УДК 004.091.043

МАРКОВСКИЙ А.П.,
ШЕВЧЕНКО О. Н.,
ФАНЬ ЧУНЬЛЭЙ

ИНТЕРАКТИВНО-ШАБЛОННЫЙ МЕТОД КОМПЬЮТЕРНОГО ПЕРЕВОДА НАУЧНО-ТЕХНИЧЕСКИХ ПУБЛИКАЦИЙ

В статье предложен подход к организации компьютерного перевода научно-технических публикаций. В основу подхода положено интерактивное использование лингвистических шаблонов, позволяющее повысить качество перевода за счет рационального разделения функций специалиста в предметной области, переводчика-лингвиста и компьютерных программ. Разработаны процедуры шаблонирования текста публикации и компьютерного перевода с использованием шаблонов. Приведен пример перевода и результаты экспериментальных исследований. Выполнена оценка эффективности предложенного подхода по сравнению с существующими системами компьютерного перевода.

The paper proposes a new approach to computer translation of technical scientific publications. The described method is based on the interactive use of translation templates which allow to improve the translation quality due to the rational distribution of functions between a domain expert, a human translator (a linguist) and computer software.The procedures for creating text templates and for template-based computer translation of publications were developed . An example of the translation and the results of experimental studies were given. The evaluation of the effectiveness of the proposed approach compared to the existing systems of computer translation was made.

Введение

Начало третьего тысячелетия знаменует динамичный процесс углубления и расширения интеграционных процессов во всех сферах человеческой деятельности. В значительной степени процесс информационной интеграции обусловлен быстрым развитием средств телекоммуникации и компьютерных сетей и, в первую очередь, Интернета, которые технически решают задачу доступа к информации.

Однако, существенным препятствием на пути обмена информацией, главным образом, связанной с научными и техническими достижениями, остается языковой барьер. Развитие технологий электронных публикаций педалировало в начале 21-го века заметный рост числа публикаций научного характера в странах азиатского региона, в первую очередь таких как Китай и Индия [1]. Это обстоятельство обостряет проблему межъязыкового взаимопонимания в процессе обмена научно-технической информацией.

Значительным потенциалом в решении этой проблемы является использование возможностей современных компьютерных технологий.

Таким образом, проблема создания эффективных средств компьютерного перевода научно-технических публикаций является важной и актуальной для расширения и углубления информационной интеграции во многих областях человеческой деятельности.

Анализ современного состояния проблемы компьютерного перевода

Проблема компьютеризованного перевода является одной из наиболее традиционных в компьютерных технологиях. Первые системы машинного перевода появились еще в 60-х годах прошлого столетия [2]. С тех пор проведено большое число исследований, опубликовано тысячи статей, создано ряд действующих систем, на порядок выросли возможности компьютерной техники.

Все системы компьютеризованного перевода, с позиций участия в их функционировании человека, принято [1] разделять на два класса: машинного (или автоматического) перевода и автоматизированного перевода (CAT – computer-aided translation). Системы первого класса осуществляют перевод текстов с одного естественного языка на другой без участия человека. Примерами подобных систем являются Systran [3] и программа-переводчик Google [4]. Первая из них строится на на основе грамматических правил (Rule-Based Machine Translation, RBMT), обеспечивающей перевод на базе встроенных словарей и грамматических правил двух языков. Основным недостатком этой системы является недостаточное для практиче-

ского использования семантическое качество перевода. Основной причиной этого является отсутствие пригодных для практического использования формальных моделей естественных языков [5]. В некоторых системах полностью машинного перевода, в частности, в упоминавшейся выше программе перевода Google, отсутствие такой модели заменяется статистическим анализом (Statistical Machine Translation, SMT) огромного количества текстов и их переводов, выполненных человеком. Системы этого типа анализируют статистику межъязыковых соответствий и используют эту информацию при выборе вариантов перевода. Их очевидным недостатком является потребность в значительном объеме ресурсов (памяти и процессорного времени). Опыт практического использования показал [4], что знамена формальной модели языка статистикой не позволяет заметно улучшить семантическое качество перевода. Такой же результат имеют и попытки заменить отсутствие формальной модели естественного языка использованием самообучающихся систем и систем на основе нейронных сетей [6]. Фактически такие системы не вышли к настоящему времени за границы экспериментальных разработок.

Таким образом, полностью компьютеризированные системы перевода, работающие без участия человека, не способны к настоящему времени обеспечить в полной мере семантическое соответствие между входным и выходным текстами.

Поэтому на практике такие системы требуют редакторской правки опытного переводчика и специалистов предметной области.

