УДК 519.767.2

Н. Ф. Хайрова[1], Н. В. Шаронова[2], Н. В. Борисова[3]
[1]НТУ «ХПИ», Харьков, Украина, nina_khajrova@yahoo.com
[2]НТУ «ХПИ», Харьков, Украина, nvsharonova@mail.ru
[3]НТУ «ХПИ», Харьков, Украина, n_borisova2004@yahoo.com

ОПРЕДЕЛЕНИЕ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ НА ОСНОВЕ КОГНИТИВНОГО ПОДХОДА

Предлагается логическая схема выделения семантических эквивалентов, основанная на формализации когнитивной и номинативной функции языка. Формализация осуществляется за счет факторизации пространства концептов. Доказывается возможность использования подходов и методов теории интеллекта для реализации данной схемы в формальных моделях на языке логики предикатов. Вводятся предикаты эквивалентности, разбивающие множество лингвистических единиц на классы общего категориального значения.

КОГНИТИВНАЯ СЕМАНТИКА, СЕМАНТИЧЕСКИЕ ЭКВИВАЛЕНТЫ, ТЕОРИЯ ИНТЕЛЛЕКТА, ЛОГИКА ПРЕДИКАТОВ

Введение

Одной из основных задач современной компьютерной лингвистики является задача автоматического определения семантических эквивалентов. Эта задача непосредственно связана с направлением когнитивной семантики, ориентированной на построение моделей процесса понимания смысла. Данные модели рассматривают функции человеческого интеллекта, связанные с использованием естественного языка. К подобным функциям относятся: реферирование, перевод, экстракция и идентификация знаний, ответы на общие и специальные вопросы, перифраз и другая интеллектуальная деятельность, связанная с пониманием текста или речи.

Семантические эквиваленты представляют собой знаковые выражения одного и того же понятия, определяемые синонимами и семантически близкими словами. На сегодняшний день когнитивная лингвистика не имеет четкого определения синонимии лингвистических единиц и четких границ выявления критериев семантической близости, что не позволяет дать непротиворечивое определение близости значений [1]. Несмотря на большое количество исследований по синонимии, задачи осмысления сущности и определения границ данного явления остаются не до конца решенными [2]. Также сегодня не существует общепринятой количественной меры для обозначения степени синонимичности значения слов. Можно только с уверенностью говорить, что семантически близкими словами являются слова с близким значением, которые встречаются в одном контексте.

Сложность определения семантически близких слов и, в частности, синонимов обусловлена рядом когнитивно предопределенных причин. Основная из них обусловлена непрерывным изменением субстанции языка, являющегося открытой системой, с особенно быстро развивающимся словообразованием и выработкой новых понятий в молодых отраслях знания. В связи со сказанным наиболее перспективным подходом к автоматизированному выявлению семантических эквивалентов является использование моделей, базирующихся на семантической обработке текстов динамически изменяющихся предметных областей.

1. Постановка задачи исследования

С точки зрения когнитивной лингвистики определение синонимии в процессе познания мира человеком представляет собой компаративное действие, т.е. устанавливание сходства и различия между лингвистическими элементами в процессе сравнения.

Благодаря основным когнитивным механизмам порождения и восприятия знаний, включающих процессы категоризации, интеллект формирует некоторые образы понимания, т.е. отождествляет элементы по некоторым существенным, общим и специфичным признакам и свойствам с известным классом или объектом. Таким образом, в сознании соединяются значения различных по формальному определению слов. Этот процесс основан на устойчивой системе обобщенных значений и относится к моменту речи, т.е. к определенной ситуации.

Использование когнитивной парадигмы выявления синонимов на основе процесса категоризации позволяет формализовать смысловую близость лингвистических единиц за счет факторизации пространства концептов, выражаемых знаками лингвистических смысловых единиц.

Данный подход полностью соответствует моделируемому объекту — деятельности интеллекта человека по пониманию смысла лингвистических объектов.

Для построения логической схемы выделения семантических эквивалентов вводятся пространство смысловых лингвистических единиц и

пространство текстов, включающих данные лингвистические единицы.

