How do Words Contribute to Sentence Semantics? Revisiting Sentence Embeddings with a Perturbation Method

Wenlin Yao Lifeng Jin Hongming Zhang Xiaoman Pan
Kaiqiang Song Dian Yu Dong Yu Jianshu Chen
Tencent AI Lab, Bellevue, WA, USA
{wenlinyao,dyu,jianshuchen}@global.tencent.com

Abstract

Understanding sentence semantics requires an interpretation of the main information from a concrete context. To investigate how individual word contributes to sentence semantics, we propose a perturbation method for unsupervised semantic analysis. We next re-examine SOTA sentence embedding models' ability to capture the main semantics of a sentence by developing a new evaluation metric to adapt sentence compression datasets for automatic evaluation. Results on three datasets show that unsupervised discourse relation recognition can serve as a general inference task that can more effectively aggregate information to essential contents than several SOTA unsupervised sentence embedding models. [1]

1 Introduction

Humans are usually able to understand sentence meaning based on a complex cognitive process — composition of words (Löbner, 2013). As main carriers of information, words generally have different levels of contribution to the final sentence semantics. For example, underlined words in *"The city of Austin is considering extending downtown parking meter hours to the weekends and later during the week"* convey the most important information of this sentence. Therefore, determining the primary semantics (or main meaning) of a sentence and estimating how sentence semantics distributes to individual words play critical roles in understanding sentence compositionality.

Recently, with the help of large pretrained Transformers, sentence representation learning has achieved great success in downstream NLP tasks (Qiu et al., 2020). However, most work either heavily relies on human annotation such as Natural Language Inference (NLI) data (Williams et al., 2018; Bowman et al., 2015) to do fine-tuning (Conneau et al., 2017a; Reimers and Gurevych, 2019) or adopt unsupervised contrastive learning to learn sentence embeddings (Gao et al., 2021; Chuang et al., 2022). They neglect one critical property of an effective sentence embedding model that essential contents should contribute to sentence semantics more than non-essential contents when encoding a sentence.

We observe that primary semantics can be acquired by learning to predict discourse relations because sentence primary semantics usually needs to support the logical relations at the discourse level. It is motivated by the distinction between asserted and projected content in semantic theory (Potts, 2003; Tonhauser et al., 2013; Venhuizen et al., 2018). Asserted content is intended to be presented for discussion and information exchange, whereas projected content represents background information that is not under discussion. Thus, asserted content should carry more weight in sentence representations than projected content, because it is at issue in the context by the speaker's intention.

Based on this observation, we apply 36 explicit discourse connectives to four big corpora and extract 9.8M sentence pairs. Acquired sentence pairs are then used to train a universal sentence encoder. Next, by perturbing the trained model, we directly estimate the contributions (importance) from individual words to the final sentence semantics. Our assumption is that the importance of a word in a sentence is proportional to how much the new sentence representation drifts from the original sentence representation if we mask that word. Figure 1 shows the overview of our approach.

To overcome the lack of evaluation data and quantitatively evaluate a sentence encoder's sensitivity to key information of a sentence, we design a new evaluation metric — important information gain, which measures model's ability to concentrate on the important words instead of randomly se-

[1] Code and data are available at https://github.com/wenlinyao/EACL23-PrimarySemantics

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 3001–3010
May 2-6, 2023 ©2023 Association for Computational Linguistics

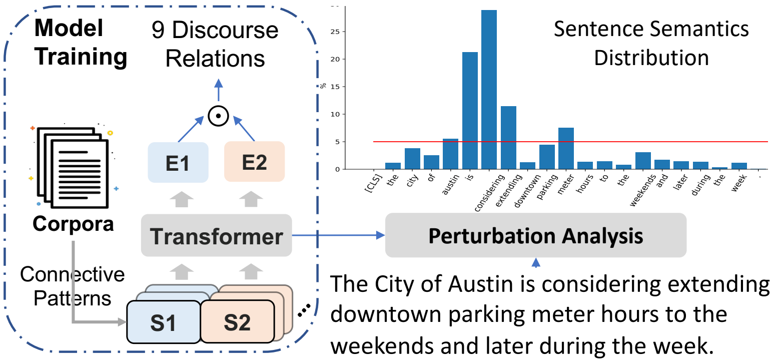

Figure 1: Overview of our approach.

lected words. Experiments on three datasets show that our model outperforms previous unsupervised and supervised sentence embedding models. Our analysis also demonstrates our model is less biased than the model trained with human-annotated data.

