УДК 519.767

ИСПОЛЬЗОВАНИЕ МАТЕМАТИЧЕСКОГО АППАРАТА ПРЕДИКАТНЫХ КАТЕГОРИЙ ДЛЯ МОДЕЛИРОВАНИЯ СЕМАНТИКИ СВЕРХФРАЗОВЫХ ЕДИНСТВ

Н.Ф. Хайрова

НТУ «ХПИ», г. Харьков, Украина, nikhayv@vlink.kharkov.ua

Проведен анализ задач лингвистического процессора современных систем машинного перевода. Показано использование математического аппарата предикатных и модифицированных категорий для моделирования отношений в универсуме всевозможных элементов языковой системы. Разработана модель снятия лексической омонимии многозначных существительных сверхфразового единства.

ТЕОРИЯ ПРЕДИКАТНЫХ КАТЕГОРИЙ, СВЕРХФРАЗОВОЕ ЕДИНСТВО, МАШИННЫЙ ПЕРЕВОД, СЕМАНТИЧЕСКИЙ АНАЛИЗ

Введение

За последние пять-семь лет произошел качественный скачок в развитии систем автоматического перевода. Практически все современные системы трансферного типа позволяют получать перевод без грамматических (морфологических и контекстно-синтаксических) ошибок. Это связано с тем, что на поверхностных уровнях моделирования естественного языка — морфологическом и синтаксическом достигнуто достаточно много практических результатов.

Этап же семантического анализа по-прежнему остается одним из актуальных направлений исследования как в прикладной лингвистике, так и в области исследований искусственного интеллекта.

Как известно, под семантическим анализом в общем случае понимается представление значения входного текста в терминах некоторого формального языка, «понятного» ЭВМ. В большинстве современных систем машинного перевода задача семантического анализа сводится к выбору правильного значения переводного эквивалента многозначного слова.

1. Постановка задачи исследования

На мировом рынке сегодня представлено около тысячи коммерческих систем автоматического перевода, большинство из которых относятся к системам трансферного типа. В такого рода системах обычно для снятия семантической омонимии используется метод семантических фильтров. Этот легко реализуемый способ выбора переводного эквивалента многозначного слова, заключающийся в привлечении к переводу знаний о тематике переводимого текста. Его реализация обычно заключается в подключении заранее определенных пользователем предметных словарей. Использование данного метода на узкоспециализированных текстах (например, руководства по использованию) позволяет получить перевод довольно высокого качества. Но так как тексты редко бывают узкоспециализированными и раскрывают, как правило, несколько микротем, подключение предметных словарей ко всему тексту нередко не только не улучшают качество перевода, но и несколько ухудшают его.

Можно показать, что привлечение к переводу знаний микротемы, раскрываемой в каждом отдельном сверхфразовом единстве, увеличивает вероятность правильного выбора переводных эквивалентов многозначного слова.

Для реализации данной задачи, для моделирования отношений в универсуме всевозможных элементов языковой системы предлагается использовать положения предикатной и модифицированной теории категорий.

2. Описание используемого метода

Обрабатываемые лингвистическим процессором объекты являются дискретными, конечными и детерминированными, что позволяет использовать при их обработке теорию категорий, включающую понятия предикатных категорий и модифицированных категорий [1, 2]. На вход системы подаются объекты различного уровня языковой системы $x_1, x_2, ..., x_n$.. (морфемы, словоформы, словосочетания, семы, и так далее). Объекты, передающие информацию, берутся из конечных множеств X_1, X_2, ..., X_n (множество морфем языка, множество словоформ словаря и так далее) причем $x_1 \in X_1$, $x_2 \in X_2$, ..., $x_n \in X_n$. В результате работы лингвистического процессора на каждом уровне обработки языковой системы на выход поступает определенное множество объектов $y_1, y_2, ..., y_n$, под которыми мы будем понимать морфемы, словоформы, словосочетания на языке перевода. Причем $y_1 \in Y_1$, $y_2 \in Y_2$, ... , $y_n \in Y_n$.. Понятно, что объекты $y_1, y_2, ..., y_n$ прямо зависят от объектов $x_1, x_2, ..., x_n$. Это обозначает, что существует класс морфизмов, отображающий совокупность однотипных объектов друг на друга (морфизмов). Морфизмы представляют собой абстрактную направленную связь между объектами.

