097400 - CAUSAL INFERENCE PROJECT
CAUSAL EFFECTS OF SOPHISTICATION IN NLP

Tal Daniel
Department of Electrical Engineering
Technion
taldanielm@campus.technion.ac.il

Eyal Ben David
Department of Industrial Engineering
Technion
eyalbd12@campus.technion.ac.il

1 INTRODUCTION

Modern neural network (NN) classifiers are hard to interpret, especially in tasks that involve text. NNs use complex functions to learn a target from input, and it is usually the case that one cannot explain the decisions made by the model. In Natural Language Processing (NLP), words and sentences are represented numerically, where this representation is usually learned by the model. We wish to analyze decisions made based on this representations, specifically on the task of sentence fusion. Sentence fusion is the task of joining several independent sentences into a single coherent text. The underlying assumption is that the two sentences are related, e.g. they have common subjects.

In this project we intend to research the causal effect of "sophistication" on text generation. To perform this task, we will use *DiscoFuse*, which is a sentence fusion dataset that holds the discourse phenomenon and marker that is used to generate each fusion. We argue that a sophistication of a fusion sentence can be inferred from its discourse phenomenon and markers, hence by generating a parallel fusion using more/less sophisticated features, we can create the counterfactual fusion and measure the total effect. The causal question we wish to answer is: what is the effect of using sophisticated words in a sentence on modern neural network models?

We try to answer this question by formulating a multi-step experiment using the latest models in NLP, including BERT and Transformer. The two main components of the experiment are a binary classifier – to classify the origin domain of a given text, and a conditional, Seq2Seq-based, generative model – to fuse sentences given a *sophistication* signal (creating counterfactuals). Having both models trained, we first evaluate the classification error on a development set, then we measure the generation quality of the generative model and finally we propose an alternative measure to the Average Treatment Effect (ATE), Model Sophistication Bias Estimation (MSBE), to measure the effect of sophistication.

Our results show that sophistication does have an effect, although it is not very significant. This aligns with our initial assumption that in order to classify the origin domain of a sentence, the classifier not only considers words from originating domain, but also considers the sophistication signal that is presented within text.

The rest of this report is structured as follows: (1) first, we define *sophistication* and give examples for sophisticated and unsophisticated words, (2) then we present the DiscoFuse dataset, (3) describe the steps we took and the models we used, (4) present the evaluation method, and finally (5) discuss the results and weaknesses.

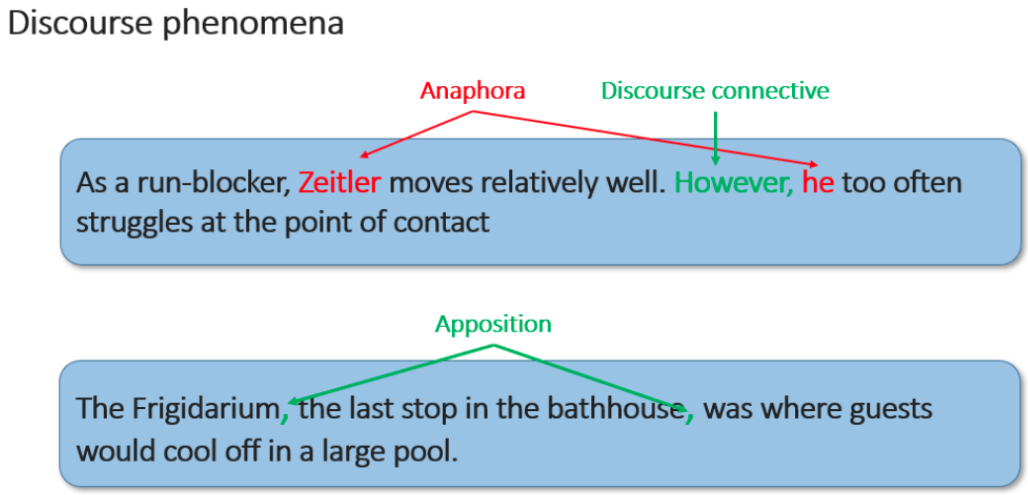

Figure 1: Discourse phenomena

Semantic Cluster	Sophisticated	Unsophisticated
Effect	*consequently, therefore*	*as a result, hence, thus*
Adding	*furthermore, moreover, plus, additionally*	*and*
Sequencing	*finally, eventually*	*in the end*
Contrasting	*whereas, nevertheless*	*on the other hand*
Qualifying	*however, although, still*	*yet, but*
Illustrating	*for instance*	*for example*

