Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 4739–4748
Marseille, 11–16 May 2020

Is Language Modeling Enough? Evaluating Effective Embedding Combinations

Rudolf Schneider*, Tom Oberhauser*, Paul Grundmann*, Felix A. Gers*
Alexander Löser*, Steffen Staab†
*Beuth University of Applied Sciences Berlin, Luxemburger Str. 10, 13353 Berlin
{ruschneider, toberhauser, pgrundmann, gers, aloeser}@beuth-hochschule.de
†University Stuttgart, Universitaetsstrasse 32, 70569 Stuttgart, Germany
†University of Southampton, University Road, Southampton, SO17 1BJ, United Kingdom
Steffen.Staab@ipvs.uni-stuttgart.de

Abstract

Universal embeddings, such as BERT or ELMo, are useful for a broad set of natural language processing tasks like text classification or sentiment analysis. Moreover, specialized embeddings also exist for tasks like topic modeling or named entity disambiguation. We study if we can complement these universal embeddings with specialized embeddings. We conduct an in-depth evaluation of nine well known natural language understanding tasks with SentEval. Also, we extend SentEval with two additional tasks to the medical domain. We present PubMedSection, a novel topic classification dataset focussed on the biomedical domain. Our comprehensive analysis covers 11 tasks and combinations of six embeddings. We report that combined embeddings outperform state of the art universal embeddings without any embedding fine-tuning. We observe that adding topic model based embeddings helps for most tasks and that differing pre-training tasks encode complementary features. Moreover, we present new state of the art results on the MPQA and SUBJ tasks in SentEval.

Keywords: representation learning, meta embedding, evaluation, neural language representation models

1. Introduction

Universal embeddings, such as BERT (Devlin et al., 2019) or ELMo (Peters et al., 2018), are an effective text representation (Nguyen et al., 2016; Conneau and Kiela, 2018). Often, they are trained on hundreds of millions of documents with an language modeling objective and contain millions to even billions of parameters. These pre-trained vectors lead to significant increases in performance in various downstream natural language processing tasks (Mikolov et al., 2013a; Joulin et al., 2017; Akbik et al., 2018; Peters et al., 2018; Radford et al., 2018). Contrary to universal embeddings, specialized embeddings for tasks like entity linking (Pappu et al., 2017; Gillick et al., 2019) or paragraph classification (Arnold et al., 2019) exist. Often, specialized embeddings are trained with objectives and training datasets different from universal embeddings. This circumstance raises the question if universal embeddings capture all useful features for downstream tasks or if specialized embeddings may provide complementary features.

Example: Clinical Decision Support Systems. Medical literature databases, such as PubMed[1] or UpToDate[2], help doctors answer their questions. These systems benefit from methods that enrich texts with semantic concepts, like entity recognition, sentence classification, topic classification, or relation extraction (Demner-Fushman et al., 2009; Berner, 2007). Medical language is highly specialized and often ambiguous in clinical documents (Leaman et al., 2015). Documents, such as medical research papers, doctors' letters or clinical notes, are heterogeneous in terms of structure, vocabulary, or grammatical correctness (Starlinger et al., 2017). We propose to complement universal embeddings with specialized embeddings to execute common downstream tasks for clinical decision support systems (CDSS). Examples are paragraph classification, subjectivity classification, question type classification, sentiment analysis and textual similarity.

Problem definition. We hypothesize that specialized neural text representations may complement universal embeddings. Given is a set of both universal and specialized embeddings with different pre-training tasks for the English language (see Table 1). These embeddings encode words, entities, or topics. We study which combination of embeddings is complementary using the *SentEval*[3] (Conneau and Kiela, 2018) benchmark. Thus, we investigate if universal embeddings capture the same features as specialized embeddings.

Probing embeddings with SentEval. We probe single embeddings and combinations with *SentEval* in a transfer-learning setting on nine different tasks. *SentEval* focuses on news and customer reviews. The language in these domains differs vastly from the medical domain. Moreover, *SentEval* concentrates on single sentence evaluation, which does not fully utilize the capabilities of contextualized embedding models (Peters et al., 2018; Devlin et al., 2019; Arnold et al., 2019).

Novel datasets. We tackle the shortcomings of *SentEval* by integrating the *WikiSection-Diseases*[4] (Arnold et al., 2019) dataset into the *SentEval* framework. WikiSection also enables an in-depth evaluation of contextualized embeddings since its paragraph classification task is multi-sentence based. As the language in CDSS resources (e.g., PubMed) differs from the Wikipedia-based *WikiSection*

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] https://www.uptodate.com/
[3] https://github.com/facebookresearch/SentEval
[4] https://github.com/sebastianarnold/WikiSection

Name	Pre-Training Task	Domain	Publication	Class
ELMo (EL)	Language Modeling	Web	(Peters et al., 2018)	Universal
BERT (BE)	Language Modeling	Web	(Devlin et al., 2019)	Universal
FastText (FT)	Language Modeling	Web	(Mikolov et al., 2018)	Universal
Pappu (PA)	Entity Linking	Wikipedia	(Pappu et al., 2017)	Specialized
SECTOR (Wikipedia) (SW)	Neural Topic Modeling	Wikipedia	(Arnold et al., 2019)	Specialized
SECTOR (PubMed) (SP)	Neural Topic Modeling	Medical	-	Specialized

Table 1: Comparison of neural text embeddings.

dataset, we propose the *PubMedSection*[5] dataset. PubMedSection is a novel medical topic classification dataset created with a method inspired by distant supervision.

In-depth experimental evaluations on 11 tasks. We study properties of single and combined text embeddings and their performance on the nine tasks from *SentEval* and on the two medical datasets, *WikiSection* and *PubMedSection*. Our focus is on examining the differences between universal and specialized embeddings and effective embedding combinations.

The remainder of this paper is structured as follows: Section 2 reviews embeddings and work on integrating embeddings. Section 3 introduces our novel datasets while Section 4 describes our setup. In Section 5 we show and discuss quantitative results from our comprehensive analysis. We conclude in Section 6 and propose future research directions.

2. Related Work

In the following we investigate universal and specialized embeddings shown in Table 1 and discuss methods for combining embeddings.

