Multilevel Text Alignment with Cross-Document Attention

Xuhui Zhou[♡] **Nikolaos Pappas**[♣] **Noah A. Smith**[♣◇]

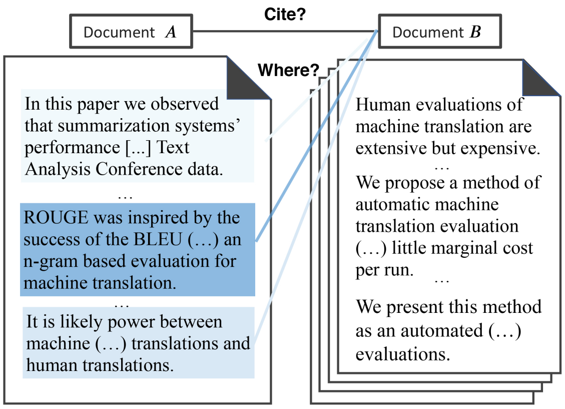
[♡]Department of Linguistics, University of Washington
[♣]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Allen Institute for Artificial Intelligence
xuhuizh@uw.edu,{npappas,nasmith}@cs.washington.edu

Abstract

Text alignment finds application in tasks such as citation recommendation and plagiarism detection. Existing alignment methods operate at a single, predefined level and cannot learn to align texts at, for example, sentence *and* document levels. We propose a new learning approach that equips previously established hierarchical attention encoders for representing documents with a cross-document attention component, enabling structural comparisons across different levels (document-to-document and sentence-to-document). Our component is weakly supervised from document pairs and can align at multiple levels. Our evaluation on predicting document-to-document relationships and sentence-to-document relationships on the tasks of citation recommendation and plagiarism detection shows that our approach outperforms previously established hierarchical, attention encoders based on recurrent and transformer contextualization that are unaware of structural correspondence between documents.

1 Introduction

Aligning texts and understanding their relationships is a common problem for NLP tasks such as citation recommendation (Bhagavatula et al., 2018; Jiang et al., 2019), comparable document mining (He et al., 2010; Peng et al., 2016; Bhagavatula et al., 2018; Guo et al., 2019), parallel sentence mining (Shi et al., 2006; Ture and Lin, 2012; Guo et al., 2018), plagiarism detection (Barrón-Cedeño et al., 2010; Forner et al., 2013; Ferrero et al., 2017), paraphrase identification (Wan et al., 2006; Das and Smith, 2009; Wang et al., 2016), and textual entailment (Dagan and Glickman, 2004; Androutsopoulos and Malakasiotis, 2010; Zhao et al., 2016). Longer texts make the problem more challenging due to the potential complexity of the underlying correspondence. Here, we develop a model to address this problem and demonstrate its applicability on three different tasks which require such understanding, namely on citation recommendation, citation localization, and plagiarism detection for general web documents.



In this paper we observed that summarization systems' performance [...] Text Analysis Conference data.

…

ROUGE was inspired by the success of the BLEU (…) an n-gram based evaluation for machine translation.

…

It is likely power between machine (…) translations and human translations.

Human evaluations of machine translation are extensive but expensive.

…

We propose a method of automatic machine translation evaluation (…) little marginal cost per run.

…

We present this method as an automated (…) evaluations.

Figure 1: A motivating example of aligning scientific documents at different levels. We consider citation recommendation (whether *A* cites *B*) and citation localization (which sentence in *A* cites) at the same time. The confidence of our model for citation localization is represented by the degree of blueness.

One key component of an NLP system for aligning documents is the encoding process. Present approaches for comparing documents rely on hierarchically structured document encoders such as hierarchical attention networks (HANs; Yang et al., 2016), which independently represent the two documents as fixed-length vectors. The vectors are fed to a classifier which makes a decision on the relation between them (Jiang et al., 2019; Guo et al., 2019). However, such methods do not provide insights about or leverage the underlying relationships across documents and are applicable only to a single, predefined level (Jiang et al., 2019; Yang et al., 2020). Importantly, when comparing documents, those methods ignore the structural correspondence between parts.

Figure 1 shows an example of predicting and localizing citations in scientific documents. To solve this problem in a cost-effective way, models need to be able to make joint predictions about these different tasks without relying on fine-grained annotations which are typically more expensive to obtain. In this paper, we propose a new approach for encoding documents that aligns document parts during the encoding process and is able to make predictions about their relationships across different levels (specifically, document-to-document and sentence-to-document). In particular, we equip a powerful class of models, namely hierarchical attention encoders (Yang et al., 2016; Liu and Lapata, 2019; Guo et al., 2019) with a cross-document attention component that "attends" to the structure of documents, enabling inferences about alignment of their parts (Section 3).

We introduce new benchmarks for joint document-to-document prediction and sentence-to-document localization of document relationships for citation recommendation and plagiarism detection (Section 4). Our experiments with variations find that cross-document attention is beneficial to strong baseline hierarchical encoders (Section 5) on these challenging tasks.

2 Comparing Documents

Many potential applications of natural language processing involve a comparative analysis of two (or more) documents. Examples include:

- recommending existing documents to be cited in a new document (Jiang et al., 2019; Yang et al., 2020);
- inferring whether one document plagiarizes another (Foltýnek et al., 2019);
- inferring whether one document is a translation of another (Guo et al., 2019); and
- multi-document summarization (Liu and Lapata, 2019) and coreference resolution (Lee et al., 2012).

Our experiments in Section 5 will consider tasks inspired by the first two applications.

Note that, in each of these examples, the most useful analysis of the document-to-document relationship will include a more fine-grained analysis: which *parts* of the source document correspond to which parts of the target document? Figure 1 illustrates an example for citation recommendation, in which the main relationship (does/should document A cite document B?) is actually composed of a number of more localized relationships between sentences in document A that contain citations and document B (or, perhaps, parts of document B). In general, whenever we seek to model relationships between documents, we believe that these relationships can be *localized* in one or both documents. We believe that automatic identification of these local correspondences is useful, both directly (e.g., mining parallel sentences for use in training a machine translation system), and for providing explanations (e.g., in plagiarism detection).

Of course, these fine-grained, localized correspondences are not typically directly observable in realistic datasets. Here, we consider scenarios where positive and negative examples of document-level relationships are available for supervision, but fine-grained correspondences between their parts are not. We exploit simple decompositions of documents (into sentences and words) but follow earlier work (Yang et al., 2016) in offering a general hierarchical model that could be extended to allow for additional levels in future work.

