Многоязычные методы выделения значений слов.

Быков Дмитрий Андреевич

Студент

Факультет ВМК МГУ имени М. В. Ломоносова, Москва, Россия

E-mail: dima13051998@gmail.com

Научный руководитель *— Арефьев Николай Викторович*

Задача выделения значений слов заключается в разбиении множества предложений с заданным целевым словом на группы так, чтобы каждому из значений соответствовала ровно одна группа.

Основным способом решения этой задачи является метод генерации лексических подстановок, который был предложен в статьях [1,2,3]. Данный метод заключается в том, чтобы значение целевого слова в контексте представить с помощью набора наиболее вероятных замен целевого слова, полученных с помощью языковых моделей. Чаще всего вместо данного слова в предложение подходят синонимы, гипонимы, гиперонимы и т. д. того значения, в котором использовалось данное слово. Поэтому такие представления помогают разрешать многозначность, так как для различных значений они обычно не совпадают.

В соревновании SemeEval-2010 Task 2 [4] была предложена многоязычная версия задачи генерации лексических подстановок. В ней от систем участников требовалось сгенерировать переводы целевых слов из предложений на английском языке на испанский язык. Данная задача похожа на задачу машинного перевода, однако отличается тем, что требуется сгенерировать несколько переводов целевого слова, указанных аннотаторами.

В этой работе предлагаются методы решения задачи выделения смыслов слов с помощью генерации подстановок на другом языке. Такие представления также должны разрешать лексическую неоднозначность, так как переводы одного и того же слова в различных значениях обычно не совпадают.

Данный способ дает несколько существенных преимуществ по сравнению с одноязычными представлениями таких как: простое создание обучающей выборки для обучения с учителем, с помощью пословного выравнивания параллельных корпусов и возможность объединения подстановок на различных языках.

Предлагаемые методы используют многоязычную маскированную языковую модель XLM-R [5]. Для генерации переводов на дру-

гой язык используются три различных приема, а также их комбинации: использование специальных шаблонов, подсказывающих модели на какой язык переводить, использование двуязычного словаря и дообучение.

Лучшие из предложенных методов превосходят существующие методы решения задач генерации лексических подстановок на другом языке SemeEval-2010 Task 2 [4,7] и задачи выделения значений слов SemeEval-2010 Task 14 [2,3,6].

Литература

1. Asaf Amrami and Yoav Goldberg. Word Sense Induction with Neural biLM and Symmetric Patterns. InProceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4860–4867, 2018.
2. Asaf Amrami and Yoav Goldberg. Towards better substitution based word sense induction.arXiveprints, pages arXiv–1905, 2019.
3. Nikolay Arefyev, Boris Sheludko, and Tatiana Aleksashina. Combining Neural Language Models for Word Sense Induction. InAnalysis of Images, Social Networks and Texts, page 105–121.Springer International Publishing, 2019.
4. Rada Mihalcea, Ravi Sinha, and Diana McCarthy. 2010. SemEval-2010 task 2: Cross-lingual lexical substitution. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 9–14, Uppsala, Sweden. Association for Computational Linguistics.
5. Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek,Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. InProceedings of the 58th Annual Meeting ofthe Association for Computational Linguistics, pages 8440–8451, 2020.
6. Suresh Manandhar and Ioannis P Klapaftis. 2009. Semeval-2010 task 14: evaluation setting for word sense induction & disambiguation systems. In Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 117–122. Association for Computational Linguistics.
7. Pierpaolo Basile and Giovanni Semeraro. 2010. UBA: Using automatic translation and Wikipedia for crosslingual lexical substitution. In Proceedings of the 5th International Workshop on Semantic Evaluation, pages 242–247, Uppsala, Sweden. Association for Computational Linguistics.