МАТЕМАТИЧЕСКАЯ МОДЕЛЬ ИДЕНТИФИКАЦИИ СЕМАНТИЧЕСКИХ ОТНОШЕНИЙ ТОЛЕРАНТНОСТИ И ЭКВИВАЛЕНТНОСТИ

Светлана Петрасова, Нина Хайрова

Аннотация*: Формализация отношения семантики является сложно реализуемой задачей автоматической обработки текстов по причине неявной выраженности в естественно-языковых конструкциях. В работе предлагается математическая модель идентификации таких семантических отношений, как толерантность и эквивалентность на базе знаний глоссария. Рассматриваются семантические отношения терминов глоссария с точки зрения возможности идентификации концептов и их отношений. Предложенная математическая модель идентификации семантических отношений позволяет выделить классы синонимичности терминов в одном из своих концептуальных значений за счет факторизации пространства концептов. Для формализации категорий межконцептуальных отношений предлагается использовать диапазон значений коэффициента семантической близости. В результате определяется эвристическая оценка эффективности разработанной модели идентификации семантических корреляций концептов.*

Ключевые слова: *автоматическая обработка текстов, естественно-языковая конструкция, идентификация семантических отношений, толерантность, эквивалентность, глоссарий, формализация межконцептуальных отношений, семантическая близость.*

ACM Classification Keywords*: H.3.3 .Information Search and Retrieval, I.2.4. Knowledge Representation Formalisms and Methods*

Введение

Формализация семантических отношений является одной из актуальных задач при автоматической обработке текстов (АОТ) естественного языка. При этом АОТ осложняется проблемой неоднозначности естественного языка, что мешает четко формализовать семантические отношения между концептами в естественно-языковых текстах.

Для решения задачи идентификации семантических корреляций ведутся исследования по разработкам интеллектуальных систем, в основе которых лежит использование баз знаний для формализации межконцептуальных отношений. Но современный уровень развития информационных технологий позволяет только частично решить эту проблему.

В работе в качестве основного источника знаний для формализации межконцептуальных отношений предлагается использовать тексты естественного языка.

Использование текстов в качестве информационной базы требует выделения в них классов элементов, играющих функциональную роль в представлении знаний. Такими классами является класс терминологических понятий и класс отношений.

Термины и межконцептуальные отношения наиболее концентрированно отражены в глоссариях, где словарные статьи представляют собой тексты с насыщенной смысловой нагрузкой. Таким образом, интеллектуальная система, построенная на основе структурированных текстов глоссария, позволит наиболее полно и точно выделять концепты и формализовать семантические отношения между концептами.

Общая постановка задачи

Возможность формализации межконцептуальных отношений глоссария обеспечивается выявлением знаний из словарных статей. При этом основной сложностью экстракции концептов и их отношений, наряду с проблемами семантического анализа слабоструктурированных текстов естественного языка, остается сложность формализации семантически близких межконцептуальных отношений.

Для выявления неявно выраженных в естественно-языковых конструкциях семантических отношений предлагается использование меры семантической близости концептов, которые обладают семантическими корреляциями и имеют определенную общность содержания, выражающую некоторое сходство обозначаемых явлений или понятий [Кобозева, 2000; Широков, 1998]. При этом виды связей между значениями концептов различаются по степени общности или эквивалентности. Математическая модель идентификации отношений толерантности и эквивалентности позволит разделить корреляции концептов по степени семантической близости.

Описание математической модели идентификации отношений толерантности и эквивалентности

Для построения логической схемы выделения семантически связанных концептов вводится метрическое пространство лингвистических единиц Θ, определяемое как множество терминов глоссария T, на котором грамматические правила задают отношения между единицами, выступающими ограничениями для корректных синтаксических структур [Jungnickel, 2008].

Для определения метрики пространства используем меру семантической близости $f(t_1, t_2)$ между двумя лингвистическими единицами (терминами) t_1 и t_2.

Меру семантической близости *f* формально определим соотношением (1) через соответствующие дефиниции глоссариев d_1 *u* d_2 как мощности множеств, образованных теоретико-множественным пересечением и объединением множеств терминов дефиниций.

$$f(t',t'') = \frac{2 \mid d_1 \bigcap d_2 \mid}{\mid d_1 \mid + \mid d_2 \mid}, \tag{1}$$

где $d_1 \bigcap d_2$ — общие термины дефиниций глоссария, а $\mid d_1 \mid + \mid d_2 \mid$ — все термины дефиниций d_1 и d_2; под термином в данном контексте мы понимаем концепт из глоссария, взятый в его канонической форме.

