УДК 004.825

М. И. СВЯТОЩИК

К ВОПРОСУ О СЕМАНТИЧЕСКОМ ТЕГИРОВАНИИ ЛЕКСЕМ

В статье демонстрируется семантическая разметка русских лексических единиц предметной области «Фотоника» в категориях частей языка (тайгенах и ёгенах) в ручном режиме. Рассматриваются конкретные приёмы и методы конвертации частей речи в части языка.

Ключевые слова: комбинаторная семантика, части речи, части языка, конвертация, искусственный интеллект, естественный язык.

M.I. SVYATOSHCHIK

REGARDING THE ISSUE OF SEMANTIC TAGGING OF LEXEMES

The article demonstrates semantic markup of Russian lexical units in the field of 'Photonics' within language parts (taigens and yogens) in manual mode. Specific techniques and methods for converting parts of speech to language parts are considered.

Keywords: combinatory semantics, parts of speech, language parts, conversion, artificial intelligence, natural language.

Современное развитие искусственного интеллекта в области анализа интернет-контента и работы с естественным языком происходит в трех ключевых аспектах:

1. Интеллектуализация Интернета и веб-платформ, превращение его в, своего рода, "глобальный мозг" (Giant Global Graph, GGG), что подразумевает интеграцию интеллектуальных функций на глобальном уровне.

2. Применение методов машинного обучения, нейросетей и статистических алгоритмов для обработки контента, включая работу с большими объемами данных (Big Data), что является основой цифровой трансформации экономики и переход к седьмому технологическому укладу. Примером служит создание чат-ботов на базе генеративного ИИ, таких как ChatGPT от OpenAI.

3. Использование машинного обучения с нейросетями и статистическими алгоритмами для анализа больших языковых моделей (Large Language Model, LLM).

Данные за последние годы показывают, что в ближайшее десятилетие основное внимание международных научных и финансовых центров будет уделено созданию графов знаний. Это позволит автоматически извлекать значимую информацию с веб-страниц, разрабатывая формализованный язык, сопоставимый по мощности с естественным языком.

В этом контексте вклад Минской школы вычислительной семантики, включая разработку Универсального семантического кода (УСК) профессором В. В. Мартыновым и интеграцию данных в Теорию автоматического порождения архитектуры знаний (ТАПАЗ), предоставляет инструменты для понимания смысла предметных областей и решения задачи "понимания" машиной текстов на естественном языке [1].

Когда правила языка имеют меньше исключений, чем правил с этими самыми исключениями, неизбежно возникает путаница до тех пор, пока не будут установлены единые критерии для различения частей речи в предложении. Так, наряду со словами, обозначающими предметы (существительные), их свойства (прилагательные), действия и процессы (глаголы) и способ их осуществления (наречия), выделяются местоимения и числительные как самостоятельные части речи. Они могут быть аналогами существительных (например, "он", "восемь") или прилагательных ("его", "восьмой"). Слово "столовая" классифицируется как существительное, хотя обозначает место принятия пищи, а слово "бег" считается существительным с глагольным значением. Междометиями называют слова, выражающие эмоции и желания (например, "ох", "ах"), но во фразе "а девица хи-хи-хи да ха-ха-ха!" эти же слова рассматриваются как междометия, хотя они описывают процесс и выполняют функцию сказуемого. Предлоги, союзы, частицы также считаются частями речи, несмотря на то что они отражают языковые конструкции, а не реальность мира. Например, предлог "на" в выражении "на столе" не обозначает объект, а служит локативом для построения предложения. Если изменить структуру предложения так, чтобы она соответствовала модели мира (например, "поверхность стола поддерживает книгу"), предлог исчезнет. Вспомогательные синтаксические средства (предлоги и т.д.) помогают связывать элементы языковых конструкций, но сами по себе не являются членами предложения [2].

Существующих противоречивых моментов достаточно, чтобы поддержать мнение Фердинанда де Соссюра и Луи Есперсена о том, что категории частей речи не являются абсолютно точной языковой категорией, а их описание значительно отличается от идеала Евклидовой геометрии в своей строгости и совершенстве [3, с. 154-170]. В этом контексте стоит вспомнить слова Генриха фон Рейхенбаха о том, что геометрия была представлена как завершенная система, основанная на аксиомах и служащая образцом научной строгости [4, с. 16]. Однако в рамках классификации частей речи по греколатинской традиции, устранить противоречия не представляется возможным из-за отсутствия одно-однозначного соответствия между морфологией, синтаксисом и семантикой. Л. Теньер выражал сомнения относительно этой классификации, указывая на ее недостатки и необходимость четкого разделения основных и второстепенных признаков для создания иерархии критериев [5, с. 55].