In this paper, we investigate how words, as main carriers of information, contribute to the final sentence semantics with different levels of significance. Our main contributions are summarized as follows: 1) We propose a perturbation method to estimate the contributions from individual words to sentence semantics to understand sentence compositionality. 2) We design a new automatic metric, Important Information Gain, to overcome the lack of evaluation. We find discourse relation recognizer can more effectively aggregate information to essential contents than several SOTA unsupervised sentence embedding models.

2 Related Work

DisSent (Nie et al., 2019) also uses discourse prediction to train sentence embeddings, but their goal is to achieve SOTA fine-tuning results on SentEval tasks (Conneau and Kiela, 2018) and the PDTB task instead of investigating primary semantics. Our work is closely related to sentence compression (Filippova et al., 2015; Kamigaito and Okumura, 2020) and extractive text summarization (Xu and Durrett, 2019; Mendes et al., 2019) that mainly focus on extracting the salient text spans in an end-to-end manner. In contrast, we aim to estimate the semantic saliency distribution to better understand sentence compositionality. Perturbation methods have been also used in data augmentation (Das and Sarkar, 2019), robustness analysis (Niu et al., 2020; Prabhakaran et al., 2019), and textual adversarial attack (Li et al., 2021; Feng et al., 2018). We are the first to use the perturbation method to estimate sentence semantics distribution.

3 Discourse Relation Recognition as a General Inference Task

3.1 Data Collection

To have a broad coverage on different types of texts, we consider four large-scale corpora. **News Articles**. We use English Gigaword 5th edition (Napoles et al., 2012), which contains 10M news articles. **Wikipedia**. We use the Wikipedia dump of 5/20/2021 which consists of 54M web pages. **Novel Books**. BookCorpus (Zhu et al., 2015) contains 11,038 novel books of 16 different genres. **Blogs**. We use the Blog Authorship Corpus (Schler et al., 2006) which consists of 680K blog posts. We applied the Stanford CoreNLP tools[2] to obtain sentence boundaries.

We next select 36 explicit connectives (corresponding to 9 discourse relations)[3] from the PDTB annotation manual (Prasad et al., 2008) and summarize them into extraction patterns. After applying them to four corpora, we extract 9.8M sentence pairs in total. Statistics of acquired sentence pairs are summarized in Table 1.

3.2 Model Architecture

Inspired by Reimers and Gurevych (2019), our model first uses an encoder to get two sentence representations and then compares them to predict the discourse relations. Specifically, we apply a pretrained BERT$_{\text{Base}}$ (Devlin et al., 2019) model to the two sentences and select the model outputs at the [CLS] tokens as the corresponding sentence representations u and v. Next, by feeding u, v and $|u - v|$ into a 9-class softmax layer, i.e.,

[2] https://stanfordnlp.github.io/CoreNLP/.

[3] We only consider connectives that are less ambiguous in relations. See Appendix for the full list of connectives.

	CT	RS	CJ	EQ	GN	IN	CA	PR	DJ
Pairs	8M	430K	800K	29K	12K	285K	188K	26K	26K

Table 1: Statistics of sentence pairs extracted from four text corpora. 9 discourse relations are Contrast (CT), Result (RS), Conjunction (CJ), Equivalence (EQ), Generalization (GN), Instantiation (IN), Chosen Alternative (CA), Precedence (PR), and Disjunctive (DJ).

$p = \mathrm{softmax}(W_t[u, v, |u - v|])$, our model predicts what is the discourse relation (e.g., Contrast, Result, Equivalence, etc.) between them.

4 Estimating Sentence Semantics Distribution via Perturbation

To estimate the semantic distribution of the input sentence $S = \{w_i\}_{i=1}^N$, we mask each word w_i individually to construct a new sentence and apply the sentence embedding model (Section 3.2) to calculate the new representation. Next, we calculate the cosine distance between the new and the original representation. The distance can quantitatively tell us how much the new sentence (after masking one word) semantically drift from the original sentence, which indicates the contribution of that word to the sentence semantics.[4] Finally, we normalize the distances by their summation so that the importance scores of all words sum to 1.

5 Evaluation

5.1 Evaluation Metric

Ideally, if we know the gold standard importance distribution in sentence semantics, we can directly compare a model's prediction with the gold standard. However, it is infeasible for humans to annotate/assign a continuous importance score to each word in a sentence. To address the lack of evaluation data, we adopt two sentence compression datasets and one summarization dataset for evaluation. The main idea is that the compressed sentence, as a shorter version of the original sentence, specifies the most important words in the original sentence. Therefore, we can score a model by its ability of concentrating on important words.