Table 1: Examples of sophisticated and unsophisticated connective words

2 Sophistication in Text

We define ”sophistication” based on two concepts: (1) the phenomena which are used to frame the text and (2) the connective words that are used to fuse sentences together. We argue that a sophistication of a fusion sentence can be inferred from its discourse phenomenon and markers as shown in Figure 1, hence by generating a parallel fusion using more/less sophisticated features, we can create the counterfactual fusion and measure the total effect. In Table 1 we give examples of sophisticated and unsophisticated connective words that are commonly used in sentence fusion. Note that it is subjective and based on the authors decision.

3 The DiscoFuse Datset

In this project we use the DiscoFuse dataset (Geva et al., 2019). It is a large scale dataset for discourse-based sentence fusion which is based on a set of rules for identifying a diverse set of discourse phenomena in raw text, and decomposing the text into two independent sentences. The examples are sentences from two domains: Sports and Wikipedia. It includes 60 million fusion examples annotated with discourse information required to reconstruct the fused text. An example from DiscoFuse is shown in Figure 2. Geva et al. (2019) used this dataset to develop a sequence-to-sequence model and thoroughly analyze its strengths and weaknesses with respect to the various discourse phenomena, using both automatic as well as human evaluation. Geva et al. (2019) also

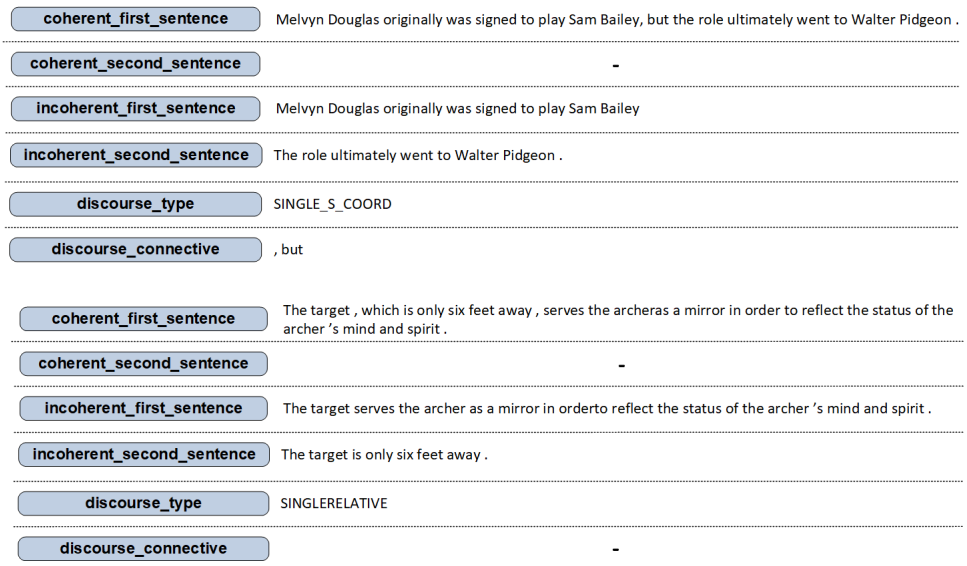

Figure 2: DiscoFuse dataset example

proposed a balanced subset of DiscoFuse, which includes 16 million examples. We take 10% of this balanced subset to serve as our development set–on which we will generate counterfactuals, and the rest is used for training both the classifier and the generative model.

4 Project Steps

We propose and follow the following pipeline:

- Preparing the data: dividing the data to "sophisticated" and "unsophisticated" examples and building training, development and test sets.
- Training a neural network model to classify the origin domain of fused sentences ("Sports", "Wikipedia").
- Training a conditional generative models to fuse sentences given a sophistication flag.
- Analyze the effect on the classifier: Does the model make decisions based on contextual words (e.g. "football" $\rightarrow$ Sports)? Or does it make decisions based on the use of "more" or "less" sophisticated words (e.g. "whose" $\rightarrow$ Wikipedia)?