В [3] дается определение объектной категории C как совокупности однотипных математичес-

ких объектов X, Y (множеств, пространств, групп и так далее), для каждой пары из которых задано множество морфизмов (или стрелок) $Hom_C(X,Y)$, причем каждому морфизму соответствуют единственные X и Y. Категория C включает правило композиции морфизмов: для пары морфизмов $f \in Hom(X,Y)$ и $g \in Hom(Y,Z)$ определена композиция $f \circ g \in Hom(X,Z)$ и задает для каждого объекта X тождественный морфизм $id_x \in Hom(X,X)$.

В объектной категории множество объектов категории C обозначается $Ob\ C$. Если X является объектом категории C (C-object), то записывают $X \in ObC$. Если f морфизм из объекта в объект Y: $X \xrightarrow{f} Y$, то объект X называют началом морфизма f, а объект Y — его концом. Так как для каждого морфизма $f \in Hom(X,Y)$ существует единственная пара объектов X, Y, такая что X, $Y \in Ob\ C$ и $f \in Hom_C(X,Y)$, то можно интерпретировать морфизмы как некоторые функции, областью определения которых X, а область значений — Y.

В качестве однотипных математических объектов категории будем использовать переменные предикаты, заданные на множествах X и Y. Предикатная категория *Pred* [4] задается на некотором универсуме U. Из элементов универсума образуются подмножества X, Y, Z,... категории *Pred*, связанные с конкретной задачей аналитико-синтетической обработки лингвистического процессора системы. В роли множества объектов *Ob Pred* используется система всех подмножеств универсума U. Предикат $P(x)$, заданый на множестве X, рассматривается как экземпляр объекта X. А предикат $Q(y)$, задан на множестве Y, рассматривается как экземпляр объекта Y. Морфизм $f \in Hom(X,Y)$, преобразующий экземпляры объекта X в экземпляры объекта Y, может быть представлен в виде линейного логического оператора $F_f(P) = Q$. Каждый такой оператор преобразует одноместные предикаты P в одноместные предикаты Q:

$$\exists x \in X\,(C_f(x,y)\ P(x)) = Q(y) \qquad (1)$$

Предикат $C_f(x, y)$ задан на множестве $X \times Y$ и полностью определяет вид преобразования (1). В категории *Pred* каждому морфизму $f \in Pred$ взаимно однозначно соответствует предикат $C_f(x, y)$ преобразования (1). Каждый морфизм категории *Pred* можно задать, указав соответствующий ему предикат $C_f(x, y)$ на множестве $X \times Y$. Множество $Hom_{Pred}(X, Y)$ представляет собой совокупность преобразований вида (1) со всевозможными предикатами $C_f(x,y)$, заданными на всевозможных декартовых произведениях $X \times Y$ множеств X, $Y \subseteq U$.

3. Описание модели

В современных системах машинного перевода используется глубинный лингвистический процессор, задача которого на этапе семантической обработки правильно определить переводной эквивалент многозначного слова [5].

Под универсумом элементов U будем понимать все возможные элементы языковой системы, используемые на различных этапах обработки глубинного лингвистического процессора (словоформы, морфемы, семы, словарные статьи переводных словарей и так далее.). Предикаты, определенные на декартовых произведениях множества объектов *Ob Pred,* характеризуют этапы аналитико-синтетической обработки естественно-языкового текста, поступающего на вход системы.

Под множеством объектов X будем понимать множество словоформ существительных N сверхфразового единства входного текста. Множество N, которое мы будем рассматривать, представляет собой некоторую, достаточно четко очерченную, совокупность существительных $N=\{x_i\}$, $1 \le i \le n$, полученную в результате обработки морфологических структур словоформ, выявленных на предыдущих этапах работы лингвистического процессора [6]. Например результатом графемного и морфологического анализа английской фразы: *after selling some of the surplus blood to hospitals* будет выделение следующих морфологических структур существительных:

6, {<(surplus), S, SG> },

7, {*<(blood) S, SG>*}, ...