Furthermore, we propose a new evaluation metric — **Importance Information Gain** — to score how much a predicted distribution is superior to a uniform distribution (all words are equally important). Specifically, given a sentence $S = \{w_i\}_{i=1}^N$ consisting of $|S| = N$ words, suppose the compressed sentence by human is $S' = \{w_j\}_{j \in 1 \sim N}$, where S' is a subset of S. Let $[v_1, v_2, \cdots, v_N]$ be the importance scores over words (normalized such that $\sum_{i=1}^N v_i = 1$). We calculate the information gain g to be the average importance score on importance words $\frac{\sum_{i \in S'} v_i}{|S'|}$ over the score that every word is equally important $\frac{1}{|S|}$: $g = \frac{\sum_{i \in S'} v_i}{|S'|} - \frac{1}{|S|}$.

We next normalize g for each test sentence instance by the upper bound g^*, where g^* is defined as the information gain of a model that can perfectly distinguish important words from non-important words (concentrate all information on only important words): $g^* = \frac{1}{|S'|} - \frac{1}{|S|}$. Then, the final Importance Information Gain = $g/g^* \in [0, 1]$.

To validate whether the proposed metric is able to measure the quality of a semantic distribution, we randomly sample 100 pairs from baseline models' predictions in Table 2 and analyze our metric's consistency with human preferences. Specifically, each sampled pair consists of predictions (semantic distributions) of two separate baseline models on the same sentence. Three expert annotators are asked to judge which one better characterizes the importance/contribution of each word to the sentence meaning. We next compare our metric's preferences with the gold human preferences[5]. Our metric agreed on 87/100 with human gold labels, achieving a substantial (Cohen, 1968) kappa agreement score of 0.74.

5.2 Evaluation Datasets

For evaluation, we experiment on three datasets. Google sentence compression dataset (GGL) (Filippova and Altun, 2013) contains 10K test sentence compression pairs and BNC written dataset (Clarke and Lapata, 2008) contains 1.5K test compression pairs. Additionally, we go through the Gigaword summarization dataset (GGW) and compare the first sentence with the title sentence in the news article. We select (first sentence, title) pairs as our testing data if words[6] in the title sentence is a subset of words in the first sentence. We collect the first 10K sentence pairs as our test data.

[4] If the target masked word is tokenized into multiple subword pieces by the tokenizer, we mask each subword piece in turn and calculate the summation score as the word importance.

[5] Gold labels are generated based on majority voting. The average Cohen's kappa inter-agreement between three annotators is 0.67.

[6] We use lemma matching in practice.

5.3 Baseline Systems

Rule-Based Model. We use a rule-based primary semantics extraction system (Zhang et al., 2020), which first parses the input sentence into a dependency parsing tree and extracts the sentence skeleton.[7] Words in the sentence skeleton are considered equally important.

GloVe Embedding. We calculate the sentence embeddings by averaging GloVe embeddings (Pennington et al., 2014) among all words except w_i.

Original BERT. We apply the original BERT-base model[8] that uses next sentence prediction as the training objective.

BERT-CT (Carlsson et al., 2021) introduces an unsupervised learning method called Contrastive Tension (CT) that only requires raw sentences and achieves SOTA performance on Semantic Textual Similarity (STS) tasks. CT tries to maximize the dot product between sentence representations for identical sentences and minimize the dot product for differing sentences. BERT-CT-STSb and BERT-CT-NLI[9] are two supervised CT models fine-tuned on STSb and NLI, respectively.

SimCSE (Gao et al., 2021) is a contrastive sentence embedding framework that achieves SOTA performance on sentence similarity tasks. The unsupervised model (SimCSE-unsup) takes an input sentence and predicts itself in a contrastive objective with dropout as the data augmentation method. The supervised model (SimCSE-sup) incorporates sentence pairs from the NLI data (Williams et al., 2018; Bowman et al., 2015) into a contrastive learning framework, by using *entailment* pairs as positives and *contradiction* pairs as negatives.

DiffCSE (Chuang et al., 2022) augments SimCSE contrastive learning with the edited sentences sampled from a masked language model.

