

Разработка автоматизированной системы пополнения таксономии на текстах конкретной предметной области

М. М. Тихомиров¹

В работе рассматривается вопрос применения разработанного метода на основе использования мета-векторных представлений слов для автоматизированного (с использованием аннотаторов) расширения таксономии на конкретную предметную область. Рассматривается область информационной безопасности, которая используется для обогащения Онтологии Естественных Наук и Технологий (ОЕНТ).

Ключевые слова: тезаурус, пополнение таксономии, метавекторное представление, интерфейс.

1. Введение

Таксономии - это полезный инструмент, который применяется в различных задачах обработки естественного языка. Таксономии обычно состоят из набора сущностей, называемых концептами или понятиями, и связей между сущностями, выражающими отношения класс-подкласс (или гипероним-гипоним) между концептами [1, 2, 3]. Построение подобных ресурсов происходит вручную и требует существенных человеческих затрат.

Для упрощения построения таксономии были предложены различные подходы для извлечения отношений гиперонимии для новых терминов из текстов, использующие шаблоны, информацию о со-встречаемости слов, характеристики распределения слов и другие подходы [4]. В настоящее время важным компонентом извлечения отношений гиперонимии из текстов являются векторные представления слов, которые могут служить дополнительным свидетельством семантического сходства между словами [5, 6, 7].

Векторные модели слов могут быть обучены с использованием различных текстовых коллекций и методов, что ведет к тому, что разные векторные представления захватывают контекст по-разному. Отсюда можно предположить, что некоторые комбинации векторов, так называемые мета-векторные представления [8], могут улучшить векторные

¹Тихомиров Михаил Михайлович — младший научный сотрудник НИВЦ МГУ, e-mail: tikhomirov.mm@gmail.com.

Tikhomirov Mikhail Mikhailovich — junior researcher, Lomonosov Moscow State University, Research Computing Center.

представления слов, что позволяет добиться лучшего предсказания семантического сходства между словами.

В недавней работе [9] было показано, что комбинации векторных представлений, обученных на общей предметной области, рассчитанные на больших текстовых коллекциях из сети Интернет, оказывают существенное влияние на качество пополнения таксономий, таких как WordNet [3] и RuWordNet [10].

В данной работе рассматривается вопрос применения разработанного метода на основе использования мета-векторных представлений слов для автоматизированного (с использованием аннотаторов) расширения таксономии на конкретную предметную область. Рассматривается область информационной безопасности, которая используется для обогащения Онтологии Естественных Наук и Технологий (ОЕНТ) [11, 12].

2. Описание подхода

В основе подхода лежит использование мета-векторных представлений слов, которые позволяют добиться более высокого качества предсказания гиперонимов за счет комбинирования в векторной модели как предметно ориентированные векторные представления, так и модели обученные на общей предметной области.

Метод состоит из двух основных компонентов: алгоритм подготовки мета-векторных представлений и алгоритм предсказания гиперонимов для целевых слов. Мета-векторные представления строятся с использованием автокодировщиков [13] (AAEME, SAEME) в комбинации с дополнительной функцией потерь, добавляющей информацию о тезаурусе [9] (triplet loss). Предсказание гиперонимов реализовано через алгоритм машинного обучения, где входные признаки формируются на основании метрик близости между сущностями в тезаурусе и целевым словом, используя векторную модель. В качестве алгоритма машинного обучения использовалась логистическая регрессия, которая обучалась на задаче классификации, то есть, предсказания, является ли обрабатываемый концепт гиперонимом целевого слова. Результаты автоматической оценки качества работы подхода представлены в Таблице 1.

