СИНТЕЗ ТУРЕЦКОГО ПРЕДОЖЕНИЯ В МНОГОЯЗЫЧНОЙ СИСТЕМЕ МАШИННОГО ПЕРЕВОДА

Слезкина О.Ю.

РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ

Работа посвящена описанию многоязычной (английский, русский, испанский и турецкий) системе машинного перевода Кросслейтор. Основное внимание уделяется описанию двух этапов синтеза турецкого предложения. На этих этапах деревья, прошедшие этап семантического анализа, трансформируются сначала в соответствии с конструкциями турецкого языка, а затем преобразуются в линейную структуру (дающую после этапа морфологического синтеза грамматически правильное предложение турецкого языка).

В данной работе мне хотелось бы описать принципы работы двух этапов[1] синтаксического синтеза турецкого предложения в рамках действующей многоязычной двусторонней системы машинного перевода *Кросслейтор*.

Краткая характеристика переводчика

Описываемая в данной работе система машинного перевода (МП) *Кросслейтор* позволяет осуществлять перевод с каждого на каждый из следующих языков: русский, английский, испанский и турецкий. Спецификой данной системы МП является наличие в ней семантического уровня и языка-посредника, облегчающего перевод на несколько языков.

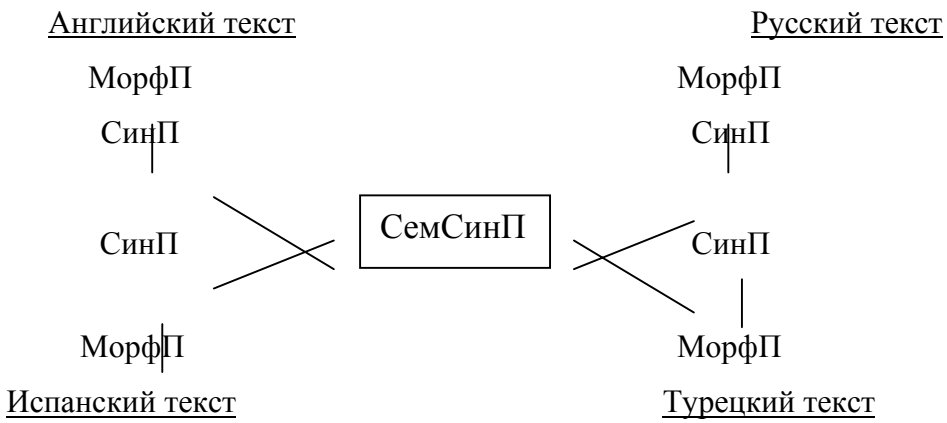
Процесс перевода разбивается на 10 основных этапов: морфологический, синтаксический и семантический анализ, фильтрацию, нахождение эквивалентов с исходного языка на язык-посредник и с языка-посредника на язык перевода, а также этапы посттрансляции, синтаксического и морфологического синтеза.

Модульность используемой методики позволяет осуществлять перевод с неограниченного количества языков. Добавление нового языка никак не затрагивает уже существующую часть переводчика, то есть для того, чтобы перевести с этого языка на любой из уже имеющихся, необходимо написать анализ данного языка, а также создать словарь «входной язык – язык перевода». Для того чтобы перевести на этот новый язык, нужно лишь написать его синтез.

Схематически процесс перевода можно изобразить следующим образом[2]:

[1] Этапы посттрансляции и собственно синтаксического синтеза.

[2] На схеме не показано, но между соседними структурами (соединенными стрелками) при анализе, а иногда и при синтезе строится цепь промежуточных структур.



Этапы синтеза

При синтезе предложения на язык перевода используются следующие этапы: этап выбора эквивалентов (с языка посредника на язык перевода); этап посттрансляции; этап синтаксического синтеза предложения; образование сложных слов[3]; этап морфологического синтеза.

Предыдущие этапы приводят к построению для фразы одного или нескольких деревьев зависимостей, в узлах которых стоят частично характеризованные турецкие лексемы: уже известны значения, во-первых, номинативных (например, число существительного), а во-вторых, словарных или грамматических признаков, полученные соответственно при переводе категорий и в процессе лексической подстановки.

Таким образом, для получения морфологического представления необходимо выполнить две задачи:

- приписать лексемам недостающие (а именно, чисто синтаксические, согласовательные) признаки *и*
- построить линейную структуру фразы выходного языка.

Первая из этих задач решается преимущественно на этапе посттрансляции, а вторая – на этапе синтаксического синтеза. То есть, на этапе посттрансляции каждое дерево, пришедшее с этапа выбора эквивалентов, преобразуется (трансформируется) в связи с необходимостью учета специфики конструкций выходного языка.

В дальнейшем трансформированное дерево служит исходным материалом для этапа синтаксического синтеза выходного языка и будет преобразовано на этом этапе в линейную структуру. Каждый из элементов этой линейной цепочки будет уже иметь при себе все параметры, необходимые для этапа морфологического синтеза.

