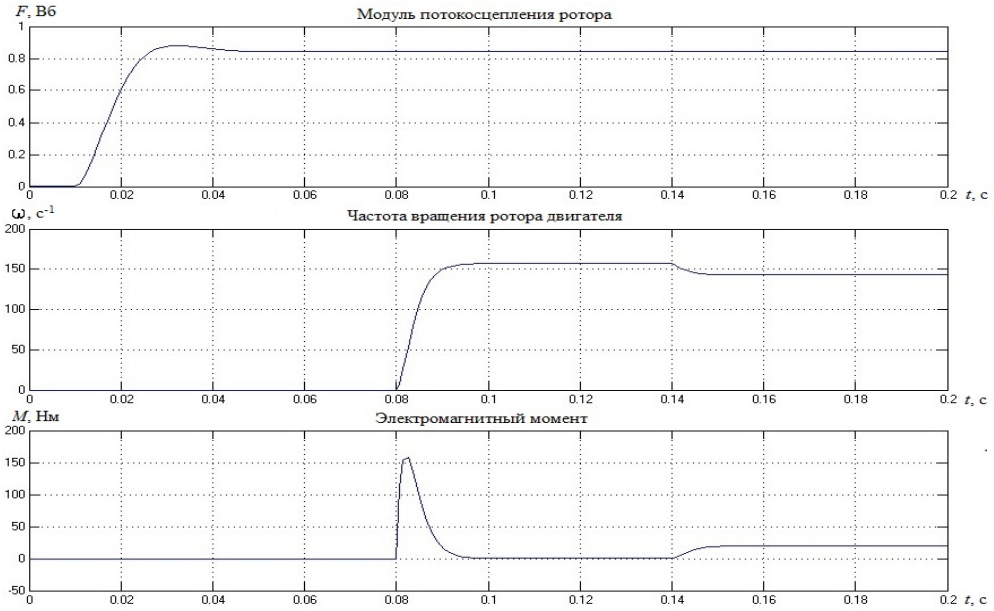

Рис. 2. Графики переходных процессов в неподвижной системе координат

A. N. Pakhomov, M. F. Korotkov, A. A. Fedorenko

ALTERNATING CURRENT ELECTRIC DRIVE MODAL CONTROL

The technique of synthesis of modal regulators of coordinates of vector system "the converter of frequency-asynchronous engine" is resulted by a method of the standard equations. The estimation of quality of processes of regulation of coordinates by the analysis of results of simulation modeling of the system in the environment of MatLab is given.

Keywords: modal regulator, alternating current electric drive, vector system.

УДК 004.912

К. В. Полянский

ПОСТРОЕНИЕ ФРЕЙМОВОЙ МОДЕЛИ ПЕРЕВОДА С ИСПОЛЬЗОВАНИЕМ КЛАСТЕРИЗАЦИИ ТЕРМОВ

Рассмотрена фреймовая модель представления знаний в IP-системах машинного перевода. Предложен алгоритм сегментации исходного и целевого текста через связь. Проанализированы различные методы кластеризации термов, определены наиболее эффективные из них для разбиения текста на кластеры.

Ключевые слова: машинный перевод, сегментация текста, кластеризация термов, фреймовая модель.

Важным этапом в IP-переводе (машинном переводе, использующем ресурсы информационно-поисковых систем) на стадии синтеза является сопоставление исходного текста (ИЯ-текста) и релевантных текстов на целевом языке (ЦЯ-текстов), выявление в них схожих сегментов. Процесс такого сопоставления выполняется в несколько шагов:

1) сегментация текста;

2) кластеризация сегментов;

3) построение фреймовой модели структуры текста.

Рассмотрим каждый шаг подробнее.

Сегментация текста. Для анализа структуры предложений ИЯ- и ЦЯ-текстов необходимо поделить эти предложения на логические сегменты, где каждый сегмент будет семантически самостоятельной единицей. Сегментом назовем непрерывный фрагмент тек-

ста, состоящего из термов одного языка, обозначающих связанную по некоторому критерию группу понятий. Составными частями сегмента могут быть термы следующих видов:

– объект (*obj*);
– субъект (*sub*);
– действие (*do*);
– свойство (*pro*);
– связь (*con*).

Идентификация составных частей сегмента осуществляется после проведения стемминга, когда установлена принадлежность термов к тем или иным частям речи. Определяется, что объект (*obj*) и субъект (*sub*) являются существительными, действие (*do*) – глаголом, свойство (*pro*) – прилагательным, а связь (*con*) включает в себя все знаки пунктуации, предлоги, союзы и частицы.

