Система смыслового машинного перевода PEREVEDI.RU

М. В. Калинин

Группой лингвистов и программистов института математики им. С. Л. Соболева СО РАН был проведен анализ существующих в мире систем машинного перевода, как российских, так и зарубежных. Проведенный анализ позволяет утверждать, что в настоящее время в мире существуют две основных разновидности систем машинного перевода:

1. Полностью автоматизированный машинный перевод (Promt, Sokrat, Systran).
2. Перевод, осуществляемый человеком-переводчиком при поддержке механизма ассоциативной памяти, английский термин Translation Memory (Trados).

Кратко охарактеризовать историческую преемственность систем машинного перевода, по нашему мнению, можно следующим образом.

1-е поколение. Перевод слово в слово без различия семантики и тематики текста.

2-е поколение. Перевод с учетом устойчивых словосочетаний и с использованием тематическим словарей. Возможность выбора тематики текста.

3-е поколение. Семантический анализ контекста в котором находится то или иное словосочетание. Например, Иван вызвал пожар или Иван вызвал пожарных. Здесь определение смысла глагола вызвать и его перевод происходит по семантическому классу аргумента производимого действия «вызвать». В первом случае пожар относится к семантическому классу явление природы, в другом – персона.

4-е поколение. Учет смысловых структур текста. Широкое использование семантических классов в этих структурах. Наследование семантических классов один от другого по родовому признаку: человек – мужчина – служащий – милиционер. При этом во фразе «брюки милиционера» слово «милиционер» должно восходить к родительскому семантическому классу человек, т.к. этот предмет одежды можно отнести к самой вершине иерархии классов: «брюки человека», т.к. брюки могут быть как и мужчины, так и у женщины. Кроме того, в данном поколении систем будет производиться оценка так называемого «коэффициента смысла», далее будут отбрасываться синтаксически некорректные предложения или их группы, и из нескольких синтаксически правильных вариантов предложения или группы предложений выбираться фраза, имеющая наибольший «коэффициент смысла». Коэффициент смысла должен выбираться с учетом тематики текста, которая будет выбираться автоматически, при этом оставляя пользователю по его желанию возможность задать тематику текста самому.

5-е поколение. Создание модели обстановки и персонажей, о которых говорится в контексте данного текстового материала. Учет логичности фраз в соответствии с созданной моделью обстановки, предметов и персонажей, и выбор наиболее логичной фразы из нескольких вариантов. Учет логичности фраз будет основан на знании свойств имеющихся в данной обстановке персонажей, предметов и совершаемых ими или над ними действий.

Дальнейшие поколения уже будут связаны с разработками систем искусственного интеллекта. Системы такого рода обязательно будут самообучающимися.

Существующие в настоящие время системы машинного перевода можно отнести в основном к 2-му поколению. Есть попытки некоторых систем учитывать семантические классы, что относится к 3-му поколению, но (особенно это относится к российским системам, за исключениям систем ЭТАП и их аналогов) данное направление развито недостаточно.

Итогом анализа стала выработка принципиально нового похода к созданию систем автоматического перевода. В результате нами была создана *действующая система интеллектуального перевода* с английского языка на русский, обладающая свойствами,

которые дает возможность применять ее пользователям, не знакомым с английским языком.

Интереснейшей особенностью в создании системы PEREVEDI.RU явилась идея применить теоретические основы ООП (Объектно-ориентированного программирования) к проблемам разработки систем анализа и синтеза естественных языков. Данный подход при его практической реализации дал интересный результат в виде принципиального нового уровня качества перевода для определенных тематик текста по сравнению с существующими системами МП. Основная идея, полученная в результате экспериментов последних лет, состоит в том, что для получения высококачественного МП требуется отказаться от конструирования отдельных слов и фрагментов текста согласно заданным алгоритмам и перейти к оперированию макроструктурами, имеющими отношение не к морфологии, а к семантике. От перевода слов и словосочетаний мы перешли к переводу смысла. Хочется подчеркнуть, что упор на смысл переводимого текста является актуальным именно для систем МП, так как людей-переводчиков уже с давних пор обучают передавать именно СМЫСЛ (то есть тому, как бы это написал тот или иной автор, если бы думал и писал на вашем языке). Принцип же пословного перевода был принят в Древней Руси по отношению к священным текстам, и к нам он попал из Византии. Думаем, что настала пора обратить свой взор в будущее и отойти от древних традиций пословного буквального перевода в системах машинного перевода!