[3] На этом этапе происходит в основном образование турецких форм возможности, невозможности, быстроты действия и некоторых других. Эти формы образуются при помощи объединения в рамках одной словоформы деепричастия смыслового глагола и основной формы (имеющей показатели времени, лица, числа) вспомогательного. Например: «он смог сделать» = yapabildi «сделав-знал»: yapmak – делать, bilmek - знать.

Предложения с неверной структурой или другими ошибками могут быть просто не разобраны на первом этапе, в результате чего система не выдаст ни одного варианта перевода. Однако и при верно построенных предложениях пока существует вероятность того, что из многих вариантов трактовки предложения система выберет неправильный вариант.

Задачи, решаемые на этапах синтеза

Этапы синтеза представляют собой переход от дерева зависимостей (в его нормализованном виде) к линейной структуре - цепочке информаций, содержащей данные о форме слов и об их порядке в турецком предложении. Таким образом, синтез включает в себя решение двух уже упоминавшихся ранее задач:

1. задачи *установления сведений о форме турецких слов*[4] (эти сведения в основном устанавливаются на этапе посттрансляции);
2. задачи *установления порядка слов* в турецкой фразе (решение этой задачи целиком относится к этапу синтаксического синтеза).

На этапе посттрансляции нормализованная структура получает специфические черты, свойственные данному языку. Нормализованные морфологические характеристики превращаются в соответствующие им падежи, времена, залоги и т.д. турецкого языка. На этапе же синтаксического синтеза осуществляется линеаризация деревьев зависимостей. Сведения о форме турецкого слова в большинстве своем либо поступают с этапов анализа, либо определяются на этапе посттрансляции (например, время, число, показатели принадлежности, падеж), однако отдельные параметры приписываются лишь на этапе синтаксического синтеза (где происходит, например, согласование сказуемого с подлежащим)

Содержание этапа посттрансляции

На этапе посттрансляции деревья зависимостей изменяются в соответствии с конструкциями турецкого языка. *На вход* этого этапа с семантического уровня поступают «нормализованные» деревья зависимостей, прошедшие этап выбора эквивалентов (с ЯП на турецкий язык), в вершинах которых стоят турецкие лексемы, имеющие при себе некоторый (в большинстве случаев неполный) набор параметров. Однако дерево зависимостей имеет еще «нормализованный» вид – общий для всех четырех языков. Но в каждом языке для выражения одной и той же конструкции в большинстве случаев используются разные средства. Возьмем для примера следующую именную группу *«человек, сын которого учится*

[4] Сюда относится информация о падеже, которым оформляется в языке данный актант, оформлении нужными падежными аффиксами изафетной группы и др.

в Москве» и покажем, какими разными конструкциями она выражается в английском и турецком языках.

Дерево зависимостей для этой именной группы будет выглядеть следующим образом:

ЧЕЛОВЕК (субъект)
|__ ЧЕЙ (определение)
|__ СЫН (субъект)
|__ УЧИТЬСЯ (предикат)
|__ МОСКВА (локатив)

Это дерево зависимостей должно быть просинтезировано в следующие структуры:

АНГЛИЙСКИЙ ЯЗЫК:

the	**man,**	**whose**	**son**	**studies**	**in**	**Moscow**
опр. арт.	человек	чей	сын	учится	в	Москве

ТУРЕЦКИЙ ЯЗЫК:

oğl-u	**Moskova'da**	**okuyan**	**adam**
сын-его	Москва - в	учащийся	человек

После сравнения этих конструкций в двух языках становится понятно, что из дерева зависимостей, соответствующего этим именным группам, за один этап нельзя построить линейную структуру предложений в обоих языках. Если для дальнейшего синтеза английского предложения необходимы минимальные преобразования (выбор нужного времени глагола, предлога, выражающего локатив и т.п.), то для синтеза турецкого предложения необходимо кардинальное изменение дерева[5]:

ЧЕЛОВЕК (субъект)
|__ УЧИТЬСЯ (определение, причастие наст. вр.)
|__ СЫН (субъект, показатель принадлежности 3л.ед.ч.)
└─── МОСКВА (объект, мест. падеж)

Для этого используются следующие правила трансформаций:

1. первое ставит глагол «учиться» в форму причастия, приписав соответствующие параметры: настоящего времени, действительного залоги и т.д.
2. Существительное «сын» в результате применения другого правила получает показатель принадлежности 3 л. ед.ч.(oğl-u - сын-его).

[5] В дереве зависимостей, приведенном ниже, указаны лишь те параметры вершин, которые важны для понимания описываемых трансформаций.

3. Еще одно правило приписывает существительному «Москва» местный падеж, обозначающий в турецком языке местонахождение предмета (Moskova'da – в Москве).