The problem we aim to solve is: (i) given two documents (each decomposed, e.g., into sentences and words), automatically categorize whether a particular relationship holds between them, and (ii) which parts between them should be "aligned" in support of the relationship in (i). We will refer to these tasks respectively as document-to-document alignment (D2D) and sentence-to-document alignment (S2D), and will conduct experiments on tasks of both kinds in Section 5.

3 Approach

We next describe our solution to this problem, starting with a high-level overview (Section 3.1). We build on a family of widely used models for document representation, known as hierarchical attention networks (HANs; Section 3.2), which is sensitive to predefined notions of hierarchy (here, sentences; Yang et al., 2016). We augment the HAN with cross-document attention (Section 3.3).

3.1 Overview

The training data assumed in our setup is a collection of labeled document pairs. In this work, the labels are binary (either the relationship of interest exists or does not). Let $\langle A, B, y\rangle$ denote a training tuple of two documents with their label y. We apply a familiar “Siamese” architecture (Mueller and Thyagarajan, 2016; Jiang et al., 2019): A and B are encoded using the same function (which we call the “document encoder”), the outputs are concatenated, and then passed through a fully-connected relu layer and a sigmoid function to yield a score. The network is trained to minimize cross-entropy.

This model relies heavily on the document encoder to learn representations relevant to the relationship of interest. As discussed in Section 2, we desire an encoder that can align parts of either or both texts, localizing the relationship to particular sentences, but any encoding function for a document can be used. Our baselines, based on the encoder we present next, do not have any notion of alignment, while our new model does (Section 3.3).

3.2 Hierarchical Attention Networks

Yang et al. (2016) introduced a family of document encoding models that are based on a word/sentence/document hierarchy, known as hierarchical attention networks (HANs). They have been shown superior to earlier hierarchical encoding models based on convolutional networks (Collobert et al., 2011; Kim, 2014; Zhang et al., 2016), they are competitive for tasks involving long documents (Choi et al., 2016; Pappas and Popescu-Belis, 2017; Sun et al., 2018; Miculicich et al., 2018; Liu and Lapata, 2019; Guo et al., 2019),[1] and they can be used orthogonally to other design decisions (e.g., word embeddings and the use of pretraining).

For document X, a HAN builds a vector representation $\mathbf{d}_X$ using the (given) structure of X: typically, the document vector is derived from sentence vectors, which are derived from (contextualized) word vectors. Working in the order that the computation proceeds, the encoding procedure is:

1. Each word in the document is mapped (by lookup) to its type embedding.[2]

2. Each word’s vector is contextualized, i.e., a new word *token* vector is derived from the word and the other words in the sentence. In this work, we consider two contextualizers: pretrained BERT (Devlin et al., 2019) and a GRU (Cho et al., 2014) whose parameters are trained only for the end task.

3. Each sentence in the document is encoded by aggregating the contextualized word vectors. Letting $\mathbf{x}_i$ denote the ith word vector and $\mathbf{y}$ denote the sentence vector, the layer that performs this aggregation has the form:

$$\mathbf{y} = \sum_i \overbrace{\frac{\exp\left[\mathbf{u}^\top \tanh\left(\text{affine}(\mathbf{x}_i)\right)\right]}{\sum_j \exp\left[\mathbf{u}^\top \tanh\left(\text{affine}(\mathbf{x}_j)\right)\right]}}^{\text{attention}_i} \mathbf{x}_i, \tag{1}$$

where i and j range over the words within the sentence. We suppress the parameters of the affine transformation but not the attention parameters $\mathbf{u}$. Note that, when using pretrained BERT, we instead take the average of word token vectors to obtain sentence vectors, following Reimers and Gurevych (2019).

4. Analogous to the two steps above, the sentence vectors are contextualized using a bidirectional GRU for both word-level contextualizers (the default encoder for HAN at the sentence level; Yang et al., 2016) and then aggregated. Aggregation is exactly as in Equation 1, but $\mathbf{x}_i$ now denotes a contextualized sentence vector and $\mathbf{y}$ is the document vector $\mathbf{d}$. A separate set of parameters is used at this level of the hierarchical model. Note that the contextualization can be done with transformers too as by Pappagari et al. (2019); we leave this alternative option as future work.

HANs handle long documents by imposing a simple notion of hierarchy and compositionality; they restrict the dependence of one part’s representation on the representations of its neighbors. They have been used effectively for semantic comparison tasks between documents (Jiang et al., 2019), but they do not offer a way to localize correspondences between parts of the two documents.

3.3 Cross-Document Attention

We augment HANs with a cross-document attention (CDA) mechanism that attends to their document parts, allowing them to reason over structural correspondences between documents. Illustrated

[1] Transformers have emerged as a successful tool across NLP (Vaswani et al., 2017), but they are not yet well suited for long sequences without an hierarchical configuration because their costs scale quadratically with sequence length. When more efficient variants of transformers become available, they will be an appealing option to consider in this setting as well.

[2] “Type embedding” refers to traditional word (subword) vectors; we use the term to contrast with contextualized embeddings associated with specific tokens.