Смысловой перевод позволяет достичь следующих эффектов:

1. Присутствует детальный учет контекста, в котором употребляется слово или словосочетание;
2. Наша система выдает только один вариант перевода, максимально близкий по смыслу, в то время как перевод слов во всех ведущих системах МП достаточно часто дается в двух взаимоисключающих вариантах, и пользователь должен сам догадаться, какой из них верный (при этом иногда ни один из двух вариантов не является правильным), что невозможно без знания языка. Например, “в газете сообщалось о танках (резервуарах)”.

Мы намерены в нашем подходе к МП произвести скачок «через ступеньку», т.е. изначально система PEREVEDI.RU разрабатывается как система машинного перевода 4-го поколения. При этом в архитектуру системы закладываются элементы для максимально упрощенного перехода к системам 5-го поколения.

Система PEREVEDI.RU является симбиозом систем ассоциативной памяти и машинного перевода. Причем эти две технологии в системе идеологически неразрывно связаны друг с другом и являются следствием использования двух особенностей: семантических сетей и иерархических классов. Использование данной структуры переводчика позволяет избежать как свойственных технологии ассоциативной памяти переводов недостатков, так и недостатков свойственных существующим на данный момент алгоритмам машинного перевода. Широко известно, что для ассоциативной памяти свойственно запоминание предложений, либо неизменяемых фрагментов текста, что может оказывается для многих применений эффективным в случае нефлективных языков, но показывает абсолютную беспомощность для флективных, к которым относиться русский язык. Этот недостаток исправляется запоминанием в базе данных ассоциативной памяти не просто жестких, с точки зрения ООП полностью константных, фрагментов текста, а запоминанием семантической сети предложения или словосочетания, в котором могут быть задействованы неконстатные, варьирующиеся ветви семантической сети. Эти ветви могут сами по себе являться вложенными семантическими сетями, либо иерархическими классами. В иерархию объединяются универсальные слова – аналоги Universal Words языка UNL, которые упорядочены в базе данных в виде древовидной структуры. Древовидная структура формируется по родовому

признаку, пользуясь терминологией ООП мы имеем здесь дело с иерархией наследованных друг от друга классов.

Семантические сети

Довольно часть в текстах можно встретить структуры предложения, которые довольно четко можно выделить структуру, которая одинакова для разных видов предложений: «если не будет дождя, мы пойдем на прогулку». Семантические сети – «если <условие>, то <действие>», причем условием может являться какое-либо событие: «Если <условие>, то <действие>». Возможен расширенный вариант данной структуры: «если <условие>, то <действие1>, а в противном случае <действие2>. Здесь мы видим четкое сходство с подобной структурой “if-then-else” объектно-ориентированного программирования. Можно привести примеры и других структур, таких, как either <предложение1>, or <предложение2>, [or <предложение3>]

neither <предложение1>, nor <предложение2>, [nor <предложение3>]

<adj> yet <adj>

Например, «powerfull yet flexible application», что переводится как «мощное, и в то же время гибкое приложение». В то же время структуры, встречающиеся в реальной речи, могут задаваться без явного указания ключевых слов (термин объектно-ориентированного программирования), таких, как if, then, else, and, or и т.д. Например, в текстах на естественных языках можно выделить такую универсальную структуру, как, например, «действие», который является аналогом однородного предложения. Данная структура свойственна для всех без исключения языков. Представим «действие» в виде объекта и следующих его компонентов:

Список подлежащих;

Сказуемое или список сказуемых

Обстоятельства: «Где, когда, как?»

И другие.

Допустим «кони и дети резвились, плескались, веселились в воде». Здесь кони и дети являются единым объектом, выполняющим одновременно список действий, которые тоже являются, по сути, единым объектом – списком действий, которые выполняет в список объектов. Таким образом получаем объект-действие, содержащий два объекта. Первый объект – список подлежащих, второй – список сказуемых. В примере: «В России в 1905 году произошла революция» четко видно обстоятельства места «В России» и времени «в 1905 году», которые входят в состав объекта «действие». И так далее.