SBERT (Reimers and Gurevych, 2019). Sentence-BERT (SBERT) uses siamese and triplet network structures to derive sentence embeddings. SBERT-NLI is the original model introduced by Reimers and Gurevych (2019) that is fine-tuned on NLI. Moreover, we also consider four more recent SBERT models that are fine-tuned with more sentence pair data.[10] Specifically, SBERT-MSMarco is fine-tuned on the MSMarco Passage Ranking Dataset containing 500K (query, relevant passage) pairs from Bing search. SBERT-Paraphrase is fine-tuned on NLI and 11 paraphrase datasets. SBERT-MultiQA is fine-tuned on 214M (question, answer) pairs from 17 QA datasets. MPNet-All is fine-tuned on all available sentence pair tasks including 32 tasks with 1,170M sentence pairs.

	Models	GGL	BNC	GGW	Avg.
	Supervised				
1	BERT-CT-STSb	7.4	19.3	8.3	11.7
2	BERT-CT-NLI	7.5	10.0	6.5	8.0
3	SimCSE-sup	15.8	25.1	10.2	17.0
4	SBERT-NLI	14.3	23.2	14.2	17.2
5	SBERT-Paraphrase	23.7	15.4	31.8	23.6
6	SBERT-MultiQA	33.0	28.6	35.6	32.4
7	SBERT-MSMarco	31.7	28.3	32.7	30.9
8	MPNet-All	**52.6**	**33.3**	**58.2**	**48.0**
	Unsupervised				
9	Parsing Tree	18.7	19.8	7.2	15.2
10	GloVe Embedding	10.2	25.3	12.1	15.9
11	Original BERT	2.0	0.3	0.8	1.0
12	BERT-CT	12.0	16.3	12.4	13.5
13	SimCSE-unsup	13.9	14.9	10.1	13.0
14	DiffCSE	21.6	**28.7**	20.8	23.7
15	Our Model-disc.	**36.9**	27.2	**36.6**	**33.6**

Table 2: Importance Information Gain (%) on three evaluation datasets. We group models based on whether it requires human-labeled data in training or not.

Table 2 shows the results of all models. We group models based on whether they use human-labeled data or not. Line 15 is our model trained on 9.8M sentence pairs that are extracted by discourse connectives. Our unsupervised discourse recognition model achieves the highest overall information gain among all unsupervised approaches, which is even higher than several supervised models that are trained using millions of human-annotated data (Lines 3-7). Surprisingly, even SimCSE and DiffCSE report better performance than MPNet-All on STS tasks, they are much worse than MPNet-All on capturing sentence primary semantics.

5.4 Visualization and Analysis

While previous studies (Conneau et al., 2017b) empirically show that fine-tuning towards NLI data yields good sentence embedding models, we visualize and compare the model trained on the NLI data with the model trained on discourse pairs. We find the NLI data model sometimes is biased to specific numbers, time expressions, etc (Figure 2). It may be mainly due to the bias introduced during the NLI data construction process when human annotators sometimes just replace the numbers or time

[7] We select the whole sentence as the prediction when the system fails on complex sentences.

[8] Select the output at the `[CLS]` token.

[9] `https://github.com/FreddeFrallan/Contrastive-Tension`.

[10] `https://www.sbert.net/docs/pretrained_models.html`.

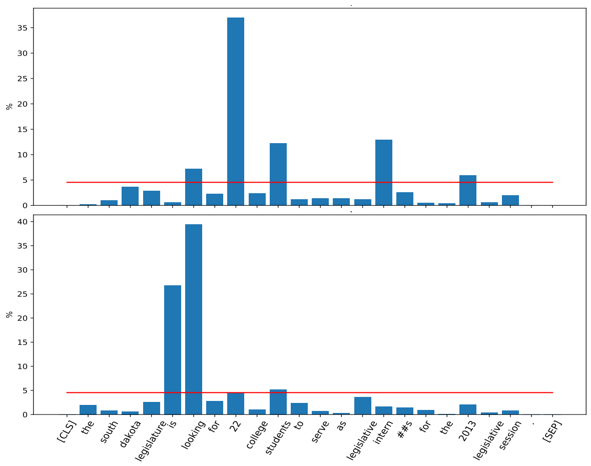

Figure 2: Semantic distribution of the model trained on NLI (upper) v.s. discourse data (lower) on the same sentence (important words are underlined). *"The South Dakota legislature is looking for 22 college students to serve as legislative interns for the 2013 legislative session."* The red line indicates $1/|S|$.

expressions to produce a contradictory sentence. In contrast, the discourse relation recognizer is less sensitive to those specific expressions. Appendix II contains more examples.