Более широко используются системы автоматизированного перевода, в которых основная роль принадлежит переводчику, труд которого автоматизируется компьютерными средствами. В простейшем случае такие средства представляют собой компьютерные словари. Более сложные системы предоставляют переводчику ряд вариантов перевода отдельных предложения и средства интерактивного редактирования. Существенным недостатком таких систем является недостаточная скорость перевода. Фактически системы автоматизированного перевода применительно к научным и техническим публикациям ориентированы на специалистов в предметной области и предполагают знание ими базовых основ входного языка. Практика использования подобных систем для перевода публикаций научного и технического характера показала значительную зависимость семантического качества перевода от квалификации специалиста и знания им входного языка. Это существенно ограничивает сферу их эффективного использования. В частности, автоматизированные системы практически не пригодны для перевода с китайского, японского и корейского языков для широкого круга европейских специалистов.

Таким образом, несмотря на достаточной широкий фронт проводимых исследований, решенной проблему эффективного перевода информации научного и технического характера считать нельзя [1].

Целью настоящей работы является повышение эффективности компьютерных систем многоязыкового перевода научно-технической информации.

Шаблонный метод представления и перевода языковых конструкций

Как отмечалось выше, фундаментальной причиной относительной низкой эффективности существующих систем компьютерного перевода научно-технических публикаций является отсутствие в достаточной мере адекватной модели естественных языков, которая позволяла бы семантически точно транслировать предложения входного языка в выходные. Одной из возможностей обойти это препятствие является введение ограничений на количество конструкции предложений входного языка.

Основная идея предлагаемого подхода состоит в том, что публикации научного и технического характера, без заметного ущерба для понимания, могут быть оформлены с использованием ограниченного набора конструкций предложений. Анализ текстов таких публикаций [6,7] показал, что число наиболее употребляемых в них синтаксических конструкций предложений не превышает 200-600. Такое ограниченное число синтаксических конструкций предложений входного языка вполне может переведено специалистами-лингвистами в адекватные конструкции одного или нескольких выходных языков. Таким образом, ограничение возможных вариантов синтаксических конструкций текстов научно-технических публикаций позволяет в определенной мере обойти проблему отсутствия моделей естественных языков.

Реализация такой идеи потребует от автора оформления текста научного или технического

характера изначально в виде предложений, синтаксические конструкции которых выбираются из определенного множества. В известном смысле можно говорить о том, на современном уровне межнациональной информационной интеграции написание и оформление текста научно-технических публикаций должно быть изначально ориентировано для перевода его на иные языки.

Фактически, при реализации предлагаемой идеи две наиболее сложные операции процесса перевода - распознавание входных синтаксических конструкций и синтаксически корректная трансформация конструкций входного языка в конструкции выходного языка выполняются человеком. Причем первая из операций выполняется автором текста непосредственно в процессе его оформления, а вторая - единоразово специалистом-лингвистом.

Для того, чтобы облегчить автору выбор синтаксической конструкции, позволяющей наиболее точно выразить семантический смысл предложения предлагается использовать специальные шаблоны предложений.

Шаблон представляет собой семантико-синтаксическую конструкцию, доминантой которой является семантическая составляющая с синтаксически изменяемыми компонентами. При переводе шаблона семантическая составляющая, являющаяся основой предложения, переводится специалистом-лингвистом, а синтаксические составляющие - заменой слов входного языка на слова выходного языка с использованием компьютерных словарей.

Автор осуществляет выбор одного из прототипов шаблонов. После этого, автором выполняется замена слов шаблона на слова, соответствующие семантическому смыслу формируемого предложения. При этом номер используемого шаблона и список заменяемых слов записываются в спецификацию предложения. Фактически автором выполняется только подходящая по смыслу замена слов, а спецификация оформляется автоматически.

Практически шаблон предъявляется автору в виде предложения естественного языка иной предметной области, в котором, сохраняя семантическую основу, необходимо заменить слова. Например, если семантический смысл предложения состоит в том, что для повышения эффективности исправления “пачки” ошибок в каналах со спектральной модуляцией автором предлагается использовать взвешенные контрольные суммы, то запрос на поиск наиболее подходящего шаблона по базовой семантической составляющей может иметь вид: “**Для повышения эффективности ... предлагается использовать ...** ”.

В ответ на данный запрос система предъявляет ряд пронумерованных шаблонов, один из которых, наиболее близкий по версии автора к желаемому имеет вид: “**Для повышения эффективности** идентификации (1) абонентов (2) в (3) интегрированных (4) базах (5) данных (6) **предлагается использовать** необратимые (7) булевы (8) преобразования (9)”. В приведенном шаблоне выделена семантическая основа и пронумерованы компоненты, которые могут быть заменены.

