***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

Состояние современных систем машинного перевода с русского языка на арабский

© Альотаиби Султан Маджед (Королевство Саудовская Аравия), 2011

Ввиду огромного прогресса в области коммуникативных связей и информационной технологии, машинный перевод стал необходимым средством в современное время и важным компонентом в различных отраслях жизни. Он особо востребован в коммерческой сфере для продажи или рекламирования товаров. Будучи окном, сквозь которое мы смотрим на достигнутую современную технологию и научные знания у других народов, машинный перевод требуется также в области научных исследований. Мы не будем говорить о важности машинного перевода, мы будем говорить о поддержке научной практики на системы машинного перевода с арабского языка на русский и наоборот. Мы попытаемся в этом докладе проанализировать эти системы, понять и оценить их, затем предложить необходимые рекомендации для поддержки данных систем, с тем, чтобы повысить качество работы этих систем.

Что такое машинный перевод

Термин «машинный перевод» многозначен. За долгую историю использования он приобрел множество интерпретаций. Сначала этот термин подразумевал только автоматические системы, работающие без участия человека [Sager 1994: 326]. Европейская ассоциация машинного перевода дала следующее определение: «использование компьютера для перевода текста с одного естественного языка на другой язык» [EAMT]. А Международная ассоциация машинного перевода (IAMT) определяет машинный перевод как «единовременный ввод полного предложения и генерирование соответствующего ему полного предложения (не обязательно хорошего качества)» [Hutchins 2000].

Ни одно из этих определений не предполагает вмешательства человека. Другие авторы, например Arnold и др., допускают некоторое участие человека: «как попытку автоматизировать полностью или частично процесс перевода с одного естественного языка на другой». Определение, в котором упоминается какая-либо форма участия человека, чаще всего воспринимается как «расплывчатое» [Balkan 1992: 408]. Эта точка зрения перекликается с мнением Аркера, который считает, что академические ученые и исследователи до сих пор расходятся во взглядах на

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

определение машинного перевода в отношении участия человека в этом процессе. Но поскольку ничего нового пока не ожидается, этот термин продолжает использоваться для обозначения полностью автоматизированных систем пусть даже и с участием человека [Somers 2003: 1-11].

Какова цель перевода?

Цель перевода заключается в передаче знаний носителям определенного естественного языка представителями других культур. Перевод представляет собой открытое для всех окно, сквозь которое мы можем взглянуть на современные технологии во всем мире. Коммерческие компании пользуются этим для достижения первого места в конкуренции в плане продажи товаров. Перевод есть важнейший шаг к культурной коммуникации.

Некоторые факторы, влияющие на процесс перевода

Рассмотрим некоторые факторы, влияющие на качество машинного перевода в рамках рассматриваемой языковой пары, а именно русского и арабского языков.

Синтаксическая конструкция: синтаксис естественного языка играет большую роль в анализе входного предложения и построении выходного. Во- первых нужно знать, есть ли сходство между языком оригинала и языком перевода? Например, персидский язык сходится с арабским языком во многих конструкциях, но не сходится, например, с китайским языком. Между русским и арабским языками также существуют определенные различия в синтаксисе.

Культура: каждый народ имеет свое представление о реальности. Наш язык организует наш взгляд на все то, что нас окружает. Языковая картина мира рисует нам окружающий нас мир. Мы видим то, что на самом деле видит наш язык.

Особую роль в машинном переводе играют некоторые обстоятельства, которые с точки зрения человека вроде бы не столь существенны. Так, велика роль знаков препинания в машинном анализе и синтезе.

Знаки препинания: Знаки препинания в исходном тексте: если бы система машинного перевода учитывала существующие в исходном тексте знаки препинания, то вопрос перевода стал бы проще. Знаки препинания должны учитываться при составлении алгоритмов анализа и перевода.

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

Знаки препинания в тексте перевода: нужно использовать знаки препинания в тексте перевода. Они способствуют разделению текста на сегменты, что облегчает понимание переведенного текста.

Семантика, или содержание переводимого текста. В основном можно считать, что она отражается в теме текста. Понятие «тема» достаточно широко, поэтому в целях упрощения можно оперировать понятиями «определенный» текст или «неопределенный», т. е. с достаточно широким содержанием.

Тема: под этим понимается характер текста, т. е. какой аспект внешнего мира затрагивает содержание данного текста? Определенный или не определенный?

Определенный: например, текст о прогнозе погоды, о медицине и т. д.

Неопределенный: тут возникает проблема так называемая «перевод текстов неопределенного характера» – «Unrestricted domain translation». К такого рода текстам можно, например, отнести некоторые высказывания политических лидеров, экономические тексты, критические статьи и т.п.

Причины превосходства систем машинного перевода за рубежом над системами арабского производства

Ученые за рубежом удаляют особое внимание анализу родного языка в синтаксическом, морфологическом и семантическом планах. Они прошли долгий путь в этой области и получили значительные результаты. В число возможных причин превосходства некоторых систем могут входить следующие моменты:

1. Их исследования в области машинного перевода опередили арабские на 40 лет.
2. Они имеют высокий уровень в плане алгоритмизации языковых анализаторов.