6 Conclusion

In this paper, we have introduced a perturbation method for estimating sentence semantic distribution and designed a new metric to achieve automatic evaluation. We find discourse relation recognition can serve as a general inference task to train an unsupervised sentence embedding model that estimates such distribution meaningfully than several SOTA sentence embedding models.

7 Limitations

To benefit from large-scale training data, we train the discourse relation recognizer on data that are automatically generated by matching explicit connectives. Even we only select explicit connectives that have one dominant discourse relation, some of them may still reflect a different relation in some contexts. For example, connective *in other words* indicates a "Equivalence" relation most of the time, but sometimes it can also indicate a "Generalization" relation. In this regard, our method shares the same limitations as the broad class of weakly supervised methods where training data are automatically generated. Considering discourse parser is not the main focus of this paper, we leave how to generate cleaner discourse relation data and train a better discourse relation recognizer for future work.

References

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

Fredrik Carlsson, Amaru Cuba Gyllensten, Evangelia Gogoulou, Erik Ylipää Hellqvist, and Magnus Sahlgren. 2021. Semantic re-tuning with contrastive tension. In *International Conference on Learning Representations*.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljacic, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. DiffCSE: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218, Seattle, United States. Association for Computational Linguistics.

James Clarke and Mirella Lapata. 2008. Global inference for sentence compression: An integer linear programming approach. *Journal of Artificial Intelligence Research*, 31:399–429.

Jacob Cohen. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

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

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

Ayan Das and Sudeshna Sarkar. 2019. A little perturbation makes a difference: Treebank augmentation by perturbation improves transfer parsing. In *Proceedings of the 16th International Conference on Natural*

Language Processing, pages 75–84, International Institute of Information Technology, Hyderabad, India. NLP Association of India.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3719–3728.

Katja Filippova, Enrique Alfonseca, Carlos A Colmenares, Łukasz Kaiser, and Oriol Vinyals. 2015. Sentence compression by deletion with lstms. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 360–368.

Katja Filippova and Yasemin Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1481–1491, Seattle, Washington, USA. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hidetaka Kamigaito and Manabu Okumura. 2020. Syntactically look-ahead attention network for sentence compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8050–8057.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Dianqi Li, Yizhe Zhang, Hao Peng, Liqun Chen, Chris Brockett, Ming-Ting Sun, and William B Dolan. 2021. Contextualized perturbation for textual adversarial attack. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5053–5069.

Sebastian Löbner. 2013. *Understanding semantics*. Routledge.

Alfonso Mendes, Shashi Narayan, Sebastião Miranda, Zita Marinho, André FT Martins, and Shay B Cohen. 2019. Jointly extracting and compressing documents with summary state representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3955–3966.

Courtney Napoles, Matthew Gormley, and Benjamin Van Durme. 2012. Annotated gigaword. In *Proceedings of the Joint Workshop on Automatic Knowledge Base Construction and Web-scale Knowledge Extraction*, pages 95–100. Association for Computational Linguistics.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Xing Niu, Prashant Mathur, Georgiana Dinu, and Yaser Al-Onaizan. 2020. Evaluating robustness to input perturbations for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8538–8544.

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

Christopher Gerard Potts. 2003. *The logic of conventional implicatures*. University of California, Santa Cruz.

Vinodkumar Prabhakaran, Ben Hutchinson, and Margaret Mitchell. 2019. Perturbation sensitivity analysis to detect unintended model biases. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5740–5745, Hong Kong, China. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, pages 1–26.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Judith Tonhauser, David Beaver, Craige Roberts, and Mandy Simons. 2013. Toward a taxonomy of projective content. *Language*, pages 66–109.

Noortje J Venhuizen, Johan Bos, Petra Hendriks, and Harm Brouwer. 2018. Discourse Semantics with Information Structure. *Journal of Semantics*, 35(1):127–169.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Jiacheng Xu and Greg Durrett. 2019. Neural extractive text summarization with syntactic compression. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3292–3303.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020. ASER: A large-scale eventuality knowledge graph. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 201–211. ACM / IW3C2.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.

Appendix I

Experiment details. We generally follow the hyper-parameter setting of previous work to train our model. In experiments, we train our model using cross-entropy loss and Adam (Kingma and Ba, 2015) optimizer with initial learning rate 5e-5, dropout rate 0.5, and batch size 256 for 5 training epochs. The training of the discourse relation recognizer was run on one machine with 8 NVIDIA P40 GPUs, taking about 6 hours per epoch and three epochs in total. We simply use the standard hyper-parameters to train our model without any hyper-parameter search.