5 Model and Technical Details

5.1 Contextualized word embedding Models

Contextualized word embedding (CWE) models produce word and sentence representations that take into account the context in which the word or the sentence appear. This is unlike type-level embeddings such as word2vec (Mikolov et al., 2013), and Glove (Pennington et al., 2014) where each word type is assigned a unique, context-independent vector.

Recently, much research has been focusing on training CWE models from massive corpora, aiming to capture broad world knowledge (Peters et al., 2018; Radford et al., 2019). These models typically employ a language modeling objective or a closely related variant (Peters et al., 2018; Ziser &

Reichart, 2018; Devlin et al., 2018; Yang et al., 2019), although in some recent papers the model is trained on a mixture of basic NLP tasks (Zhang et al., 2019; Rotman & Reichart, 2019). The contribution of such models to the state-of-the-art in a variety of NLP tasks is already well-established.

In this work we utilize a CWE model to extract good representation for input text. We use the CWE model as an encoder, in an encoder-decoder scheme that aims to generate text. In the following section we elaborate on the CWE model we use, which is BERT.

5.2 Bidirectional Encoder Representations from Transformers (BERT)

BERT (Devlin et al., 2018) is a CWE model that leverages the strengths of self attention architectures (Vaswani et al., 2017) in order to achieve high quality contextualized text representations. It is trained on large broad corpora, Wikipedia and Books, with two training tasks, Masked Language Model (MLM) and Next Sentence Prediction (NSP).

The representation which can be extracted from this large model (over 105 Million trainable parameters) have proven effective to many downstream NLP tasks, such as text classification (sentiment analysis), sequence tagging (Named Entity Recognition, Dependency Parsing), Question Answering, and Machine Translation (MT).

5.3 Attention Is All You Need

Transformer (Vaswani et al., 2017) is a Seq2Seq, encoder-decoder, model which is based on self attention architecture. This model has shown significant performance improvements on top of recurrent neural networks (RNNs), such as LSTMs and GRUs, for text-to-text tasks, such as MT.

Furthermore, while Transformers rely on fully attention architecture, they also improve the model's parallelism ability by avoiding the dependency between token representations, in contrast to recurrent based encoder-decoder. In this work we use (1) the BERT model as the encoder in our scheme, which is based on Transformers architecture; and (2) a Transformer decoder to be the decoder.

5.4 Classifier

We train a DNN to classify the origin domain of *fused* sentences. In our experiment, there are two domains, Wikipedia and Sports, thus this is a binary classification task. The classifier is a BERT-based uncased model, with 12 attention layers, hidden dim of 768, max sequence length is set to 60 and a vocabulary size of 30K tokens. Overall, this state-of-the-art classification model sums up to more than 105 Million parameters. The classifier's architecture is illustrated in Figure 3. We trained the classifier for 4 epochs and the results are reported in Section 7.

5.5 Generative Model for Fusions

In order to answer the causal question we presented, we need to generate good counterfactuals. These counterfactuals, along with the original data, are used to measure the effect on the trained classifier from the previous section. In order to generate these counterfactuals, we propose an encoder-decoder generative model that can fuse sentences with respect to a condition, which is the sophistication treatment signal in our case. We use a pre-trained BERT as the encoder and a trainable Transformer as the decoder. The sophistication conditional (binary) is concatenated to encoder input, which is how we control whether the output sentence will use sophisticated word for