Для решения данной проблемы предлагается парадигма частей языка А. Н. Гордея [6], имеющая корни в китайско-японской лингвистической традиции [7], [8]. В отличие от традиционного морфологического подхода, эта теория акцентирует внимание на содержании знака и его процедуральном определении. Обратимся к рассмотрению основных понятий данной теории:

Стереотип – повторяющийся элемент представления.

Модель мира – архитектура стереотипов, т. е. упорядоченное множество стереотипов и упорядоченное множество преобразований одних стереотипов в другие.

Языковая картина мира – частично упорядоченное множество стереотипов и частично упорядоченное множество преобразований одних стереотипов в другие.

Индивид – отдельная сущность в выделенном фрагменте модели мира.

Знаки алфавита синтаксиса – средства метаязыка (предлоги, послелоги, союзы, частицы и проч.), используемые для построения и преобразования языковых структур.

Части языка – подмножества языковой системы, элементами которых являются знаки с общим предельно абстрактным значением.

Тайген – часть языка, обозначающая индивид: *стол*, *восемь, мы.*

Ёген – часть языка, обозначающая признак индивида: *бежать*, *синий*, *смело.*

Лексема – тайген или ёген конкретного естественного языка.

Члены предложения – роли частей языка в предложении [6].

По происхождению тайгены и ёгены разделяются на производные от ёгенов или тайгенов (столовая, деревянный) и непроизводные (дерево, белый). Информационными называются производные и непроизводные тайгены и ёгены, обозначающие индивидов или признаки индивидов в информационном фрагменте модели мира *(мысль, думать),* физическими - в физическом *(топор, рубить)*. Информационные и физические тайгены и ёгены квалифицируются как постоянные или переменные в зависимости от того, обозначают ли они постоянных индивидов *(вера, небо)* и постоянные признаки индивидов *(духовный, небесный)* или переменных индивидов *(ты, то)* и переменные признаки индивидов *(мечтать, бежать)*. Постоянные и переменные тайгены и ёгены подразделяются на качественные *(душа, добрый, читать)* и количественные *(деньги, тройной, вычислять)*, далее – на одноместные, то есть обозначающие один индивид или один признак индивида *(шкаф, целый, открыть),* и многоместные, т. е. обозначающие множество индивидов или множество признаков индивидов *(мебель, пестрый, открывать)* [2, с. 24].

Части языка различаются четырьмя параметрами:

1) *семантически* – интуитивное распознавание тайгенов и ёгенов по общему предельно абстрактному значению проверяется процедурно: если в результате подстановки диагностируемого элемента слева от *...вызывает...* получается отмеченное предложение, то он – ёген, если неотмеченное, то тайген, напр. *Бег вызывает усталость,* но **Город вызывает...?* [9, с. 17–19] для повышения надежности теста и исключения метафор, типа *Город вызывает восхищение* рекомендуем использовать встречную процедуру: если справа от *...тогда, когда* диагностируемый элемент в роли сказуемого превращает предложение в правильное и семантически тождественное предложению с *...вызывает...,* то он – ёген, в противном случае – тайген,

причем допускается любое переписывание аффиксов, напр. *Я устаю тогда, когда бегаю*, но **Я восхищаюсь тогда, когда горожу?* – правильное предложение: *Я восхищаюсь тогда, когда вижу город* с диагностируемым элементом в роли дополнения, следовательно, *бег* – ёген, а *город* – тайген;

2) *синтаксически* – в развернутом предложении ёгены занимают центральные позиции, тайгены – маргинальные [9, с. 24];

3) *синтагматически* – в тайгенах на первом месте модификатор, на втором – актуализатор (часто свёрнут в суффикс или стерт: *белый заяц —> бел-як, булочная лавка —► булочная-);* в ёгенах, наоборот, на первом месте актуализатор (часто свернут в префикс или стёрт: *покрыться морщинами —► с-морщиться, бежать галопом —> -галопировать),* на втором – модификатор [10, с. 24];

4) *парадигматически*– ёгены имеют степень, тайгены – нет.