Введение отношений эквивалентности между лингвистическими единицами и элементами связного текста позволяет факторизовать данные пространства. А использование моделей и методов теории интеллекта позволяет перейти от логической схемы выявления семантических эквивалентов к ее реализации на языке логики предикатов в формальных моделях смысловой идентификации синонимов.

Так как понятие синонимии и семантической эквивалентности определяется не для слов, а для концептов слов, т.е. синонимия неразрывно связана с контекстом, то можно построить логическую схему, позволяющую формализовать отношения между концептом и инсайтным смыслом элемента связанного текста. Предлагаемая модель позволяет перейти от отношений между концептами и инсайтным пониманием к отношениям между лексическими единицами, являющимися знаковым выражением данного концепта, и элементом связного текста.

Понятие «связный текст» как объект когнитивной семантики допускает множество определений и интерпретаций, которые обусловлены сложностью и многоаспектностью подходов к изучению объекта. Будем понимать под фрагментом связного текста законченное информационное и структурное целое, семантически и синтаксически объединяющее смысловой связью последовательность языковых единиц в единый фрагмент. Связный текст представляет собой целостный объект знаковой смысловой единицы верхнего уровня иерархической языковой системы [3]. Фрагмент связного текста может быть представлен высказыванием (реализованным предложением) или межфразовым единством (ряд высказываний в едином фрагменте) [4].

2. Логическая схема выделения семантических эквивалентов

Введем метрическое пространство лингвистических смысловых единиц Θ, определяемое как множество лингвистических единиц лексикона T, на котором грамматические правила задают отношения между единицами, выступающими ограничениями для корректных синтаксических структур.

Для определения метрики пространства в качестве расстояния $\beta(t', t'')$ между двумя лингвистическими единицами t' и t'' используем меру семантической близости $f(t', t'')$ такую, что:

$$\beta(t', t'') = 1/f(t', t'').$$

Меру семантической близости f формально определим соотношением (1) через соответствующие дефиниции глоссариев x_1 и x_2 как мощности множеств, образованных теоретико-множественным пересечением и объединением множеств терминов дефиниций.

$$f(t',t'') = \frac{|X_1 \bigcap X_2|}{|X_1 \bigcup X_2|} \quad (1)$$

Здесь $x_1 \bigcap x_2$ — общие термины определений, а $x_1 \bigcup x_2$ — все термины определений x_1 и x_2; под термином в данном контексте мы понимаем понятие из глоссария, взятое в его канонической форме.

Так как для определения семантической близости между понятиями будем использовать несколько словарей, в которых существуют допустимо различные дефиниции одних и тех же лингвистических смысловых единиц, расстояния между двумя лингвистическими единицами удобнее переписать в виде:

$$f(t',t'') = \frac{\sum_{i=1}^{n_1} \frac{\sum_{j=1}^{n_2} f(x_{1i}, x_{2j})}{n_2}}{n_1},$$

где n_1— количество определений первого термина, взятых из обрабатываемых глоссариев; n_2 — количество определений второго термина, взятых из обрабатываемых глоссариев, x_{1i} — i-е определение первого термина; x_{2j} — j-е определение второго термина.

Фрагмент связного текста, включающий лингвистические смысловые единицы $t \in \Theta$, обозначим как d. Иерархия отношений элементов связного текста многоуровневой языковой системы наглядно представляется соответствующей теоретико-множественной структурой, в которой D представляет граф конечного множества фрагментов связных текстов $\{D_1, D_2, ..., D_m\}$, принадлежащих пространству исследуемых связных текстов Ω [5]. Здесь текст $D_i \in \Omega$, $i = 1,..., m$.. При этом текст D_i более высокого уровня иерархии языковой системы можно формально определить через элементы D^j_i ($D^j_i \subset D_i$, $j = 1, 2,.....n$) связного текста предыдущего уровня иерархии (сверхфразовое единство определяется через фразу, тогда как связный текст документа можно определить через сверхфразовые единства):

$$D_i = \bigcup_{j=1}^{n} D_i^j, \quad \bigcap_{j=1}^{n} D_i^j = \varnothing$$

В рассматриваемом пространстве Ω вершина D_i графа D будет родительской для вершин множества $\{D^1_i, D^2_i, ..., D^n_i\}$.