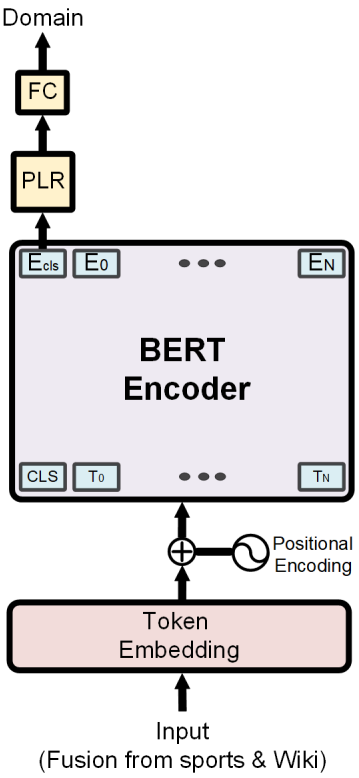

Figure 3: The architecture of the origin domain BERT-based classifier

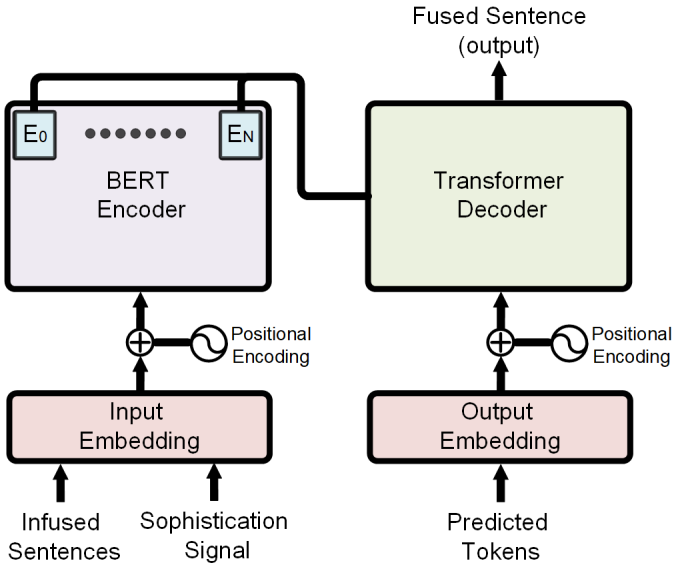

Figure 4: The architecture of the BERT-based encoder-decoder Seq2Seq model

the fusion or not. This signal takes the form of two tokens: <*soph*> and <*nsoph*> which represent generating a sophisticated fusion and an unsophisticated fusion, respectively. The architecture of the model is illustrated in Figure 4.

Our training pipeline differs from inference, where we generate counterfactual fusions, in two ways: (1) In train time we use a teacher forcing method, where a token prediction is based on golden previous token presented to the model. On the other hand, in inference time, a prediction is made on top of the previously predicted token, which were actually predicted by the model. (2) During training we use the golden sophistication signal to generate a fused output while in inference we use the opposite signal, to create a counterfactual.

As for hyper-parameters, we used embedding size of 768 for the words, batch size of 32 and ran a total of 5 epochs. We used the Adam optimizer with a learning rate of 1e-4. Our implementation was done in PyTroch (Paszke et al., 2019) and we ran our experiments on Nvidia 1080 GTX GPU.

6 EVALUATION

In order to evaluate the sophistication (the treatment) effect, we calculate an alternative measure to the Average Treatment Effect (ATE), which we name Model Sophistication Bias Estimation (MSBE). Also, in order to measure the generation quality of the generative model, we calculate the Generation Reliability (GR) score which is defined below. For each domain, we generate two versions of each sample, one with sophistication ($T = 1$) and one without ($T = 0$). We then calculate the MSBE by considering the scores of the classifier for each sample. For example, we take an infused sentence from the sports domain and generate a sophisticated fusion of it, and an unsophisticated version of it. As we know it is from the sports domain, we calculate the difference of confidences that the sentences are indeed from the sports domain.

We denote:

- $d \in \{0, 1\}$ - the domain (i.e., sports-0 or Wikipedia-1)
- $i \in \{1, ..., N\}$ - the sample
- $f_d(X_i) \in [0, 1]$ - classifier output for sample i. Higher score means the classifier assigns high probability to being from domain d.
- $T = 1$ - sophistication treatment.
- $T = 0$ - unsophistication (de-sophistication) treatment.
- Y_1 - score of the classifier for samples from that were originally sophisticated. If $T = 1$, the sample remains the same and if $T = 0$, the sample goes through unsophistication.
- Y_0 - score of the classifier for samples that were originally unsophisticated. If $T = 1$, the sample goes through sophistication and if $T = 0$, the sample remains the same.

At this point, the reader might wonder why we do not calculate the ATE. The ATE for domain d is defined as:

$$ATE_d = \mathbb{E}\left[\mathbb{E}\left[Y_1 - Y_0 \mid T\right]\right]$$

$$= P(T = 1)\mathbb{E}\left[Y_1 - Y_0 \mid T = 1\right] + P(T = 0)\mathbb{E}\left[Y_1 - Y_0 \mid T = 0\right]$$

$$= P(T = 1)\left[\mathbb{E}\left[Y_1 \mid T = 1\right] - \mathbb{E}\left[Y_0 \mid T = 1\right]\right] + P(T = 0)\left[\mathbb{E}\left[Y_1 \mid T = 0\right] - \mathbb{E}\left[Y_0 \mid T = 0\right]\right]$$

Notice that the left-hand part (ATT) measures how well are the generated counterfactuals, while the right-hand part measures the same but in the opposite direction. Thus, the ATE is not informative to answer the causal question and we formulate evaluation measures that capture the essence of this work in a more appropriate way.