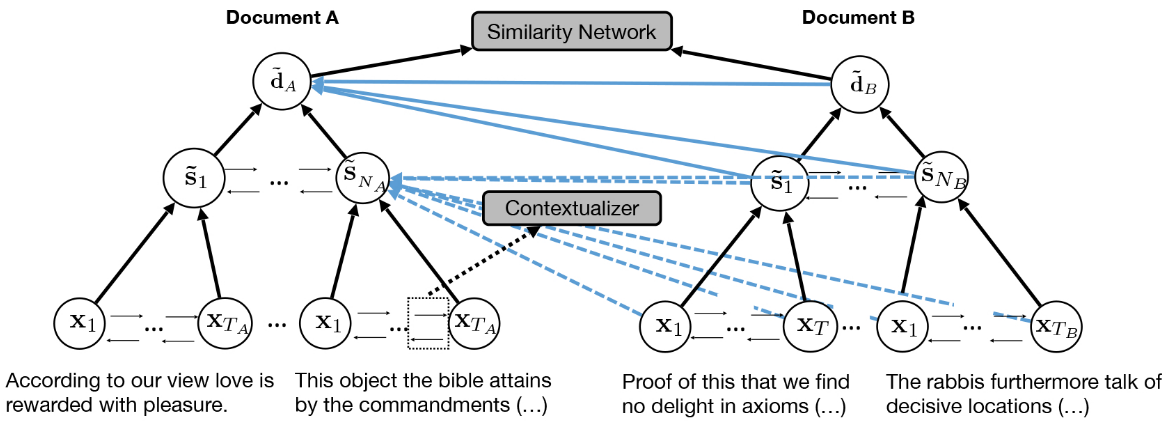

Document A
Similarity Network
Document B
$\tilde{\mathbf{d}}_A$
$\tilde{\mathbf{d}}_B$
$\tilde{\mathbf{s}}_1$
$\tilde{\mathbf{s}}_{N_A}$
$\tilde{\mathbf{s}}_1$
$\tilde{\mathbf{s}}_{N_B}$
Contextualizer
$\mathbf{x}_1$
$\mathbf{x}_{T_A}$
$\mathbf{x}_1$
$\mathbf{x}_{T_A}$
$\mathbf{x}_1$
$\mathbf{x}_T$
$\mathbf{x}_1$
$\mathbf{x}_{T_B}$
According to our view love is rewarded with pleasure.
This object the bible attains by the commandments (...)
Proof of this that we find no delight in axioms (...)
The rabbis furthermore talk of decisive locations (...)

Figure 2: Illustration of our models. In the exposition, SHALLOW considers cross-document attention without the dashed line, while DEEP considers all levels' cross-document attention. Only part of attention is shown in the figure for clarity. The similarity network is for predicting the binary label. The blue arrows indicate the cross-document attention for two example nodes, namely $\tilde{\mathbf{d}}_A$ and $\tilde{\mathbf{s}}_N$.

in Figure 2, the main idea is to allow the representation of a sentence or word in document A to be influenced by the representations of sentences and words in document B, and vice versa.

Consider the document vector $\mathbf{d}_A$ for document A. In the HAN, the aggregation function considered only the (contextualized) vectors for sentences within the document (Equation 1). We inject another layer that "attends" to document B and its sentences. Let $\mathcal{B}$ denote the set containing all contextualized sentence vectors[3] from B and B's document vector. We have:

$$\tilde{\mathbf{d}}_A = \text{affine}\left([\mathbf{d}_A; \sum_{\mathbf{v}\in\mathcal{B}} \frac{\exp \mathbf{v}^\top \mathbf{d}_A}{\sum_{\mathbf{v}'\in\mathcal{B}} \exp \mathbf{v}'^\top \mathbf{d}_A} \mathbf{v}] \right) \tag{2}$$

Again, we suppress the parameters of the affine transformation for clarity. The new vector $\tilde{\mathbf{d}}_A$ is now used as the document representation for A. The same process is repeated in the other direction for obtaining the document vector $\tilde{\mathbf{d}}_B$ which is used as the document representation for the candidate document B.

The modification above is a variant of our model we call SHALLOW; it modifies only the final layer of the encoding so that A's document vector depends on B's document and sentence vectors, and vice versa. A similar layer can optionally be added to update each *sentence* vector in A, using attention over the sentence and word vectors in B; this is illustrated in Figure 2, and we refer to it as the DEEP variant of our model, because CDA is used to modify both sentence and document vectors.

More generally, CDA could be applied with additional levels in a HAN's hierarchy (e.g., paragraphs) and with different design choices about attention across levels.

Relation to previous models. Our idea is related to prior work which has encoded shorter texts such as sentences using attention over their syntactic structures (Liu et al., 2018) to better align texts at the sentence-level. We go beyond sentences by encoding longer texts such as documents at multiple levels using attention over their document structures. An approach similar to ours is due to Li et al. (2019), who used cross-graph attention to compute alignment between computer programs. However, they only evaluated document-to-document alignment (not other levels). They also rely on a graph representation of the documents, which may be costly both in terms of annotation and in computational cost for the required graph-based encoder; semantic and discourse graph structures for natural language are an interesting opportunity to explore in future work.

4 A Benchmark for Document Relation Prediction and Localization

While many tasks and datasets focus on understanding the relationships between sentences or documents separately, to the best of our knowledge, there are no joint publicly available English benchmark for both D2D and S2D tasks. Annotating document correspondences is expensive and time-

[3] Note that we have different strategies for different models here. Details are included in Appendix A.4.2.

	Pairs	Docs	Words		Sentences	
Dataset	count	count	avg	std	avg	std
AAN	132K	13K	122.7	11.2	4.9	2.7
OC	300K	567K	190.4	16.3	7.0	3.5
S2ORC	190K	270K	263.7	19.2	9.3	5.9
PAN	34K	23K	1569.7	90.4	47.4	66.1

Table 1: Dataset statistics, namely the number of unique documents (count), average (avg) number of words per sentence and sentences per document along with their standard deviations (std).

consuming, especially at a fine-grained level like sentences. Therefore, we introduce a new benchmark consisting of six tasks (four D2D and two S2D). This benchmark is shared publicly to encourage continued research.[4]

Datasets. Our data resources of citation recommendation come from the ACL Anthology Network Corpus (AAN; Radev et al., 2009), the Semantic Scholar Open Corpus (OC; Bhagavatula et al., 2018), and the Semantic Scholar Open Research Corpus (S2ORC; Lo et al., 2020). For plagiarism detection, we use the PAN plagiarism alignment task (Potthast et al., 2013). We downsample OC and S2ORC, which are very large. All of our datasets are preprossessed similarly: we filter out characters that are not digits, letters, punctuation, or white space in the texts.

AAN. Contains computational linguistics papers published on ACL Anthology from 2001 to 2014, along with their metadata. For each paper, we extract its abstract and the abstracts of its citations and treat them as positive pairs without including full texts. For each positive pair's source paper, an (uncited, presumed irrelevant) negative paper is sampled at random to create a negative instance. Since the dataset is not complete, we filter out pairs where either document lacks an abstract, but otherwise include all positive citation pairs.

OC. Contains about 7.1M papers in computer science and neuroscience. We follow a similar procedure to that for AAN. Here we only select one citation per source paper, for wider coverage.

S2ORC. A large contextual citation graph of 8.1M open access papers across broad domains of science. The papers in S2ORC are divided into sections and linked by citation edges. We select one section with at least one citation edge provided and the abstract of the cited paper to obtain a positive pair. To obtain a negative pair, we randomly select a paper from S2ORC which is not cited by the source section. Pairs with incomplete abstract or text are filtered out. To obtain the S2D ground-truth for a positive pair (which we will use only in evaluation, not as supervision), we use the citation span stored in the citation edge to identify where the citation appears in the citing document. Specifically, the information implied by the edges (that a paper cites another) for any pair is used to localize the ground-truth sentences which contain the citation. That citation in the sentence is removed from the text to prevent leakage, and the citing sentence is recorded. Note that not every pair has a citation edge that contains relevant sentence-level information, in which case the pair is discarded.