Appendix II

Here is the full list of explicit discourse connectives for extracting sentence pairs (Section 3.1).

Contrast (CT): although, but, by comparison, by contrast, however, in contrast, nevertheless, nonetheless, on the contrary, on the other hand, though
Result (RS): accordingly, as a result, because, consequently, hence, therefore, thus
Conjunction (CJ): additionally, besides, furthermore, in addition, in fact, indeed, moreover, overall, similarly
Equivalence (EQ): in other words
Generalization (GN): in short, in sum
Instantiation (IN): for example, for instance
Chosen Alternative (CA): instead
Precedence (PR): ultimately
Disjunctive (DJ): otherwise, unless

Appendix III

More examples of semantic distribution of the model trained on NLI (upper) v.s. general inference knowledge (lower) on the same sentence.

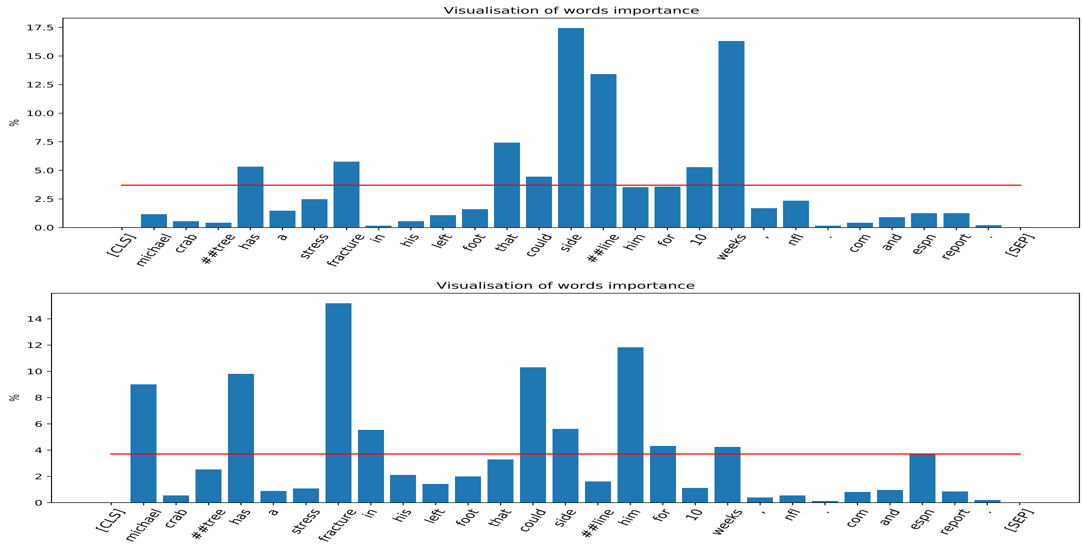

Figure 3: *Michael Crabtree has a stress fracture in his left foot that could sideline him for 10 weeks , NFL.com and ESPN report .*

