

ЛИНГВИСТИЧЕСКИЕ СТРАТЕГИИ В ДЕКОДИРОВАНИИ СЕМАНТИКИ СЛОЖНЫХ СЛОВ ДЛЯ РАЗВИТИЯ КОМПЬЮТЕРНЫХ МОДЕЛЕЙ ОБРАБОТКИ ТЕКСТА

С. Г. Милько

Белорусский государственный университет

Минск, Беларусь

e-mail: milkosg@bsu.by

В статье рассматриваются особенности машинного перевода идиоматичных сложных существительных, представляется анализ примеров и проблемы, с которыми сталкиваются современные модели обработки текста. Особое внимание уделяется лингвистическим стратегиям и перспективам развития алгоритмов для точного понимания и передачи смысла таких конструкций.

Ключевые слова: идиоматичность; идиоматичность сложных существительных; Google Translate; машинный перевод; семантический анализ; лингвистические стратегии; нейронные сети.

LINGUISTIC STRATEGIES IN DECODING THE SEMANTICS OF COMPLEX WORDS FOR ADVANCING COMPUTER TEXT PROCESSING MODELS

S. G. Milko

Belarusian State University

Minsk, Belarus

e-mail: milkosg@bsu.by

The article explores the intricacies of machine-translating idiomatic complex nouns, presenting an analysis of examples and the challenges faced by contemporary text processing models. Special attention is given to linguistic strategies and the prospects for algorithm development to accurately understand and convey the meaning of such constructions.

Keywords: idiomaticity; idiomaticity of complex nouns; Google Translate; machine translation; semantic analysis; linguistic strategies; neural networks.

В современном мире, где международное взаимодействие становится все более интенсивным, вопросы машинного перевода и обработки сложных слов представляют собой актуальную проблему. Точность перевода, особенно в контексте идиоматичных сложных существительных, остается вызовом для компьютерных моделей обработки текста.

Актуальность данного аспекта обусловлена необходимостью понимания, как компьютеры обрабатывают и переводят идиоматичные выражения и сложные существительные. Особый интерес представляет их перевод, поскольку часто машинные системы предлагают перевод частей сложных слов, игнорируя их идиоматический контекст. Например, как машина анализирует слово *flatbed*, и какой перевод предлагается?

Это слово состоит из двух частей: *flat* ‘плоский’ и *bed* ‘кровать’. Однако, в контексте техники или транспорта *flatbed* означает ‘платформа’ или ‘грузовик с плоской платформой’. Компьютер, не обладая контекстом или знанием специфики области, может буквально перевести части слова и предложить ‘плоская кровать’, что в данном контексте абсолютно неверно.

Существующие исследования в области машинного перевода охватывают разнообразные аспекты, включая оптимизацию алгоритмов и анализ проблем семантического понимания. Многие из работ фокусируются на развитии эффективных методов для улучшения точности перевода, адаптации к контексту и обработке сложных языковых структур.

Одной из ключевых работ в этой области является "Neural Machine Translation by Jointly Learning to Align and Translate" Д. Богданова (D. Bahdanau), в которой была представлена новая модель нейронного машинного перевода, которая опирается на механизм внимания. Этот подход значительно улучшил качество машинного перевода по сравнению с традиционными методами. Нейронная сеть в данной модели способна сосредотачиваться на различных частях входного текста при переводе, что позволяет более точно передавать смысл и улучшать качество перевода [2].

Другой значимой разработкой является модель BERT, описанная в статье Дж. Девлина (J. Devlin) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", которая была опубликована в журнале arXiv в 2018 году. Модель BERT широко применяется в задачах машинного перевода и обработки естественного языка, позволяя модели учитывать контекст и добиваться более точного понимания значения слов и фраз в тексте [1].

Рассмотрение идиоматичных сложных существительных в контексте машинного перевода открывает важные аспекты, препятствующие автоматизированным системам в полной интерпретации и переводе таких языковых единиц. Представим более детальный анализ нескольких примеров с целью продемонстрировать многогранность и особенности таких лингвистических конструкций в рамках машинного перевода.