Note that for all the citation-related datasets above, the examples are counted as "negative" as long as they are uncited by the relevant paper. It is possible to use a heuristic approach to avoid treating similar documents as negative examples but we chose not to because the constraint is already largely satisfied with the random sampling procedure and the hypothesis that a paper should be cited by another one when they have high lexical overlap which may not always be true.

PAN. A collection of web documents which contain several kinds of plagiarism phenomena. Human annotations show the segments of texts that are relevant to the plagiarism both in the source and suspicious documents. We construct a positive pair by extracting the relevant segment in the source document and a span (continuous) of text containing the relevant segment in the suspicious document. A negative pair is subsequently constructed by replacing the source segment in the afore-created positive pair with a segment from the corresponding source document which is not annotated as being plagiarised. For the S2D task, the sentences on the suspicious side that are not relevant to the plagiarism are treated as negative candidates in the positive pair. Note that the mapping between sentences is missing from the annotation, which prevents us from creating a sentence-to-sentence task.

[4] Relevant details such as train/dev./test splits are included in Appendix A.1.

Evaluation scores. For D2D tasks, we report accuracy and F_1 score. For S2D tasks, we report the mean reciprocal rank (MRR) and precision-at-N (P@N). In the plagiarism case, multiple sentences can be seen as positive instances; MRR only considers the rank of the first relevant sentence, while P@N reports the number of relevant sentences in the top N.

5 Experiments

Our experiments are performed on the above benchmark to test the benefit of cross-document attention. We first evaluate our model on scientific document citation recommendation (Section 5.2) followed by citation localization (Section 5.3). Then, we evaluate our model on web document plagiarism detection and localization (Section 5.4).

5.1 Settings

Baselines. We compare our method with previous established hierarchical document methods adapted for the task of similarity learning described in Section 3.2. For baseline selection, we considered only methods that could deal with documents of arbitrary length on all of the examined datasets. In particular, we focus on two types of hierarchical attention networks (HANs), namely the first is using pretrained transformer representations from BERT and the other bidirectional GRU trained end-to-end:

- **BERT-AVG**: represents each sentence with the average embedding of its tokens from BERT (Devlin et al., 2019). The representation of the document is computed as the average of the sentence representations.
- **BERT-HAN**: uses BERT to represent sentences with the average embedding. Following Pappagari et al. (2019), the representation of the document is computed by the HAN network starting from the sentence-level representations. The model does not have direct access to word-level representations.
- **GRU-HAN**: encodes documents with a hierarchical attention network with word-level and sentence-level abstractions based on GRU (Yang et al., 2016; Jiang et al., 2019).

For our augmentation, we equip both types of hierarchical encoders with a SHALLOW or a DEEP CDA component, keeping the base setup exactly the same. GRU and BERT are widely used contextualizers in NLP, while each can be viewed as a strong representative of the family of recurrent neural networks and transformers respectively. Note that BERT-HAN only trains a model over sentence-level representations, thus DEEP does not apply to BERT-HAN. For the S2D task, we extract sentence representations from the candidate document $\mathbf{v} \in \mathcal{B}$ per model and rank them according to their similarity with the the target document vector $\mathbf{d}_A$ using an attention function:

$$\text{AttScore} = \frac{\exp \mathbf{v}^\top \mathbf{d}_A}{\sum_{\mathbf{v}' \in \mathcal{B}} \exp \mathbf{v}'^\top \mathbf{d}_A}, \qquad (3)$$

or a cosine similarity function:

$$\text{CosScore} = \frac{\mathbf{v}^\top \mathbf{d}_A}{\|\mathbf{v}\| \|\mathbf{d}_A\|}. \qquad (4)$$

We will refer to them as *attention alignment* and *cosine alignment* respectively. The best scores for each metric and encoder type are marked in **bold**.

Note that the goal of our experiments was not to compare to state-of-the-art document models but to make a controlled experiment using various hierarchical configurations and provide some initial estimates of the difficulty of our benchmark for multilevel document alignment.

Configuration. All the models are implemented in PyTorch. Our code is available on Github.[5] We use Adam to optimize the parameters with an initial learning rate of 10^{-5}. The dimensions of hidden state vectors in GRUs and other hidden layers are set to 50 as in the original HAN (Yang et al., 2016). For word embeddings, we use 50-dimensional GloVe embeddings, which are updated during the training phase.

For pretrained contextualized embeddings, we use BERT-large implemented by HuggingFace.[6] Note that, due to budget constraints, we keep BERT frozen in all experiments except for the finetuning experiment in Section 5.2. Unless otherwise noted, we perform early stopping based on the validation loss if there is no improvement for 5 consecutive epochs. The size of parameters of HAN models with GRU and BERT (kept frozen) are 20M and 1M respectively. Our corresponding models with

[5] https://github.com/XuhuiZhou/CDA

[6] https://huggingface.co/transformers/model_doc/bert.html

Encoder	CDA	AAN Acc	AAN F_1	OC Acc	OC F_1	S2ORC Acc	S2ORC F_1	PAN Acc	PAN F_1
BERT-AVG	–	53.54	53.89	84.72	84.99	77.78	76.92	79.62	76.60
BERT-HAN	–	67.32	64.97	85.96	86.33	90.67	90.76	**87.57**	**87.36**
	Shallow	**71.57**	**69.08**	**87.81**	**87.89**	**91.92**	**92.07**	86.23	86.19
GRU-HAN	–	68.01	67.23	84.46	82.26	82.36	83.28	75.70	75.88
	Shallow	74.51	74.81	88.71	88.96	88.91	89.92	**77.04**	**78.23**
	Deep	**75.08**	**75.18**	**89.79**	**89.92**	**91.59**	**91.61**	75.77	76.71

Table 2: Comparison of our models with the HAN baseline using different encoders on document-to-document alignment over AAN, OC, and S2ORC datasets in terms of accuracy and F_1 score.

Encoder	CDA	Acc	F_1
BERT-HAN	–	73.36	73.51
	Shallow	**82.03**	**82.08**

Table 3: Comparison with BERT-HAN using finetuning on document-to-document alignment on AAN.

a cross-document alignment component increase the number of parameters marginally, namely by 20K parameters. The networks with hierarchical configuration have $\mathcal{O}(T \log D)$ complexity with GRU and $\mathcal{O}(T^2 \log D)$ with BERT (T: sequence length, D: number of layers). Shallow (Deep) adds one (two) more linear and quadratic terms respectively to these complexities, hence the asymptotic complexity remains the same. In practice, adding CDA negligibly impacts decoding speed.[7] For more training details, see Appendix A.2.