The Model Sophistication Bias Estimation (MSBE), is defined as follows:

$$MSBE = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_1 \mid T = 0]] + [\mathbb{E}[Y_0 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 0]]$$

which indicates how much the classifier determines its domain prediction on the sophistication signal.

The Generation Reliability (GR) is defined as:

$$GR = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 1]] + [\mathbb{E}[Y_0 \mid T = 0] - \mathbb{E}[Y_1 \mid T = 0]]$$

Notice that the first term is the Average Treatment Effect (ATT).

Intuitively, both right and left terms represent the model's domain confidence difference between original examples and generated examples from the same sophistication group (where generated examples are the ones that are the counterfactuals of original examples from the opposite sophistication group). To give a more specific example, given that we evaluate examples from sports, a large difference in the left part of the equation indicates that while the model has high sports-confidence for sophisticated examples, this is not the case for examples that were originally non-sophisticated and were treated with the sophistication treatment.

7 RESULTS

The first module in our experiment is the origin-domain classifier, which takes in a fused sentence from one of the domains (Sports or Wikipedia) and predicts the domain of the sentence. We evaluated the classifier performance on the development set and achieved 94% accuracy. Thus, we can safely assume that we can rely on it when we use to measure the causal effect of sophistication.

We argue that an ideal model is one that relies much on the context presented within text, which highly differentiates between these domain, and hence can yield accurate predictions. However, we would like to test this hypothesis.

In what follows, we evaluate the causal effect using the defined evaluation measures. To get numerical values, we define the score of the classifier to be the Softmax activation term in order to bound the results, such that 1 is the highest confidence score and 0 is the lowest confidence score.

7.1 SPORTS

$$MSBE = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_1 \mid T = 0]] + [\mathbb{E}[Y_0 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 0]] =$$

$$= (0.917 - 0.902) + (0.937 - 0.943) = 0.01$$

Hence there a very small positive sophistication effect.

$$GR = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 1]] + [\mathbb{E}[Y_0 \mid T = 0] - \mathbb{E}[Y_1 \mid T = 0]] =$$

$$= (0.917 - 0.937) + (0.943 - 0.902) = 0.021$$

7.2 WIKIPEDIA

$$MSBE = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_1 \mid T = 0]] + [\mathbb{E}[Y_0 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 0]] =$$

$$= (0.842 - 0.853) + (0.829 - 0.852) = -0.034$$

Hence the sophistication effect is negative.

$$GR = [\mathbb{E}[Y_1 \mid T = 1] - \mathbb{E}[Y_0 \mid T = 1]] + [\mathbb{E}[Y_0 \mid T = 0] - \mathbb{E}[Y_1 \mid T = 0]] =$$

$$= (0.842 - 0.829) + (0.852 - 0.853) = 0.012$$

As presented above, we see that the sophistication signal has some impact on the classifiers prediction. To be more specific, in Sports examples, the sophistication signal has small positive impact (1%) on the classifier's confidence. This small impact of the signal can be attributed to the fact that Sports is a diverse corpus, written by professional writers and addressed to a wide range of readers, hence containing many writing styles. This makes it difficult for the classifier to build on top of the signal correlation to this domain label. However, a more optimistic view to this can be that the model actually relies on the context within text and hence is not affected by the sophistication signal.

The Wikipedia results are a bit less optimistic. In this case we observe notably high negative signal affect (3.5%) on the model's confidence. This indicates that the model tends to reject the hypothesis that an example's origin is from Wikipedia once it notices a sophisticated text, regardless of what the text is about.

Finally, we observe a very promising GR results (especially in Wikipedia), indicating that our generative model yields high quality generations, that don't introduce high biases into our measures.

In Table 2 we demonstrate the model ability to generate fusions and counterfactuals.

