УДК 004.8

Анализ технологии искусственного интеллекта для адаптации социальных диалектов разных языковых групп

Е. В. Долженко[1], В. С. Вольвач[1], А. П. Иванова[1], А. А. Шарапов[1]*

[1] Сибирский государственный университет геосистем и технологий, г. Новосибирск, Российская Федерация

* e-mail: ekaterina70302@mail.ru

Аннотация. В статье рассматривается машинный перевод и принцип его работы для адаптации социальных диалектов разны языковых групп. В ходе работы авторы выяснили и показали на примере, как именно искусственный интеллект распознает слова в предложении и переводит их.

Ключевые слова: машинный перевод, искусственный интеллект, жаргон, социолекты, анализ, жаргонизмы

Analysis of artificial Intelligence technology for adaptation of social dialects of different language groups

E. V. Dolzhenko[1], V. S. Volvach[1], A. P. Ivanova[1], A. A. Sharapov[1]*

[1] Siberian State University of Geosystems and Technologies, Novosibirsk, Russian Federation

* e-mail: ekaterina70302@mail.ru

Abstract. The article discusses machine translation and the principle of its operation for the adaptation of social dialects of different language groups. In the course of the work, the authors found out and showed by example exactly how artificial intelligence recognizes words in a sentence and translates them.

Keywords: machine translation, artificial intelligence, jargon, sociolects, analysis, jargonisms

Введение

Все мы сталкиваемся с такой проблемой, как жаргонизмы. Они встречаются и в нашей повседневной жизни, например, согласитесь, Вам становится трудно построить диалог, когда Вы не до конца понимаете собеседника и информацию, которую он хочет Вам передать.

Поэтому целью работы стало исследование и анализ системы машинного перевода (МП) для адаптации социальных диалектов разных языковых групп.

Задачи, которые перед собой поставили авторы работы: найти диалектизмы, употребляемые той или иной языковой группой; расшифровать их; проанализировать и рассмотреть, как искусственный интеллект распознает диалектизмы.

Методы и материалы

Как уже было сказано выше, цель исследования – рассмотрение принципа работы МП. Поэтому давайте рассмотрим, что же такое МП и социалекты?

Социолект или социальный диалект – это групповые речевые особенности, характерные для какой-либо социальной группы – профессиональной, возрастной, субкультуры.

Одним из видов социалектов являются жаргонизмы.

Жаргонизмы – это слова, составляющие жаргон.

Жаргон – искусственно созданный специфический словарь, понятный только отдельной социальной группе.

Примерами жаргонизмов являются «посудина» (в сфере судоходства означает корабль), «самоделкин» (в медицине – врач-травматолог) и другие. (рис. 1).

Сфера применения	Жаргонизм	Значение жаргонизма
Судоходство	Посудина	Корабль
Программирование	Линк	Гиперссылка
	Баг	Недоработка в компьютерной программе
	Бета	Пробная версия программы
Медицина	Самоделкин	Травматолог
	Гармошка	Аппарат искусственной вентиляции легких (ИВЛ) с ручным приводом

Рис. 1. Примеры жаргонизмов

Как видно из примеров жаргонизмы используются в самых разных сферах деятельности.

Как уже авторы упомянули выше, цель работы – рассмотреть системы машинного перевода для адаптации социальных диалектов разных языковых групп. Поэтому давайте рассмотрим работу МП, основанного на лингвистическом анализе:

Шаг 1. Сначала загружаются исходные предложения текста из файла или из буфера в памяти.

Шаг 2. Система разбивает предложения на слова и определяет границы предложения.

Этот шаг обманчив, несмотря на кажущуюся простоту, разбиение текста на слова и предложения в общем случае далеко не банально. Слова распознаются с помощью определенных шаблонов. Они описывают различные буквенные, цифровые и буквенно-цифровые группы и символы пунктуации, которые затем выделяются в качестве отдельных слов. В результате анализа выделенных слов, некоторым из них (инициалам, сокращениям и т. п.) будут присвоены специальные маркеры, которые помогают разрешить многозначность при распознавании границ предложений. Также на этом этапе происходит подготовка слов для поиска в словаре [2].

Шаг 3. Проводит морфологический анализ исходного текста.

Решение данной задачи основывается на словаре исходного языка.