2.1. Universal Text Embeddings

Recently, researchers explore universal text embeddings trained on extensive Web corpora, such as the *Common Crawl*[6] (Mikolov et al., 2018; Radford et al., 2019), the billion word benchmark (Chelba, 2010; Peters et al., 2018) and *Wikipedia* (Bojanowski et al., 2017). Universal text embeddings often perform language modeling tasks where the model is asked to predict a missing word given a small window of neighboring words (Mikolov et al., 2013b; Joulin et al., 2017; Mikolov et al., 2018; Pennington et al., 2014). Another common task is to predict the next, or masked word of a sentence given previously predicted words as context (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019). For the encoder-decoder architecture, Kiros et al. (2015) propose an encoder network that encodes a sequence of words in such a way that the decoder can predict the previous and the next sentence given the encoder's vector representation.

Universal embeddings vary in their granularity at the sub-word, word, or sentence level. For example, Bojanowski et al. (2017) improve the model of Mikolov et al. (2013b) by adding sub-word information to handle ambiguous spelling or typos. This sub-word embedding takes advantage of the fact that similarly spelled words often also have a similar meaning.

Universal text embeddings encode the meaning of frequent words (Peters et al., 2018; Devlin et al., 2019; Radford et al., 2019). However, they perform worse in comparison with domain adapted representations in specialized domains (Sheikhshabbafghi et al., 2018; Lee et al., 2019). Furthermore, universal text embeddings might miss essential aspects about named entities. The reason is that most training methods base on the co-occurrence of words in relatively short local contexts. This hinders the models to capture more global features of texts such as genre, topic, receiver, the authors' intention or they miss to learn the precise meaning of a word in special domains such as medicine (Sheikhshabbafghi et al., 2018; Lee et al., 2019). Also, computing embedding models for highly regulated domains is often hard and not feasible due to the lack of training data (Berner, 2007; Starlinger et al., 2017) or high computational costs.

2.2. Specialized Text Embeddings

Neural topic modeling. Arnold et al. (2019) introduce a specialized embedding using a coherent topic modeling task for pre-training. This model encodes both structural and topical facets of documents (see work of MacAvaney et al. (2018)) and assigns each sentence in a document a dense distributed representation of occurring latent topics (Blei, 2012). For this purpose, the model consolidates the topical structure and context over the entire document. It leverages sequence information on the granularity of paragraphs and sentences using a Bidirectional LSTM architecture (Graves, 2012) with forget gates (Gers et al., 2000). In addition, this model captures long-range topical information. However, it does not focus on disambiguating single words. Therefore, we suggest complementing universal text embeddings (disambiguation task) with neural topic models (paragraph classification task).

Neural entity embeddings. Pappu et al. (2017) and Gillick et al. (2019) encode meanings of entities for entity candidate retrieval and entity disambiguation tasks. The model of Pappu et al. (2017) builds on ideas of Le and Mikolov (2014) and models an entity using local token context. It generalizes over multiple documents as well as co-occurrences of entities in a document with a shared neural representation. This joint approach enables the model to capture world knowledge regarding entities from training data. This approach delivers a vector representation for each entity mention, encodes its relatedness to other entities and takes local context into account. However, such entity

[5] https://pubmedsection.demo.datexis.com

[6] https://commoncrawl.org/

embeddings capture facets of named entities but might fail to encode topical structure or non-entity words. Hence, we hypothesize entity embeddings might benefit from a combination with topical embeddings.

Biomedical domain specialization. Sheikhshabbafghi et al. (2018) show a domain adapted version of ELMo (Peters et al., 2018). Their contextualized word representation performs better than a general-purpose variant even with a smaller training set. However, this model cannot generalize to out of domain contexts. Therefore, Lee et al. (2019) propose BioBERT, which is a BERT model adapted to the biomedical domain. They initialize this model with pre-trained weights of the original BERT. This method prevents shortcomings of Sheikhshabbafghi et al. (2018) and preserves the ability to generalize to other domains than biomedical text.

2.3. Combining Embeddings

Multi-modal combinations. Previous research reports that combining embeddings with training objective is effective, such as combining representations of data with different modalities into a single shared vector space. For example, Heinz et al. (2017) integrate customer and image data in a shared vector space and show its effectiveness for recommending products. Wang et al. (2017) combine text and image embeddings in the field of computer vision. They employ a neighborhood preserving ranking loss to learn a non-linear mapping between image and word embeddings for image captioning tasks.

Combining neural text embeddings. To the best of our knowledge, we are the first investigating effective combinations of universal with specialized text embeddings in an extensive study on 11 tasks. In contrast, most related work focuses on novel combination methods.

Kiela et al. (2018) and Rettig et al. (2019) study methods to automatically select universal purpose word embeddings that are best suited for a particular task. Kiela et al. (2018) use an attention mechanism to learn a task-specific mixture mapping between multiple word embeddings dynamically. In contrast, Rettig et al. (2019) report a method to compare and rank word embeddings regarding their relevance to a given domain. Muromägi et al. (2017) learn a linear mapping to combine various word embeddings trained on the same dataset with the same method but with different random initialization into an ensemble. They use the *ordinary least squares problem* and the *orthogonal Procrustes problem* in their objective function. The method of Yin and Schütze (2015) is similar to Muromägi et al. (2017) but employs no orthogonality constraint on the objective function. Bollegala et al. (2018) introduce a local linear mapping method that takes local neighborhoods into account when projecting source embeddings into a combined vector space. This method has similarities to the work of Wang et al. (2017). Coates and Bollegala (2018) presents a surprisingly effective method to combine universal embeddings by averaging word vectors and padding them with zeros to compensate dimensionality mismatches. However, our focus lies in studying effective embedding combinations for medical documents.

3. Medical Topic Classification Dataset

The capabilities of contextualized embeddings cannot be measured with the SentEval framework because all of its natural language understanding tasks are single sentence-based. None of the tasks in SentEval evaluates the domain independence of the tested embeddings. To measure such embeddings and their combinations, we extend SentEval with tasks that require to track contexts that span over multiple sentences. Detecting coherent topics on document passages is a challenging task that requires to keep track of the overall context of a paragraph or even whole document.