Выделение сегментов можно производить несколькими методами. Рассмотрим наиболее эффективный метод сегментации – сегментацию через связь (*con*). В основе данного метода лежит предположение о том, что семантические скопления термов в ИЯ- и ЦЯ-предложениях отделены друг от друга связями (*con*) – знаками препинания, предлогами, союзами и частицами [1].

Таким образом, при каждом возникновении связи (*con*) происходит трансформация семантической структуры, возникает новый сегмент текста, несущий новую семантику. Следовательно, для осуществления сегментации текста необходимым и достаточным является наличие словаря служебных частей речи и словаря знаков препинания. Механизм сегментации через связь (*con*) для фрагмента предложения «*The goal of integrating syntactic information into translation model…*» приведен на рис. 1.

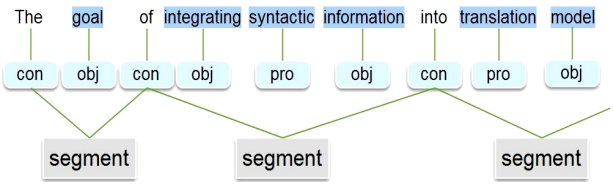

Рис. 1. Сегментация через связь (*con*)

Кластеризация. Для управления полученными сегментами применяется фреймовая модель представления знаний, где каждый терм сегмента описывается соответствующим фреймом. Однако для формирования такой модели предварительно необходимо сгруппировать имеющиеся в сегментах термы в кластеры – группы термов со схожими свойствами. Рассмотрим несколько алгоритмов кластеризации.

Для каждого вида термов (*obj*, *sub*, *pro*, *do*, *con*) определен ряд характеризующих их признаков. Так, для термов *obj*, *sub* и *pro* такими признаками являются «род», «число» и «падеж», для термов *do* – это «время», «вид» и «залог», а для термов *con* отличительными признаками являются свойства «предлог», «союз» и «пунктуация». Каждый из этих признаков, в зависимости от типа терма, принимает определенные значения. Например, свойство «род» может принимать одно из трех значений [мужской, женский, средний], а свойство «вид» – всего два значения [совершенный, несовершенный] и т. д. Данные значения берутся в качестве критериев кластеризации – деления на группы в зависимости от принимаемых значений. Для формализации значений термов сопоставим каждому значению числовую меру. Так, например, значениям [мужской, женский, средний] сопоставим значения [1, 2, 3], а значениям [совершенный, несовершенный] – значения [1, 2] и т. д. Таким образом, данные числовые значения играют роль расстояний между свойствами термов. Функция расстояния между двумя свойствами x_i и x_j записывается как $L(x_i, x_j)$ и обладает следующими признаками.

Неотрицательность расстояния:

$$L(x_i, x_j) \geq 0 . \quad (1)$$

Симметрия:

$$L(x_i, x_i) = 0 . \quad (2)$$

Неразличимость тождественных свойств:

$$L(x_i, x_j) = L(x_j, x_i) . \quad (3)$$

Неравенство треугольника:

$$L(x_i, x_j) \leq L(x_i, x_k) + L(x_k, x_j) . \quad (4)$$

Если все свойства термов $x_1, x_2, ..., x_n$ представить в виде матрицы данных X размером $p \times n$

$$X \begin{pmatrix} x_{11} & x_{12} & ...x_{1n} \\ x_{21} & x_{22} & ...x_{2n} \\ x_{p1} & x_{p2} & ...x_{pn} \end{pmatrix} = (x_1, x_2, ..., x_n) , \quad (5)$$

то расстояния $L(x_i, x_i)$ могут быть представлены в виде матрицы расстояний, имеющей симметричный вид:

$$L = \begin{pmatrix} 0 & L_{12} & ...L_{1n} \\ L_{21} & 0 & ...L_{2n} \\ L_{n1} & L_{n2} & ...0 \end{pmatrix} . \quad (6)$$

Чем больше мера $L(x_i, x_i)$, тем больше отличие в свойствах термов, а, следовательно, возрастает и вероятность принадлежности термов к разным кластерам. И наоборот, чем меньше значение $L(x_i, x_i)$, тем больше вероятность того, что термы принадлежат одному кластеру.

Расстояние $L(x_i, x_i)$ может быть вычислено несколькими способами.

Общая формула геометрического расстояния в многомерном пространстве, т. е. расстояния Минковского, определяется по формуле

$$L_p(x_i, x_j) = \left(\sum_{k=1}^{d} \left| x_{k,i} - x_{k,j} \right|^p \right)^{1/p} , \quad (7)$$

где d – размерность пространства; p – количество значений, принимаемое признаками.

