УДК 81’33

К ОСНОВНЫМ ТРУДНОСТЯМ ПРИ СОЗДАНИИ RULE-BASED И CORPUS-BASED СИСТЕМ МАШИННОГО ПЕРЕВОДА

Гудков В.В., магистрант, тел. 8(969)702-66-15, vadim0006@gmail.com
СПбГУ, Санкт-Петербург, Россия

Ключевые слова: *машинный перевод, статистический машинный перевод, перевод на основе правил.*

Работа освещает основные проблемы при создании систем машинного перевода на основе корпусов и правил, возникших в ходе их разработки. Предлагаются примеры и предложены варианты решения возникших проблем.

Введение. Среди прикладных задач современной компьютерной лингвистики особое значение имеет разработка наиболее эффективной системы машинного перевода: в силу многозначности и многоаспектности человеческого языка, данные системы очень сложны для реализации. По мере разработки систем машинного перевода (Rule-based MT, Corpus-based MT) были выявлены трудности, препятствующие созданию безупречных систем, нашедшие отражение в данной статье.

Цель работы – выявить и формализовать неочевидные трудности при создании систем машинного перевода на основе правил и на основе статистики.

Материал и методика исследований. Работа была выполнена, основываясь на опыт разработки подобных систем на кафедре математической лингвистики Санкт-Петербургского государственного университета, а также документальный анализ и наблюдение

В ходе исследования был сделан упор на теоретический и практический анализ принципов создания RBMT и CBMT систем.

Так, в основе систем машинного перевода, основанных на правилах (Rule-based MT), лежат двуязычные словари лексических единиц и грамматик, охватывающие основные семантические, морфологические и синтаксические закономерности языков, что обусловливается принципом работы таких систем: анализом связи структуры входного и выходного предложения на всех релевантных данной задаче уровнях языка. Помимо очевидных недостатков подобных систем (высокая стоимость, долгие годы трудоёмкой работы лингвистов, неспособность показать высокие результаты в узких областях) были выявлены нетривиальные ситуации: неоднозначность перевода полисемантических синтаксических структур.

Рассмотрим конкретный пример из проведенного экспериментального исследования: герундий может в определенных ситуациях одновременно включать в себя особенности глагола и существительного, что в значительной

степени влияет на механизм перевода [5]. Структуры русского языка категориально похожие на английский герундий в таких ситуациях – это отглагольные существительные, обозначающие какую-либо деятельность (singing → пение, reading → чтение). Однако, если верить результатам исследования систем, основанных на правилах, прямой машинный перевод герундия в русское отглагольное существительное в системах RBMT является наименее вероятным вариантом перевода из трех возможных схем: герундий → предложение со спрягаемой формой глагола; герундий → предложение с инфинитивом; герундий → отглагольное существительное.

Кроме того, возникали препятствия на лексическом уровне: в языках существуют идиоматические выражения, смысл которых не может быть предсказан из значения и расположения составляющих элементов. Например, выражение “to be hand in glove” не могло быть корректно переведено без дополнительного словаря идиоматических выражений. Подробные результаты исследования машинного перевода идиом представлены в исследовании Microsoft [4].

Исследование проблем при создании Corpus-based систем так же начиналось с теоретического обоснования принципа их работы: в её основе лежит сравнение большого объема языковых пар в форме параллельных корпусов, из которых и происходит извлечение машинного перевода: он генерируется на основе статистических моделей, выводящихся из теоремы Байеса (каждое предложение на одном языке является одним из возможных переводов любого предложения в другом языке, и наиболее подходящим является тот перевод, которому система присваивает наивысшую вероятность).

В ходе практических экспериментов были выявлены следующие препятствия: во-первых, на практике количество примеров в крупномасштабном двуязычном корпусе может быть неограниченно (они могут быть дополнительно разложены более чем одним возможным способом на более короткие примеры и взаимно накладываться друг на друга [3]). Во-вторых, в параллельных корпусах перевод некоторых предложений исходного языка зачастую осуществлялся методом дробления или объединения, что сбивало статистическую модель [4]. В-третьих, в основополагающих моделях могли присутствовать статистические аномалии, они могли не принимать во внимание, к примеру, имена собственные в тех случаях, когда в обучающей выборке наличествовал избыток похожих предложений с другими именами. Например, «I booked a hotel in Budapest» могло неверно перевестись как «Я забронировал отель в Париже», если сочетание «a hotel in Paris» встречалось в выборке чаще [5].

Результатом исследований стало создание новых критериев оценки качества систем машинного перевода на основе правил и на основе корпусов.

Кроме того, стало понятно, что гибридизация подходов может помочь в решении вышеуказанных проблем.

Заключение. Таким образом, нетривиальность задачи создания полноценной и эффективной системы машинного перевода многогранного человеческого языка, в некотором роде, осложняется статистическим барьером корпусных исследований. С учётом подобных ограничителей, система машинного перевода должна принимать решение с учётом неполноты знаний. Однако, гибридизация подходов и использование лингвистических модулей для предварительной обработки текста в будущем позволит нам в решении некоторых проблем, однако перевод, выполненный человеком, всё ещё остаётся недостижимым эталоном.

Библиографическое описание

1. Козеренко Елена Борисовна Синтаксическая многозначность и неоднозначность в перспективе машинного перевода // Rhema. Рема. 2016. №1.
2. Smets M., Pentheroudakis J., Menezes A. Translation of Verbal Idioms // Microsoft Research. – 2002. 10 p.
3. Pan H. Example-Based Machine Translation: A New Paradigm. // Department of Chinese, Translation and Linguistics City University of Hong Kong. – 2002. 22 p.
4. Brown P. A Statistical Approach to Machine Translation // Computational Linguistics. – 1990. – Vol.16 №2. – P. 79-85.
5. Okpor, M. Machine Translation Approaches: Issues and Challenges // IJCSI International Journal of Computer Science Issues. – 2014. – Vol. 11 №5 (2). – P. 159 – 165.

THE MAIN CHALLENGES IN THE DEVELOPMENT OF RULE-BASED AND CORPUS-BASED MACHINE TRANSLATION SYSTEMS

Gudkov V.

Key words: *machine translation problems, corpus-based machine translation, rule-based machine translation.*

The paper investigates the main problems in the creation of corpus-based and rule-based machine translation systems which appeared during their development. Examples of occurring confusions are shown and possible solutions to the arisen problems are proposed.