9, {*<(hospital), S, PL>*}, где *6*, *7*, *9* — порядковый номер словоформы во фразе; *S* — часть речи (существительное); *SG*, *PL* — соответственно единственное и множественное число.

Предикат *P(x)*, заданный на множестве *N*, рассматривается как экземпляр объекта *N*. Так как между множеством предикатов *P(x)* и множеством элементов *N* существует взаимно однозначное соответствие, то эти два множества взаимозаменяемы. Взяв множество предикатов в роли объекта, можно использовать элементы этого множества в роли экземпляров объектов.

Введем множество Y — множество семантических областей, включающих понятия, выражаемые существительными обрабатываемого сверхфразового единства. Под понятием понимается совокупность суждений о каком-либо объекте, отражающем его сущность; мысль, являющаяся результатом познания объекта, выделяющая предметы некоторого класса по определенным общим и совокупным специфичным для них признакам. Понятие формируется в сфере мышления и имеет внеязыковую природу. Но поскольку мысль не может существовать вне слова, под понятием будем подразумевать лексическую единицу, представляемую значением переводного эквивалента данного существительного и выражающую определенное понятие.

Семантические области Y возможных переводных эквивалентов многозначных существительных

сверхфразовых единств являются подмножеством предметных областей специализированных словарей, используемых при переводе $Y \subset D$. Множество используемых специализированных словарей является конечным и достаточно четко очерченным $D = \{y_j\}$, $1 \le j \le m$. Предикат $Q(y)$, заданный на множестве D, рассматривается как экземпляр объекта D. Так как между множеством предикатов $Q(y)$ и множеством элементов D существует взаимно однозначное соответствие, то эти два множества взаимозаменяемы.

Два множества $N = \{x_i\}$ и $D = \{y_j\}$ являются базовыми при использовании метода идентификации элементов смысла многозначных существительных сверхфразовых единств с использованием предикатных модифицированных категорий.

Ядром морфизма категории *Pred* будет служить предикат $C_f(x,y)$ (1), заданый на множестве $N \times D$. В предикатной категории *Pred* каждому морфизму $f \in$ *Pred* взаимно однозначно соответствует ядро $C_f(x,y)$, задающее отношение между существительными сверхфразового единства N и предметными словарями, отображающими семантические области понятия переводного эквивалента, D.

Множество N представляет собой множество всех многозначных существительных анализируемого сверхфразового единства входного текста. Под многозначным существительным будем понимать словоформу, которой на этапе морфологического анализа присвоен параметр S (часть речи существительное) и перевод данной словоформы есть в более чем одном подключенном словаре. Множество D представляет собой некоторую четко очерченную совокупность подключенных к переводу предметных, базовых и пользовательских словарей. Предполагается, что подключенные к переводу словари охватывают область семантического пространства существительных анализируемого сверхфразового единства переводимого текста. Наличие в подключенном словаре словарной статьи существительного множества N устанавливает морфизм $f \in$ *Pred*. Рассмотрев все возможные пары из множества $N \times D$, получаем множество морфизмов $Hom_{Pred}(N,D)$, включающее все возможные отношения $C(x_i,y_j)$, $1 \le i \le n$, $1 \le j \le m$.

4. Пример реализации модели

Рассмотрим сверхфразовое единство:

After selling some of the surplus blood to hospitals, the Red Cross has begun to destroy thousands of pints that have outlasted their shelf life. Directors of several Red Cross blood centers said their sites may discard as many as 1 of every 5 donations and the national total could easily reach tens of thousands.

На вход семантического анализа глубинного лингвистического процессора поступает множество $N = \{x_i\}$, $1 \le i \le 8$, многозначных существительных: x_1= *blood;* x_2= *hospital;* x_3= *red;* x_4= *cross;* x_5= *shelf;* x_6= *director;* x_7= *center;* x_8= *site.*

Множество подключенных переводных словарей $D = \{y_j\}$, $1 \le j \le m$: y_1= *General;* y_2= *Law;* y_3=*Biology;* y_4= *Medical;* y_5= *Economics;* y_6= *Education;* y_7= *Telecoms;* y_8= *Polytechnic.*

На рис. 1 показана категорная диаграмма, представляющая двудольный граф предиката $C(x, y)$.