Автор выполняет замену слов в шаблоне в соответствии со смыслом формируемого предложения, которое после этого приобретает вид: “**Для повышения эффективности** исправления (1) пачки (2) ошибок (3) в (4) каналах (5) со (6) спектральной (7) модуляцией (8) **предлагается использовать** взвешенные (9) контрольные (10) суммы (11)”.

Если предположить, что в рамках рассмотренного выше примера выбранный автором шаблон имеет номер 172, то его спецификация может быть представлена в виде: {**172**, <1-1>, <2-2,3>, <3-4>, <4,5,6-5,6,7,8>, <7,8,9-9,10,11>}. Это означает, что используемая шаблонная конструкция имеет номер 172, и в процессе ее наполнения новым семантическим смыслом первое слово шаблона заменено первым словом предложения, второе слово шаблона заменено вторым и третьим словами предложения и так далее.

При написании каждого предложения публикации автором формируется по семантической основе запрос на выбор шаблона, выбирается наиболее подходящий из предъявляемых по запросу и корректируется путем заменой слов. В результате формируется предложение текста публикации и спецификация предложения, которые сохраняются в одном файле.

Наиболее сложным этапом при компьютерном переводе является синтаксический анализ входного предложения. Это этап предлагается выполнять в интерактивном режиме путем шаблонирования - замены произвольных предложений входного текста на идентичные в семантическом плане предложения-шаблоны. При этом одному предложению исходного текста могут соответствовать несколько предложений шаблонированного текста. Шаблонирование входного текста публикации может вы-

полняться как самим автором, так и квалифицированным специалистом в данной предметной области. Процесс шаблонирования предполагает создание двух синхронизируемых документов: самого текста, предложения которого принадлежат набору шаблонов и спецификации текста, которая каждое из предложений входного текста описывает номером используемого шаблона и списком слов, которые постанавливаются в шаблон.

Схематично процесс шаблонирования показан на рис.1. Вначале при написании предложения интерактивно выбирается базовая языковая конструкция, подходящая для выражения семантического смысла формируемого предложения. Технологически для выбора базовой конструкции автором вводится последовательность ее базовых слов.

Система выполняет по введенной последовательности поиск ряда отвечающих запросу шаблонов, которые предъявляются автору.

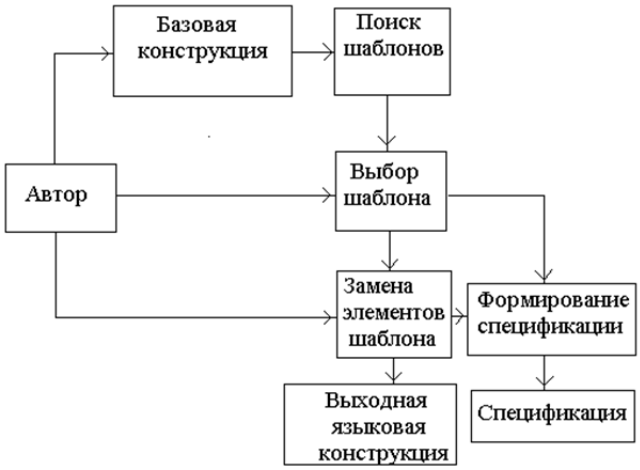

Рис. 1. Организация формирования текста на основе шаблонов

Шаблоны состоят из элементов базовой семантической конструкции и пронумерованных заменяемых элементов. Как указывалось выше, список прототипов шаблонов состоит из тысяч предложений на входном языке. Список шаблонов формируется путем выбора наиболее часто встречающихся в научных публикациях языковых конструкций, которые в своей совокупности достаточны для адекватной передачи положений научно-технических публикаций. На практике список шаблонов может пополняться авторами публикаций в случае, если ни один из шаблонов не способен адекватно отразить смысл предложения.

Выбрав наиболее подходящий по смыслу формируемого предложения шаблон на родном языке, автор выполняет замену в нем заменяемых слов, формируя языковую конструкцию предложения и соответствующему ему спецификацию. Фактически результатом описанного процесса является текст публикации и соответствующий ему список спецификаций, используемых только при переводе.

