XXXI Неделя науки СПбГПУ. Материалы межвузовской научной конференции. Ч. VI: С.59-60, 2003.

УДК 662.642.621.926.7

А.А. Ивашкин (асп. каф. ИИТ)

ВИЗУАЛИЗАЦИЯ ПРОЦЕССА ПЕРЕВОДА ТЕКСТА В СИСТЕМАХ МАШИННОГО ПЕРЕВОДА

Системы автоматического (машинного) перевода (далее – МП) – это одно из направлений компьютерной лингвистики, задачей которого является перевод текстов с одного естественного языка на другой. Развитие этого направления идет уже достаточно давно (первые системы МП появились в конце 50-х – начале 60-х годов XX века, а идея о возможности автоматизации перевода была высказана еще Лейбницем), и определенные результаты к настоящему моменту уже достигнуты. Тем не менее, до достижения момента, когда системы МП смогут заменить переводчиков-людей, еще очень далеко.

Основными проблемами при создании систем МП являются многообразие форм естественного языка (одна и та же мысль может быть передана различными способами) и их неоднозначность (одна и та же конструкция может означать разные вещи).

В качестве примера можно взять две одинаковые по структуре и имеющие сходные по семантике элементы фразы – «научные книги и журналы» и «научные книги и комиксы». При анализе таких фраз должен быть решен вопрос, является ли прилагательное «научные» определением только к существительному «книги» или к обоим однородным существительным.

Человек при анализе такого неоднозначного текста, как правило, опирается на собственный опыт, используя ту или иную его часть по мере возникновения такого рода проблем. Поэтому переводчик-человек, практически не задумываясь, скажет, что в первой фразе прилагательное относится к обоим существительным, а во второй – нет, поскольку сочетание «научные комиксы» является бессмысленным. В то же время в системе МП каждая конкретная ситуация должна быть заранее предусмотрена разработчиками системы, что по сути является эквивалентным созданию системы искусственного интеллекта. Поэтому в настоящий момент две приведенные фразы будут, скорее всего, разобраны одинаково. Как именно – зависит от конкретной системы.

Для решения задачи получения качественного перевода на настоящем этапе развития систем МП существует только один действенный метод. Это – использование человека в процессе перевода.

Наиболее часто применяемый подход – это пред- и постредактирование переводимого текста. Смысл предредактирования – в устранении из переводимого оригинального текста имеющихся там неоднозначностей для упрощения структуры текста и соответственно более точного разбора системой МП. Однако примеры, подобные приведенному выше, мало кто из людей сочтет неоднозначными. Постредактирование – исправление уже переведенного текста. Однако для того, чтобы заметить ошибку в нашем примере, человек, занимающийся корректурой, должен знать и исходный текст (поскольку с точки зрения грамматики перевод будет верным), т.е. являться переводчиком. Таким образом, при таком подходе система МП используется в качестве одного из компонентов АРМ переводчика.

Для устранения появления такой же ошибки в будущем можно провести подстройку системы. Для пользователя системы МП практически единственным способом подстройки будет добавление в словарь словосочетания с правильным переводом.

Другой способ – исправление алгоритма анализа исходного текста – требует активного вмешательства команды разработчиков и является, по сути, элементом разработки, а не эксплуатации системы. Таким образом, область возможного воздействия пользователя на систему весьма ограничена.

Следствием этих двух проблем (требование знания исходного и целевого языков для обнаружения ошибки и сложность воздействия на систему для ее устранения) является ограничение круга лиц, которые могут полноценно пользоваться системой МП. В среде остальных пользователей возникает определенное предубеждение против систем МП. Достаточно вспомнить многочисленные анекдотические истории про переводы, выполненные компьютером, корни которых лежат в невозможности обнаружения и устранения ошибки пользователем системы.

В качестве метода, который может помочь решению проблемы обнаружения ошибок, совершенных системой МП в ходе анализа текста, предлагается использовать визуализацию процесса перевода. Имеется ввиду поэтапное, шаг за шагом, отображение принимаемых в ходе анализа текста решений. Способов такого отображения может быть множество, и для каждой системы МП наиболее подходящим может оказаться свой собственный. Наиболее универсальным среди этого множества является представление разбираемого текста в виде дерева, корнем которого является анализируемый независимо от остального текста фрагмент (в идеале – весь текст целиком, реально в существующих системах МП в роли такого фрагмента выступает предложение), узлами – синтаксические элементы (простые предложения в составе сложного, глагольные группы в составе простого предложения и т.д.), а листьями – минимальные элементы (отдельные слова, или, как вариант, отдельные морфемы).

При таком подходе для обнаружения ошибки анализа пользователю будет достаточно владеть только языком исходного текста. Даже если пользователь владеет обоими языками, то, имея перед глазами структуру осуществленного разбора, ошибку обнаружить значительно проще.

При этом если отображение происходит поэтапно (в начале – плоская структура, в которой все элементы равноправны, в конце – готовый разбор), то можно определить именно то решение, которое послужило источником ошибки. Если в сочетании с этим ввести механизм обратной связи, то тем самым будет решена вторая проблема систем МП – у пользователя появится возможность адресного воздействия на конкретные алгоритмы системы МП.

Возможный источник проблем визуализации процесса перевода кроется в огромном числе решений, принимаемых в ходе анализа даже небольшого фрагмента. Число их может исчисляться десятками тысяч. Однако эта проблема имеет решение – в качестве примера можно привести системы отладки программ, когда местоположение ошибки шаг за шагом локализуется.

Вывод. Включение в состав системы МП механизма визуализации процесса перевода в сочетании с механизмом обратной связи может существенно повысить эффективность использования системы.