for research paper recommendations is costly and usually limited to small quantities (Beel et al., 2016). The small dataset size hampers the application of learning algorithms. To mitigate the data scarcity problem, researchers rely on citations as ground truth, i. e., when a citation exists between two papers, the two papers are considered similar (Jiang et al., 2019; Cohan et al., 2020). Whether one or no citation exists corresponds to a label for a binary classification. To make the similarity aspect-based, we transfer this idea to the problem of multi-label multi-class classification. As ground truth, we adopt the title of the section in which the citation from paper *A* (seed) to *B* (target) occurs as label class (Figure 2a). The classification is multi-class because of multiple section titles, and multi-label because paper A can cite B in multiple sections. For example, paper *A* citing *B* in the *Introduction* and *Discussion* section would correspond to one sample of the dataset.

ACL Anthology We use the ACL Anthology Reference Corpus (Bird et al., 2008) as a dataset. The corpus comprises 22,878 research papers about computational linguistics. Aside from full-texts, the ACL Anthology dataset provides additional citation data. The citations are annotated with the title of the section in which the citation markers are located. This information is required for our experiments.

CORD-19 The COVID-19 Open Research Dataset (CORD-19) is a collection of papers on COVID-19 and related coronavirus research from several biomedical digital libraries (Wang et al., 2020). The citation and metadata of all CORD-19 papers are standardized according to the processing pipeline of Lo et al. (2019). Citations in CORD-19 are also annotated with section titles.

3.2 Data Preprocessing

Label class	Count	Label class	Count
Introduction	16,279	Conclusion	1,158
Related Work	12,600	Discussion	1,132
Experiment	4,025	Evaluation	971
Background	1,365	Methods	719
Results	1,181	*Other*	22,249

(a) ACL Anthology

Label class	Count	Label class	Count
Introduction	15,108	Background	454
Discussion	13,258	Materials	420
Conclusion	1,003	Virus	218
Results	910	Future work	171
Methods	523	*Other*	43,154

(b) CORD-19

Table 1: Label class distribution as extracted from the citations' section titles in the two datasets. We report the top nine section-classes in decreasing order, and group the remaining as *Other*.

Considering the ACL Anthology and CORD-19, we derive two datasets for pairwise multi-label multi-class document classification. The section titles of the citations, i. e., the label classes, are presented in Table 1. We normalize sections titles (lowercase, letters-only) and resolve combined sections into multiple ones (*Conclusion and Future Work* to *Conclusion*; *Future Work*). We query the API of DBLP (Ley, 2009) and Semantic Scholar (Lo et al., 2019) to match citation and retrieve missing information from the papers such as abstracts. Invalid papers without any text or duplicated ones are removed. We divide both datasets, ACL Anthology and CORD-19, in ten classes according to their number of samples, so that the first nine compose the most popular section titles and the tenth (*Other*) groups the remaining ones. Even though the selection of our ten classes might neglect section title variations in the literature, our model still doubles the number of research aspects Huang et al. (2020) and Chan et al. (2018) defined. The resulting class distribution is unbalanced but it reflects the true nature of the corpora as Table 5 shows. Scripts for reproducing the datasets are available with our source code.

3.3 Negative Sampling

In addition to the ten positive classes (Table 1), we introduce a class named *None* that works as a negative counterpart for our positive samples in the same proportion (Mikolov et al., 2013). The *None* document pairs are randomly selected and are dissimilar. A random pair of papers is a negative sample when the papers do not exist as a positive pair, are not co-cited together, do not share any authors, and are not published in the same venue. We generate 24,275 negative samples for ACL Anthology and 33,083 for CORD-19. These samples let the models distinguish between similar and dissimilar documents.

3.4 Systems

We focus on sequence pair classification with models based on the Transformer architecture (Vaswani et al., 2017). Transformer-based models are often used in text similarity tasks (Jiang et al., 2019; Reimers and Gurevych, 2019). Moreover, Ostendorff et al. (2020) found vanilla Transformers, e. g., BERT (Devlin et al., 2019), XLNet (Yang et al., 2019), outperform Siamese networks (Bromley et al., 1993) and traditional word embeddings (e. g., GloVe (Pennington et al., 2014), Paragraph Vectors (Le and Mikolov, 2014)) in the pairwise document classification task. Hence, we exclude Siamese networks and pretrained word embedding models in our experiments. Instead, we investigate six Transformer variations and an additional baseline for comparison. The titles and abstracts of research paper pairs are used as input for the model, so that the the `[SEP]` token separates seed and target paper (Figure 2b). This procedure is based on our prior work (Ostendorff et al., 2020). We do not use full-texts in our experiments as many papers are not freely available and the selected Transformer models impose a hard limit of 512 tokens.

