ОБРАБОТКА МАЛЫХ ЯЗЫКОВ ПУТЁМ МОДЕЛИРОВАНИЯ УСВОЕНИЯ ЯЗЫКА

Д. И. Качков

Белорусский государственный университет, г. Минск;
dmitriydikanskiy@gmail.com;
науч. рук. – М.К.Буза, д-р тех. наук, проф.

В данной работе изучается проблема моделирования малых языков, для которых доступно ограниченное количество ресурсов для обучения. В настоящее время основное направление работы в рассматриваемой области – адаптация существующих моделей, разработанных для крупных языков, к условиям дефицита ресурсов. В данной работе рассматривается другой подход – моделирование усвоения языка ребёнком. Упор сделан на исследование усвоения лексики.

Ключевые слова: обработка естественного языка; малые языки; моделирование онтогенеза; векторное представление слов; нейронные сети

ВВЕДЕНИЕ

Туканские или туканоанские языки — семейство языков коренных жителей Южной Америки, которое насчитывает около 20 — 25 языков. Туканские племена проживают на северо-восточных территориях бассейна Амазонки, в южных и юго-восточных регионах Колумбии, а также прилегающих территория Бразилии, Перу и Эквадора. Общая численность носителей туканских языков — около 20 тысяч человек [1].

Обосновать актуальность проблемы моделирования малых языков – в частности, туканской языковой семьи, – можно с помощью следующих четырёх аргументов.

Социальный аргумент. В настоящее время большинство материалов в глобальной сети Интернет доступно на английском, русском, китайском, испанском, турецком языках. Носители этих языков имеют доступ к бытовым, медицинским, экономическим, юридическим материалам. Носителям малых языков эти материалы не доступны. Возникает информационное неравенство. С помощью автоматических переводчиков с больших языков на малые можно дать людям равный доступ к знаниям. Показательно, что весной 2020 года появилось много работ о переводе медицинских текстов на языки Азии, Африки и Южной Америки: было необходимо как можно скорее распространить информацию о пандемии коронавируса и мерах борьбы с ним [2][3].

Лингвистический аргумент. Разработка языковой модели позволит использовать её в качестве дополнительного источника информации о малом языке. Исследования малых языков, в свою очередь, актуальны для

лингвистики, поскольку каждая черта каждого естественного языка даёт дополнительную информацию об устройстве языка вообще. Кроме того, отдельные характеристики малых языков являются необычными, то есть редко встречающимися в мире. Например, примечательность категории эвиденциальности в туюка отмечалась лингвистами [4, p. 240].

Культурный аргумент. Языковая модель может поспособствовать сохранению языка. С одной стороны, с её помощью можно порождать уникальный тексты и материалы на целевом языке. С другой стороны, она может использоваться как вспомогательный инструмент при обучении языку. Важность сохранения языков неоднократно обсуждалась в литературе [5]. Язык и мышление тесно связаны, и поэтому можно утверждать, что каждый язык отражает уникальный взгляд на мир. Следовательно, сохранение языков актуально не только для лингвистики, но также для этнографии, этнологии и культурной антропологии.

Научно-технический аргумент. Для компьютерной лингвистики моделирование малых языков — это вызов, который требует новых идей. После создания архитектуры нейронных сетей Transformer [6] развитие моделейкрупных языковв основном идёт в направлении увеличения числа параметров и обучающей выборки. Например, в [7] был подготовлен корпус на триллион слов, а в [8] исследуется возможность обучить модель, содержащую триллион параметров. Для малых языков подобные подходы не применимы в силу малого числа доступных материалов. Исследователям надо предложить новые архитектуры, которые смогут работать в условиях дефицита ресурсов.

МОДЕЛИРОВАНИЕ УСВОЕНИЯ ЯЗЫКА

Среди разработанных подходов к обработке малых языков можно выделить следующую тенденцию: приёмы, выработанные для крупных языков, адаптируются под условия дефицита ресурсов. Чаще всего адаптация осуществляется одним из двух путей: автоматическое расширение корпуса размеченных данных и совместное обучение, то есть трансфер знаний, полученных при обучении крупному языку [9].

В данной работе выдвигается другая гипотеза: возможно разработать модель малого языка на основе имитации усвоения языка ребёнком.

Важно понимать, что механизмы, которые помогают ребёнку эффективно усвоить родной язык, достоверно не изучены, однако стадии изучения языка детьми исследованы подробно [10, с. 75-85].

Вопрос моделирования онтогенеза языка исследовался в литературе. Как правило, целью таких исследований преимущественно является проверка психолингвистических гипотез о процессе изучения языка [11, p. 92].

В данной работе представлено несколько экспериментов, касающихся усвоения лексики малого языка.

ИЗУЧЕНИЕ МЕТОК ДЛЯ «ВИДИМЫХ» ПРЕДМЕТОВ

Изучение языка не происходит в отрыве от изучения мира: это два взаимно обусловленных процесса. Советский психолог Л. С. Выготский сближал факт развития значения слова с фактом развития сознания. Для него слово — это средство, которое отражает внешний мир в его связях и отношениях [12, с. 42].

Первый эксперимент предполагает, что каждому высказыванию сопоставлена некоторая «сцена», а слова интерпретируются как метки для событий, представленных на «сцене». Задача обучающейся системы – соотнести предметы и метки.

Данный эксперимент был основан на работе [13]. Источником высказываний служила база CHILDES [14], представляющая записи диалогов маленьких детей и их родителей. Компоненты фразы лемматизировались и переводились с помощью словаря на язык туюка. В качестве «сцены» использовался тот же набор лемм, с точки зрения программы представляющий собой набор объектов иной природы. Задачей системы было построение вектора вероятности того, что данное слово соответствует тому или иному объекту «сцены».