Частным случаем геометрического расстояния между несколькими значениями свойств того или иного терма является евклидово расстояние. Его формула приведена ниже:

$$L_2(x_i, x_j) = \left[\sum_{k=1}^{d} \left(x_{k,i} - x_{k,j} \right)^2 \right]^{1/2}. \quad (8)$$

Следующий тип расстояния – манхэттенское (сити-блок, хэмминговское) расстояние:

$$L_1(x_i, x_j) = \sum_{k=1}^{d} \left| x_{k,i} - x_{k,j} \right|. \quad (9)$$

Однако манхэттенское расстояние обычно применяют при наличии дихотомических свойств (свойств, имеющих два значения). А так как некоторые свойства термов могут принимать более чем два значения, то такой тип расстояния является непригодным для кластеризации термов ИЯ- и ЦЯ-текстов.

Еще одним типом расстояния является супремум-норма (расстояние Чебышева):

$$L_\infty(x_i, x_j) = \sup \left\{ \left| x_{k,i} - x_{k,j} \right| \right\}. \quad (10)$$

Анализ рассмотренных типов расстояния показал, что для задачи кластеризации сегментов ИЯ- и ЦЯ-текстов пригодными являются расстояние Чебышева и евклидово расстояние [2].

Построение фреймовой модели структуры текста. После того как исходный и целевые тексты разбиты на сегменты и проведена кластеризация термов для всех сегментов, строится фреймовая модель представления полученной структуры (см. таблицу). Для каждого вида термов – *obj*, *sub*, *pro*, *do, con* – определяется одноименный вид фрейма, хранящий информацию о свойствах связанного с ним терма.

Структура фреймов, используемых при построении шаблонов

Имя фрейма	Идентификатор, присваиваемый фрейму, уникальный в данной фреймовой системе (*obj*, *sub*, *pro*, *do*, *con*)
Слоты	Свойства фрейма, принимающие значения из некоторого диапазона
Демоны	Автоматически запускаемые процедуры. Выполняются при осуществлении каких-либо действий над слотом: *IF-NEEDED* – указывает, какое действие необходимо выполнить если значение вставляется в пустой слот. *IF-ADDED* – указывает, какое действие необходимо выполнить при добавлении в слот значения. *IF-REMOVED* – указывает, какое действие необходимо выполнить при удалении значения из слота

Все свойства термов хранятся в слотах фрейма и имеют строковый тип данных. Так, например, для фрейма *do*, описывающего термы-глаголы определено три слота: время, вид и залог. При обработке терма-глагола формируется экземпляр фрейма *do*, а свойства терма записываются в слоты. Результат записи может выглядеть следующим образом: время – прошедшее, вид – несовершенный, залог – активный.

