УДК 81'33

АНАЛИЗ НЕКОТОРЫХ ПРОБЛЕМ И ЗАДАЧ КОМПЬЮТЕРНОЙ ЛИНГВИСТИКИ.

Куликова Ю.Д., Лешукова К.А.
научный руководитель ст. преподаватель Лабушева Т. М.
Сибирский федеральный университет

В наше компьютеризированное время люди любого возраста предпочитают искать всю необходимую информацию в интернете, пользуясь специально разработанной программой. При этом они не всегда получают то, что ищут. Так, например, при попытке перевести текст или фразу с одного языка на другой с помощью онлайн переводчика частенько можно получить неадекватный и некорректный перевод. Поиск правильных путей решения на запрос, составленный на естественном языке, является целой наукой. И эта наука – компьютерная лингвистика, которая занимается составлением подобных программ.

Создание программ, обрабатывающих запрос на естественном языке, - трудоемкий процесс, требующий слаженной работы программистов и лингвистов, и он происходит следующим образом. Лингвисты передают материал по описанию естественного языка программистам, а они в свою очередь, основываясь на этой информации, разрабатывают решения задач, выявленных из практики использования языка. Получается, что инженеры-программисты нуждаются в том, чтобы лингвисты обеспечили их моделями языка. Но, к сожалению, на сегодняшний день лингвистами не выявлено универсальной модели языка, которою можно было бы приложить к любому языковому материалу. Один и тот же запрос можно интерпретировать различными способами, вложить в него различный смысл. Для обработки языка необходим последовательный анализ, этапы которого соответствуют уровням структуры языка. И все, что нужно программисту – это воплотить в жизнь последовательность этапов анализа. Для наглядного представления этой последовательности сделаем пояснение к каждому из её уровней, который нужно пройти программисту.

1. Лексический анализ. На данном этапе нужно разобраться, что следует принимать за слово. Любой текст представляет собой последовательность символов. Одним словом, принято считать некоторую последовательность символов, ограниченную справа и слева какими-либо разделителями. Но не все так просто. Не во всех языках присутствуют пробелы между словами. В китайском языке, например, их нет. Могут существовать элементы языка, о которых раньше никто и не задумывался. К ним можно отнести те же номера телефонов, различные формулы, адреса сайтов или почты, и т. д... И, кроме того, за слово программисту желательно принимать устойчивые словосочетания и фразеологизмы.

2. Морфологический анализ. Здесь следует определить грамматические характеристики лексем. После того, как мы выделили элементы языка, необходимо определить их статус. Например, узнать форму слова, и приписать ему какие-либо параметры (падеж, склонение, род, от какого слова образован и т. д.). На данном этапе встречаются свои проблемы. Допустим, у нас есть две формы слова. С точки зрения поиска нужно понять это, то самое слово или уже другое. Например, если в запросе есть слово «идти», то нужно решить, выдавать ли тексты, в которых есть другие формы этого слова – «шедший», «идя». Не всегда следует выдавать информацию с порожденной формой слова. Так, в случае слов «рыть – роем». Скорее всего, «рыть» - это рыть что-то, и вряд ли будет корректно вывести информацию со словом «роем», ведь это может быть формой слова «рой» (рой пчел). Еще одна задача состоит в том,

чтобы научиться различать фрагменты разных языков в одном запросе. Понять, о чем идет речь можно по контекстам или, привлекая знания об окружающем мире. И конечно, отдельная большая проблема – «синонимы». Их можно рассматривать как экзотическое расширение морфологии. В поиске «синонимов» помогают стандартные паттерны. К ним относятся варианты написания слов (нокиа – нокия), транслитерация (apple - эппл), слитно-раздельно (ярко-красный/яркокрасный). Также, люди часто делают ошибки в написании слов, эту проблему тоже нужно как-то решать.

3. Синтаксический анализ. На данном этапе следует установить структуру предложения – системы связей между словами. В тексте, предложении есть система связей, которая создается разными способами – вспомогательными словами, пунктуацией, порядком слов. Бывают такие специфические запросы, в которых необходимо учитывать стоп-слова (это предлоги, которые встречаются довольно часто). Например, «машина без крыши», в данном запросе предлог важен, а вот предлог «в» встречается почти в каждом документе, и поэтому его можно проигнорировать.

4. Семантический анализ. Этап, на котором нужно разобраться со смыслом текста, понять, о чем идет речь. Буквально сопоставить слову его значение в толковом словаре. Чтобы глубже разобраться с этим, программа классифицирует текст по жанру, тематике, «взрослости» текста, естественности. Задача усложняется еще и тем, что существуют так называемые лакуны. Нормы языка, принятые в какой-то области, специализации, регионе, за пределами которых о существовании таких слов могут и не знать.