5.2 Citation Recommendation

We evaluate the ability of our models to predict whether one document cites another, given citing signal at the document level, and specifically to quantify the effect of augmenting a model with CDA. From the results shown in Table 2 (left), we see a consistent benefit from CDA across AAN, OC, and S2ORC, on accuracy and F_1. Further, the Deep version of our model consistently outperforms the Shallow one on these tasks.

Finetuning BERT. To further evaluate our method, we finetune the BERT- HAN and Shallow with BERT models, on the AAN dataset (where GRU models show an advantage).[8] Table 2 shows that finetuning improves both models' performance, and the benefit of CDA is still present.

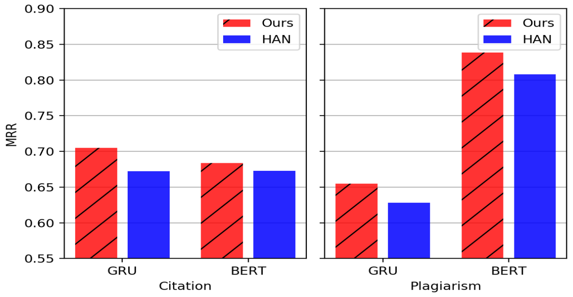


Figure 3: Citation localization results in terms of MRR given oracle document-to-document alignments. *Left*: S2ORC. *Right*: plagiarism detection. For GRU and BERT, "ours" refers to adding a Deep and Shallow CDA component, respectively.

Comparison to state-of-the-art models. our finetuned BERT-HAN with CDA (Shallow) is stronger than the SMASH model of Jiang et al. (2019), which achieves 80.68% accuracy and 80.84% F_1 on AAN. Yang et al. (2020) introduce a method similar to our BERT-HAN baseline and achieve 85.36% accuracy and 85.43% F_1. Note that both models carried out training on full texts; we only use abstracts, using a much smaller computational budget. As reported in Jiang et al. (2019), the baseline HAN trained on full texts achieves 78.13% accuracy, while HAN only achieves 68.01% accuracy on our AAN task.

5.3 Citation Localization

The same models as those in the D2D experiments above can be used to extract S2D alignments for evaluation on the second S2ORC task. If a model fails to predict the document alignment, the

[7] See Appendix A.7 for details.

[8] Maximum 8 epochs, batch size 8; other hyperparameters set according to HuggingFace's GLUE task, https://huggingface.co/transformers/examples.html.

		Attention Alignment				**Cosine Alignment**			
Encoder	**CDA**	MRR	P@10	P@5	P@1	MRR	P@10	P@5	P@1
Random	–	0.3611	46.03	39.30	28.83	0.3611	46.03	39.30	28.83
BERT-AVG	–	0.5596	72.67	64.24	47.22	0.5470	74.18	64.25	45.07
BERT-HAN	–	0.6068	82.75	70.67	51.20	0.5481	63.71	56.73	48.46
	Shallow	**0.6240**	**84.95**	**73.14**	**52.51**	**0.6152**	**81.21**	**71.16**	**52.27**
GRU-HAN	–	0.5430	74.94	63.38	44.75	0.4742	52.59	47.91	41.68
	Shallow	0.6225	84.17	73.11	52.42	0.5013	54.19	48.93	45.31
	Deep	**0.6474**	**86.37**	**76.11**	**54.90**	**0.6252**	**81.93**	**72.04**	**53.35**

Table 4: Comparison of baselines and our models at the sentence-level S2ORC task. The *Random* baseline assigns a random alignment score between 0 and 1 for each sentence.

sentence-level alignment is counted as incorrect.

As shown in Table 3, the deep variant of CDA here shows a consistent advantage over the shallow one, suggesting that explicitly modeling word-level correspondences helps localize citations. We also find that *attention alignment* is consistently better than *cosine alignment* in S2D tasks.

We also consider an oracle evaluation, where the trained models are given the correct D2D prediction (recall that above, the D2D and S2D alignments are jointly predicted). Figure 3 (left) illustrates that CDA is beneficial in this setting as well, for both encoders. The other evaluation metrics show similar trends.

5.4 Plagiarism Detection

For plagiarism detection, the input consists of a source document and a suspicious document; the D2D task is to predict whether the suspicious document plagiarizes the source document, and the S2D (localization) task is to identify which sentences in the suspicious document plagiarize. The dataset here (PAN) is considerably smaller than those we explored for citations (Table 1).

D2D results are shown in Table 2 (right). CDA is not helpful to the BERT-HAN model and only Shallow CDA helps the HAN GRU model on the D2D task, which could be attributed to the small size of the plagiarism dataset.

S2D performance is shown in Table 5 with *attention alignment*; here we see a consistent benefit from CDA across encoders and evaluation scores.

Enc	**CDA**	MRR	P@10	P@5
Random	–	0.4215	44.23	43.28
BERT-AVG	–	0.7864	58.24	64.69
BERT-HAN	–	0.8072	60.36	68.94
	Shallow	**0.8386**	**60.47**	**69.07**
GRU-HAN	–	0.6205	50.72	51.90
	Shallow	**0.6479**	51.71	53.05
	Deep	0.6378	**52.07**	**53.82**

Table 5: Sentence-to-document plagiarism detection.

6 Other Related Work

Latent Alignment. Attention has been previously used to align word sequences based on their intermediate hidden states for summarization (Rush et al., 2015) and machine translation (Bahdanau et al., 2015). The alignment is typically softly learned and does not consider alternative alignments in a probabilistic sense. Hard attention (Luong et al., 2015) is an alternative approach which selects only one word at a time but it is non-differentiable and requires more complicated techniques such as reinforcement learning to train. Deng et al. (2018) considered an alternative attention network for learning latent variable alignment models based on amortized variational inference. Others modified attention to attend to partial segmentations and subtrees (Kim et al., 2017) or trees (Liu et al., 2018), while Yang et al. (2018) cast the problem as latent graph learning to capture dependencies between pairs of words from unlabeled data. Orthogonal to these studies, we use attention to compare documents represented by hierarchical document encoders at multiple levels.