Baseline LSTM As a baseline, we use a bidirectional LSTM (Hochreiter and Schmidhuber, 1997). To derive representations for document pairs, we feed the title and abstract of two papers through the LSTM, where the papers are separated with a special separator token. We use the SpaCy tokenizer (Honnibal and Montani, 2020) and word vectors from fastText (Bojanowski et al., 2017). The word vectors are pretrained on the abstracts of ACL Anthology or CORD-19 datasets.

BERT, Covid-BERT & SciBERT BERT is a neural language model based on the Transformer architecture (Devlin et al., 2019). Commonly, BERT models are pretrained on large text corpora in unsupervised fashion. The two pretraining objectives are the recovery of masked tokens (i. e., mask language modeling) and next sentence prediction (NSP). After pretraining, BERT models are fine-tuned for specific tasks like sentence similarity (Reimers and Gurevych, 2019) or document classification (Ostendorff et al., 2019). Several BERT models pretrained on different corpora are publicly available. For our experiments, we evaluate three BERT variations. (1) The BERT model from Devlin et al. (2019), trained on English Wikipedia and the BooksCorpus (Zhu et al., 2015). (2) SciBERT (Beltagy et al., 2019), a variation of BERT tailored for scientific literature, which is pretrained on computer science and biomedical research papers. (3) Covid-BERT (Chan, 2020) is the original BERT model from Devlin et al. (2019) but fine-tuned on the CORD-19 corpus.

BioBERT (Lee et al., 2019) is another BERT model specialized in the biomedical domain. Nonetheless, we exclude BioBERT as SciBERT even outperforms it on biomedical tasks (Beltagy et al., 2019). We also omit the BERT variations from Cohan et al. (2020) since they use citation during pretraining risking data leakage into our test set. All three models, i. e., BERT, SciBERT, and Covid-BERT, are similar in their structure, except for the corpus used during the language model training.

RoBERTa Liu et al. (2019) propose RoBERTa, which is a BERT model trained on larger batches, longer training time, and drops the NSP task from its objective. Moreover, RoBERTa uses additional corpora for pretraining, namely Common Crawl News (Nagel, 2016), OpenWebText (Gokaslan and Cohen, 2019), and STORIES (Trinh and Le, 2018).

XLNet Unlike BERT, XLNet (Yang et al., 2019) is not an autoencoder but an autoregressive language model. XLNet does not employ NSP. We use a XLNet model published by its authors, which is pretrained on Wikipedia, BooksCorpus (Zhu et al., 2015), Giga5 (Parker et al., 2011), ClueWeb 2012-B (Callan et al., 2009), and Common Crawl (Elbaz, 2007).

ELECTRA ELECTRA (Clark et al., 2020) has in addition to mask language modelling the pretraining objective of detecting replaced tokens in the input sequence. For this objective, Clark et al. (2020) use a generator that replaces tokens and a discriminator network that detects the replacements. The generator and discriminator are both Transformer models. ELECTRA does not use the NSP objective. For our experiments, we use the discriminator model of ELECTRA. The pretrained ELECTRA discriminator model is pretrained on the same data as BERT.

Hyperparameters & Implementation We choose the LSTM hyperparameters according to the findings of Reimers and Gurevych (2017) as follows: 10 epochs for training, batch size $b = 8$, learning rate $\eta = 1^{-5}$, two LSTM layers with 100 hidden size, attention, and dropout with probability $d = 0.1$. While the LSTM baseline uses vanilla PyTorch, all Transformer-based techniques are implemented using the Huggingface API (Wolf et al., 2019). Each Transformer model is used in its `BASE` version. The hyperparameters for Tranformer fine-tuning are aligned with Devlin et al. (2019): four training epochs, learning rate $\eta = 2^{-5}$, batch size $b = 8$, and Adam optimizer with $\varepsilon = 1^{-8}$. We conduct the evaluation in a stratified k-fold cross-validation with $k = 4$ (i.e., the class distribution remains identical for each fold). This results, on average, in 54,618.75/18,206.25 train/test samples for ACL Anthology, and 74,436/24,812 train/test samples for CORD-19. The source code, datasets, and trained models are publicly available on GitHub[1] and Zenodo[2]. We provide a Google Colab to try out the trained models on any papers from Semantic Scholar [3].