Тогда расстояние между двумя связными текстами можно определить как длину пути между соответствующими контекстами $||\alpha(D_i, D_j)||$, определяемую количеством несовпадающих листьев вершин D_i и D_j.

Пара элементов $(t, d) \in (\Theta, \Omega)$ представляет собой одну лингвистическую смысловую единицу и один фрагмент связного текста, где Θ — пространство лингвистических единиц рассматриваемого лексикона T, а Ω — пространство рассматриваемых фрагментов связных текстов.

Если рассмотреть все возможные пары декартового произведения Θ*Ω, то можно построить отображение F: (Θ*Ω) → ϑ, где ϑ — пространство смысловых полей связных текстов. Схема появления пространства смысловых полей из рассматриваемых фрагментов связных текстов и привлекаемых лингвистических смысловых единиц представлена на рис. 1.

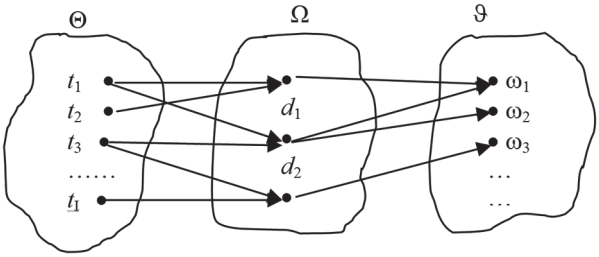

Рис. 1. Схема отображения пространств $(\Theta,\Omega) \xrightarrow{F} \vartheta$

В данном случае единым категориальным значением синонимичных лингвистических единиц выступает единое смысловое поле рассматриваемых фрагментов текста. Выбрав лингвистическую смысловую единицу t и связный текст d, включающий данную лингвистическую единицу, мы определяем смысл элемента связного текста ω через отображение F.

Например:

F (spring, "I made a spring towards a boat") = осуществление человеком прыжка.

F (spring, " He was in the spring of his years") = период жизни человека.

F (spring, "I was in my five and twentieth spring") = период жизни человека.

Будем говорить, что две лингвистические единицы связаны в одном смысле (или синонимичны в своем сигнификативном значении), и писать $(t_i, d_i) \sim (t_j, d_j)$, если только $F(t_i, d_i) = F(t_j, d_j)$.

F("application", "the most Internet applications for the Web are XML-applications") = "программное обеспечение";

F("application", "application for admission to a university") = "заявление";

F("software", "using commercial computer-based software") = "программное обеспечение";

F ("application", "the most Internet applications for the Web are XML-applications") = F("software", "using commercial computer-based software").

Можно показать, что отношение ~, устанавливаемое между лингвистическими смысловыми единицами t и элементами связного текста d, выражает эквивалентность и факторизует пространства лингвистических смысловых единиц Θ и исследуемых связных текстов Ω, разбивая их на классы эквивалентности. Для этого достаточно показать, что отношение ~ является рефлексивным, транзитивным и симметричным [6].

Отношение $(t_i, d_i) \sim (t_j, d_j)$ является рефлексивным отношением. Одна лингвистическая единица в одном своем сигнификативном значении связано само с собой, ибо

$$(t_i, d_i) \sim (t_i, d_i) \leftrightarrow F(t_i, d_i) = F(t_i, d_i).$$

Отношение $(t_i, d_i) \sim (t_j, d_j)$ является симметричным отношением: если одна лингвистическая единица в одном своем сигнификативном значении связана с другой (в одном из ее значений), то вторая лингвистическая единица связана с первой (в вышеназванных значениях):

$$(t_i, d_i) \sim (t_j, d_j) \leftrightarrow F(t_i, d_i) = F(t_j, d_j) \equiv$$
$$\equiv F(t_j, d_j) = F(t_i, d_i) \leftrightarrow (t_j, d_j) \sim (t_i, d_i).$$