Для более точного определения семантической близости между концептами будем использовать несколько глоссариев. В таком случае возможны различные дефиниции одних и тех же терминов, при этом расстояние между двумя сроками будет иметь вид:

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

где n_1 – количество дефиниций первого термина, взятых из обрабатываемых глоссариев, n_2 – количество дефиниций второго термина, взятых из обрабатываемых глоссариев, d_{1i} – i-я дефиниция первого термина, d_{2j} – j-я дефиниция второго термина.

Дефиницию, включающую термины глоссария $t \in \Theta$, 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 [Jungnickel, 2008]. Здесь текст $D_i \in \Omega$, $i = 1,\ldots, m$. При этом текст D_i более высокого уровня иерархии языковой системы можно формально определить через элементы D_i^j ($D_i^j \subset D_i$, $j = 1, 2,\ldots, n$) связного текста предыдущего уровня иерархии (сверхфразовом единство определяется через фразу, тогда как связный текст дефиниции 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 $\{D^1_i, D^2_i, \ldots, D^n_i\}$.

Пара элементов $(t, d) \in (\Theta, \Omega)$ представляет собой один термин и один фрагмент связного текста дефиниции глоссария (фраза, предложение, полное определение), где Θ — пространство

лингвистических единиц рассматриваемого глоссария T, а Ω — пространство рассматриваемых фрагментов связных текстов глоссария.

Будем говорить, что две лингвистические единицы связаны в одном семантическом поле и писать $(t_i, d_i) \sim (t_j, d_i)$, если только $F(t_i, d_i) = F(t_j, d_j)$.

Например, F("*спецификация*", d_1); F("*формат*", d_2); F("*спецификация*", d_1) = F("*формат*", d_2).

При этом для термина t_1 d_1={*"Повний опис вимог по складанню початкової програми з урахуванням обмежень на використовувані засоби і представлення даних, зв'язків з іншими програмними модулями та ін."*} [Півняк, 2010];

а для термина t_2 d_2={*«Специфікація та спосіб розташування і представлення даних у пам'яті, в базі даних або на зовнішньому носієві"*} [Півняк, 2010].

Можно показать, что отношение ~, устанавливаемое между терминами t и элементами связного текста d, выражает толерантность и факторизует пространства лингвистических смысловых единиц Θ и исследуемых связных текстов Ω, разбивая их на классы толерантности.

Толерантностью называется отношение, обладающее свойствами рефлексивности и симметричности [Шрейдер, 1971]. Можно показать, что отношение $(t_i, d_i) \sim (t_j, d_i)$ является рефлексивным отношением, т.е. один термин глоссария в одном своем сигнификативном значении связан сам с собой, и симметричным, один термин глоссария в одном своем сигнификативном значении связан с другим (в одном из его значений) и одновременно второй термин связан отношением ~ с первым (в вышеназванных значениях).

Например,

$$(\text{“формат”}, d_2) \sim (\text{“структура”}, d_3) \leftrightarrow F(\text{“формат”}, d_2) = F(\text{“структура”}, d_3) \equiv$$
$$\equiv F(\text{“структура ”}, d_3) = F(\text{“формат”}, d_2) \leftrightarrow (\text{“структура”}, d_3) \sim (\text{“формат”}, d_2).$$

При этом для термина t_3 d_3={*"Організація даних, що характеризується спеціальним описом посилань на зв'язки між елементам"*} [Півняк, 2010].

Пространство толерантности S_p, где p – число классов толерантности, состоящее из множеств номеров вида $N=\{n_1, n_2, \ldots, n_k\}$, при этом все $n_i \leq p$, причем элементы (t_i, d_i) и (t_j, d_i) толерантны, если они содержат общий номер. Множество K_h является классом толерантности, если K_h состоит из всех множеств вида $\{i, n_1, \ldots, n_k\}$ и число элементов множества K_h равно 2^{p-1} – число всех подмножеств множества из оставшихся $p-1$ номеров.

Смысл этого утверждения состоит в том, что отношение $(t_i, d_i) \sim (t_j, d_i)$ выполняется тогда и только тогда, когда существует класс K_i, содержащий одновременно (t_i, d_i) и (t_j, d_i). Если $(t_i, d_i) \sim (t_j, d_i)$, то (t_i, d_i) и (t_j, d_i) содержат некоторый общий номер h, и тем самым входят в класс K_h.