5. Прагматический анализ. Интерпретация семантической структуры в контексте модели текста и знаний о мире.

Итак, допустим, что программист воплотил все эти этапы в жизнь, но все же нужный ответ он так и не получил. Дело в некоторых неудобных свойствах языка.

1) Неоднозначность – одно из самых неприятных свойств. Существуют выражения, которые можно интерпретировать по-разному, например, анекдоты, иронии построены на этом. Оказывается, это свойство можно наблюдать на всех уровнях языка, но мы часто этого не замечаем, так как обычно наше понимание языка опирается на контекст. Если не учитывать это свойство, то системы автоматического анализа текста могут выдавать тысячи вариантов разбора, причем для каждого варианта можно будет найти объяснение, языковое основание. «Его семью хлебами не прокормишь». В данном примере программа должна сделать выбор между существительным и числительным.

2) Несимметричность языков. В разных языках разные способы кодирования одинакового смысла. Например, в русском языке существует конструкция, которая передает значение приблизительности путем перестановки слов. («В классе человек 50»). Системе дословного перевода не справиться с этой проблемой. Также есть смыслы, которые в разных языках передаются с помощью каких-то грамматических значений. Например, при переводе с английского языка на русский мы не переводим артикли, при переводе с русского на китайский опускается число существительного.

3) Вариативность. В естественном языке существует множество способов передать один и тот же смысл. Способов перифразов много: на вас лежит ответственность за мероприятие, вы отвечаете за успех мероприятия, успех, с которым пройдет мероприятие, зависит от вас, и т. п..

4) Конвенциальность. Существуют конструкции допустимые теоретически, но не употребляемые практически. То есть существуют конвенции, которые нужно соблюдать. В русском языке нельзя сказать «потерпеть удачу» или «роем ласточек».

5) Эллиптичность. В естественном языке действует множество умолчаний. Случается, некоторые слова опускаются, так как они понятны из контекста. Но это в

естественном языке, а для машинного понимания необходимо восстанавливать опущенную информацию.

6) Непрозрачность. Сложные средства референции. Вместо повторения одних и тех же слов, мы зачастую заменяем их различными способами (Петя – мой друг, он, мой товарищ).

Говоря о вышеприведенных свойствах языка, необходимо отметить, что все они до конца не изучены. Из-за этого на пути инженера-программиста постоянно возникают все новые и новые трудности. В ход идет альтернативный способ, который основывается на статистическом подходе. Программы-переводчики полагаются на закономерности, выявленные при анализе миллиардов текстов, вследствие чего переведенный текст может выглядеть абракадаброй. Вот так из-за несостоятельности лингвистов создать универсальную модель языка страдает пользователь, получая «не переводимую игру слов» на требуемом ему языке.

Рассмотрим два наиболее ярких примера таких «шедевров» перевода, и приведем универсальный совет от лингвистов, как этого избежать.

Английскую фразу «He told me he had already had a letter from Mary which he would have enjoyed answering but he had to ignore it». Google переведет так: «Он сказал мне, что он уже получил письмо от Марии, который он наслаждался бы ответить, но он должен был игнорировать её». Во избежание такого перевода лингвисты рекомендуют конкретнее указывать взаимосвязи между словами, соблюдать пунктуацию при обращениях, подчинениях, приложениях и т. п.

Русское предложение «У нашей кошки родились котята – белый, рыжий и черный» может быть переведено на английский: «У нашей кошки родились котята – белый, рыжий и афроамериканец». Этот переводной ляп наглядно демонстрирует то, что нельзя писать так, как привыкли говорить. Необходимо использовать такие конструкции, которые понимаются однозначно.

И основная рекомендация пользователю программ машинного перевода заключается в следующем. Необходимо помнить, что нельзя опускать какие-либо слова в предложении, как это часто происходит в разговорной речи. Желательно перепроверять текст на орфографию, соблюдать устоявшиеся конвенции.

В заключение необходимо заметить, что главная задача компьютерной лингвистикой состоит, конечно же, не в том, чтобы дать универсальный совет пользователю, а в том, чтобы научить компьютер понимать смысл текста так, как это делает человек. На сегодняшний день эту проблему относят к разряду неразрешимой, так как на протяжении полсотни лет ученые всего мира тщетно бьются над её решением. Но все-таки громадный прогресс науки и техники в жизни современного общества позволяет надеяться на успех.