[1] `https://github.com/malteos/aspect-document-similarity`
[2] `https://doi.org/10.5281/zenodo.4087898`
[3] `https://ostendorff.org/r/coling2020-colab`

4 Results

Our results are divided into three parts: overall, label classes, and qualitative evaluation[4].

4.1 Overall Evaluation

The overall results of the quantitative evaluation are presented in Table 2. We conduct the evaluation as 4-fold cross validation based on our datasets. We report micro and macro average for precision, recall, and F1-score to account for the unbalanced label class distribution (see Section 3.1).

Dataset	ACL Anthology						CORD-19					
	macro avg			micro avg			macro avg			micro avg		
	F1(std)	P	R	F1(std)	P	R	F1(std)	P	R	F1(std)	P	R
$LSTM_{baseline}$	.063 ±.001	.069	.058	.290 ±.004	**.761**	.179	.128 ±.001	.137	.121	.579 ±.005	.758	.469
BERT	.256 ±.002	.317	.238	.641 ±.002	.719	.578	.387 ±.011	**.619**	.357	.822 ±.002	.840	.806
Covid-BERT	.270 ±.006	.404	.253	.648 ±.005	.715	.592	.394 ±.010	.578	.364	.818 ±.001	.836	.802
SciBERT	**.326 ±.005**	**.458**	**.303**	**.678 ±.002**	.725	**.637**	**.439 ±.010**	.560	**.401**	**.833 ±.003**	**.846**	**.820**
RoBERTa	.250 ±.003	.285	.232	.626 ±.003	.703	.564	.332 ±.008	.473	.316	.820 ±.001	.840	.801
XLNet	.263 ±.011	.372	.250	.645 ±.011	.705	.595	.362 ±.025	.523	.345	.817 ±.002	.832	.804
ELECTRA	.245 ±.005	.287	.228	.616 ±.021	.693	.554	.280 ±.001	.306	.276	.820 ±.002	.840	.801

Table 2: Overall F1-score (with standard deviation), precision, and recall for macro and micro average of seven methods for ACL Anthology and CORD-19. SciBERT yields best results in both datasets.

Given the overall scores, SciBERT is the best method with 0.326 macro-F1 and 0.678 micro-F1 on ACL Anthology, and with 0.439 macro-F1 and 0.833 micro-F1 on CORD-19. All Transformer models outperform, in all metrics, the $LSTM_{baseline}$ except for the micro-precision on ACL Anthology. The gap between macro and micro average results is due to discrepancies between the label classes (see Section 4.2). BERT, SciBERT, and Covid-BERT perform better, on average, for ACL Anthology and CORD-19 when compared to the baseline and the other Transformer-based models. For ACL Anthology, the methods are ranked equal for both macro and micro. SciBERT presents the highest scores with a large margin, followed by Covid-BERT, XLNet, and BERT. The lower performers are RoBERTa (0.626 micro-F1) and ELECTRA (0.616 micro-F1). In terms of macro average, the methods present the same ranking for CORD-19 and ACL Anthology except for BERT which outperforms XLNet. Only for micro average on CORD-19 the outcome is different, i. e., ELECTRA and RoBERTa achieve higher F1-scores than Covid-BERT and XLNet. Even though Covid-BERT is fine tuned on CORD-19 its performance yields a 0.818 micro-F1.

4.2 Label Classes Evaluation

We divide both datasets, ACL Anthology and CORD-19, into 11 label classes between positive and negative examples (Section 3.2 and 3.3). Each class represents a different section in which a paper is cited. The section indicates in what aspects two papers are similar. The aspects can also be ambiguous making the label classification a hard task. The following section investigates the classification performance with respect to the different label classes. Table 3 presents F1-score, precision, and recall of SciBERT for all 11 labels. Additionally, we include the overall results for single and multi-label samples (i. e., 2, and ≥ 3). The remaining methods from Table 2 present lower but proportionally similar scores[5].

The *None* has the highest F1-score (0.942 for ACL Anthology, 0.980 for CORD-19) with a large margin. *Other* shows the second-best F1-score, which in a similar-dissimilar classification scenario can be interpreted as an opposite class to the *None* label. The remaining positive labels yield lower scores

[4]In the label and qualitative evaluations, we had to exclude one of the two datasets due to space constraints but they are available on GitHub.

[5]The detailed data on the remaining methods is available together with the trained models in our GitHub repository.