Частным случаем отношения толерантности является отношение эквивалентности. Чтобы показать, что отношение ~, устанавливаемое между терминами глоссария 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, достаточно показать, что отношение ~ является не только рефлексивным и симметричным, но и транзитивным [Бондаренко, 2007].

Отношение ~ является транзитивным отношением, если один термин глоссария в одном из своих значений имеет тот же сигнификативный смысл, что и второй термин в одном из своих значений, и второй термин в уже обозначенном значении имеет тот же сигнификативный смысл, что и третий в одном из своих смыслов, и тогда первый термин в определенном сигнификативном значении связан с третьем:

Например,

$$(\text{“спецификация”}, d_1) \sim (\text{“формат”}, d_2) \text{ и } (\text{“формат”}, d_2) \sim (\text{“структура”}, d_3) \leftrightarrow$$

$$\leftrightarrow F(\text{“спецификация”}, d_1) = F(\text{“формат”}, d_2) = F(\text{“структура”}, d_3).$$

Данное отношение эквивалентности позволяет организовать различные пары терминов и фрагментов связных текстов, включающих данные единицы, (t, d), в классы эквивалентности, которые определяют один и тот же сигнификативный смысл, тем самым, позволяя факторизовать пространство концептов, выражаемых знаками лингвистических смысловых единиц, на классы синонимичных в каком-то из своих смыслов концептов [Хайрова, 2013].

Эвристическая оценка эффективности разработанной модели

В качестве кортежа объективно измеряемых показателей, характеризующих эффективность работы разработанной модели идентификации семантически связных концептов, были использованы показатели эффективности, утвержденные межгосударственным стандартом по информации, библиотечному и издательского дела: коэффициенты точности (precision) и полноты (recall).

Для определения коэффициентов точности – *precision* и полноты – *recall* необходимо по результатам выделения множества межконцептуальных отношений определить n_{yy} – число выделенных элементов, релевантных семантическому полю с точки зрения эксперта, n_{yn} – число выделенных элементов, нежелательных с точки зрения эксперта, n_{ny} – число релевантных элементов, невыделенных системой и n_{nn} – число нежелательных элементов, невыделенных системой.

При определении эффективности работы системы коэффициент точности определяется следующей формулой:

$$precision = \frac{n_{yy}}{n_{yy} + n_{yn}}, \quad (3)$$

коэффициент полноты определяем как:

$$recall = \frac{n_{yy}}{n_{yy} + n_{ny}}. \quad (4)$$

Коэффициент шума вычисляем по формуле:

$$error = \frac{n_{ny} + n_{yn}}{n_{yy} + n_{ny} + n_{yn} + n_{nn}}. \quad (5)$$

Для определения качества разработанной математической модели исследовалась выборка из тысячи концептов на украинском языке. Проведенный ранее эвристический анализ коэффициентов семантической близости концептов показал, что концепты с коэффициентом семантической близости 0,28-0,35 связаны отношением толерантности. Семантическими эквивалентами могут считаться те концепты, мера семантической близости которых входит в диапазон от 0,35 до 1 [Хайрова, 2014].

Например, t_1 = *"спецификация"*, t_2 = *"формат"* и t_3 = *"структура"*. Согласно формуле (1) определяем меру семантической близости *f* для (t_1, d_1), (t_2, d_2) и (t_3, d_3):

$$f(t_1, t_2) = 0{,}37;$$

$$f(t_2, t_3) = 0{,}35;$$

$$f(t_1, t_3) = 0{,}46.$$

Данный показатель семантической близости подтверждает, что концепты t_1 = *"спецификация"*, t_2 = *"формат"* и t_3 = *"структура"* связаны отношением эквивалентности. Следовательно, можно показать, что отношение, которое устанавливается между данными терминами, удовлетворяет как свойствам симметричности и рефлексивности, так и свойства транзитивности:

(“спецификация”, d_1) ~ (“формат”, d_2) и

*(“формат”, d_2) ~ (“*структура*”, d_3)* ↔

↔ *F(“спецификация”, d_1) = F(“формат”, d_2) = F(“*структура*”, d_3)*.

В результате данного эксперимента было определено 750 концептов, связанных отношением толерантности, и 400 семантических эквивалентов.