3.1. The WikiSection Dataset

The *WikiSection* dataset (Arnold et al., 2019) consists of 38k comprehensively annotated Wikipedia articles $D = (S, L, H)$ with section and topic labels L and naturally contained headings H with respect to all of its sentences S. The dataset covers up to 30 topics about diseases (e.g., symptoms, treatments, diagnosis) or cities (e.g., history, politics, economy, climate). The task is to split Wikipedia articles d_w into a sequence of distinct topic sections $L = [l_1, ..., l_n]$, so that each predicted section $l_n = (S_k, t_j, h_i)$ contains a sequence of coherent sentences $S_k = s_1, ..., s_m$, and is associated to a heading h_i, and a topic label t_j that describes the common topic in these sentences.

3.2. Creating the PubMedSection Dataset

We introduce PubMedSection, a topic classification dataset based on medical research articles. This task requires to detect and classify structural topic facets in plain text and is inspired by the WikiSection dataset. The PubMedSection dataset consists of 51,500 PubMed articles section-wise annotated with topic labels. We construct PubMedSection similar to the WikiSection dataset. Our focus is on the disease subset of WikiSection with section-wise annotated medical topics, which we aim to transfer to PubMed articles. Our initial PubMed collection consists of 2,142,050 articles with 29,522,566 headings. Steps to create the dataset include learning a classifier for detecting articles in PubMed similar to WikiSection and assigning labels.

Learning to classify relevant articles. Labeling such a large dataset is time-consuming and costly. Following this, we annotate the PubMedSection articles using distant supervision (Mintz et al., 2009; Morgan et al., 2004) with WikiSection as ground truth. For this purpose, we filter the open-access subset of PubMed[7] D_p for articles that exhibit a high textual similarity to WikiSection for a successful label transfer. We model a neural network based non-linear binary classifier for this task[8]. First, we encode all headlines of the WikiSection diseases subset $H_w = \{h_{w1}, ..., h_{wn}\}$ as well as the headlines of the PubMed articles $H_p = \{h_{p1}, ..., h_{pn}\}$ with a fastText (Mikolov et al., 2018) embedding model. For this step we train a domain-specific fastText model on the full corpus of the open-access subset of PubMed. Next, we use concatenated fastText encoded word vectors of each article's headlines

[7] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

[8] https://github.com/DATEXIS/pubmedsection

Name	Embedding A	Embedding B
FT+PA	fastText	Pappu
FT+SW	fastText	SECTOR (Wikipedia)
FT+SP	fastText	SECTOR (PubMed)
EL+FT	ELMo	fastText
EL+PA	ELMo	Pappu
EL+SW	ELMo	SECTOR (Wikipedia)
EL+SP	ELMo	SECTOR (PubMed)
SW+PA	SECTOR (Wikipedia)	Pappu
SP+PA	SECTOR (PubMed)	Pappu

Table 2: Surveyed embedding combinations.

H_w, H_p as input for our model. We choose a one-layer neural network with ReLu activation (Glorot et al., 2011) and softmax output over more complex architectures to minimize computational complexity. We train on 3200 human-labeled examples for the headline structure similarity task. We use the Xavier weight initialization (Glorot and Bengio, 2010), and employ Adam (Kingma and Ba, 2015) with stochastic gradient descent as optimizer and a multi-class cross-entropy loss. Our hyperparameter search suggests an L2 regularization (Ng, 2004) of 10^{-4}, a learning rate of 10^{-5}, a batch size of 128 and we set the training duration to 60 epochs.

Assigning WikiSection labels to PubMedSection. After training, we sample the top 51,500 articles by their similarity score from the filtered PubMed collection. Next, we calculate the cosine similarity between every headline of each article set (D_p, D_w) to estimate the probability that a topic label for the PubMed headline could be generated from the Wikisection labels. Next, we transfer the best fitting topic labels from the best matching headline's section in WikiSection to the sampled PubMed article's corresponding section. After that, we split the dataset in a training subset with 50.000, a validation subset with 1000, and a test subset with 500 labeled articles. Finally, we validate the PubMedSection dataset with two human judges that evaluate 100 randomly sampled articles for correctness.

4. Evaluation Methodology

Methodology overview. We evaluate the performance of embeddings as well as their combinations. Our methodology follows the paradigm of probing tasks (van Aken et al., 2019; Weston et al., 2015): We test combined embeddings on overall nine natural language understanding tasks and data sets from SentEval as well as two tasks from the WikiSection and the PubMedSection dataset. For probing these eleven tasks, we train a linear classifier with single or combined embeddings as input and observe properties of different embedding types and their combinations. As combination method we chose *concatenation*. Despite its simplicity, concatenating embeddings has been shown to be a strong baseline (Yin and Schütze, 2015; Coates and Bollegala, 2018; Kiela et al., 2018; Rettig et al., 2019). Other combination methods are subject to our future research.

4.1. Text Embeddings and Combinations

We select a variety of universal and specialized embeddings as shown in Table 1 for our experiments. Our evaluation setting is sentence-based. Some of the surveyed embeddings are word vector oriented. Therefore, we follow Arora et al. (2017) and Perone et al. (2018) and average word vectors in a sentence for each of those word embeddings to obtain a sentence embedding vector. The embeddings employed are:

Random (RND) As a baseline, we compute random vectors.

fastText (FT) (Mikolov et al., 2018) is word vector oriented and trained on a language modeling task with word and sub-word tokens.

ELMo (EL) by Peters et al. (2018) train a bi-directional language modeling task with two stacked LSTMs that use a Character-CNN to capture sub-word information.

BERT (BE) (Devlin et al., 2019) bases on the transformer architecture (Vaswani et al., 2017) and masked language modeling task pre-training.

Entity embedding (PA) Pappu et al. (2017) train an embedding for a named entity disambiguation task with a knowledge base as target, like Wikidata[9] or UMLS[10].

SECTOR Wikipedia (SW) Arnold et al. (2019) propose a contextual topic embedding which is trained with section headings from Wikipedia articles. They show that the latent topic information contained in their SECTOR embedding can be utilized to segment documents and to classify these segments into up to 30 topics.

SECTOR PubMed (SP) Same as above but trained on our novel PubMedSection dataset.