Similarity Learning. There are three types of similarity learning in NLP. The supervised paradigm differs from typical supervised learning in that training examples are cast into pairwise constraints (Yang and Jin, 2006), as in cross-lingual word embedding learning based on word-level alignments (Faruqui and Dyer, 2014) and zero-shot utterance/document classification (Yazdani and Henderson, 2015; Nam et al., 2016; Pappas and Henderson, 2019) based on utterance/document-level annotations. The unsupervised paradigm aims to learn an underlying low-dimensional space where the relationships between most of the observed data are preserved, as in word embedding learning (Collobert et al., 2011; Mikolov et al., 2013; Pennington et al., 2014; Levy and Goldberg, 2014). The weakly supervised paradigm is the middle ground between the two, as in cross-lingual word embedding learning based on sentence-level alignments (Hermann and Blunsom, 2014; Gouws et al., 2015). Our approach is weakly supervised and operates at the document-level, making use of structural correspondence between documents.

7 Conclusion

We augment hierarchical attention networks with cross-document attention, allowing their use in document-to-document and sentence-to-document alignment tasks. We introduce benchmarks, based on existing datasets, to evaluate model performance on such tasks. In controlled experiments, we observe a benefit from cross-document attention on three out of the four document-to-document tasks and two out of two sentence-to-document tasks.

Acknowledgments

The authors thank Efstathios Stamatatos, Phillip Keung, Lucy Lin, and the anonymous reviewers for their helpful feedback. Nikolaos Pappas was supported by the Swiss National Science Foundation under the project UNISON, grant number P400P2_183911. This work was supported in part by US NSF grant 1562364.

References

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *JAIR*, 38(1):135–187.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. of ICLR*.

Alberto Barrón-Cedeño, Paolo Rosso, Eneko Agirre, and Gorka Labaka. 2010. Plagiarism detection across distant language pairs. In *Proc. of COLING*.

Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-based citation recommendation. In *Proc. of NAACL*.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. of EMNLP*.

Eunsol Choi, Daniel Hewlett, Alexandre Lacoste, Illia Polosukhin, Jakob Uszkoreit, and Jonathan Berant. 2016. Hierarchical question answering for long documents. *CoRR*. Abs/1611.01839.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, page 2493–2537.

Ido Dagan and Oren Glickman. 2004. Probabilistic textual entailment: Generic applied modeling of language variability.

Dipanjan Das and Noah A. Smith. 2009. Paraphrase identification as probabilistic quasi-synchronous recognition. In *Proc. of AFNLP*.

Yuntian Deng, Yoon Kim, Justin Chiu, Demi Guo, and Alexander Rush. 2018. Latent alignment and variational attention. In *Proc. of NeurIPS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. of NAACL*.

Manaal Faruqui and Chris Dyer. 2014. Improving vector space word representations using multilingual correlation. In *Proc. of EACL*.

Jérémy Ferrero, Laurent Besacier, Didier Schwab, and Frédéric Agnès. 2017. Using word embedding for cross-language plagiarism detection. In *Proc. of EACL*.

Tomás Foltýnek, Norman Meuschke, and Bela Gipp. 2019. Academic plagiarism detection: A systematic literature review. *CSUR*, 52:1 – 42.

Pamela Forner, Roberto Navigli, Dan Tufis (eds, Martin Potthast, Matthias Hagen, Tim Gollub, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th international competition on plagiarism detection. *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*.

Stephan Gouws, Yoshua Bengio, and Gregory S. Corrado. 2015. BilBOWA: Fast bilingual distributed representations without word alignments. In *Proc. of ICML*.

Mandy Guo, Qinlan Shen, Yinfei Yang, Heming Ge, Daniel Cer, Gustavo Hernandez Abrego, Keith Stevens, Noah Constant, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Effective parallel corpus mining using bilingual sentence embeddings. In *Proc. of WMT*.

Mandy Guo, Yinfei Yang, Keith Stevens, Daniel Cer, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2019. Hierarchical document encoder for parallel corpus mining. In *Proc. of WMT*.

Qi He, Jian Pei, Daniel Kifer, Prasenjit Mitra, and Lee Giles. 2010. Context-aware citation recommendation. In *Proc. of WWW*.

Karl Moritz Hermann and Phil Blunsom. 2014. Multilingual models for compositional distributed semantics. In *Proc. of ACL*.

Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *Proc. of WWW*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proc. of EMNLP*.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured attention networks. *CoRR*, abs/1702.00887.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proc. of EMNLP*.

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proc. of NeurIPS*.

Yujia Li, Chenjie Gu, Thomas Dullien, Oriol Vinyals, and Pushmeet Kohli. 2019. Graph matching networks for learning the similarity of graph structured objects. *CoRR*, abs/1904.12787.

Yang Liu, Matt Gardner, and Mirella Lapata. 2018. Structured alignment networks for matching sentences. In *Proc. of EMNLP*.

Yang Liu and Mirella Lapata. 2019. Hierarchical transformers for multi-document summarization. In *Proc. of ACL*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proc. of ACL*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proc. of EMNLP*.

Lesly Miculicich, Dhananjay Ram, Nikolaos Pappas, and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. In *Proc. of EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proc. of ICLR*.

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Proc. of AAAI*.

Jinseok Nam, Eneldo Loza Mencía, and Johannes Fürnkranz. 2016. All-in text: Learning document, label, and word representations jointly. In *Proc. of AAAI*.

Raghavendra Pappagari, Piotr Żelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. ArXiv:1910.10781.

Nikolaos Pappas and James Henderson. 2019. GILE: A generalized input-label embedding for text classification. *TACL*, 7:139–155.

Nikolaos Pappas and Andrei Popescu-Belis. 2017. Multilingual hierarchical attention networks for document classification. In *Proc. of IJCNLP*.

Hao Peng, Jing Liu, and Chin-Yew Lin. 2016. News citation recommendation with implicit and explicit semantics. In *Proc. of ACL*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proc. of EMNLP*.

Martin Potthast, Tim Gollub, Matthias Hagen, Martin Tippmann, Johannes Kiesel, Paolo Rosso, Efstathios Stamatatos, and Benno Stein. 2013. Overview of the 5th International Competition on Plagiarism Detection. In *Working Notes Papers of the CLEF 2013 Evaluation Labs*.

Dragomir R. Radev, Pradeep Muthukrishnan, Vahed Qazvinian, and Amjad Abu-Jbara. 2009. The acl anthology network corpus. *ELRA*, 47:919–944.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proc. of EMNLP*.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proc. of EMNLP*.

Lei Shi, Cheng Niu, Ming Zhou, and Jianfeng Gao. 2006. A DOM tree alignment model for mining parallel data from the web. In *Proc. of ACL*.

Qingying Sun, Zhongqing Wang, Qiaoming Zhu, and Guodong Zhou. 2018. Stance detection with hierarchical attention network. In *Proc. of COLING*.