Отношение ~ является транзитивным отношением: если одна лингвистическая единица определяет тот же сигнификативный смысл, что и вторая, а вторая лингвистическая единица имеет тот же сигнификативный смысл, что и третья, то первая лингвистическая единица в одном из своих сигнификативных значений связана с третьей:

$$(t_i, d_i) \sim (t_j, d_j) \text{ и } (t_j, d_j) \sim (t_k, d_k) \leftrightarrow F(t_i, d_i) = F(t_j, d_j) \text{ и}$$
$$F(t_j, d_j) = F(t_k, d_k) \Rightarrow$$
$$F(t_i, d_i) = F(t_k, d_k) \leftrightarrow (t_j, d_j) \sim (t_k, d_k).$$

Например,

("application", "the most Internet applications for the Web are XML-applications") ~ ("software", "using commercial computer-based software") и *("software", "using commercial computer-based software") ~ ("program", "everything done on a computer is done by using a program") ↔ F("application", "the most Internet applications for the Web are XML-applications")= F ("software", "using commercial computer-based software")= F("program", "everything done on a computer is done by using a program") ="программное обеспечение".*

Данное отношение эквивалентности позволяет организовать различные пары лингвистических единиц и фрагментов связных текстов, включающих данные единицы, (t, d), в классы эквивалентности, которые определяют один и тот же сигнификативный смысл. Это, тем самым, факторизует пространство концептов, выражаемых знаками лингвистических смысловых единиц.

Отношение эквивалентности ~ делает F однозначным отображением, в котором два концепта имеют одинаковое синонимичное значение, если они являются одним и тем же классом, что позволяет нам выбрать одну репрезентативную лингвистическую единицу, представляющую подходящее значение из каждого класса эквивалентности.

3. Введение контекстно-знакового предиката

Для реализации данной логической схемы в формальной модели идентификации семантических

отношений синонимии используем подходы теории интеллекта [6].

На декартовом произведении элементов множеств $T*D$ вводим контекстно-знаковый предикат $L(t_i, d_j)$, задающий отношения между лингвистическими единицами лексикона и контекстом. Если $L(t_i, d_j) = 1$, то это значит, что лингвистическая единица t_i из множества T однозначно соответствует обрабатываемому контексту $d_j \in D$. Если $L(t_i, d_j) = 0$, то t_i не соответствует d_j.

Предикат L должен удовлетворять постулату существования: предикат $L(t_i, d_j)$ реально существует в том и только в том случае, если при повторном предъявлении любой пары (t_i, d_j) из множества $T*D$ всегда будет получен тот же ответ, что и в первый раз.

Таким образом, контекстно-знаковый предикат $L(t_i, d_j)$ для каждой пары t_i и d_j объективно отображает отношение включения знака лингвистической смысловой единицы в элемент связного текста.

Отношение эквивалентности ~, факторизующее пространство лингвистических смысловых единиц Θ посредством разбиения его на классы эквивалентности, однозначно определяется контекстно-знаковым предикатом $L(t_i, d_j)$. Это позволяет ввести предикат категориальных семантических признаков лингвистических единиц G_t, заданный на декартовом квадрате $T*T$. Можно показать, что предикат G_t является предикатом эквивалентности:

$$G_t(t', t'') = \forall d \in D\,(L(t', d) \sim L(t'', d)).$$

Предикат $G_t(t', t'')$ можно использовать для объективного определения общих категориальных семантических признаков лингвистических единиц.

Действительно, если $G_t(t', t'') = 1$, то $L(t', d) = L(t'', d)$ и если $G_t(t', t'')=0$, то $L(t', d) \neq L(t'', d)$ для любого связного текста $d \in D$. Таким образом, две лингвистические единицы в одном контексте либо имеют общие категориальные семантические признаки (один или более), либо не имеют таковых.

Например, если G_t (*"application", "software"*)=*1*, то *L (application, "using commercial computer-based application") = L ("software", "using commercial computer-based software")*.

Предикат G_t определяет разбиение Ψ множества T на слои лингвистических единиц. Все лингвистические единицы, принадлежащие одному слою разбиения, относятся к концептам, обладающим некоторым категориальным признаком, т.е. к синонимичным концептам, а любые лингвистические смысловые единицы, взятые из разных слоев разбиения Ψ, относятся к концептам, не имеющим общих категориальных признаков или элементов смысла.