Ядро морфизма категории *Pred*, заданного на декартовом произведении $N \times D$ множеств $N=\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$ и $D = \{y_1, y_2, y_3, y_4, y_5, y_6, y_7, y_8\}$, отображает отношения между элементами каждой пары x_i, y_j:

$$C(x,y) = (x^{x1} \vee x^{x2} \vee x^{x3} \vee x^{x4} \vee x^{x5} \vee x^{x6} \vee x^{x7})y^{y1} \vee$$

$$\vee x^{x5} y^{y2} \vee (x^{x3} \vee x^{x4})y^{y3} \vee \vee (x^{x3} \vee x^{x4})y^{y5} \vee x^{x8} y^{y6} \vee$$

$$\vee (x^{x4} \vee x^{x5} \vee x^{x6} \vee x^{x8})y^{y7} \vee (x^{x5} \vee x^{x6} \vee x^{x7})y^{y8}. \quad (2)$$

Тогда согласно преобразованию (1) соответствующий линейный логический оператор будет записан в виде:

$$Q(y) = \exists x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}(((x^{x1} \vee x^{x2} \vee x^{x3} \vee$$

$$\vee x^{x4} \vee x^{x5} \vee x^{x6} \vee x^{x7})y^{y1} \vee x^{x5} y^{y2} \vee (x^{x3} \vee x^{x4})y^{y3} \vee$$

$$\vee y^{y4}(x^{x1} \vee x^{x2} \vee x^{x8}) \vee (x^{x3} \vee x^{x4})y^{y5} \vee x^{x8} y^{y6} \vee (x^{x4} \vee$$

$$\vee x^{x5} \vee x^{x6} \vee x^{x8})y^{y7} \vee (x^{x5} \vee x^{x6} \vee x^{x7})y^{y8})P(x)). \quad (3)$$

Используя предикат узнавания предмета a по переменной x_i [7]:

$$x_i^a = \begin{cases} 1, x_i = a \\ 0, x_i \ne a \end{cases}, \quad (4)$$

где $i = \{1, 2, \ldots n\}$; a — любой элемент универсума, определяем реакцию морфизма на предикат $P(x) = x^{x1} \vee x^{x2}$, включающий существительные первой фразы сверхфразового единства *After selling some of the surplus blood to hospitals*:

$$Q(y) = \exists x \in \{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}(((x^{x1} \vee x^{x2} \vee x^{x3} \vee$$

$$\vee x^{x4} \vee x^{x5} \vee x^{x6} \vee x^{x7})y^{y1} \vee x^{x5} y^{y2} \vee (x^{x3} \vee x^{x4})y^{y3} \vee$$

$$\vee (x^{x3} \vee x^{x4})y^{y3} \vee y^{y4}(x^{x1} \vee x^{x2} \vee x^{x8}) \vee (x^{x3} \vee x^{x4})y^{y5} \vee$$

$$\vee x^{x8} y^{y6} \vee \vee (x^{x4} \vee x^{x5} \vee x^{x6} \vee x^{x8})y^{y7} \vee \vee (x^{x5} \vee x^{x6} \vee$$

$$x^{x7})y^{y8})(x^{x1} \vee x^{x2})) = y^{y1} \vee y^{y4}. \quad (5)$$

Для получения множества Q на графе (рис. 1) собираются вместе те элементы D, которые связаны с ребрами двудольного графа предиката $C(x, y)$ с элементами N, образующими множество $P(x) = x^{x1} \vee x^{x2}$. Множество Q= $\{y1, y2\}$ в рассматриваемом нами примере – это подключенные сло-

вари y_1= *"General dictionary"* и y_4= *"Medical dictionary"*.

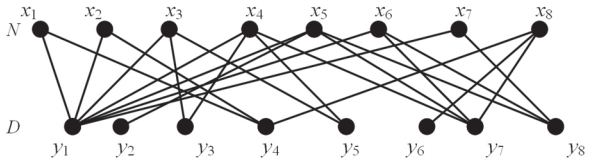

Рис. 1. Категорная диаграмма

Соответствующие переводные эквиваленты существительного x_1 — *кровь* (в базовом словаре *"General dictionary"* и в медицинском словаре — *"Medical dictionary")*; существительного x_2 — *госпиталь* (в базовом словаре *"General dictionary"* и в медицинском словаре — *"Medical dictionary")*. Используя полученные значения многозначных переводных эквивалентов, получаем наиболее правильный перевод рассматриваемой фразы: *"После сбыта некоторых излишков крови в госпитали..."*.