ACL Anthology					CORD-19				
Label	**Samples**	**F1 (Std)**	**P**	**R**	**Label**	**Samples**	**F1 (Std)**	**P**	**R**
Background	341	0.436 ± 0.045	0.651	0.329	Background	113	0.617 ± 0.042	0.655	0.588
Conclusion	289	0.000 ± 0.000	0.000	0.000	Conclusion	250	0.274 ± 0.039	0.563	0.182
Discussion	283	0.000 ± 0.000	0.000	0.000	Discussion	3314	0.636 ± 0.008	0.641	0.631
Evaluation	242	0.008 ± 0.007	0.396	0.004	Future work	42	0.032 ± 0.064	0.150	0.018
Experiment	1006	0.360 ± 0.008	0.491	0.284	Introduction	3777	0.644 ± 0.004	0.669	0.620
Introduction	4069	0.527 ± 0.005	0.576	0.486	Materials	105	0.241 ± 0.038	0.552	0.157
Methods	179	0.014 ± 0.028	0.208	0.007	Methods	130	0.205 ± 0.030	0.519	0.130
Related work	3150	0.638 ± 0.012	0.660	0.617	Results	227	0.322 ± 0.021	0.558	0.227
Results	295	0.015 ± 0.011	0.475	0.008	Virus	54	0.000 ± 0.000	0.000	0.000
Other	5562	0.645 ± 0.005	0.646	0.645	Other	10788	0.876 ± 0.002	0.872	0.879
None	6068	0.942 ± 0.002	0.934	0.951	*None*	8270	0.979 ± 0.001	0.980	0.977
1 label	15652	0.721 ± 0.002	0.717	0.726	1 label	22885	0.860 ± 0.003	0.844	0.876
2 labels	1968	0.540 ± 0.003	0.738	0.425	2 labels	1632	0.656 ± 0.004	0.849	0.535
$\geq$ 3 labels	585	0.492 ± 0.015	0.857	0.345	$\geq$ 3 labels	295	0.590 ± 0.010	0.925	0.433

Table 3: Results of SciBERT on ACL Anthology and CORD-19 datasets per label class, number of samples available (test set), F1-score (with standard deviation), precision, and recall.

but also a lower number of samples. Since we conduct a 4-fold cross validation the ratio of train and test samples is 75/25. In CORD-19, 10,788 *Other* test samples exist compared to 3,777 *Introduction* samples, which is the most common section title (Table 1). Still, the lower number of samples does not necessarily correlates with low accuracy. In ACL Anthology, the label *Related work* (3,150 samples) yields higher scores when compared to *Introduction* (4,069 samples) with a F1-score of 0.638 and 0.527 respectively. The label *Background* in CORD-19 has a F1-score of 0.617 despite having only 113 samples. The results in Table 3 show an impact from the label classes on the overall performance. Six labels (ACL Anthology - *Conclusion*, *Discussion*, *Evaluation*, and *Methods*; CORD-19 - *Future work* and *Virus*) have F1-scores between zero and 0.05. The discrepancy in the number of samples and difficulty in uncovering latent information from aspects contribute for the decrease in some labels' accuracy. Even for domain experts, the location of whether one paper cites another, e. g., in *Introduction* or *Experiment*, is not trivial to predict.

The bottom rows in Table 3 illustrate the effect of multi-labels. F1-scores decrease on both datasets as the number of labels increases. This is due to decreasing recall. The precision increases with more labels. Table 4 shows a portion of the distribution of multi-label samples in CORD-19 and corresponding SciBERT predictions (the list is incomplete due to space restrictions). When two or more labels are present, SciBERT often correctly predicts one of the labels but not the others. For example, the label pair of *Discussion* and *Introduction* (D,I) has only 22% test samples correct. Still, SciBERT correctly predicts for the remaining samples one of the two labels, i. e., either *Discussion* (35%) or *Introduction* (31%). We see comparable results for other multi-labels such as *Discussion*, *Introduction*, and *Other* (D,I,O).

4.3 Qualitative Evaluation

To validate the quantitative findings, we qualitatively evaluate the prediction from SciBERT on ACL Anthology. For each example in Table 5, SciBERT predicts whether the seed cites the target paper and in which the section the citation should occur. We manually examine the predictions on their correctness.