Специалистами-лингвистами выполняется перевод ограниченного списка шаблонов на несколько языков. Если предположить, что система ориентирована на перевод публикаций с использованием трех языков: русского, английского и итальянского, то в рамках рассмотренного выше примера, шаблон с номером 172 представляет собой совокупность семантически одинаковых предложений на трех языках:

Русском: “**Для повышения эффективности** идентификации (1) абонентов (2) в (3) интегрированных (4) базах (5) данных (6) **предлагается использовать** необратимые (7) булевы (8) преобразования (9)”. Базовыми элементами этой конструкции являются: “**Для повышения эффективности** ... **предлагается использовать**”.

Английском: “**To improve the efficiency of** subscriber (2) identification (1) in (3) integrated (4) databases (5-6) **it is proposed to use** irreversible (7) Boolean (8) manipulations (9)”. с базовыми элементами: “**To improve the efficiency of….. it is proposed to use…**”.

Итальянском: “**Per migliorare l’efficienza di** identificazione (1) dell’abbonato (2) nei (3) database (5-6) integrati (4), **si propone di utilizzare** manipolazioni (9) booleane (8) irreversibili (7)”. Базовые элементы семантики конструкции имеют вид: “**Per migliorare l’efficienza di….. si propone di utilizzare…”.**

Структурно процесс перевода предложения с использованием спецификации показан на рис.2.

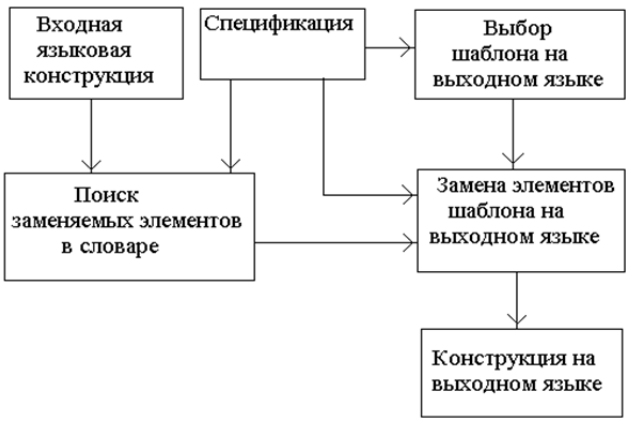

Рис. 1. Организация перевода с использованием языковых шаблонов

В отличие от формирования текста, его перевод на один из языков выполняется в автоматическом режиме.

При формировании текста научно-технической публикации на выходном языке каждая языковая конструкция (предложение) переводится отдельно. С использованием спецификации входной языковой конструкции выполняется выделение в ней заменяемых элементов. Для таких элементов в выполняется поиск в словаре соответствующих слов выходного языка. Но номеру из спецификации выбирается шаблон языковой конструкции на выходном языке.

В выбранном шаблоне на основе ссылок подстановок осуществляется замена переменных элементов на перевод слов входной языковой конструкции. В результате такой замены формируется языковая конструкция (предложение) на выходном языке, которая семантически идентичная входной.

В рамках рассматриваемого примера при переводе с русского на английский конструкции: "**Для повышения эффективности** исправления (1) пачки (2) ошибок (3) в (4) каналах (5) со (6) спектральной (7) модуляцией (8) **предлагается использовать** взвешенные (9) контрольные (10) суммы (11)" с использованием словаря осуществляется перевод заменяемых слов, отмеченных цифрами. Далее, по номеру 172, указанном в спецификации выбирается англоязычный шаблон "**To improve the efficiency of** subscriber (2) identification (1) in (3) integrated (4) databases (5-6) **it is proposed to use** irreversible (7) Boolean (8) manipulations (9)" в котором, в соответствии со спецификацией выполняется замена пронумерованных слов, в результате чего формируется предложение: "**To improve the efficiency of** bust (2) error (3) correction (1) in (4) channels (5) with (6) spectral (7) modulation (8) **it is proposed to use** weighted (9) checksums (10-11)".

Аналогично, при переводе на итальянский замена слов в шаблоне формирует выходное предложение в виде: "**Per migliorare l'efficienza di** correzione (1) di errori (3) del cluster (2) in (4) canali (5) con (6) modulazione (8) spettrale (7)**, si propone di utilizzare** checksum (10-11) ponderate (9)".

Как показал проведенный анализ, основная часть семантических ошибок при переводе публикаций научного и технического характера связаны с неправильным пониманием конструкции предложения на неродном для специалиста языке. Реализация предложенного подхода позволяет в значительной мере свести к минимуму такие ошибки перевода.