Embedding combinations. We choose the combinations of embeddings presented in Table 2 for our experiments. We assume that the most compelling improvements are obtained when we combine combine specialized with universal text embeddings. We verify this assumption by evaluating if combining the two universal embeddings ELMo and fastText is as effective as combinations with specialized embeddings. Additionally, we conduct experiments with the combination of the entity embedding with both SECTOR models.

Embedding models. We evaluate the following models as provided by their authors: *BERT Large* (BE), *ELMo Original 5.5B* (EL), *fastText crawl-300d-2M-subword* (FT), *Pappu* (PA) and *SECTOR SEC>H+emb@fullwiki* (SW). These models cover a wide variety of domains and topics. This is contrary to our SECTOR PubMed (SP) model that we train exclusively on medical research articles.

4.2. Tasks and Parameter

We use SentEval (Conneau and Kiela, 2018) to perform an analysis of the effectiveness of each embedding combination of natural language understanding tasks. We integrate the WikiSection dieseases and PubMedSection task

[9] https://www.wikidata.org

[10] https://www.nlm.nih.gov/research/umls/index.html

into SentEval[11] to obtain comparable results for our evaluation. Overall, we conduct our survey on the following nine plus two medical tasks WikiSection and PubMedSection with ten-fold cross-validation.

Textual similarity: MRPC Paraphrase detection (Dolan et al., 2004) on the Microsoft Research Paraphrase Corpus which consists of sentences pair extracted from *news* sources. It is a binary classification task of deciding whether a sentence paraphrases the other or not.

Textual similarity: SICK-E Sentences Involving Compositional Knowledge-Entailment (Marelli et al., 2014) is a 3 classes natural language inference classification task based on sentences collected from Flickr *image captions* and the Microsoft Research *Video Description* Corpus.

Sentiment analysis: MPQA Multi-Perspective Question Answering (Wiebe et al., 2005) is a binary sentiment classification task on a *news dataset* from the world press.

Sentiment analysis: SST-2 Stanford Sentiment Analysis (Socher et al., 2013) is a binary sentiment classification task on a *movie review* data set.

Sentiment analysis: SST-5 Stanford Sentiment Analysis (Socher et al., 2013) is fine-grained 5 class sentiment analysis task based on the same corpus as SST-2 (*movie review*).

Sentiment analysis: CR Customer Reviews (Hu and Liu, 2004) is a binary sentiment analysis task based on *product reviews*.

Sentiment analysis: MR Movie Reviews (Pang and Lee, 2005) is a binary sentiment analysis data set on *movie reviews*.

Classification: SUBJ Subjectivity vs. Objectivity (Pang and Lee, 2004) is a classification task of subjectivity and objectivity on *movie reviews*.

Classification: TREC Text Retrieval Conference Question Answering (Voorhees and Tice, 2000) 6 class question type classification. The corpus consists mostly of *newswire and newspaper* articles.

Coherent topic classification: WikS WikiSection diseases (Arnold et al., 2019) is a 27 class topic classification task sourced from the *medical subset of Wikipedia*.

Coherent topic classification: PubS PubMedSection is a novel 27 class topic classifcation task based on medical research articles from *PubMed*. We randomly sample the PubMedSection training set down to 2200 articles since evaluating with the whole training set is prohibitively time-consuming.

Evaluation parameters. We use the parameters provided by Conneau and Kiela (2018), as shown in Table 3.

5. Experimental Results and Discussion

Table 4 overviews the results of seven single embeddings as well as nine embedding combinations on eleven evaluation tasks. Table 5 shows accuracy scores for single model performance and of combined embedding models. Finally, Table 6 reveals the delta of each surveyed embedding combinations' score regarding their source embedding scores.

[11] https://github.com/DATEXIS/SentEval-k8s

Parameter	Value
KFOLD	10
CLASSIFIER_NHID	0
CLASSIFIER_OPTIM	Adam
CLASSIFIER_BATCHSIZE	64
CLASSIFIER_TENACITY	5
CLASSIFIER_EPOCHSIZE	4
CLASSIFIER_DROPOUT	0

Table 3: Parameters used in evaluation with SentEval as suggested by Conneau and Kiela (2018).

Model	Strong +	Minor +	Minor -	Strong -
Language Model combined with Topic Model				
EL+SW	**7**	2	1	1
EL+SP	**5**	3	2	1
FT+SW	**6**	3	0	2
FT+SP	**7**	2	2	0
Language Model combined with Entity Embedding				
EL+PA	0	**6**	4	1
FT+PA	**6**	5	0	0
Topic Model combined with Entity Embedding				
SW+PA	**5**	2	1	3
SP+PA	**6**	4	0	1
Language Model + Contextualized Language Model				
EL+FT	0	**8**	3	0

Table 4: This table shows the effectiveness classification in tasks for each surveyed embedding combination. We count a model combination as "Strong+" if it advances in more than one percentage point in accuracy compared to both of its base models. Accordingly, we count a result as "Minor+" if the improvement is smaller than one percentage point. "Minor-" and "Strong-" are similarly defined for performance decreases.

5.1. Language Models plus Topic Models

We observe an significant increase in accuracy scores in 35 out of 44 experiments (see Table 4) which qualify in 25 cases for the "Strong+" category when combining a language modeling based embedding with a topic embedding. Moreover, we report EL+SP as the overall best performing model with a macro accuracy across all tasks of 75.83. We conclude that language modeling and topic modeling pretraining tasks capture complementary information.