The first example of Bär et al. (2012) and Agirre et al. (2012) is a correct prediction. Given the ground truth, the aspect is *Other* (the citation occurs in a section called "Results on Test Data"). We assess *Introduction* as a potential valid prediction since Bär et al. (2012) is a submission to the shared-task described in Agirre et al. (2012). Therefore, one could have cited it in the introduction. All predictions in example 2 are correct. Compared to the other examples, we consider example 2 a simple case as both papers mention their topic (i. e., query segmentation) in the title and in the first sentence of the

Ground Truth		Predictions														
Sections	**Sample**	***N***	**B**	**C**	**D**	**I**	**O**	**R**	**C,O**	**D,I**	**D,O**	**D,R**	**I,O**	**O,R**	**D,I,O**	**D,O,R**
C,D	21	-	-	-	1	6	7	-	-	1	-	-	1	-	-	-
C,O	79	-	-	2	1	2	58	-	13	-	-	-	3	-	-	-
D,I	**459**	**1**	-	-	**163**	**146**	**17**	-	-	**103**	**7**	**2**	**9**	-	**10**	-
D,O	351	1	2	-	102	30	120	1	-	15	59	1	4	1	4	-
D,R	65	1	-	-	6	10	10	-	-	1	3	28	-	-	-	1
I,O	453	2	1	-	15	114	215	1	-	12	16	1	62	-	9	-
D,I,O	142	1	1	-	28	31	11	-	-	33	8	-	12	-	14	-
D,O,R	23	-	-	-	5	-	7	-	-	-	5	2	-	1	-	1

Table 4: Confusion matrix of selected multi-labels for SciBERT on CORD-19 (N=None, C=Conclusion, O=Other, D=Discussion, I=Introduction, R=Results). For example **(in bold)**, 459 test samples are assigned to *Discussion* and *Introduction* (D,I), of which 103 are correctly classified. The remaining samples are mostly classified as single-label, i. e., either *Discussion* (163) or *Introduction* (146).

abstract (hint for *Introduction* label). Both abstracts of example 2 also refer to "mutual information and EM optimization" as their methods. In example 3, Zhang and Clark (2009) and Xi et al. (2012) do not share any citation. Hence, the paper pair is assigned with the *None* label according to the ground truth data even though they are topically related. Zhang and Clark (2009) and Xi et al. (2012) are both about Chinese machine translation. Still, we disagree with the model's prediction of *Experiment* since the two papers conduct different experiments making *Experiment* an invalid prediction. Example 4's predictions are correct. Polifroni et al. (1992) is published before Winterboer and Moore (2007) and, therefore, a citation cannot exist. Nonetheless, the two papers cover a related topic. Thus, one could expect a citation of Polifroni et al. (1992) in Winterboer and Moore (2007) in the introduction section as SciBERT predicted. The model finds this semantic similarity given their latent information on the topic. Example 5-6 present two pairs for which *None* was correctly predicted according to the ground truth. Agarwal et al. (2011) and Gandhe et al. (2006) from Example 6 are topically unrelated as their titles already suggest. However, Karov and Edelman (1998) and Wang et al. (2012) on Example 5 share the topic of *disambiguation*. Thus, we would agree with the prediction of a positive label.

In summary, the qualitative evaluation does not contradicts the quantitative findings. SciBERT distinguishes documents at a higher level and classifies which aspects makes them similar. In addition to traditional document similarity, the aspect-based predictions allow to asses how two papers relate to each other at a semantic level. For instance, whether two papers are similar in the aspects of *Introduction* or *Experiment* is valuable information, especially in literature reviews.

5 Discussion

In the experiments, SciBERT outperforms all other methods in the pairwise document classification. We observe in-domain pretraining and NSP objective often lead to higher F1-scores. Transferring generic language models to a specific domain usually decreases the performance in our experiments. A possible explanation for this is the narrowly defined vocabulary in ACL Anthology or CORD-19. Beltagy et al. (2019) and Lee et al. (2019) have also explored the transfer learning between domains with similar findings. Covid-BERT seems to be an exception as it yields lower results (micro-F1) than BERT on CORD-19 even though Covid-BERT was fine-tuned on CORD-19. We observe the language model fine-tuning in Covid-BERT does not guarantee a higher performance compared to pretraining from scratch in SciBERT. However, Covid-BERT's authors provide too little information to give a proper explanation for its performance. Apart from in-domain pretraining, the NSP objective has a positive effect on the models. All BERT-based systems, which use NSP, outperform the models that excluded NSP (XLNet, RoBERTa, and ELECTRA). We attribute the positive effect of NSP to its similarity to our task since both