В семантических цепях в качестве частей семантических цепей, которые могут варьироваться, используется имя семантического класса. Мы организуем семантические классы в иерархию наподобие того, как это делается в объектно-ориентированном программировании. Например класс «город», от которого могут наследоваться отдельные конкретные города. Или «научный работник» – гуманитарий, естественник, математик. При этом точно также, как в объектно-ориентированном программировании применимо наследование классов.

Идеи базировались во многом на языке UNL, но, в то же время, подобно квазиязыку при использовании объектной модели появляется возможность детализировать семантические модификаторы, указывать вероятность выбора различных вариантов, а также указывать альтернативные варианты.

Основной отличительно особенностью нашей системы смыслового перевода является отход от общепринятого перевода с использованием языковых пар в существующих системах. В системе PEREVEDI.RU производится перевод не одной языковой пары на другую, то есть, например, с английского языка на русский и с русского на английский, я осуществляется перевод с любого естественного языка на созданный нами специальный смысловой язык – интерлингву. Этот смысловой язык основан на

реально существующих в нашей жизни объектах и явлениях, а так же общечеловеческих понятиях, свойственных большинству современных естественных языков. Для реализации смыслового языка нами была создана база данных, содержащая в себе реально существующие в природе объекты, явления, понятия, и различные варианты действий и свойств объектов. Отметим, что идея и наполнение базы во некоторой степени основывается на стандартах международного сетевого языка UNL и является надстройкой над этим стандартом, обеспечивающая преобразование из структур нашего метаязыка в UNL, таким образом обеспечивая совместимость с ним в большей части случаев, но при этом с потерей некоторых смысловых и языковых оттенков и особенностей. Главным критерием отбора объектов, свойств и действий была их независимость от конкретики какого-либо человеческого языка. Действительно, языковые различия в нашем обществе очень сильны, языковой барьер является самым серьезным препятствием для обмена информацией. В то же время все мы живем в одном и том же мире и, соответственно имеется значительная общая база понятий, который в различных культурах обладают значительной степенью родства. При этом предлагаемый нами подход подразумевает добавление нового понятия, в том случае, если оно появилось только лишь в каком-то одном из поддерживаемых системой PEREVEDI.RU языков или в языковой группе.

Интерлингва построена по принципу максимального уточнения смысла текста – каждая смысловая фраза совершенно однозначно интерпретируется на любой из языков. Для каждого языка необходимо написание специального адаптера, который переводит текст на этом языке на интерлингву. Адаптер включает в себя мощный анализатор пунктуации и синтаксиса, смысловой анализатор, и работает по принципу максимально автономного от пользователя выбора наиболее адекватного по смыслу варианта. В данное время нами реализованы адаптеры для перевода с английского и русского языков на интерлингву. Об обратном преобразовании мы расскажем далее.

Рассмотрим примеры:

Иван вызвал пожар.

Иван вызвал пожарных.

Иван вызвал подчиненных.

В этих трех примерах одно и то же слово «вызвал» является «действием», слова «пожар», «пожарных», «подчиненных» –аргументом этого действия.

В зависимости от семантического класса аргумента происходит преобразование в соответствующее действие в интерлингве. Назовем действие такого вида «методом» по аналогии с обозначающим то же самое явление термином в объектно-ориентированном программировании. Также по аналогии того, как это определяется в объектно-ориентированном программировании назовем семантический класс или несколько семантических классов аргументами метода. Для данного случая в составе словаря интерлингвы имеется три метода:

вызвать/явиться-причиной (явление)	to cause
вызвать/удаленно-через-средства-коммуникации (персона)	to call
вызвать/официальное-лицо (официальная должность)	to summon

Здесь указаны значения этих методов при синтезе английского выходного текста. Заметим, что количество методов, например, для глагола «вызвать» в действительности может составлять значительно большее число.

Оценка «коэффициента смысла».