ELMo plus SECTOR yields a substantial increase in accuracy. The combination of EL and SW yields "Strong+" results (see Table 4) for 7 of the 11 downstream tasks. We observe only a considerable performance loss of 3.8 percentage points for the TREC task. For the three other measurements of this model, the performance increases slightly for two tasks by less than 0.33 accuracy points and decreases for the SICK-E task by 0.06 accuracy points (see also Table 6). EL and SP also yields strong results, with five tasks for which the source models encode complementary information. We observe only one "Strong-" loss in performance of the SST-5 task of 1.62 and report for all

	Textual Similarity		Sentiment Analysis					Classification			
Model	MRPC	SICK-E	MPQA	SST-2	SST-5	CR	MR	SUBJ	TREC	WikS	PubS
RND	66.49	56.69	68.77	49.92	23.39	63.76	49.48	49.60	20.60	15.24	24.99
FT	69.86	74.35	86.69	78.69	39.68	72.00	74.68	90.22	76.00	39.11	28.15
PA	72.17	76.62	85.31	77.43	40.00	74.09	72.90	89.65	78.80	39.84	28.04
SW	67.71	67.30	84.30	65.68	34.80	80.34	77.56	*97.84*	69.60	29.82	29.27
SP	67.19	56.48	*96.85*	71.28	37.65	76.19	82.91	95.61	66.20	*46.84*	*39.43*
BE	69.16	75.75	86.91	***89.57***	*49.37*	***90.07***	*84.84*	95.83	***93.20***	44.94	31.12
EL	*73.68*	*79.54*	90.00	85.01	47.19	83.39	80.66	94.56	92.40	43.09	30.85
Language Model combined with Topic Model											
EL+SW	73.86	79.48	92.58	85.34	**49.59**	87.23	86.25	**99.17**	88.60	45.05	32.11
EL+SP	**74.61**	78.87	96.14	86.66	45.57	84.53	87.03	97.26	92.80	**50.86**	39.76
FT+SW	70.78	74.51	90.35	76.28	40.18	82.91	83.49	98.13	73.60	42.64	31.24
FT+SP	70.96	74.33	**97.34**	77.81	43.71	80.69	**87.11**	97.27	78.60	49.60	**39.85**
Language Model combined with Entity Embedding											
EL+PA	73.45	79.81	90.27	85.94	45.97	83.47	80.91	94.40	92.80	42.67	30.70
FT+PA	72.99	79.07	87.03	81.11	41.95	76.42	75.54	91.17	84.80	41.36	28.78
Topic Model combined with Entity Embedding											
SW+PA	71.65	74.79	89.92	75.40	40.68	82.89	83.60	98.43	77.60	43.18	31.16
SP+PA	72.70	77.15	97.16	75.95	44.34	81.14	87.05	97.36	85.80	50.13	39.63
Language Model combined with Contextualized Language Model											
EL+FT	73.33	**79.91**	90.19	85.78	46.65	83.76	80.79	94.46	93.00	43.24	30.97
SOTA	93.00[c]	87.80[d]	93.30[b]	96.80[c]	64.40[e]	87.45[a]	96.21[c]	95.70[f]	98.07[a]	56.70[g]	-

Table 5: This table shows the accuracy score of single model approaches and the best embedding combinations for each task. We highlight the overall best score with **bold** numbers while numbers in *italic* denote the best single model results. Additionally, we gathered recent results on our surveys tasks in the SOTA row, which are reported by the following publications: (Cer et al., 2018)[a], (Zhao et al., 2015)[b], (Yang et al., 2019)[c], (Subramanian et al., 2018)[d], (Patro et al., 2018)[e], (Tang and de Sa, 2018)[f] and (Arnold et al., 2019)[g] on section-wide topic classification. We do not take SOTA results into account when highlighting best results since they are obtained with specialized models.

remaining tasks a fluctuation in performance between "Minor+" and "Minor-".

fastText and SECTOR encode complementary features. Results for fastText plus SECTOR are nearly analogue to ELMo, except that we observe an even higher performance increase on average as shown in Table 6. We note that tasks MRPC, SICK-E, SST-2 do not benefit from the features captured in SW and SP. Surprisingly, the situation for the fine-grained sentiment classification task SST-5 is different compared to results of the binary sentiment analysis task SST-2. We observe a considerable accuracy increase for the model combination EL+SW and FT+SP, a small increase for FT+SW, and a performance decrease for EL+SP.

EL+SP outperforms EL+SW in the medical domain. Corresponding to the differing training domains of the SW and SP model, we can observe a more substantial increase in performance for the combination of EL and SP in both medical tasks WikiSection-Diseases and PubMedSection compared to the combination of EL and SW. Likewise, we observe a similar situation for the combination of fastText (FT) with SW and SP. This can be explained by the fact that SP is trained on medical research articles and therefore closer to the target domain than SW.

New SOTA for MPQA and SUBJ tasks. Table 5 shows that embedding combinations EL+SP (96.14 acc) and FT+SP (97.34 acc) outperform the current state of the art in the MPQA task (see Zhao et al. (2015) 93.30 acc). An analogue EL+SW, drastically outperforms the current state of the art (see (Tang and de Sa, 2018) 95.70 acc) in the SUBJ task with 99.17 accuracy measure. Following this result, we conclude that the differing pre-training task captures complementary features that lead to improved evaluation results.

Different pre-training tasks capture complementary features. We verify the complementary nature of the pre-training tasks with an additional experiment. We evaluate the SUBJ task again with a fastText model that was, similar to SW, exclusively trained on Wikipedia (Bojanowski et al., 2017). With this setting, we control if the objective writing style in Wikipedia is the cause of our good results. We observe only a small increase in accuracy for the Wikipedia based model (90.98 Acc) compared to the FastText model trained on the Common Crawl (90.22 Acc). Following this result, we conclude that different pre-training tasks of FT and SW capture complementary features that lead to improved evaluation results. We explain the complementary nature of these combinations with the document-wide context that topic models encode. Topic models need to keep track of the context coherently over whole documents while respecting local topic shifts. That is contrary to language modeling based embeddings that often focus mainly on local context spanning over nearby sentences.