	Seed Paper	Target Paper	Citation	Prediction
1	UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures (Bär et al., 2012)	SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity (Agirre et al., 2012)	Other	Introduction×
2	Query segmentation based on eigenspace similarity (Zhang et al., 2009)	Unsupervised query segmentation using generative language models and wikipedia (Tan and Peng, 2008)	Introduction, Experiment	Introduction✓, Experiment✓
3	Transition-Based Parsing of the Chinese Treebank using a Global Discriminative Model (Zhang and Clark, 2009)	Enhancing Statistical Machine Translation with Character Alignment (Xi et al., 2012)	None	Experiment×
4	Experiments in evaluating interactive spoken language systems (Polifroni et al., 1992)	Evaluating information presentation strategies for spoken recommendations (Winterboer and Moore, 2007)	None	Introduction×, Other×
5	Similarity-based Word Sense Disambiguation (Karov and Edelman, 1998)	Targeted disambiguation of ad-hoc, homogeneous sets of named entities (Wang et al., 2012)	None	None✓
6	SciSumm: A Multi-Document Summarization System for Scientific Articles (Agarwal et al., 2011)	Improving question-answering with linking dialogues (Gandhe et al., 2006)	None	None✓

Table 5: Example labels of research paper pairs (seed and target) as defined by citations and as predicted by SciBERT. Based on the test set, correct predictions are marked with ✓, invalid ones with ×.

are sequence pair classification tasks. Table 2 and 3 show variance among labels and both datasets. The larger number of training samples in CORD-19 (36%) may have contributed to a higher performance in comparison to ACL Anthology. An unbalanced class distribution and different challenges of the labels cause the performance to differ between the label classes. The high F1-scores of above 0.9 for negative samples are expected since the *None* label is essentially an aspect-free similarity or citation prediction problem. Transformer models have been shown to perform well in these two problems (Reimers and Gurevych, 2019; Cohan et al., 2020). Besides the unbalanced distribution of training samples, we attribute the differences among positive labels to their ambiguity and to the different challenges posed by the label classes. Authors often diverge when naming their section titles (e. g., *Results*, *Evaluation*), thus, increasing the challenge of labeling the different aspects of a paper. This also contributes to the high number of *Other* samples. Some sections are also content-wise more unique than others. An *Introduction* section usually contains different content than a *Results* section. The content difference makes some sections and the corresponding label classes easier to distinguish and predict than others. We suspect the poor performance for *Future work* is due to little or no information about them in the titles or abstracts.

Our main research objective in this paper is to explore methods that are capable to incorporate aspect information into the traditional similar-dissimilar classification. In this regard, we deem the results as promising. In particular, the micro-F1 score of 0.86 of SciBERT for the CORD-19 dataset is encouraging. Our qualitative evaluation indicates that SciBERT's predictions can correctly identify similar aspects of two research papers. In order to verify if our first indication generalizes, a large qualitative survey needs to be conducted. Furthermore, we observe that label classes with little training data performed poorly. For example, *Conclusion* and *Discussion* have a zero F1-score for ACL Anthology whereas for the larger CORD-19 dataset *Discussion* yields 0.636 F1. We anticipate that more training data leads to more correct predictions.

6 Conclusion

We applied pairwise multi-label multi-class document classification on scientific papers to compute aspect-based document similarity scores. We used section titles as aspects of paper and labeled citations occurring in these sections accordingly. The investigated models were trained to predict citations and the respected label based on the paper's title and abstract. We evaluated the Transformer models BERT,

Covid-BERT, SciBERT, ELECTRA, RoBERTa, and XLNet and a LSTM baseline over two scientific corpora, i. e., ACL Anthology and CORD-19. Overall, SciBERT performed best in our experiments. Despite the challenging task, SciBERT predicted the aspect-based document similarity with F1-scores of up to 0.83. SciBERT's performance motivates further research in this direction. It seems reasonable to include the aspect-based document similarity task as a new pretraining objective in the Transformers architecture. This new objective could be integrated in similar fashion as the binary citation prediction objective Cohan et al. (2020) proposed. As future work, we plan to integrate the aspect-based document similarity into a recommender system. Thus, enabling a large user study to confirm our first indications that aspect-based document similarity indeed helps users to find more relevant recommendations. However, our extensive empirical analysis already demonstrates that Transformers are well-suited to correctly compute the aspect-based document similarity for research papers. Our datasets, code, and trained models are publicly available.

Acknowledgements

We would like to thank all reviewers and Christoph Alt for their comments and valuable feedback. The research presented in this article is funded by the German Federal Ministry of Education and Research (BMBF) through the project QURATOR (Unternehmen Region, Wachstumskern, no. 03WKDA1A).

References

A. Adhikari, A. Ram, R. Tang, J. Lin, and D. R. Cheriton. 2019. DocBERT: BERT for Document Classification. *arXiv:1904.08398v1*.