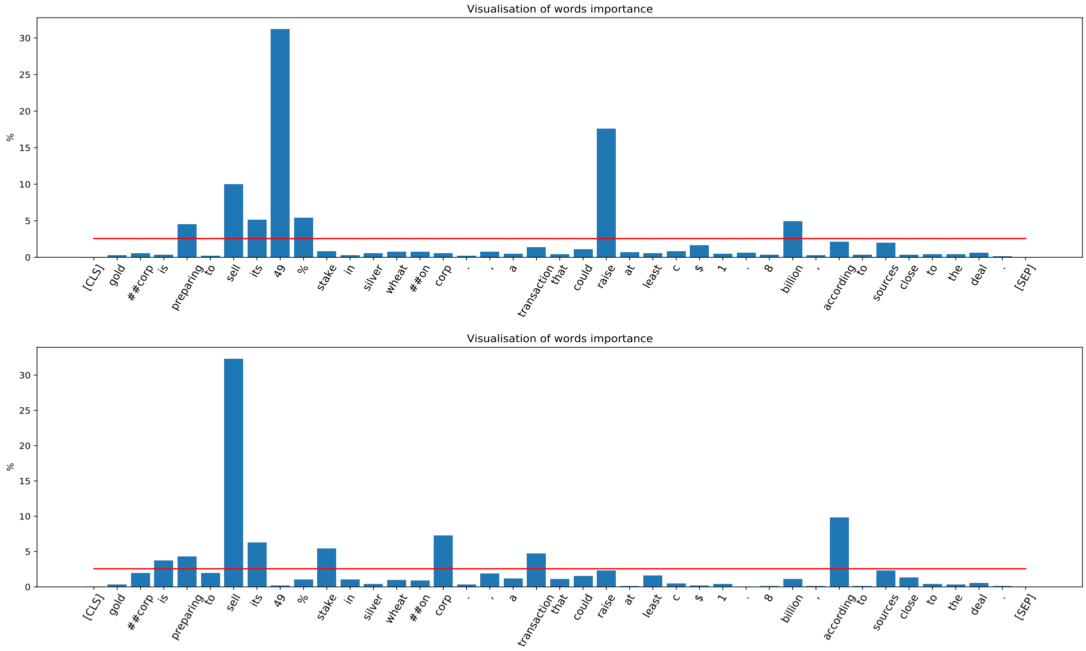

Figure 4: *Goldcorp is preparing to sell its 49 % stake in Silver Wheaton Corp. , a transaction that could raise at least $ 1.8 billion , according to sources close to the deal .*

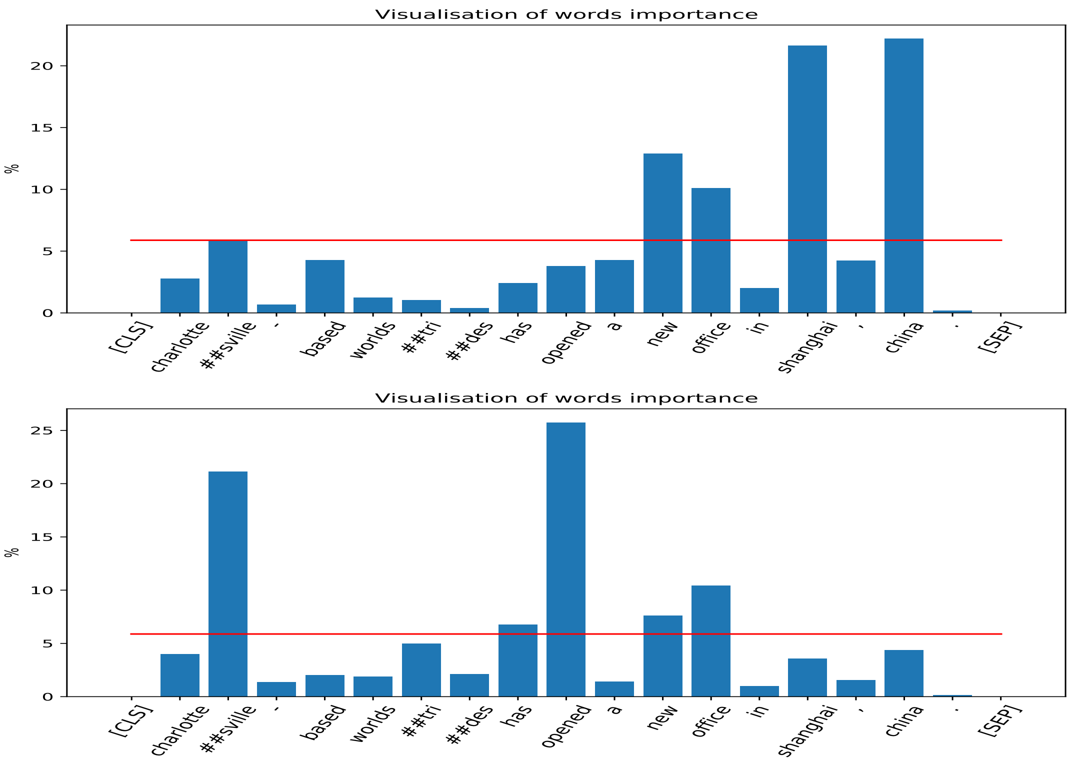

Figure 5: *Charlottesville - based WorldStrides has opened a new office in Shanghai , China .*