Общая характеристика правил трансформаций

К дереву последовательно применяется целый ряд правил трансформации. На данный момент имеется 430 правил трансформации, каждое из которых имеет свою очередь выполнения. Сначала применяются правила трансформации 1-ой очереди выполнения, затем (уже после того, как выяснилось, что ни одно из правил 1-ой очереди выполнения не применимо) правила трансформации 2-ой, 3-ей и т.д. очередей выполнения. Всего очередей выполнения около 20. Если в правилах одной очереди выполнения оказались два (или более) правил, применимых к одной и той же ветке дерева, то они применяются оба, из-за чего количество деревьев удваивается. Поэтому чтобы не увеличивать количество деревьев, такая дублетность используется очень редко: лишь в конструкциях, не различимых в других языках, и при этом влияющая на смысл всего предложения. В более простых случаях обычно используются допущения. Например, при переводе с русского языка в большинстве случаев невозможно достоверно узнать, является существительное определенным или нет. В турецком же языке определенные и неопределенные существительные оформляются (в винительном падеже) разными падежными аффиксами. Поэтому при синтезе используется допущение, что все эти объекты определенные.

Цель этапа синтаксического синтеза

Основной задачей данного этапа синтеза турецкого предложения является линеаризация деревьев зависимости и восполнение всех параметров, необходимых для морфологического синтеза. То есть, на этом этапе дерево зависимостей, например:

KARDEŞ /БРАТ/ (subject; N; number=sg; def=def;person=2;num_pers=sg; case=osn[6])

└──── GITMEK /ИДТИ/(predic; V; tense1=def;tense2=simp; norm_con=norm)

└──── OKUL /ШКОЛА/ (object3; N; case=dat; number=sg; person=n; num_pers=n)

преобразуется в линейную структуру, каждый элемент которой имеет все параметры, необходимые для морфологического синтеза турецкого предложения:

GITMEK (number=sg;person=3; ...[7]) OKUL (...)KARDEŞ (…)

[6] Параметры означают, что существительное БРАТ является субъектом предложения и имеет следующие параметры: ед.ч., определенное, в основном падеже,

[7] Многоточием обозначаются параметры, уже имевшиеся у слова при входе на этап синтаксического синтеза (указаны в приведенном выше дереве зависимостей).

Сведения о форме турецкого слова, вырабатываемые на этапе синтаксического синтеза, присоединяются к тем, которые поступили от предыдущих этапов и остаются в синтезе без изменения (например, число, показатели принадлежности, определенность, падеж, уже установленные на этапе посттрансляции). Эти параметры, уже имевшиеся у слова при входе на этап синтаксического синтеза (указаны в приведенном выше дереве зависимостей), обозначаются многоточием. На этом этапе определяются лишь значения параметров лица и числа глагола, которые передаются от подлежащего к сказуемому.

После этапа морфологического синтеза предложение имеет следующий вид:

Kardeş-in okul-a gitti.

Брат-твой школа-в пошел. – «Твой брат пошел в школу».

Всего на этапе синтаксического синтеза используется порядка 180 контекстно-свободных правил, имеющих вид структуры непосредственных составляющих. Каждое правило имеет вид X $\rightarrow$ У, где У может состоять из нескольких компонентов, каждый из которых, в свою очередь, может замещаться каким-либо У’ из другого правила подстановки.

В данной работе была дана краткая характеристика системы машинного перевода Кросслейтор и несколько подробнее описаны принципы работы этапов синтеза турецкого предложения.

ЛИТЕРАТУРА

Апресян 1984 – Апресян Ю.Д. Лингвистическое обеспечение автоматической системы французско-русского автоматического перевода ЭТАП-1, М.: 1984.

Бакулов и др. 1990а – Бакулов А.Д., Леонтьева Н.Н. Теоретические основы машинного перевода//Искусственный интеллект: в 3 кн. Кн. 1 Системы общения и экспертные системы: Справочник/ под ред. Э.В.Попова, М.: Радио и связь, 1990.

Клышинский и др. 2000 – Клышинский Э.С., Андреев А.С., Ёлкин С.В. Метод машинного перевода текстов// Сб. трудов 3-го научно-практического семинара "Новые информационные технологии". М.: МГИЭМ, 2000.

Клышинский и др. 2002 – Клышинский Э.С., Слезкина О.Ю. Применение модифицированных бэкусовских нормальных форм для задач анализа и синтеза естественных языков//Новые информационные технологии: материалы пятого научно-практического семинара, М., Моск. гос. институт электроники и математики, 2002.

Synthesis of the Turkish sentence in the multilingual system of machine translation

Oksana Slezkina

The work is devoted to the description of the multilingual (English, Russian, Spanish and Turkish) system of machine translation CROSSLATOR. The principle attention is paid to the two stages of synthesis of the Turkish sentence developed by the author. On these stages the semantic-synthesis tree first transformed into a synthesis tree and then into a linear structure ready to be converted into grammatically right Turkish sentence.