СРАВНЕНИЕ ПРЕДЛОЖЕНИЙ С ПОМОЩЬЮ МОДЕЛИ SBERT

Курбонов К.С.
Национальный исследовательский Томский политехнический университет, аспирант, А1-36 ИШИТР
e-mail: ksk32@tpu.ru

Аннотация

В работе использована предобученная модель Sentence-BERT (SBERT), которая использует сиамские и триплетные сетевые структуры для получения семантически значимых включений предложений. Проведено сравнение текстовых данных и построена матрица косинусной меры сходства.

Ключевые слова: семантическое текстовое сходство, Sentence-BERT, SBERT.

Введение

Цель работы – провести семантическое сходство текстов с использованием моделей глубокого обучения. Для достижения поставленной цели мы решили следующие задачи: собрали текстовые данные, провели обучение моделей, выполнили тестирование на вне выборочных данных.

В качестве объекта исследования мы использовали текстовые формулировки компетенций, указанные в регламентирующих документах российских вузов для магистрантов, обучающихся по направлению 01.04.02 «Прикладная математика и информатика». В выборку были включены матрицы компетенций следующих вузов: Национальный исследовательский Томский политехнический университет (ТПУ), Национальный исследовательский университет «Высшая школа экономики» (ВШЭ), Пермский национальный исследовательский политехнический университет (ПНИПУ), Казанский федеральный университет (КФУ), Дагестанский государственный технический университет (ДГТУ). Созданный набор данных мы разделили: 80 % на обучение и 20 % на тестирование. В качестве критерия семантического текстового сходства было использовано косинусная мера сходства.

Наш основной вклад в область исследований является сравнение матриц компетенций магистерского направления 01.04.02 «Прикладная математика и информатика» разных университетов с помощью модели SBERT.

Модель

Использование SBERT для поиска наиболее похожей пары предложений в коллекции из 10 000 предложений требует около 5 секунд. Благодаря использованию оптимизированных индексных структур поиск наиболее похожего вопроса на Quora может быть сокращен с 50 часов до нескольких миллисекунд [1].

Для решения задачи семантического сходства текста (semantic textual similarity, STS) мы использовали предобученную модель SBERT, которая показала преимущество по сравнению с InferSent [2] и Encoder Universal Sentence [3]. SBERT дополняет операцию соединения с выходом BERT/RoBERTA для получения вложений фиксированных размеров. Мы провели эксперименты с тремя стратегиями объединения и остановились на MEAN стратегии. Для настройки BERT/RoBERTA мы создали сеть сиамских и трехплетных сетей [6] для изменения весов, чтобы получаемые вкрапления предложения были значительными с целью дальнейшего вычисления меры косинусного сходства [7]. На рис. 1 приведены две архитектуры SBERT.

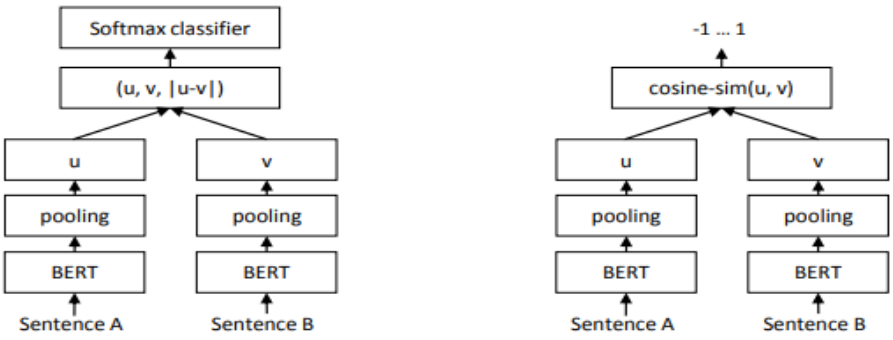

Рис. 1. Архитектура SBERT:
с целевой функцией классификации (слева), для вычисления оценок сходства (справа)

Целевая функция классификации. Мы сопоставляем вкрапления предложений u и v с разницей между элементами $|u–v|$ и умножаем ее на вес $W_t \in R^{3nk}$:

$$O=\text{softmax}(W_t(u, v, |u–v|)),$$

где n – размерность векторного представления предложений (эмбеддинг), а k – количество меток [8]. Мы оптимизируем потери перекрестной энтропии. Данная архитектура изображена на рисунке 1 (слева).

Регрессионная целевая функция. Вычисляется косинусное сходство между двумя вкраплениями предложений u и v (рис. 1, справа). В качестве целевой функции мы используем среднее квадратичное отклонение [9].