Некоторые оценочные тесты для систем машинного перевода

В большинстве языков мира предложения разделяются на 2 группы:

1. Простые предложения, которые несут в себе один семантический смысл, несмотря на количество их составляющих. Например, «*вчера Иван поужинал дома*». Тут в предложении говорится об одном факте – это действие, которое совершил Иван – «поужинал», несмотря на те

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

компоненты, которые указывают на время «вчера» и пространство «дома».

2. Сложные предложения, которые содержат больше двух семантических высказываний. Например, «Мой друг, который прилетел вчера вечером, поехал встретить своего старого друга».

Оценочные тесты имеют разные типы. Мы назовем 3 из них:

1. **Тесты на уровне уяснения значения слова:** некоторые слова имеют много значений. Выбор подходящего по смыслу значения определяет контекст, в котором оно употребляется. На правильный выбор значения слова влияют следующие факторы:

- Соседние слова в предложении. Нет сомнения в том, что соседние слова представляют собой важный фактор для определения значения слова. Перечисленные в словарях лексические единицы представляют собой лишь начальное звено бесконечной цепочки многочисленных значений и имеют лишь потенциальные значения (неактуальные). Эти значения становятся актуальными в определенном контексте [كريم زكي 2000: 101]. И только синтаксическим или семантическим контекстом определяется понимание значения того или иного многозначного слова [Панич 2007]. Как считает ученый лингвист Р. Д. Альталхи значения многозначных слов определяются лингвистическим и\или ситуативным контекстом [ردة الله 2003]. Современные системы машинного перевода, к сожалению, пока неспособны подбирать более подходящее к контексту значение. Решение проблемы лексической многозначности известный ученый-лингвист Ю.Н. Марчук видит в точном учете специфических особенностей предметного поля и лингвистического состава конкретных подъязыков (т.е. областей, заведомо существенно меньших, чем вся система естественного языка). Для таких языковых общностей как подъязыки возможно определять значение отдельных языковых единиц таким образом, что их совокупность (линейная комбинация) не противоречит идее цельного текста как такового. На этой теоретической основе была впервые выдвинута Ю.Н. Марчуком в 1976 г. идея создания контекстологических словарей для определенных типов текстов в рамках определенных семантических полей и подъязыков, которые сегодня помимо контекстов употребления включают и толкование [Марчу

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

2007]. Идея контекстологического словаря и сам словарь описаны в работе Ю.Н. Марчука «Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский» [Марчук 1976].

- Синтаксическая структура, которая не позволяет перенести одну часть предложения в другое место.

2. **Тесты на уровне безличных предложений**: например, «Лодку сносит к берегу». Это предложение было протестировано на разных системах онлайнового перевода, которые поддерживают русский и арабский языки. К сожалению, таких систем в интернете мало. Результаты были следующими (данный эксперимент был проведен 21.10.2010):

ImTranslator переводит данное предложение на арабский язык как: *ضربات القارب إلى الضفة* (бук. перевод: *удары лодки к берегу*). Очевидно, что перевод далек от смысла.

Google-переводчик выдает нам тот же самый результат: *ضربات القراب إلى الضفة* □ (бук. перевод: *удары лодки к берегу*).

Bing translator выдает нам более удачный результат, хотя эта программа еще находится на этапе испытаний. Результат был таков: *يهدم القارب إلى الضفة* □ (бук. перевод: он сносит лодку к берегу). Тут два замечания: 1) указание на производителя действия, которого на исходном тексте не существует; 2) неудачный выбор значения, более подходящего к данному контексту, чем все остальные значения у глагола «сносить». В исходном тексте глагол «сносить» употребляется в значении *сдвижения с места какого-л. объекта с силой*, что соответствует в арабском языке слову جرف – *[джарафа]*, однако система Bing translator в нашем случае использовала для перевода 4-ое значение: *разрушать*, *ломать* (هدم – *[хадама]*), что привело к искажению смысла в данном предложении. Приведем еще два примера: *Молнией ударило дерево* и *Вечереет*.

ImTranslator переводит данные предложения на арабский язык соответственно как: *ضربت صاعقة شجرة* (бук. перевод: *молния ударила в дерево*) и *المساء* (бук. перевод: *вечер*). Первое предложение было удачно переведено на арабский язык, а второе надо было перевести как *حل المساء* (бук. перевод: *вечер наступил*), а не как существительное.

Google-переводчик выводит нам те же самые результаты, которые система ImTranslator вывела на выход.

Bing translator переводит первое предложение на арабский язык как: *البرق اصطدمت بشجرة* (бук. перевод: *молния врезался в дерево*). Система здесь, как очевидно, не учитывает мужского рода слова «برق», а рассматривает его как слово женского рода, что привело к неправильному

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

склонению глагола «اصطدم». Что касается второго предложения, то система, оказывается, не может найти соответствующий ему эквивалент в арабском языке.