Ferhan Ture and Jimmy Lin. 2012. Why not grab a free lunch? mining large corpora for parallel sentences to improve translation modeling. In *Proc. of NAACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *In Proc. of NeurIPS*.

Stephen Wan, Mark Dras, Robert Dale, and Cécile Paris. 2006. Using dependency-based features to take the 'para-farce' out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*.

Zhiguo Wang, Haitao Mi, and Abraham Ittycheriah. 2016. Sentence similarity learning by lexical decomposition and composition. In *Proc. of COLING*.

Liu Yang and Rong Jin. 2006. Distance metric learning: A comprehensive survey. *Michigan State Universiy*, 2.

Liu Yang, Mingyang Zhang, Cheng Li, Michael Bendersky, and Marc Najork. 2020. Beyond 512 tokens: Siamese multi-depth transformer-based hierarchical encoder for document matching. ArXiv:2004.12297.

Zhilin Yang, Junbo Jake Zhao, Bhuwan Dhingra, Kaiming He, William W. Cohen, Ruslan Salakhutdinov, and Yann LeCun. 2018. GLoMo: Unsupervisedly learned relational graphs as transferable representations. *CoRR*, abs/1806.05662.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proc. of NAACL*.

Majid Yazdani and James Henderson. 2015. A model of zero-shot learning of spoken language understanding. In *Proc. of EMNLP*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. In *Proc. of NeurIPS*.

Kai Zhao, Liang Huang, and Mingbo Ma. 2016. Textual entailment with structured attentions and composition. In *Proc. of COLING*.

A Supplementary Material for "Multilevel Text Alignment with Cross-Document Attention"

We provide more details about the datasets in the new benchmark, experimental setup, choices of hyperparameters, model design choices, and validation performances corresponding to the ones reported in the main paper. Moreover, we provide qualitative examples that visualize the cross-document attention component.

A.1 Benchmark Details

In Table 6, we list the statistics about training, development and test splits of the four datasets that are part of our benchmark. The exact splits used in our experiments are released along with the datasets of the benchmark.[9] A sample of our datasets are attached along with our submission. Note that we have not included the whole benchmark because its size exceeds the limit allowed in the submission portal.

Dataset	Training	Validation	Test
AAN	106,592	13,324	13,324
OC	240,000	30,000	30,000
S2ORC	152,000	19000	19000
PAN	17,968	2,908	2,906

Table 6: Dataset statistics regarding the number of examples for the training/validation/test splits.

A.2 Experimental Setup Details

For our experiments, we used the following computing infrastructure: 1 GeForce 960, 1 GeForce 1080, and 1 Titan Xp for the model training. The batch size is set to 128 for GRU-HAN (including our augmentation) experiments on AAN task, and is set to 256 for all other experiments. The running time ranges from 36 hours to 48 hours for the GRU-based models, and from 1 to 2 hours for BERT-frozen models, and about 24 hours for BERT-finetuning models.

A.3 Development Scores

We report validation performance for all the reported test results for the document-to-document alignment tasks in Tables 8–9. Note that for the sentence-to-sentence alignment tasks there is no validation taking place, hence, there are no development scores to report here.

A.4 Model Design Choices

In this section, we describe the set of model design choices that were made based on development performance before running our main experiment.

A.4.1 Word Embedding Dimension

To decide what embedding size to use for our main experiments, we experimented with 50-dimensional and 200-dimensional GloVe embeddings by training the GRU-HAN model on the AAN task. When keeping other settings exactly the same as the aforementioned GRU-HAN models on the AAN task, the performance of GRU-HAN with a larger word embedding size is lower than the 50-dimensional model as shown in Table 7. Therefore, we stick with 50-dimensional GloVe embeddings for the other experiments.

Encoder	Dim	Acc	F_1
HAN	50	68.01	67.23
	200	66.94	66.24

Table 7: Influence of the dimensionality of word embeddings to the baseline model HAN.

A.4.2 Sentence Contextualization

Preliminary experiments show that one could obtain better performance on the AAN D2D task by using sentence vectors before contextualization in Equation 2 for GRU-based models. Therefore, the experiments for GRU-based models above use sentence vectors before contextualization for CDA. For BERT-based models, we use two GRU layers to contextualize sentence vectors, the sentence vectors after the first GRU layer are used in CDA. Practitioners can be flexible in deciding how CDA is used for different tasks and encoders.

Encoder	CDA	Acc	F_1
BERT-HAN	–	75.41	74.25
	SHALLOW	**83.72**	**82.57**

Table 8: Development set results corresponding to Table 3.

[9] https://xuhuizhou.github.io/Multilevel-Text-Alignment/

Encoder	CDA	AAN Acc	AAN F_1	OC Acc	OC F_1	S2ORC Acc	S2ORC F_1	PAN Acc	PAN F_1
BERT-AVG	–	54.72	54.12	84.58	84.67	80.52	81.14	82.73	82.42
BERT-HAN	–	67.34	64.69	85.73	86.23	90.46	90.49	**88.85**	**88.77**
	SHALLOW	**73.20**	**70.80**	**87.73**	**87.91**	**91.66**	**91.54**	86.76	86.79
GRU-HAN	–	69.68	69.15	83.06	83.84	82.65	83.41	76.40	76.77
	SHALLOW	77.46	75.41	89.31	89.41	89.78	89.89	**77.02**	**78.15**
	DEEP	**78.17**	**75.94**	**91.10**	**91.11**	**92.01**	**92.02**	76.99	78.01

Table 9: Development set results corresponding to Table 2.

A.5 Integrating Cross-Document Attention

For the integration of the cross-document attention representations with the representations of the hierarchical attention network we experimented with two options, namely concatenation and addition of vectors. We found that our method is more competitive with concatenation. However, integrating with concatenation involves a linear projection to match the original hidden size of the network which increases slightly the number of parameters. Here, we evaluate the performance of our model when it uses addition, that is when the number of parameters remains exactly the same with that of the base network.

We evaluated the performance of our SHALLOW augmentation on BERT-HAN with finetuning BERT end-to-end. The results are displayed in Table 10. With SHALLOW (addition), BERT-HAN achieves 79.02% accuracy and 79.08% F_1, which still improves over the BERT-HAN baseline. Interestingly, using addition instead of concatenation performs quite well and its performance is still better than the hierarchical attention network baseline. Hence, we conclude that the additional number of parameters is not the only factor responsible for the superior performance of our model.