Выводы

Работа современных систем машинного перевода невозможна без реализации глубинного лингвистического процессора, осуществляющего аналитико-синтаксическую обработку текстов на естественном языке, использование которого на этапе семантического анализа требует формализовать выявление предметной области, раскрываемой в каждом сверхфразовом единстве или абзаце.

Использование предикатных и модифицированных категорий для моделирования отношений между многозначными существительными сверхфразовых единств и специализированными словарями позволяет создать легко реализуемую модель, снимающую многозначность переводных эквивалентов систем автоматического перевода. Так как математическим аппаратом данной модели является не базовая, а модифицированная предикатная теория категория, то данная модель может быть практически реализована не только средствами программного обеспечения, но и средствами аппаратного обеспечения, что в свою очередь обеспечит максимальное быстродействие.

Список литературы: 1. *С. Мак Лейн* [Maclane S.] Категории для работающего математика [текст]. — М.: Физматлит, 2004 [1998]. 2. *Булкин, В.И.* Использование метода декомпозиции бинарных предикатов при формализации интеллектуальной деятельности / *В.И. Булкин, Н. Ф. Хайрова, Н. В. Шаронова* // Вестн. Херсон. гос. техн. ун-та., Херсон, 2005.— N 1. С. 78-82. 3. *D.E. Rydeheard, R.M. Burstall* Computational Category Theory. — New York: Prentice Hall. — 1988. — XIII, 257 p. 4. *Бондаренко, М.Ф.* О модифицированных категориях [текст] / *М.Ф. Бондаренко, З.В. Дударь, А.А.Иванилов, В.В. Маникин, Ю.П. Шабанов-Кушнаренко* // Радиоэлектроника и информатика. – Х.: Изд-во ХНУРЭ. – 2005.– № 1 – С. 87-99. 5. *Хайрова, Н.* Модель разбиения множества элементов смысла многозначных слов переводимого предложения в системах автоматического перевода [текст]/ *Н. Хайрова, Н. Шаронова* // Бионика интеллекта: науч.-техн. журнал. – 2007. – № 2 (67). – С. 37-40. 6. *Хайрова,* Н.Ф. Машинный перевод [текст]: Учеб. пособие / *Н.Ф. Хайрова, И.В. Замаруева*. – Х.: Око, 1998. — 82 с. 7. *Шабанов-Кушнаренко, Ю.П.* Теория интеллекта: математические средства [текст] / Ю.П. Шабанов-Кушнаренко — Х.: Вища шк., 1984.— 143 с.

Поступила в редколлегию 18.09.2009

УДК 519.767

Використання математичного апарату предикативних категорій для моделювання семантики зверхфразових єдностей / Н.Ф. Хайрова // Біоніка інтелекту: наук.-техн. журнал. – 2009. – № 2 (71). – С. 36-39.

У пропонованій статті розглядаються завдання роботи лінгвістичного процесора в сучасних системах машинного перекладу. Показано використання математичного апарату предикативних і модифікованих категорій для моделювання зв'язків в універсумі всіляких елементів мовної системи. Запропонована модель дозволяє істотно зменшити лексичну омонімію багатозначних іменників зверхфразових єдностей.

Іл. 1. Бібліогр.: 7 найм.

UDC 519.767

Use of a mathematical tools of predicate categories for modelling semantics of superphrases unities / N. Khairova // Bionics of Intelligence: Sci. Mag. – 2009. – № 2 (71). – P. 36-39.

The present article is on working out algorithms of linguistic processor of machine translation systems. This article reviews the method of modeling relations between every possible elements of linguistic system by mathematical tools of predicate categories. Proposed model allows to decrease lexical ambiguity of polysemantic nouns of superphrases unities.

Fig. 1. Ref.: 7 item.