3. Тесты на уровне соответствия слов в предложении по роду и числу:

- **Единственное число жен. и муж. рода:** приведем несколько примеров для тестирования: «*Мальчик идет в школу*», «*Девушка идет в школу*», «Мой брат читает газету». При переводе первых двух предложений на арабский язык все вышеуказанные системы перевода учитывают род и число существительных и спрягают глаголы согласно правилам арабского языка. Что касается последнего предложения, то обе системы (ImTranslator и Google-переводчик) выводят на экран следующий перевод: *أخي قراءة الصحف* (бук. перевод: *Мой брат чтение газет*), т.е. вместо глагола «читать» представляется отглагольное существительное «чтение», а слово «газета» переводится во множественном числе. Система Bing translator предлагает лучший вариант, хотя для арабского слуха он считается окрашенным высоким стилем, что не соответствует тому стилю, на котором было построено предложение в исходном тексте: *أخي يقوم بقراءة الصحف* (бук. перевод: *Мой брат совершает чтение газет*). Слово «газета» осталось без изменения. Все предлагаемые здесь системы переводят его во множественном числе.
- **Множественное число жен. и муж. рода:** например, «Женщины устали» и «мужчины устали». При переводе данных двух предложений система Bing translator рассматривает глагол «устать» как существительное «усталость», в результате чего не только выдается неправильный эквивалент, но и сам смысл искажается: *النساء هن التعب* (бук. перевод: *Женщины есть усталость*), *الرجال هم التعب* (бук. перевод: *Мужчины есть усталость*). Системы ImTranslator и Google-переводчик предлагают другой вариант: *المرأة تعبوا* (бук. перевод: *Женщина устали*), *الرجال متعبون* (бук. перевод: *Мужчины устали*). Первое предложение было переведено без учета множественного числа слово «женщина». Кроме того, обе эти системы склоняют глагол «устать» в отношении к мужскому роду, а не женскому. Что касается второго предложения, то данные системы его удачно перевели на арабский язык.

***Язык, сознание, коммуникация**: Сб. статей / Отв. ред. В. В. Красных, А. И. Изотов. – М.: МАКС Пресс, 2011. – Вып. 42. – 144 с. ISBN 978-5-317-03566-2*

З а к л ю ч е н и е

Машинный перевод является важной технологией в самых различных планах: социально-политическом, коммерческом и научном. Однако системы машинного перевода на арабском рынке до сих пор не получили подробной и детальной оценки, которая могла бы быть полезной для пользователей или самих разработчиков. Поэтому надо было оценить данные имеющихся систем. Правда, рассмотрение данного вопроса в нашем докладе было как рассмотрение черного ящика в некоторой степени, т.е. нам неизвестно, какие алгоритмы и анализаторы используются в этих системах. Находятся ли они на этапе испытаний? Несмотря на некоторые представления об успехах или неудачах систем машинного перевода, мы высоко оцениваем прилагаемые усилия в этой области, без которых было бы невозможно достичь высокого уровня эффективности работы данных систем. Настоящее исследование является только первым и начальным этапом исследования основных характеристик действующих систем машинного перевода в данной языковой паре. По мере накопления и более детального анализа ошибок действующих систем можно будет сформулировать более подробные рекомендации относительно учета и алгоритмического построения соответствующих алгоритмов.

Л и т е р а т у р а

1. *Balkan L.* Translation Tools // META. – 1992. 37(3). – P. 408-420.
2. European Association for Machine Translation EAMT http://www.eamt.org/mt.html.
3. Hutchins W. J. The IAMT Certification Initiative and Defining Translation System Categories // Proceedings of 5th EAMT Workshop, Slovenia, 2000. http://ourworld.compuserve.com/homepages/WJHutchins/IAMTcert.html.April 2002.
4. *Sager J.C.* Language Engineering and Translation: Consequenses of Automation. – Amsterdam: John Benjamins, 1994.
5. Somers H. L. Intoduction // Computers and Translation: A Translator's Guide / *H.L. Somers* (Ed.) – Amsterdam: John Benjamins, 2003. – P. 1-11.
6. *Марчук Ю.Н.* Контекстологический словарь для машинного перевода многозначных слов с английского языка на русский. – М.: ВЦП, 1976. Часть 1.
7. *Марчук Ю.Н.* Компьютерная лингвистика. – М.: Восток-Запад, 2007.
8. *Панич Ю.В.* Универсальное смысловое кодирование исходного текста и его перевод с использованием системы согласованных словарей (I,II) http://www.elektron2000.com 2007.
9. ردة الله بن ردة بن ضيف الله الطلحي. دلالة السياق. جامعة ام القرى. مكة المكرمة, 2003
10. كريم زكي حسام الدين. التحليل الدلالي: إجراءاته ومناهجه. الجزء الأول. دار غريب. مصر, 2000 .□