Encoder	CDA	Acc	F_1
BERT-HAN	–	73.36	73.51
	SHALLOW (concatenation)	**82.03**	**82.08**
	SHALLOW (addition)	79.02	79.08

Table 10: Comparison with BERT-HAN using finetuning on document-to-document alignment on AAN.

A.6 Qualitative Inspection

We select two examples from the test sets of GORC and PAN, where both HAN and HAN SHALLOW with BERT obtain correct D2D results. However, while HAN is confused of finding the sentence where citation or plagiarism happens, HAN SHALLOW is able to locate the relevant sentences in the document as shown in Figure 4.

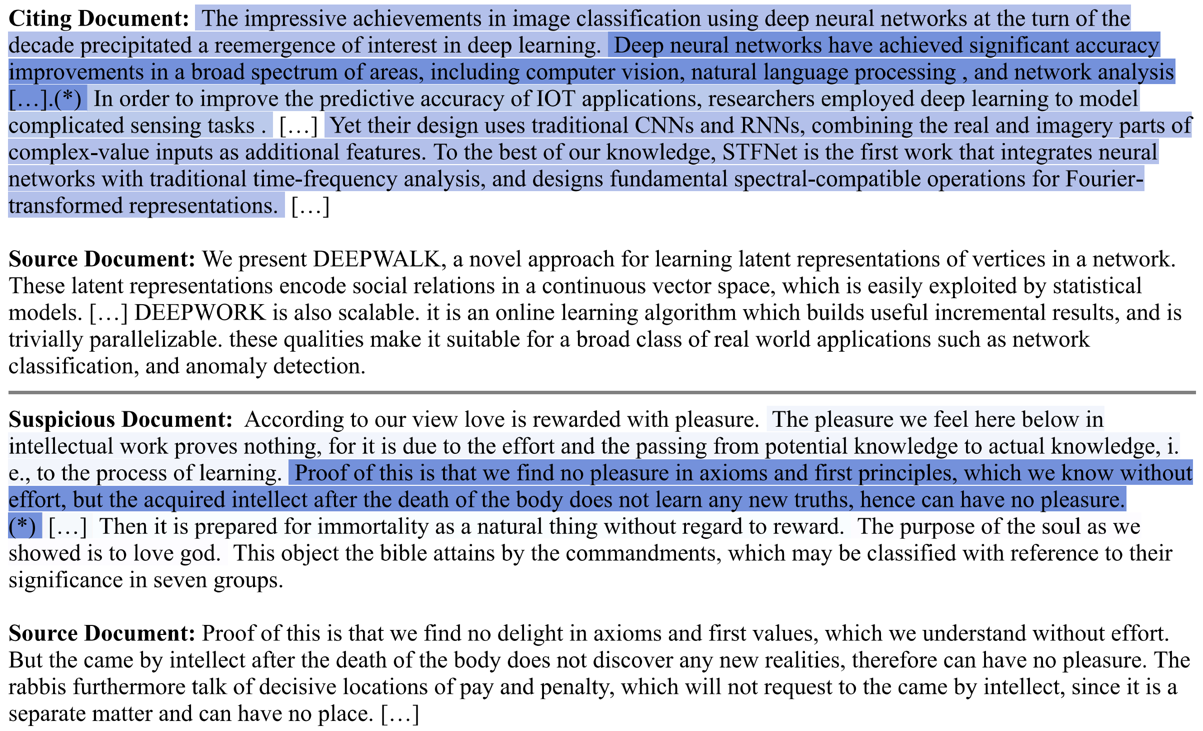
We find that the document-to-document results in these two examples are heavily dependent on the localization of the sentences. While we have difficulty in interpreting HAN's decision for the two examples, it is not hard for us to see how HAN SHALLOW, as a unified model for D2D and S2D tasks, obtains its decision on whether this document cites or plagiarizes the other one. This property should be important for future models to pursue instead of simply producing a yes or no decision.

A.7 Computational Cost of CDA

In Table 11, We show the average inference time of each epoch (256 batch size) on GeForce 1080 for tasks S2ORC and PAN, which have longer texts among our tasks. In general, the extra computational cost for CDA is negligible (1–2% extra wall time), especially for SHALLOW. Note that BERT-based models share the same property.

Encoder	CDA	S2ORC	PAN
HAN	–	0.256	0.568
	SHALLOW	0.256	0.581
	DEEP	0.258	0.637

Table 11: Comparison of average inference time (s) of each epoch for S2ORC and PAN.



Citing Document: The impressive achievements in image classification using deep neural networks at the turn of the decade precipitated a reemergence of interest in deep learning. Deep neural networks have achieved significant accuracy improvements in a broad spectrum of areas, including computer vision, natural language processing , and network analysis […].(*) In order to improve the predictive accuracy of IOT applications, researchers employed deep learning to model complicated sensing tasks . […] Yet their design uses traditional CNNs and RNNs, combining the real and imagery parts of complex-value inputs as additional features. To the best of our knowledge, STFNet is the first work that integrates neural networks with traditional time-frequency analysis, and designs fundamental spectral-compatible operations for Fourier-transformed representations. […]

Source Document: We present DEEPWALK, a novel approach for learning latent representations of vertices in a network. These latent representations encode social relations in a continuous vector space, which is easily exploited by statistical models. […] DEEPWORK is also scalable. it is an online learning algorithm which builds useful incremental results, and is trivially parallelizable. these qualities make it suitable for a broad class of real world applications such as network classification, and anomaly detection.

Suspicious Document: According to our view love is rewarded with pleasure. The pleasure we feel here below in intellectual work proves nothing, for it is due to the effort and the passing from potential knowledge to actual knowledge, i. e., to the process of learning. Proof of this is that we find no pleasure in axioms and first principles, which we know without effort, but the acquired intellect after the death of the body does not learn any new truths, hence can have no pleasure. (*) […] Then it is prepared for immortality as a natural thing without regard to reward. The purpose of the soul as we showed is to love god. This object the bible attains by the commandments, which may be classified with reference to their significance in seven groups.

Source Document: Proof of this is that we find no delight in axioms and first values, which we understand without effort. But the came by intellect after the death of the body does not discover any new realities, therefore can have no pleasure. The rabbis furthermore talk of decisive locations of pay and penalty, which will not request to the came by intellect, since it is a separate matter and can have no place. […]

Figure 4: Example of BERT HAN SHALLOW's prediction on citation recommendation (above) and plagiarism detection (below). The attention scores produced for each sentence by HAN SHALLOW are represented by the degree of blueness. The positive sentence is marked with an asterisk at the end.