Nitin Agarwal, Ravi Shankar Reddy, Kiran Gvr, and Carolyn Penstein Rosé. 2011. SciSumm: A Multi-Document Summarization System for Scientific Articles. *49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of Student Session (ACL HLT 2011)*, pages 115–120.

Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity - Google Search. *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 385–393.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A Reflective View on Text Similarity. *International Conference Recent Advances in Natural Language Processing (RANLP)*, pages 515–520.

Daniel Bär, Chris Biemann, Iryna Gurevych, and Torsten Zesch. 2012. UKP: Computing Semantic Textual Similarity by Combining Multiple Content Similarity Measures. *1st Joint Conference on Lexical and Computational Semantics (SEM 2012)*, 2:435–440.

Joeran Beel, Bela Gipp, Stefan Langer, and Corinna Breitinger. 2016. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*, 17(4):305–338.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3613–3618, Stroudsburg, PA, USA. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie J. Dorr, Bryan Gibson, Mark T. Joseph, Min Yen Kan, Dongwon Lee, Brett Powley, Dragomir R. Radev, and Yee Fan Tan. 2008. The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics. *Proceedings of the 6th International Conference on Language Resources and Evaluation, LREC 2008*, pages 1755–1759.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 632–642.

J. Bromley, J.W. Bentz, L. Bottou, I. Guyon, Y. Lecun, C. Moore, E. Sackinger, and R. Shah. 1993. Signature verification using a Siamese time delay neural network. *International Journal of Pattern Recognition and Artificial Intelligence*, 7(4).

Jamie Callan, Mark Hoy, Changkuk Yoo, and Le Zhao. 2009. Clueweb09 data set. https://lemurproject.org/clueweb09/.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo 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)*, volume 371, pages 1–14, Stroudsburg, PA, USA. Association for Computational Linguistics.

Joel Chan, Joseph Chee Chang, Tom Hope, Dafna Shahaf, and Aniket Kittur. 2018. SOLVENT: A Mixed Initiative System for Finding Analogies between Research Papers. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–21, nov.

Branden Chan. 2020. CORD-19 BERT Model. https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge/discussion/138250.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *International Conference on Learning Representations*, pages 1–18.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Andrew Collins and Joeran Beel. 2019. Document Embeddings vs. Keyphrases vs. Terms: An Online Evaluation in Digital Library Recommender Systems. In *ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 130–133.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186, Minneapolis, Minnesota, oct. Association for Computational Linguistics.

Gil Elbaz. 2007. Common Crawl. http://commoncrawl.org.

Sudeep Gandhe, Andrew S. Gordon, and David Traum. 2006. Improving question-answering with linking dialogues. *International Conference on Intelligent User Interfaces, Proceedings IUI*, 2006:369–371.

Aaron Gokaslan and Vanya Cohen. 2019. Openwebtext corpus. https://skylion007.github.io/OpenWebTextCorpus/.

Nelson Goodman. 1972. Seven strictures on similarity. *Problems and Projects*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, nov.

Matthew Honnibal and Ines Montani. 2020. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Ting-Hao 'Kenneth' Huang, Chieh-Yang Huang, Chien-Kuang Cornelia Ding, Yen-Chia Hsu, and C. Lee Giles. 2020. CODA-19: Reliably Annotating Research Aspects on 10,000+ CORD-19 Abstracts Using a Non-Expert Crowd. *arXiv:2005.02367*.

Jyun-yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic Text Matching for Long-Form Documents. In *The World Wide Web Conference on - WWW '19*, pages 795–806, New York, New York, USA. ACM Press.

Anshul Kanakia, Zhihong Shen, Darrin Eide, and Kuansan Wang. 2019. A Scalable Hybrid Research Paper Recommender System for Microsoft Academic. In *The World Wide Web Conference on - WWW '19*, pages 2893–2899, New York, New York, USA. ACM Press.

Yael Karov and Shimon Edelman. 1998. Similarity-based Word Sense Disambiguation. *Computational Linguistics*, 24(1).

Yuta Kobayashi, Masashi Shimbo, and Yuji Matsumoto. 2018. Citation Recommendation Using Distributed Representation of Discourse Facets in Scientific Articles. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 243–251, New York, NY, USA, may. ACM.

Q. V. Le and T. Mikolov. 2014. Distributed Representations of Sentences and Documents. *Proceedings of the 31st International Conference on Machine Learning*, 32:1188–1196.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, pages 1–8, sep.