Анализ эффективности и результаты экспериментальных исследований

Вводимое в рамках подхода ограничение на конструкции входного текста позволяет достаточно просто, без использования значительных вычислительных ресурсов реализовать на приемлемом для практики уровне семантическое соответствие синтаксических конструкций различных языков. Фактически, в рамках предлагаемого подхода задача перевода разделяется на две составляющие: перевод языковой конструкции с одного естественного языка на другой и перевод отдельных слов заполнения этой конструкции. В рамках предложенного подхода первая, наиболее сложно формализуемая часть перевода, выполняется квалифицированным переводчиком, владеющим как входным, так и выходным языками. Вторая, существенно более простая часть: замена слов одного языка словами другого - осуществляется компьютерными средствами. Это позволяет оптимизировать разделение человеческой и компьютерной составляющих в процессе перевода: плохо формализуемая часть выполняется человеком, а более трудоемкая, но относительно простая часть реализуется компьютерными средствами.

При этом важным является то, что наиболее сложная часть работы, требующая труда высококвалифицированных переводчиков, выполняется одноразово в процессе инициализации системы. После инициализации, в процессе которой подбираются и переводятся шаблоны, система переводит научно-технические публикации автоматически.

Для практического исследования эффективности предложенного подхода была разработана экспериментальная система, ориентированная на перевод научно-технических текстов ограниченной тематики: компьютерные технологии. Это позволило ограничить количество шаблонов трехстами и использовать терминологически более качественный тематический словарь. В разработанной системе использованы три языка: русский, английский и итальянский. В перспективе создание экспериментального образца системы, ориентированной для перевода публикаций на китайский язык. Экспериментальные исследования показали, что система обеспечивает достаточно высокое

для адекватного понимания качество перевода для ограниченной предметной области.

Достоинствами предложенного подхода к компьютерному переводу текстов научно-технического характера по сравнению с известными системами являются:

1. Ориентация на использование относительно небольших ресурсов компьютерных систем; это позволяет реализовать систему перевода на простых, в том числе мобильных, вычислительных платформах с высокой производительностью.

2. Возможность перевода с языков, которыми пользователь совершенно не владеет, что особенно важно для преодоления языкового барьера между восточными и западными специалистами.

3. Изначальная ориентация на многоязыковый перевод.

Вместе с тем, необходимо отметить, что высокий уровень эффективности компьютерного перевода достигнут за счет ограничений на возможность использования языковых конструкций при написании публикации автором или редактировании готовой статьи специалистом. Экспериментальное исследование показало, что такие ограничения не сказываются существенным образом на качестве публикации, а скорость подготовки текста несколько возрастает. Фактически процесс написания публикации в значительной степени подобен современным технологиям интерактивного программирования, поэтому психологически не вызывает дискомфорта, сопровождается снижением уровня грамматических и стилистических ошибок.

Заключение

Таким образом, в работе предложен подход к повышению эффективности компьютерного перевода. Разработанный подход позволяет в значительной мере обойти узловую для компьютерных систем перевода проблему отсутствия как формальной модели естественного языка, так и моделей трансформации конструкций с одного языка на другой за счет ограничений на возможный набор языковых конструкций. Проведенные теоретически и экспериментальные исследования показали, что указанные ограничения вполне оправданы для определенной сферы переводимого материала, в частности, для публикаций научного и технического характера. Именно на этот, один из наиболее практически важных классов переводимых материалов ориентирован предлагаемый подход.

Литература

1. Хроменков П.Н. Современные системы машинного перевода /Хроменков П.Н. -М.: Мысль, 2005 –191 с.
2. Сдобников В.В.Теория перевода./Сдобников В.В., Петрова О.В.-М:Высшая школа: , 2008 – 254 с.
3. Марчук Ю.Н. Модели перевода /Марчук Ю.Н.- М.: Академия, 2010 - 188 с.
4. Briscoe T. Lexical Issues in Natural Language Processing //Natural Language and Speech, Springer-Verlag, 1991, - P. 39-40.
5. Кисленко Ю.И. Системна організація мови / Кисленко Ю.І.-К: Український літопис. 1997.- 217 с.
6. Белоногов Г.Г. Компьютерная лингвистика и перспективные информационные технологии: теория и практика построения систем автоматической обработки текстовой информации / Белоногов Г.Г. - М.: Русский мир, 2004 - 358 с.
7. Шевченко О.Н., Шевченко Д.С. Компьютеризированная система обучения иностранному языку/ Шевченко О.Н., Шевченко Д.С.//Системний аналіз та інформаційні технології: матер.наук-технічної конф. 26-30 квітня 2013: тез.допов.- К.:НТУУ “КПІ”, ІПСА.-2013- С.263.