Теггирование – маркировка, нанесение тегов на поверхность предложения. Тег – метка, которая размечает информацию и позволяет облегчить процесс поиска необходимой информации. Описанный выше порядок разграничения ЧЯ апробирован на множестве текстов выделенной предметной области «Фотоника». В качестве примера представлен текст «Наноэлектростанция», где разграничиваются категориальные признаки, в соответствии с парадигмой частей языка А. Н. Гордея [6].

«Наноэлектростанция» (тайген, развёрнутый, сложный, физический, постоянный, нарицательный, качественный, одноместный) преобразует (ёген, свёрнутый, сокращённый, физический, переменный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, одноместный) колебания (ёген, свёрнутый, сжатый, переменный, 1 степени, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) молекул (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, многоместный, экстенсивный) в (знак алфавита синтаксиса (ЗАС)) электричество (тайген, свёрнутый, сокращённый, физический, постоянный, нарицательный, качественный, многоместный, экстенсивный) .

Исследователи (тайген, свёрнутый, сокращённый, информационный, постоянный, нарицательный, качественный, многоместный, экстенсивный) создали (ёген, свёрнутый, сокращённый, физический, переменный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, одноместный) генератор (тайген, развёрнутый, сложный, физический, постоянный, нарицательный, качественный, одноместный) энергии (тайген, свёрнутый, сжатный, физический, постоянный, нарицательный, качественный, многоместный, интенсивный), работающий (ёген, свёрнутый, сокращённый, физический, постоянный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, одноместный) на (ЗАС) основе (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, одноместный) молекулярных (ёген, свёрнутый, сжатый, физический,

постоянный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) колебаний (ёген, свёрнутый, сжатый, переменный, 1 степени, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный). Физики (тайген, свёрнутый, сокращённый, информационный, постоянный, нарицательный, качественный, многоместный, экстенсивный) разработали (ёген, свёрнутый, сокращённый, физический, переменный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) устройство (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, одноместный) для (ЗАС) сбора (ёген, свёрнутый, сжатый, переменный, 1 степени, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) молекулярной (ёген, свёрнутый, сжатый, физический, постоянный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, одноместный) энергии (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, одноместный), которое (ёген, свёрнутый, сжатый, постоянный, 1 степени, качественный, произвольный, относительный, повествовательный, повествовательный, одноместный) улавливает (ёген, свёрнутый, сокращённый, физический, переменный, 1й степени, нарицательный, качественный, произвольный, относительный, повествовательный, одноместный) естественные (ёген, свёрнутый, сжатый, физический, постоянный, 2й положительной степени, качественный, многоместный, нарицательный, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) колебания (ёген, свёрнутый, сжатый, переменный, 1 степени, качественный, произвольный, относительный, повествовательный, многоместный, экстенсивный) молекул (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, многоместный, экстенсивный) в (ЗАС) жидкости (тайген, свёрнутый, сжатый, физический, постоянный, нарицательный, качественный, многоместный, интенсивный).

«После теггирования, с опорой на словарь А. А. Зализняка, нами составляются парадигмы для каждой части языка и выделяются леммы. Лемма – первоначальная словарная форма слова [11]. Лемматизация (англ. lemmatization) – метод морфологического анализа, который сводится к приведению словоформы к ее первоначальной словарной форме (лемме). В результате лемматизации от словоформы отбрасываются флективные окончания и возвращается основная или словарная форма слова. Например, в русском языке словарной формой для существительного считается именительный падеж, единственное число (*мечами - меч*); для глагола – инфинитивная форма (*читали - читать*); для прилагательного – единственное число, именительный падеж, мужской род (*заснеженными - заснеженный*) [12].

Тайгены в именительном падеже в единственном и множественном числе рассматривается как разные леммы, а ёгены во множественном числе – варианты одной леммы.

«Необходимо отметить, что семантическая категория количества не имеет своим прямым коррелятом морфологическую категорию числа. Во-первых, не все греко-латинские части речи имеют парадигму единственного и множественного числа. Во-вторых, для некоторых частей речи категория числа оказывается не связанной с их семантикой (знак *белые* имеет форму множественного числа, однако обозначает только одно свойство индивидов, знак *пестрый* имеет форму единственного числа, но обозначает некоторую совокупность свойств; *умерли* – единократный процесс, несмотря на морфологическую форму множественного числа, *стучал* – многократный процесс, несмотря на единичность участвующего в нем индивида), и для того, чтобы описать количественность обозначаемых данными частями речи явлений, требуется введение дополнительных категорий: итеративность или фреквентативность для глаголов [5, с. 450–451]. В-третьих, даже для частей языка, которые обозначают индивидов, а не их признаки, морфологическая категория числа не всегда верно отражает количественный характер этих индивидов: знак *Афины* имеет форму множественного числа, однако обозначает один индивид, в то время как знак *толпа*, имеющий форму единственного числа, служит для обозначения множества, некоторой совокупности индивидов. Таким образом, применение морфологической категории числа в анализе языка весьма ограничено и даже в рамках этой ограниченной области не всегда дает однозначные результаты относительно отражения идеи единичности/множественности в языке.