8 WEAKNESSES

Our proposed method aims to test the effect of a sophistication signal on a predictions that are given by a classification model. To this aim, we propose to generate counterfactual text for given examples, allowing us to measure the causal effect of the signal on the classifier predictions. To avoid introducing bias to our measurements, the model should be able to generate text that is syntactically correct and preserves the original information, while surpassing its origin task - generating a counterfactual example. Although it seems that in most cases our model is successful in generating a counter example, in some examples the generated fusion meaning is not identical to the origin fusion, as presented in the first example in Table 2, where the generation semantic meaning is not identical to the original fusion (*although* is replaced by *and*).

Furthermore, our method suffers from the arbitrary definition given for sophistication, which is perhaps inaccurate and hence prevents our measurements from catching the real effect of the signal. However, some assumptions that are based on world knowledge and human language understanding must be made to train our unsupervised generation model, which lacks any labeled data.

Despite the presented weaknesses, we argue that our method is able to shed a little light on the actual ways a state-of-the-art model, such as BERT, makes its decisions.

Origin	Fusion
	Wikipedia Examples
G	(N) The seeing eye is a non - profit organization **and** is funded through private donations.
P	(N) The seeing eye is a non - profit organization **and** is funded through private donations.
C	(S) The seeing eye is a non - profit organization, **although** it is funded through private donations.
G	(S) **Although** annoyed with Kralik's stubbornness, Matuschek is reluctant to ignore his judgment
P	(S) **Although** annoyed with Kralik's stubbornness, Matuschek is reluctant to ignore his judgment.
C	(N) Annoyed with Kralik's stubbornness, **but** Matuschek is reluctant to ignore his judgment.
G	(N) Lane didn't appear that night **and** psychosis defended and lost the title to disco instead.
P	(N) Lane didn't appear that night **, but** psychosis defended and lost the title to disco instead.
C	(S) Lane didn't appear that night. **however,** psychosis defended and lost the title instead.
	Sports Examples
G	(S) There is still time, **although,** time is short.
P	(S) There is still time, **although,** time is short.
C	(N) There is still time, **but,** time is short.
G	(N) The leaves are turning nicely **and** greens are rolling at a 12!
P	(N) The leaves are turning nicely **and** greens are rolling at a 12!
C	(S) The leaves are turning nicely **whereas** greens are rolling at a 12!
G	(N) James Webb well done coach**, and** now the cowboy rides into the sunset!!!
P	(N) James Webb well done coach**, and** now the cowboy rides into the sunset!!!
C	(S) James Webb well done coach**. Plus,** now the cowboy rides into the sunset!!!

Table 2: Examples from the trained generative model on the development set. **G** - ground truth, **P** - model prediction, **C** - generated prediction. **(S)** and **(N)** denotes whether the sentence is sophisticated or not, respectively.

9 Conclusion

In this work, we aimed to quantify the causal effect of sentence sophistication on a trained classifier. We conducted the experiment with modern, powerful tools from NLP. We implemented a generative model that can fuse sentences given a conditional sophistication signal, which gave us the ability to generate counterfactuals. In order to evaluate the causal effect, we formulated an alternative to the ATE , the Model Sophistication Bias Estimation (MSBE). Using the MSBE, we were able to determine that although small, the causal effect does exist in such way that sophistication yielded slightly different scores from the classifier. Specifically, on the tested domains, we have observed a small positive sophistication effect in Sports and in the Wikipedia domain we observed the opposite. Moreover, due to the small effect, we are more confident that the classifier does look at "incriminating" words to classify the domain and not only on the "sophistication" level of the sentence.

REFERENCES

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Mor Geva, Eric Malmi, Idan Szpektor, and Jonathan Berant. Discofuse: A large-scale dataset for discourse-based sentence fusion. *CoRR*, abs/1902.10526, 2019. URL http://arxiv.org/abs/1902.10526.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 3111–3119, USA, 2013. Curran Associates Inc. URL http://dl.acm.org/citation.cfm?id=2999792.2999959.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

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

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237. Association for Computational Linguistics, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8), 2019.

Guy Rotman and Roi Reichart. Deep contextualized self-training for low resource dependency parsing. *Transactions of the Association for Computational Linguistics*, 7:695–713, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding. *CoRR*, abs/1906.08237, 2019. URL http://arxiv.org/abs/1906.08237.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. ERNIE: enhanced language representation with informative entities. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 1441–1451. Association for Computational Linguistics, 2019.

Yftah Ziser and Roi Reichart. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 1241–1251. Association for Computational Linguistics, 2018.