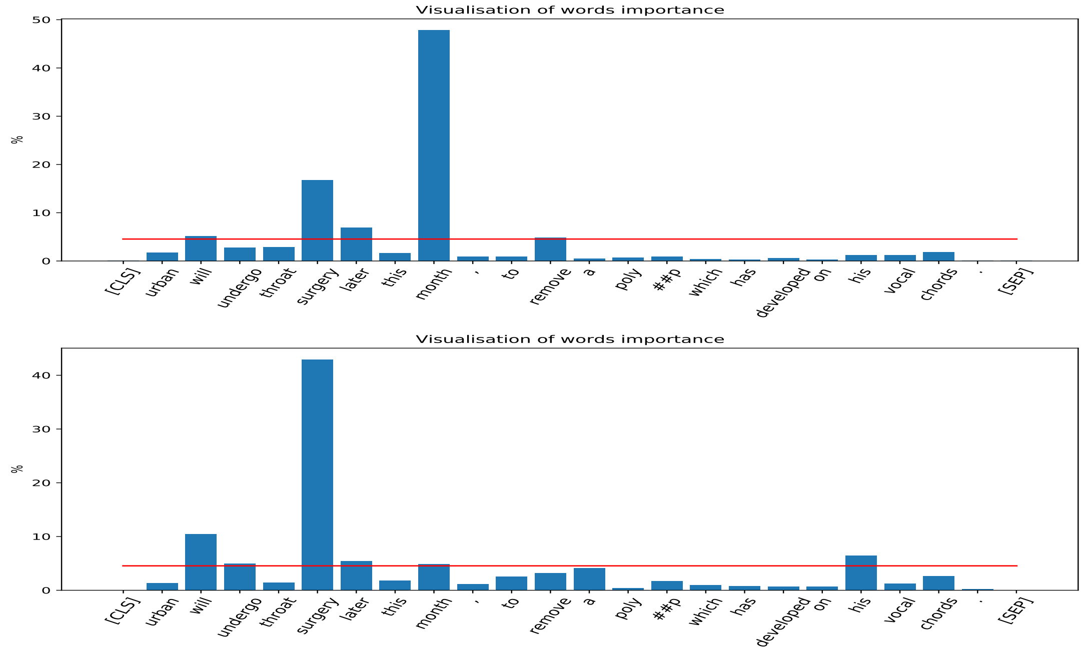

Figure 6: *Urban will undergo throat surgery later this month , to remove a polyp which has developed on his vocal chords .*

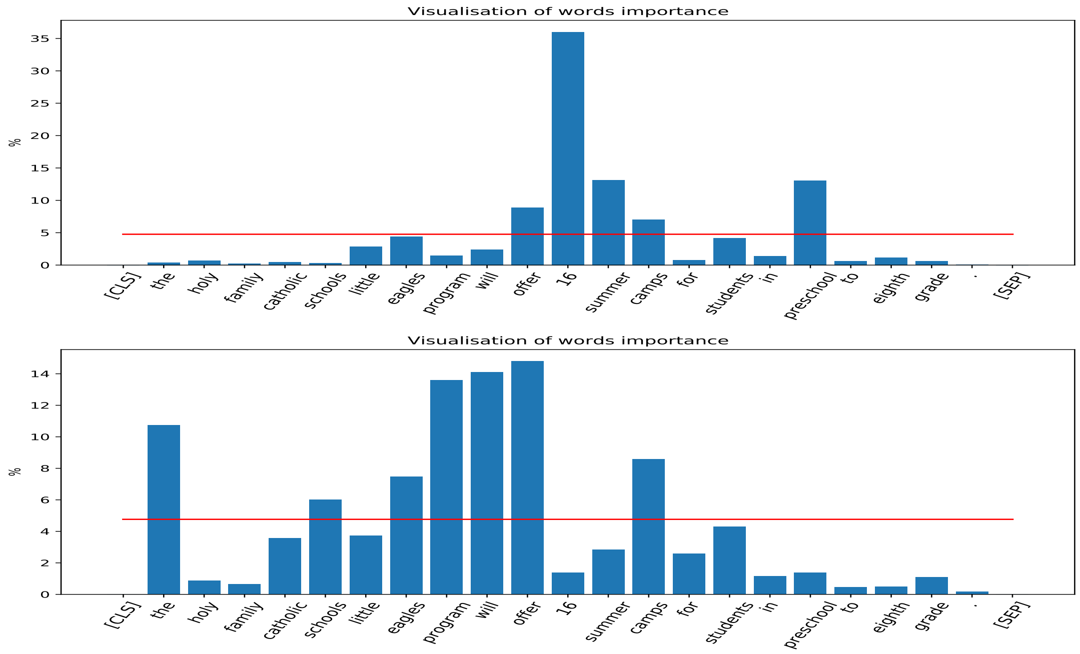

Figure 7: *The Holy Family Catholic Schools Little Eagles program will offer 16 summer camps for students in preschool to eighth grade .*