Триплетная целевая функция. В зависимости от базового предложения A, положительного предложения p и положительного предложения n, триплетная функция потерь настраивает сеть так, что расстояние между A и p меньше расстояния между A и n. [10]. Тогда необходимо минимизировать следующую функцию потерь:

$$\max(\|S_A – S_p\|) – \|S_A – S_n\| + e, 0),$$

где S_x метрика для предложения $x = \{A, n, p\}$, e – погрешность, которая обеспечивает, что S_p хотя бы на величину e ближе к S_A, чем к S_n. В качестве метрики мы используем евклидово расстояние и в наших экспериментах использовали $e = 1$ [11]. Набор данных мы разделили на два класса: 1 – название университетов и его направление, 2 – предложение для сравнения [12]. Исходя из проделанной работы, мы можем получили следующие результаты (таблица 1). В таблице в ячейке указано значение косинусной меры сходства между текстовыми формулировками.

Таблица 1

Значение косинусной меры сходства
между текстовыми формулировками для вузов из выборки

	ТПУ	ВШЭ	ПНИПУ	КФУ	ДГТУ
ТПУ	1.00	0.65	0.97	0.65	0.99
ВШЭ		1.00	0.67	0.69	0.68
ПНИПУ			1.00	0.66	0.92
КФУ				1.00	0.66
ДГТУ					1.00

Заключение

Проведено исследование семантического текстового сходства с использованием предобученной модели Sentence-BERT (SBERT) на задаче сравнения матриц компетенций российских университетов для магистерской программы по направлению 01.04.02 "Прикладная математика и информатика". Сиамские и триплетные сетевые структуры, заложенные с модели SBERT, позволили оптимизировать процесс поиска наиболее похожих пар предложений и значительно сократить время на вычисление меры косинусного сходства. В ходе экспериментов были проверены различные стратегии объединения вложений предложений и наилучший результат был достигнут с использованием стратегии MEAN. Мы использовали три разные функции потерь: для классификации, регрессии и триплетной сети. Полученные результаты продемонстрировали, что модель SBERT может быть эффективно сравнивать текстовые формулировки компетенций различных университетов с целью их дальнейшего анализа.

Список использованных источников

1. Capuozzo P., Lauriola I., Strapparava C., Aiolli F., & Sartori G. Decop: A multilingual and multi-domain corpus for detecting deception in typed text // Proceedings of the 12th Language Resources and Evaluation Conference – 2020. – P. 1423–1430.

2. Chakraborty A., Paranjape B., Kakarla S., & Ganguly N. Stop clickbait: Detecting and preventing clickbaits in online news media // IEEE conference on Advances in Social. – 2026. – P. 9–16

3. Chen D., Fisch A., Weston J., & Bordes A. (2017). Reading Wikipedia to answer open-domain questions. / In R. Barzilay, & M. Kan (Eds.) // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL. – Vancouver, Canada, July 30 - August 4. – 2017. – Vol. 1. – P. 1870–1879. (URL: http://dx.doi.org/10.18653/v1/P17-1171)

4. Clark K., Luong M., Le Q.V., & Manning C.D. ELECTRA: Pre-training text encoders as discriminators rather than generators // 8th International Conference on Learning Representations, ICLR 2020. – Addis Ababa, Ethiopia. – April 26-30. (URL: https://openreview.net/forum?id=r1xMH1BtvB).

5. Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. arXiv preprint – arXiv:1904.09675. – 2019.

6. Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. arXiv preprint – arXiv:1906.08237, – abs/1906.08237. – 2019.

7. Yinfei Yang, Steve Yuan, Daniel Cer, Sheng-Yi Kong, Noah Constant, Petr Pilar, Heming Ge, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. Learning Semantic Textual Similarity from Conversations // Proceedings of The Third Workshop on Representation Learning for NLP, Melbourne, Australia. – 2018. – P. 164– 174.

8. Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, German Rigau, Larraitz Uria, and Janyce Wiebe. SemEval-2015 Task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, Colorado. – 2015. – P. 252–263.

9. Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. SemEval-2014 Task 10: Multilingual Semantic Textual Similarity // Proceedings of the 8th International Workshop on Semantic Evaluation, Dublin, Ireland. – 2014. – P. 81–91.

10. 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.

11. Bill Dolan, Chris Quirk, and Chris Brockett. Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources // Proceedings of the 20th International Conference on Computational Linguistics, COLING ’04, Stroudsburg, PA, USA. – 2004.