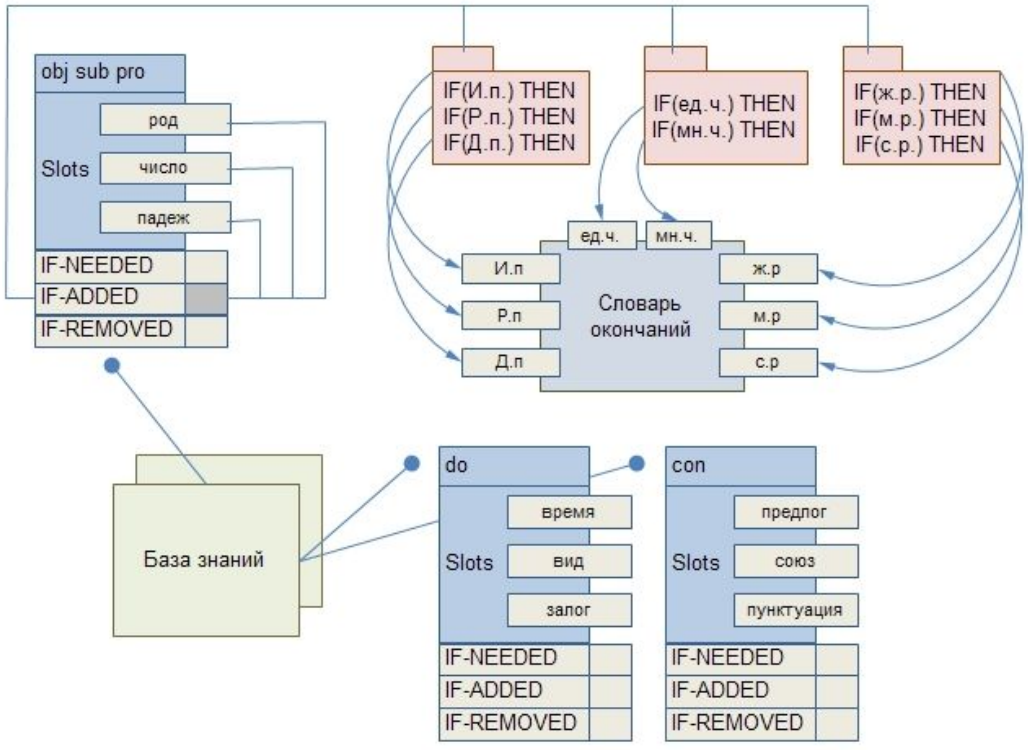

Рис. 2. Фреймовое представление знаний

Описание структуры фрейма. Сразу после добавления свойств терма в слоты соответствующих фреймов происходит вызов демона *IF-ADDED*, запуск которого осуществляется при каждой очередной вставке значения в тот или иной слот. Демон *IF-ADDED*, в зависимости от того, каким фреймом он был вызван, вызывает соответствующую процедуру приведения значения свойства терма, хранящегося в слоте к форме на целевом языке.

Так, например, при вставке в слот «число» значения «множественное» демон *IF-ADDED* вызовет процедуру преобразования формы терма из множественного числа исходного языка к множественному числу целевого языка. Для этого преобразующая процедура обращается к имеющемуся в системе словарю окончаний для пары «ИЯ–ЦЯ». Таким образом, формируется база знаний на основе фреймового представления (рис. 2), хранящая информацию о структуре сегментов текста и термов, образующих эти сегменты.

Данная модель пригодна для осуществления сопоставления исходного и целевых текстов на этапе синтеза ЦЯ-текста в IP-системе машинного перевода, а также для выполнения посегментного перевода фраз ИЯ-текста в фразы ЦЯ-текста.

Рассмотренная фреймовая модель является эффективным средством представления знаний в IP-системе машинного перевода на этапе синтеза текста, так как позволяет управлять формой термов при переходе от исходного языка к целевому, является менее громоздкой, чем представление через нейронную сеть, и более гибкой, чем продукционное представление. Приведенный алгоритм сегментации текста через связь позволяет быстро и эффективно производить разбиение текстового массива на фрагменты, что ускоряет процесс их анализа.

Библиографические ссылки

1. Мультилингвистическая модель распределенной системы на основе тезауруса / П. В. Зеленков, И. В. Ковалев, М. В. Карасева, С. В. Рогов // Вестник СибГАУ. Вып. 1(18). 2008. С. 26.

2. Заболеева-Зотова А. В., Камаев В. А. Лингвистическое обеспечение автоматизированных систем. М. : Высш. шк., 2008. С. 174–177.

K. V. Polyansky

TRANSLATING FRAME MODEL CONSTRUCTION WITH USE OF TERMS CLUSTERING

The knowledge representation frame model in IP-systems of machine translation is considered. The segmentation algorithm of the source and target text through communication is offered. Various terms clustering methods are analysed, the most suitable are offered to clusters text splitting.

Keywords: machine translation, text segmentation, terms clustering, frame model.

УДК 519.8

Е. С. Семенкин, А. А. Шабалов, С. Н. Ефимов

АВТОМАТИЗИРОВАННОЕ ПРОЕКТИРОВАНИЕ КОЛЛЕКТИВОВ ИНТЕЛЛЕКТУАЛЬНЫХ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ МЕТОДОМ ГЕНЕТИЧЕСКОГО ПРОГРАММИРОВАНИЯ*

Рассмотрены методы автоматизированного проектирования интеллектуальных информационных технологий (ИИТ) для решения сложных задач анализа данных и принятия решений. При генерации нейросетевых моделей, систем на нечеткой логике и нейро-нечетких систем применяются эволюционные алгоритмы. В проектировании коллектива ИИТ с целью повышения эффективности и надежности системы предложено применять метод генетического программирования.

Ключевые слова: нейронные сети, системы на нечеткой логике, нейро-нечеткие системы, эволюционные алгоритмы, генетическое программирование, коллективное принятие решений.

На сегодняшний день интеллектуальные системы получили широкое распространение при решении сложных задач анализа данных в различных областях человеческой деятельности. Искусственные нейронные сети [1], нечеткая логика [2], нейро-нечеткие системы [3], эволюционные алгоритмы [4] и другие методики и технологии являются популярным объектом исследования в силу их способности решать сложные интеллектуальные задачи, которые трудно решить с помощью классических методов [5].

* Работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» (НИР НК-136П/3, гос. контракт П1007) и ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007–2013 годы» (НИР 2011-1.9-519-005-042, гос. контракт 11.519.11.4002).