Полученный средний коэффициент полноты recall=1. Система выделяет все связи концептов, определенные интеллектуально экспертом. Средний коэффициент точности, показывающий правильность определения семантических связей, precision=0,9093. Коэффициент шума, определяющийся отношением числа неправильно определенных системой связей к общему числу связей, выданных системой, error=0,0898.

В дальнейшей разработке данные показатели должны быть учтены для повышения качества работы системы, осуществляющей автоматическую идентификацию семантически связных концептов.

Выводы

Результатом данного исследования является разработка математической модели идентификации семантических межконцептуальных отношений толерантности и эквивалентности из глоссария на украинском языке как естественно-языкового текста, наиболее полно концентрируемого знания о предметной области.

Модель позволяет кластеризовать концепты за счет факторизации пространства терминов глоссария, используя в качестве основы их смысловую близость. Категориальным значением синонимических лингвистических единиц при этом выступает единое смысловое поле рассмотренных дефиниций.

Эвристическая оценка эффективности разработанной модели показала достаточно высокие результаты полноты и точности и в дальнейшем данные показатели должны быть учтены для повышения качества работы системы, осуществляющей автоматическую идентификацию семантически связных концептов.

Литература:

[Jungnickel, 2008] Jungnickel D. Graph, Networks and Algorithms // Algorithms and Computation in mathematics. – Vol. 5. – Springer Berlin Heidelberg New York, 2008. – 650 p.

[Бондаренко, 2007] Бондаренко М.Ф., Шабанов-Кушнаренко Ю.П. Теория интеллекта. – Харьков : Комп. СМИТ, 2007. – 576 с.

[Кобозева, 2000] Кобозева И.М. Лингвистическая семантика. – М. : Эдиториал УРСС, 2000. – 352 с.

[Півняк, 2010] Тлумачний словник з інформатики / Г.Г. Півняк, Б.С. Бусигін, М.М. Дівізінюк та ін. – Д. : Нац. гірнич. ун-т, 2010. – 600 с.

[Хайрова, 2013] Петрасова С.В., Кочуева З.А., Хайрова Н.Ф. Метод автоматической экстракции парадигматических отношений между понятиями толкового словаря // Вестник НТУ "ХПИ". Серия : Системный анализ, управление и информационные технологии. – Х. : НТУ "ХПИ", 2013. – № 62 (1035). – С. 118–124.

[Хайрова, 2014] Хайрова Н.Ф., Петрасова С.В., Ленков С.В. Метод автоматической идентификации семантических корреляций терминов глоссария // Сборник научных трудов Военного института Киевского национального университета имени Тараса Шевченко. – К. : ВИКНУ, 2014. – № 46. – С. 128–135.

[Широков, 1998] Широков В.А. Інформаційна теорія лексикографічних систем: моногр. — К. : Довіра, 1998. — 331 с.

[Шрейдер, 1971] Шрейдер Ю.А. Равенство, сходство, порядок. – М. : «Наука», 1971. – 256 с.

Authors' Information

Нина Хайрова *– профессор кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: nina_khajrova@yahoo.com*

Научные интересы: искусственный интеллект, идентификация знаний из текстов, Text Mining, Opinion Mining, Web Mining, Natural language processing

Светлана Петрасова *– аспирант кафедры интеллектуальных компьютерных систем Национального технического университета «Харьковский политехнический институт», ул. Фрунзе, 21, Харьков, 61002, Украина e-mail: svetapetrasova@gmail.com*

Научные интересы: искусственный интеллект, интеллектуальные системы представления знаний, компьютерная лингвистика, Natural language processing.

The mathematical model of the identification of the semantic relations of tolerance and equivalence

Svetlana Petrasova, Nina Khairova

Abstract*: The formalization of semantics is a difficult task of automatic text processing by reason of its implicit expression in natural language constructions. The paper proposes the mathematical model of the identification of semantic relations of tolerance and equivalence on the basis of glossary knowledge. The semantic relations of glossary terms are considered in order to identify concepts and their relations. The mathematical model of the identification of semantic relations allows extracting classes of synonymous terms in one of their conceptual meanings by factoring the space of concepts. A range of values of a semantic similarity coefficient is used to formalize categories of relations between concepts. As a result the heuristic effectiveness evaluation of the model of the identification of semantic correlations between concepts is defined.*

Keywords*: automatic text processing, natural language construction, identification of semantic relations, tolerance, equivalence, glossary, formalization of relations between concepts, semantic similarity.*