Michael Ley. 2009. DBLP: some lessons learned. *Proceedings of the VLDB Endowment*, 2(2):1493–1500, aug.

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:1907.11692*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S. Weld. 2019. S2ORC: The Semantic Scholar Open Research Corpus. *arXiv:1911.02782*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, pages 3111–3119.

Hebatallah A. Mohamed Hassan, Giuseppe Sansonetti, Fabio Gasparetti, Alessandro Micarelli, and Joeran Beel. 2019. BERT, ELMo, USE and InferSent Sentence Encoders: The Panacea for Research-Paper Recommendation? In *CEUR Workshop Proceedings*, volume 2431, pages 6–10.

Sebastian Nagel. 2016. Common Crawl News. `http://commoncrawl.org/2016/10/news-dataset-available/`.

Federico Nanni, Simone Paolo Ponzetto, and Laura Dietz. 2018. Entity-Aspect Linking: Providing Fine-Grained Semantics of Entities in Context. In *Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries*, pages 49–58, New York, NY, USA, may. ACM.

Malte Ostendorff, Peter Bourgonje, Maria Berger, Julian Moreno-Schneider, Georg Rehm, and Bela Gipp. 2019. Enriching BERT with Knowledge Graph Embeddings for Document Classification. In *Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019)*, pages 305–312, Erlangen, Germany. German Society for Computational Linguistics & Language Technology.

Malte Ostendorff, Terry Ruas, Moritz Schubotz, Georg Rehm, and Bela Gipp. 2020. Pairwise Multi-Class Document Classification for Semantic Relations between Wikipedia Articles. In *Proceedings of the 2020 ACM/IEEE Joint Conference on Digital Libraries (JCDL'20)*.

Robert Parker, David Graff, Junbo Kong, Ke Chen, and Kazuaki Maeda. 2011. English Gigaword Fifth Edition. `https://catalog.ldc.upenn.edu/LDC2011T07`.

J. Pennington, R. Socher, and C. 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.

Joseph Polifroni, Lynette Hirschman, Stephanie Seneff, and Victor Zue. 1992. Experiments in evaluating interactive spoken language systems. In *Proceedings of the workshop on Speech and Natural Language - HLT '91*, page 28, Morristown, NJ, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2017. Reporting Score Distributions Makes a Difference: Performance Study of LSTM-networks for Sequence Tagging. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 338–348, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *The 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP 2019)*.

Bin Tan and Fuchun Peng. 2008. Unsupervised query segmentation using generative language models and wikipedia. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 347, New York, New York, USA. ACM Press.

Trieu H. Trinh and Quoc V. Le. 2018. A Simple Method for Commonsense Reasoning. *arXiv:1806.02847*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Jun.

Chi Wang, Kaushik Chakrabarti, Tao Cheng, and Surajit Chaudhuri. 2012. Targeted disambiguation of ad-hoc, homogeneous sets of named entities. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web*, pages 719–728.

Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Darrin Eide, Kathryn Funk, Rodney Kinney, Ziyang Liu, William Merrill, Paul Mooney, Dewey Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Chris Wilhelm, Boya Xie, Douglas Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier. 2020. CORD-19: The Covid-19 Open Research Dataset. *arXiv:2004.10706*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *arXiv:1704.05426*, pages 1112–1122.

Andi Winterboer and Johanna D. Moore. 2007. Evaluating information presentation strategies for spoken recommendations. *RecSys'07: Proceedings of the 2007 ACM Conference on Recommender Systems*, pages 157–160.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv:1910.03771*, oct.

Ning Xi, Guangchao Tang, Xinyu Dai, Shujian Huang, and Jiajun Chen. 2012. Enhancing statistical machine translation with character alignment. *50th Annual Meeting of the Association for Computational Linguistics, ACL 2012 - Proceedings of the Conference*, 2(July):285–290.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764.

Yue Zhang and Stephen Clark. 2009. Transition-based parsing of the Chinese treebank using a global discriminative model. In *Proceedings of the 11th International Conference on Parsing Technologies - IWPT '09*, page 162, Morristown, NJ, USA. Association for Computational Linguistics.

Chao Zhang, Nan Sun, Xia Hu, Tingzhu Huang, and Tat Seng Chua. 2009. Query segmentation based on eigenspace similarity. *ACL-IJCNLP 2009 - Joint Conf. of the 47th Annual Meeting of the Association for Computational Linguistics and 4th Int. Joint Conf. on Natural Language Processing of the AFNLP, Proceedings of the Conf.*, pages 185–188.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter:19–27.