Обучение проводилось на 18000 предложениях, содержащих около 8000 уникальных слов. Для пар «слово – метка» модель оценивала вероятность, что данная метка обозначается данным словом. Корректные пары, встречающиеся более раза, получили оценку от 0.43 до 0.85, среднее значение – 0.71.

ИЗУЧЕНИЕ ЗНАЧЕНИЯ КВАНТОРНЫХ МЕСТОИМЕНИЙ

Отдельный интерес представляет изучение таких лексем, значение которых непредметно. В частности, таковой является местоимение «все».

В эксперименте рассматривалось множество объектов, обладающих двумя характеристиками: форма и цвет. Значение характеристик выбиралось из ограниченных множеств. Ситуацией называется набор, содержащий несколько объектов, не обязательно различных.

Каждой ситуации сопоставляется высказывание, построенное по следующей схеме: КВАНТОР + ЦВЕТ + ФОРМА. Здесь квантор – одно из слов «каждый», «существует», «отсутствует». Например, «каждый синий

треугольник», т. е. Каждая фигура в «ситуации» является синим треугольником. Задача интеллектуальной системы — вычислить, является ли истинным данное утверждение для данной ситуации.

Для обучения использовалась нейронная сеть с широким внутренним слоем (512 нейронов). Пересчёт коэффициентов производился для каждых 5-10 элементов выборки. Также использовались дополнительный оптимизации (оптимизатор Adam[15]) и технология дропаута — метода регуляризации нейронной сети, заключающегося в исключении на каждом шаге обучения случайного набора нейронов [16].

Тренировочная выборка содержала 4000 случайно сгенерированных экспериментов. Тестовая выборка содержала 500 уникальных записей. На 25 эпохах обучения удалось получить качество обучения около 94% (96,2% на наиболее успешном запуске).

ЗАКЛЮЧЕНИЕ

В данной работе обоснована актуальность моделирования малых языков. Выдвинута гипотеза, согласно которой модель усвоения языка ребёнком может стать средством для автоматической работы с малыми языками. Произведено два эксперимента, касающихся усвоения лексики. Первый эксперимент включал соотношение слов-меток и объектов на «сцене». Второй эксперимент касался изучения значений непредметных слов, таких как «все» («каждый») и «никто» («отсутствует»). Результаты показали, что системы могут усваивать значения слов с помощью имитации взаимодействия ребёнка с окружающим миром.

Следующим этапом данного исследования должно стать усвоение морфологии и синтаксиса высказываний.

Библиографические ссылки

1. *Герасимов Д. В.* Туканоанские языки. // Большая российская энциклопедия. Том 32. Москва, 2016, стр. 478-479.
2. *Anastasopoulos A., Cattelan A., Dou Z.-Y., ..., Tur S.* TICO-19: the Translation Initiative for Covid-19 // Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. — https://doi.org/10.18653/v1/2020.nlpcovid19-2.5.
3. *Spangher A., Peng N., May J., Ferrara E.* Enabling Low-Resource Transfer Learning across COVID-19 Corpora by Combining Event-Extraction and Co-Training // Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020.
4. *Ferdinand De Haan.* The Interaction of Modality and Negation: A Typological Study // The University of Southern California, 1994. — 270 p. — https://doi.org/10.4324/9781315052380.
5. *Замятин К., Пасанен А., Саарикиви Я.* Как и зачем сохранять языки народов России // Хельсинки, 2012. — 181 с. — ISBN 978-952-93-0407-3.

6. *Vaswani A., Shazeer N., Parmar N., …, Polosukhin I.* Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 4–9 December 2017, pp. 6000–6010. Available at: https://arxiv.org/abs/ 1706.03762.
7. *Raffel C., Shazeer N., Roberts A., …, Liu P. J.* Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer // ArXiv preprint 2019: https://arxiv.org/abs/1910.10683.
8. *Narayanan D., Shoeybi M., Casper, J., …, Zaharia M.* Efficient Large-Scale Language Model Training on GPU Clusters // ArXic preprint 2021: https://arxiv.org/abs/2104.04473.
9. *Hedderich M. A., Lange L., Adel H., Strötgen J., Klakow D.* A Survey on Recent Approaches for Natural Language Processing in Low-Resource Scenarios // ArXiv preprint, 2020: https://arxiv.org/abs/2010.12309.
10. *Бурлак С. А.* Происхождение языка. Факты, исследования, гипотезы // «Альпина Диджитал», 2019.
11. *Freudenthal D., Alishahi A.* Computational Models of Language Development // Brooks P. J., Kempe V. (eds). Encyclopedia of Language Development. —1st ed. — 2014. — SAGE Publications, Inc. — pp. 92-96.
12. *Лурия, А. Р.* Язык и сознание. Под редакцией Е. Д. Хомской // М: Изд-во Моск. Ун-та, 1979, 320 с.
13. *Fazly A., Alishahi A., Stevenson S.* A Probabilistic Computational Model of Cross-Situational Word Learning // Cognitive Science. — 2010. — vol. 34. — iss. 6. — pp. 1017-1063. — https://doi.org/10.1111/j.1551-6709.2010.01104.x.
14. *MacWhinney, B.* The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.) // Lawrence Erlbaum Associates Publishers, 2000.
15. *Kingma D. P., Ba J. Adam.* A Method for Stochastic Optimization // Adam: A Method for Stochastic Optimization. Proceedings of the 3rd International Conference on Learning Representations, 2015.
16. *Srivastava N.* Dropout: A Simple Way to Prevent Neural Networks from Overfitting / N. Srivastava et al. // Journal of Machine Learning Research, 2014, vol. 15, no. 1. — P. 1929–1958.