Процесс поиска слов по словарю предполагает, как поиск оригинального слова в случае, если оно не было найдено в словаре, так и поиск слов с удалением префиксов. Для более эффективного поиска этих префиксов используют древовидную структуру. Её элементами являются буквы предлогов. Поиск прекращается,

когда нет дальнейшего перехода в дереве или найден предлог и слово без него существует в словаре [3].

Шаг 4. Осуществляет синтаксический анализ исходного текста.

Сначала для каждого слова находится главное слово, с которым оно согласовано в результате перевода. При этом нельзя утверждать, что многозначность полностью снята. Лишь в процессе поиска главных слов происходит полное снятие многозначности [4].

Шаг 5. Производит семантический анализ исходного текста.

Главная задача данного этапа – снятие многозначности на основе полученной древовидной структуры зависимостей. В первую очередь снимается многозначность базовых слов. Исследования показали, что лучше всего использовать попарное согласование рядом стоящих базовых слов в обратном порядке, т. е. в порядке обратном положению слов в предложении. Когда всем базовым словам присваивается в соответствие один лексико-грамматический класс, происходит «досогласование» зависимых от них слов [2].

Шаг 6. Делает перевод построенного дерева.

Шаг 7. Производит согласование переведенного дерева.

В результате перевода жаргонизмов на литературный язык получаем в некоторой мере согласованную древовидную структуру зависимостей. Чтобы получить полное согласование, используется процедура, которая аналогична процедуре окончательного разрешения многозначности, которая применялась на этапе построения дерева. Так как перевод жаргонизмов осуществлялся на основе дерева зависимостей, то данная процедура помогает получить согласованное представление предложения на литературном языке [1].

Шаг 8. Записывает переведённое предложение в файл или в буфер.

Этот шаг не окончательный, практически всегда допускается постредактирование человеком.

Рассмотрим пример работы МП.

На вход попадает предложение: «На 13 минуте ГГ рипнулся на хард левеле». Система разбивает предложение на слова и определяет его границы. Дале начинается поиск слов по словарю, используя древовидную структуру. После этого начинается перевод и согласование на основе дерева зависимостей. На выходе мы получаем переведенное и согласованное предложение (рис. 2).

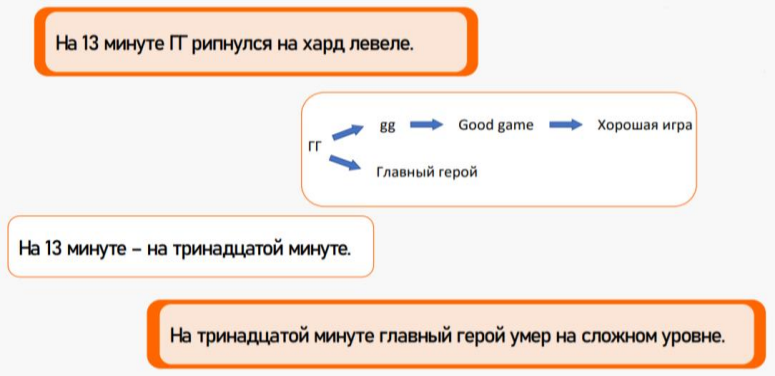

Рис. 2. Пример работы МП

Заключение

Рассмотрев работу машинного перевода на основе лингвистического анализа, авторы изучили перевод жаргонизмов на литературный язык; исследовали работу системы МП для адаптации социальных диалектов разных языковых групп; разобрались и показали на примере работу МП.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Кулагин О.В. Исследования по машинному переводу. М.: Наука, 1979. - 376 с.
2. Воронович В.В. Машинный перевод. - Минск: 2013. - 39 с.
3. Зубов А.В., Зубова И.И. Основы искусственного интеллекта для лингвистов. М., 2007.
4. Кулагина О. С. Исследования по машинному переводу. – М., 1979.
5. Козеренко Е.Б. Глагольно-именные трансформации при англо-русском машинном переводе. – URL: https://www.dialog21.ru/digests/dialog2007/materials/html/43.htm.
6. Панич Ю. В. Предварительная идентификация неоднозначного исходного текста и его перевод на другие языки с использованием системы согласованных словарей. – URL: http://www.sciteclibrary.ru/rus/catalog/pages/9402.html.
7. Машинный перевод. – URL: https://ru.wikipedia.org/wiki/Машинный_перевод.
8. Нейронный машинный перевод Google. – URL: https://ru.wikipedia.org/wiki/Нейронный_машинный_перевод_Google.
9. Глубокое обучение. – URL: https://ru.wikipedia.org/wiki/Глубокое_обучение.