Система PEREVEDI.RU осуществляет смысловой анализ с целью избежать «смыслового взрыва», являющегося основным препятствием для разработчиков систем машинного перевода: в некоторых случаях может возникнуть слишком много вариантов интерпретации предложения. Путем смыслового анализа происходит оценка коэффициента смысла – синтезируется несколько вариантов ветвей графа. Для каждой из

ветвей, на основании анализа, выбирается комбинация, имеющая наибольший коэффициент смысла. При определении коэффициента смысла происходит подсчет встречаемости словосочетаний и терминов, которые относятся к данной тематике.

Перспективы использования

В направлении с иностранного на родной для пользователя язык будет осуществляться, используя все указанные методы, наиболее адекватный, качественный перевод, настолько качественный, насколько компьютер только сможет получить за счет логики работы системы перевода. В редких случаях смысловой неоднозначности не удается избежать, и система выбирает вариант, который не будет являться самым удачным по смыслу. Для этих редких случаев предусмотрена функция выбора альтернативного варианта для отдельного слова, словосочетания и предложения целиком. Заметив явную нелогичность в тексте, пользовать может просмотреть несколько вариантов перевода из выпадающего меню, привязанному к предложению целиком, либо к его отдельной части. Данная возможность будет особенно полезна при чтении пользователем новостей, а так же во всех других случаях перевода с иностранного языка на язык родной пользователю.

В направлении с родного на иностранный система PEREVEDI.RU будет с успехом использоваться для составления писем и интерактивного общения по ICQ. При составлении сообщения или письма будет производиться анализ неоднозначностей текста наподобие того, как производится проверка грамматики в MS Word. При этом неоднозначные слова и фразы будут выделяться подчеркиванием. Пользователь сможет произвести коррекцию собственного текста, так как это текст на родном для него языке. Например, внести определенность в значение слове «все», которое может означать «все» или «всё». «Спор», которое может означать споры грибов или спор в смысле дискуссия. «Пар», которое может означать водяной пар, либо несколько пар. «Вызвать» химическую реакцию, или вызвать человека и т.д. Компьютер не будет предлагать варианты на каждое второе слово, как это происходит в некоторых существующих в настоящее время системах машинного перевода. Это будет гарантироваться использованием базы данных словосочетаний и готовых предложений по технологии ассоциативной памяти, которая основана на наших собственных алгоритмах и отличается некоторыми принципиальными усовершенствованиями от существующих систем, в том числе Trados. Система PEREVEDI.RU определит случай, когда в тексте упоминаются «танки/резервуары» и выберет подходящий вариант, в зависимости от контекста использования словосочетания. Выбор варианта для данного слова изменит, по выбору пользователя, данное слово либо во всем тексте, либо только в данном месте текста.

Чем обеспечивается принципиальная разница в качестве перевода в системе PEREVEDI.RU помимо применения смыслового перевода? Существующие системы перевода направлены на перевод текстов по самым разным тематикам. В этих системах количество слов и словосочетаний составляет в среднем порядка 100,000 слов, и лишь в Systran превыщает 1 млн. слов и словосочетаний. При этом подавляющее большинство из этого словарного набора составляют единицы, относящиеся к научной, медицинской и технической терминологии. При этом на долю некоторой конкретной тематики, допустим новостей, либо интерактивного общения в системах, подобных ICQ приходится чрезвычайно бедный набор готовых структур. Мы намерены в нашей системе исследовать, то каким образом повысится качество перевода, в случае, если набор фраз для конкретной тематики, например, «Новости», превысит некий уровень, после которого количество уже перейдет в тот вид качества, при котором оно будет приемлемым для чтения новостей среднестатистическим пользователем интернет. Заметим, что в существующих в данное время системах этот уровень далеко еще на достигнут.

Мы планируем в ближайшие несколько месяцев сделать доступным для широкого пользователя портал смыслового перевода PEREVEDI.RU для того, чтобы пользователи интернет смогли по достоинству оценить новое качество перевода для трех направлений: деловая переписка, чтение новостей и общения на форуме, пользуясь англо-русским и русско-английским смысловыми переводчиками. Далее набор языков будет постепенно наращиваться. Напомним, что в системе PEREVEDI.RU возможны любые комбинации языковых пар, а не только те, которые жестко запрограммированы в системе, как это сделано в российских системах машинного перевода. В дальнейшем планируется создание специального модуля смыслового перевода, подключаемого к системе быстрого обмена сообщениями ICQ, а также к другим подобным системам.