1. *Blackleg*. Данный термин относится к человеку, продолжающему работать во время забастовки или замещающему бастующих работников. Однако, машинный перевод Google «Несмотря на забастовку, ‘черная ножка’ вышла на работу, что вызвало напряжение среди бастующих сотрудников» упрощает содержание фразы, не учитывая общественные и этические аспекты действий *blackleg*.

2. *Stickybeak*. Это существительное определяет человека, навязчиво или грубо пытающегося узнать личную информацию о других людях. Однако, машинный перевод ‘прилипчивый человек’ не отражает негативной окраски фразы и не передает полную семантическую нагрузку данного термина.

3. *Hothead*. Данный термин описывает человека, склонного к необдуманным действиям из-за быстрого реагирования. Однако, перевод ‘горячая голова’ пренебрегает нюансами смысла и не учитывает возможные негативные последствия такого поведения.

4. *Hardtop*. Это определение автомобиля с металлической крышей. Однако, перевод ‘жесткая крыша’ не передает всего спектра технических характеристик и не учитывает особенности автомобильной терминологии.

Эти примеры демонстрируют наличие трудностей, с которыми сталкиваются разработчики систем машинного перевода в попытках создать алгоритмы, которые интерпретируют идиоматичные сложные существительные. Необходимость контекстуального и семантического понимания в данных случаях подчеркивает важность человеческого вмешательства в процессе машинного перевода для более точной и адекватной передачи смысла подобных лингвистических конструкций.

Системы машинного перевода, включая Google Переводчик, оперируют различными моделями, такими как нейронные сети и трансформеры, для обеспечения более точного и быстрого перевода. Несомненно, они значительно продвинулись в области интерпретации и перевода текста, в том числе идиоматически сложных слов. Однако, идиоматические сложные слова представляют собой особый вызов для машинного перевода из-за своей многозначности и семантической гибкости. Google Переводчик сталкивается с проблемой интерпретации идиом в контексте, так как часто предложенный перевод может упрощать или упускать идиоматический смысл фразы. Ученые стремятся улучшить возможности моделей переводить идиоматические выражения, однако они все еще имеют ограничения в точности передачи нюансов и подтекстов идиом, особенно связанных с культурными оттенками. Работы над улучшением этих моделей включают в себя адаптацию алгоритмов и обучение систем на более разнообразных и культурно насыщенных данных, чтобы улучшить их способность воспроизводить идиоматические выражения с большей точностью и адекватностью.

Неотъемлемым аспектом в сфере моделей обработки текста является важность использования и развития лингвистических стратегий для более точного понимания и передачи семантики текста. Рассмотрение

конкретных примеров идиоматичности позволило увидеть, как сложности контекста и семантики могут препятствовать точной интерпретации компьютерных моделей. Использование лингвистических подходов, таких как семантические анализы, контекстуальные модели и синтаксические разборы, помогает системам машинного перевода более эффективно декодировать сложности языка.

В заключении необходимо подчеркнуть важность необходимости постоянного совершенствования и развития моделей обработки текста для более эффективного понимания и передачи смысла сложных языковых конструкций и идиоматических выражений в контексте машинного перевода.

Библиографические ссылки

1. Devlin J. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [Electronic resource] // North American Chapter of the Association for Computational Linguistics. 2019. Mode of access: <https://www.semanticscholar.org/paper/BERT%3A-Pre-training-of-Deep-Bidirectional-for-Devlin-Chang/df2b0e26d0599ce3e70df8a9da02e51594e0e992> (date of access: 05.09.2023).
2. Bahdanau D. Neural Machine Translation by Jointly Learning to Align and Translate: papers from International Conference on Learning Representations [Electronic resource] // International Conference on Learning Representations. 2014. Mode of access: <https://www.semanticscholar.org/paper/Neural-Machine-Translation-by-Jointly-Learning-to-Bahdanau-Cho/fa72afa9b2cbc8f0d7b05d52548906610ffbb9c5> (date of access: 02.09.2023).