3. Описание реализованной системы

В рамках работы была реализована система предсказания гиперонимов и веб-сервис для работы с ней. Работа с системой происходит следующим образом:

Метод	MAP	MRR
concat	0.386	0.434
SVD	0.387	0.433
CAEME	0.385	0.434
CAEME triplet loss	0.408	0.456
AAEME	0.414	0.463
AAEME triplet loss	0.427	0.479

Таблица 1. Оценка качества расширения ОЕНТ-lite

Слово: ЛИСП

ЯЗЫК ПРОГРАММИРОВАНИЯ	0.98	Пусто	Гиперонимы Гипонимы
ПРОЦЕДУРНО-ОРИЕНТИРОВАННЫЙ ЯЗЫК ПРОГРАММИРОВАНИЯ	0.895	Пусто	Гиперонимы Гипонимы
ВЫСОКОУРОВНЕВЫЙ ЯЗЫК ПРОГРАММИРОВАНИЯ	0.856	Гипероним	Гиперонимы Гипонимы
ИСКУССТВЕННЫЙ ЯЗЫК	0.787	Пусто	Гиперонимы Гипонимы
КОМПЬЮТЕРНЫЙ ЯЗЫК	0.763	Пусто	Гиперонимы Гипонимы
ОБЪЕКТНО-ОРИЕНТИРОВАННЫЙ ЯЗЫК	0.749	Пусто	Гиперонимы Гипонимы
ФОРМАЛЬНЫЙ ЯЗЫК	0.742	Пусто	Гиперонимы Гипонимы
ЯЗЫК ПРОГРАММИРОВАНИЯ НИЗКОГО УРОВНЯ	0.736	Пусто	Гиперонимы Гипонимы
СИСТЕМНОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ	0.734	Пусто	Гиперонимы Гипонимы
ПРОГРАММА-ТРАНСЛЯТОР	0.713	Пусто	Гиперонимы Гипонимы

Следующий Прошлый Сохранить

Рис. 1. Интерфейс веб-сервиса.

- 1) Администратор системы запускает метод предсказания гиперонимов для интересующего набора слов,
- 2) Результат предсказания загружается в веб-сервис,
- 3) Аннотаторы получают доступ к системе, где для каждого слова есть набор из 10 предсказаний, которые должны разметить.

Каждое предсказание, которое отображается пользователю, также содержит инструменты отображения дополнительной информации о гиперонимах и гипонимах предсказанного концепта. В данный инструмент также входит возможность добавить в текущий список более подходящий концепт. Помимо этого выводится информация о весе предсказания, и сам список упорядочен в соответствии с этим весом. От аннотаторов требуется: просмотреть список, пополнить его при необходимости близкими концептами и разметить, связан ли каждый концепт с целевым словом некоторым отношением.

Список литературы

- [1] Berners-Lee, Tim and Hendler, James and Lassila, Ora, "The semantic web", *Scientific american*, **284**:5 (2001), 34–43.

- [2] Gómez-Pérez, Asunción and Corcho, Oscar, “Ontology languages for the semantic web”, *IEEE Intelligent systems*, **17**:1 (2002), 54–60.
- [3] Miller, George A, *WordNet: An electronic lexical database*, 1998.
- [4] Nikishina, Irina and Logacheva, Varvara and Panchenko, Alexander and Loukachevitch, Natalia, “RUSSE’2020: Findings of the First Taxonomy Enrichment Task for the Russian Language”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2020.
- [5] Fu, Ruiji and Guo, Jiang and Qin, Bing and Che, Wanxiang and Wang, Haifeng and Liu, Ting, “Learning semantic hierarchies via word embeddings”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, 1199–1209.
- [6] Levy, Omer and Remus, Steffen and Biemann, Chris and Dagan, Ido, “Do supervised distributional methods really learn lexical inference relations?”, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, 970–976.
- [7] Nikishina, Irina and Panchenko, Alexander and Logacheva, Varvara and Loukachevitch, Natalia, “Studying Taxonomy Enrichment on Diachronic WordNet Versions”, *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.
- [8] Coates, Joshua and Bollegala, Danushka, “Frustratingly Easy Meta-Embedding-Computing Meta-Embeddings by Averaging Source Word Embeddings”, *arXiv preprint arXiv:1804.05262*, 2018.
- [9] Tikhomirov, MM and Loukachevitch, NV, “Meta-Embeddings in Taxonomy Enrichment Task”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2021.
- [10] Loukachevitch, Natalia V and Lashevich, German and Gerasimova, Anastasia A and Ivanov, Vladimir V and Dobrov, Boris V, “Creating Russian wordnet by conversion”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2016, 405–415.
- [11] Dobrov, Boris V and Loukachevitch, Natalia V, “Development of Linguistic Ontology on Natural Sciences and Technology”, LREC, 2006, 1077–1082.
- [12] Tikhomirov, Mikhail and Loukachevitch, Natalia and Dobrov, Boris, “Methods for Assessing Theme Adherence in Student Thesis”, *International Conference on Text, Speech, and Dialogue*, 2019, 69–81.
- [13] Bollegala, Danushka and Bao, Cong, “Learning word meta-embeddings by autoencoding”, *Proceedings of the 27th international conference on computational linguistics*, 2018, 1650–1661.