«Для обозначения категорий языка, соответствующих категории количества в модели мира мы будем придерживаться терминологии А.Н. Гордея [2, с. 174], и вслед за ним выделять одноместные, то есть обозначающие один индивид или один признак индивида, и многоместные, то есть обозначающие множество индивидов или множество признаков индивидов, части языка» [2, с. 96–97]. Тогда *Афины* - одноместный тайген, *толпа* - многоместный тайген; *белые, стукнули* - одноместные ёгены, а *пестрый, стучал* - многоместные ёгены» [13, с. 29–36].

Процедура конвертации частей речи в части языка помогает устанавливать одно-однозначное соответствие между синтаксическим и семантическим анализом предложения, а также является перспективным исследованием и наработкой для создания синтаксического анализатора на семантической основе.

Эти разработки открывают новые возможности для преодоления барьеров в развитии ИИ, не имея аналогов среди существующих решений.

ЛИТЕРАТУРА

1. *Святощик, М. И.* Алгоритм автоматизированной семантической разметки текстов / М. И. Святощик, А. Н. Гордей, Р. С.Панащик, О. А.Стрельчонок, В. В. Ткаченко. // XXIII Международная научно-техническая конференция «Развитие информатизации и государственной системы научно-технической информации», – Минск: ОИПИ НАН Беларуси, 2024. – С. 400–404.
2. *Гордей, А. Н.* Части языка вместо частей речи / А. Н. Гордей // Язык. Глагол. Предложение. – Смоленск: СПГУ, 2000. – С. 258 –271.
3. *Соссюр, Ф. де.* Труды по языкознанию / Ф. де Соссюр. – Москва : Прогресс, 1977. — 695 с.
4. *Рейхенбах, Г.* Философия пространства и времени / Г. Рейхенбах. – М. : Прогресс, 1985. – С. 16.
5. *Теньер, Л.* Основы структурного синтаксиса / Л. Теньер ; пер. с фр. И. М. Богуславского [и др.] ; вступ. ст., общ. ред. В. Г. Гака. – М. : Прогресс, 1988. – 656 с.
6. *Гордей, А. Н.* Парадигма частей языка / А. Н. Гордей // Словообразование и номинативная деривация в славянских языках : материалы VIII Междунар. науч. конф., Гродно, 15–16 апр. 2003 г. / Гродн. гос. ун-т ; отв. ред.: С. А. Емельянова, Л. В. Рычкова, А. В. Никитевич. – Гродно, 2003. – С. 173–179.
7. *Цзин, Чжаоцзы*. Исследование грамматики национального языка / Чжаоцзы Цзин. – Шанхай.: Языки и языкознание, 1955. – С. 199.
8. *Задоенко Т. П.* Учебник китайского языка / Т. П. Задоенко, Хуан Шуин. – М., 2002. – 761 с.
9. *Мартынов, В. В.* Категории языка / В. В. Мартынов. – М. : Наука, 1982. – 192 с.
10. *Гордей, А. Н.* Принципы исчисления семантики предметных областей / А. Н. Гордей. – Минск : Белорус. гос. ун-т, 1998. – 153 с.
11. *Большакова, Е. И.* Автоматическая обработка текстов на естественном языке и анализ данных: учеб. пособие / К. В. Воронцов, Н. Э. Ефремова, Э. С. Клышинский, Н. В. Лукашевич, А. С. Сапин – М.: НИУ ВШЭ, 2017.
12. Лемматизация. [Электронный ресурс]. – Режим доступа : https://promopult.ru/library/Лемматизация.
13. *Святощик, М. И.* О конвертации частей речи в части языка / М. И. Святощик // Вестник МГЛУ. Филология. Вып. № 1(122) – Минск. –2023. – С. 29–36.

Информация об авторе:

Святощик Марина Игоревна – мл. научный сотрудник Объединённого института проблем информатики (ОИПИ) Национальной академии наук, г. Минск, Республика Беларусь.