	Textual Similarity		Sentiment Analysis					Classification			
Comb. Δ	MRPC	SICK-E	MPQA	SST-2	SST-5	CR	MR	SUBJ	TREC	WikiS	PubS
Language Model combined with Topic Model											
EL+SWΔEL	0.18	-0.06	2.58	0.33	2.40	3.84	5.59	4.61	-3.80	1.96	1.26
EL+SWΔSW	6.15	12.18	8.28	19.66	14.79	6.89	8.69	1.33	19.00	15.23	2.84
EL+SPΔEL	0.93	-0.67	6.14	1.65	-1.62	1.14	6.37	2.70	0.40	7.77	8.91
EL+SPΔSP	7.42	22.39	-0.71	15.38	7.92	8.34	4.12	1.65	26.60	4.02	0.33
FT+SWΔFT	0.92	0.16	3.66	-2.41	0.50	10.91	8.81	7.91	-2.40	3.53	3.09
FT+SWΔSW	6.15	12.18	8.28	19.66	14.79	6.89	8.69	1.33	19.00	15.23	2.84
FT+SPΔFT	1.10	-0.02	10.65	-0.88	4.03	8.69	12.43	7.05	2.60	10.49	11.70
FT+SPΔSP	3.77	17.85	0.49	6.53	6.06	4.50	4.20	1.66	12.40	2.76	0.42
Language Model combined with Entity Embedding											
FT+PAΔFT	3.13	4.72	0.34	2.42	2.27	4.42	0.86	0.95	8.80	2.25	0.63
FT+PAΔPA	0.82	2.45	1.72	3.68	1.95	2.33	2.64	1.52	6.00	1.52	0.74
EL+PAΔEL	-0.23	0.27	0.27	0.93	-1.22	0.08	0.25	-0.16	0.40	-0.42	-0.15
EL+PAΔPA	1.28	3.19	4.96	8.51	5.97	9.38	8.01	4.75	14.00	2.83	2.66
Topic Model combined with Entity Embedding											
SW+PAΔSW	3.94	7.49	5.62	9.72	5.88	2.55	6.04	0.59	8.00	13.36	1.89
SW+PAΔPA	-0.52	-1.83	4.61	-2.03	0.68	8.80	10.70	8.78	-1.20	3.34	3.12
SP+PAΔSP	5.51	20.67	0.31	4.67	6.69	4.95	4.14	1.75	19.60	3.29	0.20
SP+PAΔPA	0.53	0.53	11.85	-1.48	4.34	7.05	14.15	7.71	7.00	10.29	11.59
Language Model combined with Contextualized Language Model											
EL+FTΔEL	-0.35	0.37	0.19	0.77	-0.54	0.37	0.13	-0.10	0.60	0.15	0.12
EL+FTΔFT	3.47	5.56	3.50	7.09	6.97	11.76	6.11	4.24	17.00	4.13	2.82

Table 6: This table shows the delta in accuracy score of each model combination with respect to the respective single model accuracy score. We highlight numbers in green if an embedding combination yields improved scores compared to both source embeddings.

5.2. Combinations with Entity Embeddings

Table 4 shows "Strong+" increases in accuracy for 17 out of 44 experiments for embedding combinations that include the surveyed entity embedding (PA).

Topic plus entity embeddings outperform. We examine the combination of the topic (SW, SP) and entity embeddings (PA) in Tables 4 and 6. Intuitively, it seems reasonable to assume that topic embeddings focus more on structure than on the meaning of single words and, therefore, capture complementary knowledge. Our results prove this assumption with 17 out of 22 experiments that show an increase in performance and 11 scores, that qualify as "Strong+." Similar to the results when combining topic and language models, we explain the performance gains with the complementary nature of the entity disambiguation and topic modeling pre-training tasks. Additionally, we note that PA does not encode any contextual information at prediction time while SW and SP do. Following this, it is reasonable to assume that the combinations SP+PA and SW+PA are generally beneficial.

Combining fastText and Pappu is beneficial. For 6 out of 11 tasks is our complementary constraint in Table 4 fulfilled, the remaining tasks have a "Minor+" accuracy increase, lower than one percentage point. We observe that FT+PA is a beneficial combination since no task has a drop in accuracy.

ELMo already captures features encoded by Pappu. On the contrary, we observe no accuracy gain over one percentage point for EL+PA. We observe six times a minor increase, four times a minor decrease and one time strong decrease. As reported in Table 6 this strong decrease is accounted to the SST-5 task with a loss of 1.22 percentage point compared to the single model result of EL. Overall we observe that this combination yields results that are comparable to the single model performance of EL (see Table 4.2.). This result suggests that the contextualized nature of EL already captures the features encoded by PA.

5.3. Baseline and Domain Transfer

To validate our results, we survey if adding more semantically meaningful dimensions to a vector is sufficient to obtain results comparable to our experiments. Therefore, we evaluate combining a contextualized (ELMo) with a traditional language model (fastText). Next, we report the results of the single model evaluation of contextualized and traditional language models on WikiSection and PubMedSection. Finally, we survey if we can enrich a universal embedding (ELMo or fastText) with domain-specific features (SP) without losing its domain independence.

ELMo plus fastText has no effect. We report no result which qualifies as either "Strong+" nor "Strong-" in Table 4 for EL+FT. In six out of nine cases, we observe a slight increase in accuracy, and in three cases, minor decreases. Intuitively it is sound to assume that contextualized em-

beddings (EL) should not benefit from static word embedding (FT) methods. Correspondingly, we evaluate, on the one hand, the combination EL+FT in order to investigate this intuition and, on the other hand, to obtain a baseline. Consequently, we conclude that adding more semantically meaningful dimensions to text representations alone is not sufficient to achieve good results comparable to the other surveyed combinations.

Classical embeddings perform surprisingly well in multi-sentence tasks. Table 5 reports surprisingly well results for non-contextualized embeddings (FT and PA) in the WikiSection and PubMedSection tasks. Their best results on WikiSection (39.84 acc) and PubMedSection (28.15 acc) are quite close to the contextualized universal embeddings BE and EL (WikS: 44.94 acc, PubS: 30.85 acc). These results are contrary to our initial assumption that the contextualized embeddings would vastly outperform FT and PA on multi-sentence based tasks.

Domain specificity. We observe 18 times a "Strong+" increase in accuracy for the 33 experiments that involve SP, which is trained on PubMed abstracts (see Table 6 and Table 4). Therefore, we can confirm the observation of Lee et al. (2019) and Sheikhshabbafghi et al. (2018) that in-domain text representations perform better on biomedical texts than universal representations. Moreover, we can show that it is possible to transfer the domain adaption into a combined embedding without experiencing catastrophic forgetting since we only observe three out of the 18 "Strong+" increases in the medical tasks (WikS, PubS). For example, as shown in Table 5 the combinations of EL+SP and SP+PA deliver the best results in our evaluation for the WikiSection disease task while being in the top three surveyed embedding combinations.