Выводы

Рассмотрены когнитивные аспекты проблемы выделения синонимов и семантических эквивалентов в тексте. Номинативные и когнитивные механизмы формирования синонимов и синонимичных рядов включают процессы категоризации, отождествляющие элементы по некоторым существенным, общим и специфическим свойствам с известным классом. Использование рассмотренной когнитивной парадигмы выявления синонимов на основе процесса категоризации позволило разработать логическую схему формализации смысловой близость лингвистических единиц за счет факторизации пространства концептов. Категориальным значением синонимичных лингвистических единиц при этом выступает единое смысловое поле рассматриваемых фрагментов текста.

Показана возможность использования подходов и методов теории интеллекта для реализации данной логической схемы в формальных моделях семантической обработки текстов на языке логики предикатов. Введенные предикаты позволяют разбить множество лингвистических смысловых единиц на классы эквивалентности, соответствующие определяемому категориальному значению семантических признаков.

Разработанная модель реализована в прототипе системы автоматизированного выделения семантических эквивалентов и эффективно используется на этапе семантической обработки для выделения близких по смыслу слов в англоязычных текстах предметной области "компьютерные технологии".

Список литературы: 1. *Шумилова А.А.* Лексическая синонимия: традиционное и когнитивное видение проблемы / А. А. Шумилова// Вестн. Челяб. гос. ун-та. 2009. № 22 (160). Филология. Искусствоведение. Вып. 33. С. 144–148. **2.** *Азарова И.В.*, Компьютерный тезаурус русского языка типа WordNet / И. В. Азарова, О. А. Митрофанова, А. А. Синопальникова //Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2003». – С. 1–6. **3.** *Хайрова Н.Ф.* Концептуальная схема идентификации смысла лингвистических единиц / Н. Ф. Хайрова // Сборник научных работк Военного института Киевского национального университета имени Тараса Шевченка. – Киев: ВИКНУ, 2013.- Вып. № 39. - С.217-223. **4.** *Ерофеева Е.В.* К вопросу о соотношении понятий «текст» и «дискурс» / Е. В. Ерофеева, А. Н. Кудлаева // Пробл. социо- и психолингвистики: сб. ст. — Вып. 3. — Пермь: Перм. ун-т, 2003. — С. 28—36. **5.** *Dieter Jungnickel.* Graph, Networks and Algorithms. *Algorithms and Computation in mathematics*. Volume5. – Springer Berlin Heidelberg New York, 2008. — 650 p. **6**. *Бондаренко М Ф.* Теория интеллекта: учебник / М. Ф. Бондаренко, Ю. П. Шабанов-Кушнаренко. – Харьков: Комп. СМИТ, 2007. — 576 с.

Поступила в редколлегию 20.05.2013

УДК 519.767.2

Визначення семантичної близькості на основі когнітивного підходу / Н.Ф. Хайрова, Н.В. Шаронова, Н.В. Борисова // Біоніка інтелекту: наук.-техн. журнал. – 2013. – № 2 (81). – С. 22-26.

В статті пропонується логічна схема виділення семантичних еквівалентів, основана на формалізації когнітивної і номінативної функцій мови. Доводиться можливість використання підходів і методів теорії інтелекту для реалізації даної схеми на мові логіки предикатів. Вводяться предикати еквівалентності, що розбивають множину лінгвістичних одиниць на класи загального категоріального значення.

Іл. 1. Бібліогр.: 6 найм.

UDK 519.7

Identification of semantic proximity based on the cognitive approach / N.F. Khairova, N.V. Sharonova, N. V. Borisova // Bionics of Intelligense: Sci. Mag. – 2013. – № 2 (81). – P. 22-26.

In the paper, we suggest logic scheme for semantic equivalence identification based on the formalization of cognitive and nominative language functions. The formalization is based on the factorization of space concepts. We prove a possibility of using approaches and methods of theory of intelligence for implementation the logic scheme in formal models in the language of predicate logic. Moreover, we provide predicates of equivalence for partitioning of the set of the linguistic units with regard to the defined categorical value.

Fig. 1. Ref.: 6 items.