Development of an automated system for taxonomy enrichment based on texts of a specific domain

Tihomirov M.

The paper considers the application of the developed method based on the use of meta-vector representations of words for automated (using annotators) enrichment of a domain taxonomy. The information security domain is considered, which is used to enrich the Ontology of Natural Sciences and Technologies (OENT). *Keywords:* taxonomy, meta-embedding, vector representation, interface.

References

- [1] Berners-Lee, Tim and Hendler, James and Lassila, Ora, “The semantic web”, *Scientific american*, **284**:5 (2001), 34–43.
- [2] Gómez-Pérez, Asunción and Corcho, Oscar, “Ontology languages for the semantic web”, *IEEE Intelligent systems*, **17**:1 (2002), 54–60.
- [3] Miller, George A, *WordNet: An electronic lexical database*, 1998.
- [4] Nikishina, Irina and Logacheva, Varvara and Panchenko, Alexander and Loukachevitch, Natalia, “RUSSE’2020: Findings of the First Taxonomy Enrichment Task for the Russian Language”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2020.
- [5] Fu, Ruiji and Guo, Jiang and Qin, Bing and Che, Wanxiang and Wang, Haifeng and Liu, Ting, “Learning semantic hierarchies via word embeddings”, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, 1199–1209.
- [6] Levy, Omer and Remus, Steffen and Biemann, Chris and Dagan, Ido, “Do supervised distributional methods really learn lexical inference relations?”, *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2015, 970–976.
- [7] Nikishina, Irina and Panchenko, Alexander and Logacheva, Varvara and Loukachevitch, Natalia, “Studying Taxonomy Enrichment on Diachronic WordNet Versions”, *Proceedings of the 28th International Conference on Computational Linguistics*, 2020.
- [8] Coates, Joshua and Bollegala, Danushka, “Frustratingly Easy Meta-Embedding—Computing Meta-Embeddings by Averaging Source Word Embeddings”, *arXiv preprint arXiv:1804.05262*, 2018.
- [9] Tikhomirov, MM and Loukachevitch, NV, “Meta-Embeddings in Taxonomy Enrichment Task”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2021.
- [10] Loukachevitch, Natalia V and Lashevich, German and Gerasimova, Anastasia A and Ivanov, Vladimir V and Dobrov, Boris V, “Creating Russian wordnet by conversion”, *Computational Linguistics and Intellectual Technologies: papers from the Annual conference “Dialogue”*, 2016, 405–415.
- [11] Dobrov, Boris V and Loukachevitch, Natalia V, “Development of Linguistic Ontology on Natural Sciences and Technology”, *LREC*, 2006, 1077–1082.
- [12] Tikhomirov, Mikhail and Loukachevitch, Natalia and Dobrov, Boris, “Methods for Assessing Theme Adherence in Student Thesis”, *International Conference on Text, Speech, and Dialogue*, 2019, 69–81.
- [13] Bollegala, Danushka and Bao, Cong, “Learning word meta-embeddings by autoencoding”, *Proceedings of the 27th international conference on computational linguistics*, 2018, 1650–1661.