5.4. Discussion

Adding topic models helps for most tasks. Our results suggest that adding topic models to either language models or entity embeddings is beneficial for the overall performance of most investigated classification tasks. This observation can be explained by the topical and structural information captured in these models. Moreover, these topical models capture the coherent flow of topics across long-range dependencies while taking local topic shifts into account. Therefore, neurons in these models may be able to capture long-range dependencies from long documents. This information seems to be complementary to information from universal text embeddings or entity embeddings, with a comparably short context window.

Textual similarity tasks do not benefit much. We observe for textual similarity tasks only for very few scenarios a "Strong+" improvement when combining embeddings. We argue that existing universal embeddings, such as ELMo or fastText, already represent sufficient features from local features close to the target word.

Concatenation is simple but easily interpretable. Our study is limited to concatenation as the operator for combining embeddings. This simple operator has a significant disadvantage in raising the dimensionality. Additionally, it is not leveraging the originating correlations in combined embedding spaces. However, despite these shortcomings, this operator permits to survey for effective embedding combinations in an explainable manner.

Different pre-training tasks encode different features. Our study confirms that embeddings trained with different pre-training tasks can encode complementary features. Combinations of specialized and universal embeddings often result in domain-independent performance increases.

6. Conclusion

To the best of our knowledge, we are the first investigating effective combinations of universal with specialized text embeddings in an extensive study on 11 tasks. Our comprehensive analysis shows that combining universal and specialized embeddings yields vastly improved results in many downstream tasks. Furthermore, we set a new state of the art for two tasks in SentEval by combining embeddings. We extend SentEval to the medical domain by integrating the WikiSection-Diseases and the novel PubMedSection task, covering 51,500 labeled PubMed articles.
Future research includes investigating features covered by specialized embeddings, such as presented by Arnold et al. (2020), that universal embeddings might miss, including recent models such as BERT (Devlin et al., 2019) or GPT2 (Radford et al., 2019). A deeper and also linguistically motivated understanding might lead to better choices for embedding combinations or new directions for designing pre-training tasks. Also, we will investigate further embedding combination methods for two and more embeddings.

Acknowledgments

Our work is funded by the German Federal Ministry of Economic Affairs and Energy (BMWi) under grant agreements 01MD19003E (PLASS), 01MD19013D (Smart-MD), 01MK2008D (Servicemeister) and the German Federal Ministry of Education and Research (BMBF) under grant agreement 01UG1735BX (NOHATE).
We thank the anonymous reviewers for their valuable feedback.

7. Bibliographical References

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th COLING*, pages 1638–1649, Santa Fe, New Mexico, USA, August. ACL.

Arnold, S., Schneider, R., Cudré-Mauroux, P., Gers, F. A., and Löser, A. (2019). SECTOR: A Neural Model for Coherent Topic Segmentation and Classification. *TACL*, 7:169–184, March.

Arnold, S., van Aken, B., Grundmann, P., Gers, F. A., and Löser, A. (2020). Learning Contextualized Document Representations for Healthcare Answer Retrieval. *arXiv:2002.00835*.

Arora, S., Liang, Y., and Ma, T. (2017). A Simple But Tough-To-Beat Baseline for Sentence Embeddings. In *Proceedings of ICLR-17*.

Eta S. Berner, editor. (2007). *Clinical Decision Support Systems: Theory and Practice*. Health Informatics. Springer, New York, NY, 2nd ed edition.

Blei, D. M. (2012). Probabilistic topic models. *Commun. ACM*, 55(4):77, April.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *TACL*, 5:135–146.

Bollegala, D., Hayashi, K., and Kawarabayashi, K.-i. (2018). Think Globally, Embed Locally — Locally Linear Meta-embedding of Words. In *proceedings of IJCAI*, pages 3970–3976, Stockholm, Sweden, July. IJCAIO.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal Sentence Encoder. *arXiv:1803.11175 [cs]*, March.

Chelba, C. (2010). Statistical Language Modeling. In Alexander Clark, et al., editors, *The Handbook of Computational Linguistics and Natural Language Processing*, pages 74–104, Oxford, UK, June. Wiley-Blackwell.

Coates, J. and Bollegala, D. (2018). Frustratingly Easy Meta-Embedding – Computing Meta-Embeddings by Averaging Source Word Embeddings. In *NAACL*, pages 194–198. ACL.

Conneau, A. and Kiela, D. (2018). SentEval: An Evaluation Toolkit for Universal Sentence Representations. In *LREC*. ELRA.

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *J. Biomed. Inform.*, 42(5):760–772, October.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the NAACL, Volume 1 (Long and Short Papers)*, pages 4171–4186. ACL.

Dolan, B., Quirk, C., and Brockett, C. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of the 20th COLING*, page 350. ACL.

Gers, F. A., Schmidhuber, J. A., and Cummins, F. A. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural computation*, 12(10):2451–2471, October.

Gillick, D., Kulkarni, S., Lansing, L., Presta, A., Baldridge, J., Ie, E., and Garcia-Olano, D. (2019). Learning Dense Representations for Entity Retrieval. *arXiv:1909.10506 [cs]*, September.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In Yee Whye Teh et al., editors, *Proceedings of the Thirteenth AISTAT*, volume 9, pages 249–256, Chia Laguna Resort, Sardinia, Italy, May. PMLR.

Glorot, X., Bordes, A., and Bengio, Y. (2011). Deep sparse rectifier neural networks. In Geoffrey Gordon, et al., editors, *Proceedings of the Fourteenth AISTATS*, volume 15, pages 315–323, Fort Lauderdale, FL, USA, April. PMLR.

Graves, A. (2012). *Supervised Sequence Labelling with Recurrent Neural Networks*, volume 385. Springer, Berlin Heidelberg.

Heinz, S., Bracher, C., and Vollgraf, R. (2017). An LSTM-Based Dynamic Customer Model for Fashion Recommendation. In *Proceedings of RecTemp Co-Located with 11th RecSys*, volume 1922, Como, Italy. CEUR-WS.

Hu, M. and Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the 10th ACM SIGKDD*, pages 168–177. ACM.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of Tricks for Efficient Text Classification. In *Proceedings of the 15th EACL*, volume 2, pages 427–431, Valencia, Spain. ACL.

Kiela, D., Wang, C., and Cho, K. (2018). Dynamic meta-embeddings for improved sentence representations. In *Proceedings of the 2018 EMNLP*, pages 1466–1477. ACL.

Kingma, D. and Ba, J. (2015). ADAM: A Method for Stochastic Optimization. In *ICLR'15*.

Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Skip-thought vectors. In *Proceedings of the 28th NeurIPS - Volume 2*, NIPS'15, pages 3276–3284, Cambridge, MA, USA. MIT Press.

Le, Q. V. and Mikolov, T. (2014). Distributed Representations of Sentences and Documents. In *ICML'14*, volume 32, pages 1188–1196.

Leaman, R., Khare, R., and Lu, Z. (2015). Challenges in clinical natural language processing for automated disorder normalization. *J. Biomed. Inform*, 57:28–37, October.

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2019). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, September.

MacAvaney, S., Yates, A., Cohan, A., Soldaini, L., Hui, K., Goharian, N., and Frieder, O. (2018). Characterizing Question Facets for Complex Answer Retrieval. In *SIGIR '18*, pages 1205–1208, Ann Arbor, MI, USA, June. ACM.

Marelli, M., Menini, S., Baroni, M., Bentivogli, L., bernardi, R., and Zamparelli, R. (2014). A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC-2014*. ELRA.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *arXiv:1301.3781 [cs.CL]*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. In *NIPS'13*, pages 3111–3119.

Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., and Joulin, A. (2018). Advances in Pre-Training Distributed Word Representations. In *Proceedings of LREC 2018*, Miyazaki, Japan, May. ELRA.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th ACL and the 4th AFNLP: Volume 2 - Volume 2*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. ACL.

Morgan, A. A., Hirschman, L., Colosimo, M., Yeh, A. S., and Colombe, J. B. (2004). Gene name identification and normalization using a model organism database. *J. Biomed. Inform.*, 37(6):396–410, December.

Muromägi, A., Sirts, K., and Laur, S. (2017). Linear Ensembles of Word Embedding Models. In *Proceedings of the 21st NoDaLiDa, 22-24 May 2017, Gothenburg, Sweden*, pages 96–104. Linköping University Electronic Press.

Ng, A. Y. (2004). Feature Selection, L1 vs. L2 Regularization, and Rotational Invariance. In *Proceedings of the Twenty-First ICML*, pages 78–, New York, NY, USA. ACM.

Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., and Deng, L. (2016). MS MARCO: A Human Generated Machine Reading Comprehension Dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating Neural and Symbolic Approaches 2016 Co-Located with the 30th NIPS 2016*, volume Vol-1773, page 10, Barcelona, Spain. CEUR-WS.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd ACL*, page 271. ACL.

Pang, B. and Lee, L. (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on ACL*, pages 115–124. ACL.

Pappu, A., Blanco, R., Mehdad, Y., Stent, A., and Thadani, K. (2017). Lightweight Multilingual Entity Extraction and Linking. In *Proceedings of the Tenth WSDM*, pages 365–374, New York, NY, USA. ACM.

Patro, B. N., Kurmi, V. K., Kumar, S., and Namboodiri, V. (2018). Learning Semantic Sentence Embeddings using Sequential Pair-wise Discriminator. In *Proceedings of the 27th COLING*, pages 2715–2729, Santa Fe, New Mexico, USA, August. ACL.

Pennington, J., Socher, R., and Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of EMNLP 2014*, pages 1532–1543.

Perone, C. S., Silveira, R., and Paula, T. S. (2018). Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv:1806.06259 [cs]*, June.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of NAACL*.

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving Language Understanding by Generative Pre-Training. Technical Report, OpenAI, June.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Technical Report, OpenAI, February.

Rettig, L., Audiffren, J., and Cudré-Mauroux, P. (2019). Fusing Vector Space Models for Domain-Specific Applications. In *ICTAI 2019*. IEEE.

Sheikhshabbafghi, G., Birol, I., and Sarkar, A. (2018). In-domain Context-aware Token Embeddings Improve Biomedical Named Entity Recognition. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 160–164. ACL.

Socher, R., Wu, A. P. J., Chuang, J., Manning, C. D., Ng, A. Y., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *EMNLP'13*, pages 1631–1642. ACL.

Starlinger, J., Kittner, M., Blankenstein, O., and Leser, U. (2017). How to improve information extraction from German medical records. *it - Information Technology*, 59(4), January.

Subramanian, S., Trischler, A., Bengio, Y., and Pal, C. J. (2018). Learning General Purpose Distributed Sentence Representations via Large Scale Multi-task Learning. In *International Conference on Learning Representations*, Vancouver, BC, Canada.

Tang, S. and de Sa, V. R. (2018). Improving Sentence Representations with Multi-view Frameworks. In *IRASL Colocated at NeurIPS*, page 13, Montréal, Canada, December.

van Aken, B., Winter, B., Löser, A., and Gers, F. A. (2019). How Does BERT Answer Questions?: A Layer-Wise Analysis of Transformer Representations. In *Proceedings of CIKM '19*, pages 1823–1832, New York, NY, USA. ACM.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, June.

Voorhees, E. M. and Tice, D. M. (2000). Building a question answering test collection. In *Proceedings of the 23rd SIGIR*, pages 200–207. ACM.

Wang, L., Li, S., Lv, Y., and Wang, H. (2017). Learning to Rank Semantic Coherence for Topic Segmentation. *Proceedings of EMNLP*, pages 1340–1344.

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A., and Mikolov, T. (2015). Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. *arXiv:1502.05698 [cs, stat]*, December.

Wiebe, J., Wilson, T., and Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *arXiv:1906.08237 [cs]*, June.

Yin, W. and Schütze, H. (2015). Learning Meta-Embeddings by Using Ensembles of Embedding Sets. *arXiv:1508.04257 [cs]*, August.

Zhao, H., Lu, Z., and Poupart, P. (2015). Self-adaptive hierarchical sentence model. In *Proceedings of IJCAI*, pages 4069–4076.