УДК 004.822

С. В. ПЕТРАСОВА, аспирант НТУ «ХПИ»;
З. А. КОЧУЕВА, ст. препод. НТУ «ХПИ»;
Н. Ф. ХАЙРОВА, канд. техн. наук, доц. НТУ «ХПИ»

МЕТОД АВТОМАТИЧЕСКОЙ ЭКСТРАКЦИИ ПАРАДИГМАТИЧЕСКИХ ОТНОШЕНИЙ МЕЖДУ ПОНЯТИЯМИ ТОЛКОВОГО СЛОВАРЯ

В статье рассматривается метод автоматического выявления семантических парадигматических отношений между концептами толкового словаря. Для выявления межконцептуальных отношений принадлежности к классу, гиперонимии, гипонимии и меронимии использованы шаблоны лексических последовательностей; для определения семантической эквивалентности вычисляется мера смысловой близости. Описана разработанная программная реализация метода и приводятся экспериментально определенные показатели качества работы.

Ключевые слова: семантические отношения, межконцептуальные отношения, смысловая близость, толковый словарь.

Введение. В настоящее время, в связи со стремительным ростом объемов информационных ресурсов, все большую актуальность приобретают задачи систематизации и обработки информации. Активно проводятся работы по извлечению данных из текстов и формальному выражению знаний, что приводит к возможности эффективного использования интеллектуальных информационных систем, основанных на знаниях.

Сегодня наиболее перспективным способом формального выражения знаний является семантическая сеть. Это обусловлено, прежде всего, наглядностью представления знаний и возможностью явного выражения семантических отношений между понятиями [1]. Семантические сети используются как приложение лингвистического процессора в информационно-поисковых системах, а также в других системах обработки естественного языка для расширения электронных тезаурусов и онтологий.

Но, несмотря на достаточно широкое использование данного способа представления знаний, его распространение сдерживается как неоднозначностью выражения знаний на естественном языке, так и трудоемкостью и сложностью разработки семантических сетей.

Существующие семантические сети (например, ConceptNet, Lexipedia, WiSeNet и др.) в своем подавляющем большинстве разрабатываются экспертами той или иной предметной области, что требует значительных временных и интеллектуальных затрат. Основная проблема, не позволяющая до настоящего времени автоматизировать разработку семантической сети, заключается в сложности формализации семантических отношений между понятиями.

Сегодня для решения задачи выявления семантических отношений в слабоструктурированной текстовой информации используют:

терминологические шаблоны, индикаторы связи и профили кластеризуемости [2]; шаблоны между каждой парой объектов в сегменте текста [3]; машины опорных векторов, оснащенные языковыми ориентированными ядрами для классификации пар объектов [4]; условные случайные поля [5] и др.

Постановка задачи исследования. В работе предлагается в качестве источника выявления парадигматических семантических отношений между понятиями использовать толковый словарь. Толковые словари на сегодня являются наиболее концентрированными средствами представления и накопления знаний, выраженных в форме текста.

Поскольку словарные статьи глоссария (или толкового словаря) достаточно часто содержат отношения концептов, выраженные в явном виде, то в этом случае их можно определять с помощью шаблонов лексических последовательностей. Однако отношения семантической эквивалентности не всегда представлены в явном виде, поэтому в этом случае необходимо применять методы, основанные на измерении смысловой близости лингвистических единиц.

Цель исследования. Целью данной работы является разработка метода формализации семантических парадигматических отношений между понятиями глоссария. Использование данного метода позволяет автоматизировать построение элементов семантической сети предметной области за счет автоматической обработки слабоструктурированной текстовой информации.

Метод построения семантической сети. Предлагаемый метод автоматизированного построения семантической сети заключается в пошаговом анализе текста. На начальном этапе анализа происходит извлечение концептов, т.е. выявление ключевых слов, словосочетаний и их группировка. Процесс группировки включает:

- нормализацию (приведение каждого слова к его нормальной форме – лемме);
- фильтрацию на основе лингвистического обеспечения (устранение стоп-слов, имен собственных, цифр и т.п.);
- ранжирование с использованием статистической информации.

В данном исследовании предлагается метод автоматизированного построения фрагмента семантической сети на базе использования знаний толкового словаря английского языка. В данном информационном ресурсе ключевыми словами и словосочетаниями являются термины словаря, которые и будут представлять собой концепты семантической сети.

Вторым необходимым этапом построения семантической сети является определение межконцептуальных связей. В разрабатываемой сети выделяемыми отношениями концептов являются: семантическая эквивалентность, принадлежность к классу, отношения гиперонимии, гипонимии и меронимии.

Одним из наиболее сложных для формализации отношений между концептами является отношение семантической (смысловой) эквивалентности. Проведенный анализ существующих статистических методов выявления семантических эквивалентов (метод формирования многомерных векторных представлений слов [6], латентно-семантический анализ [7], метод Клайнберга [8]), а также семантических методов, основанных на использовании онтологий, показал недостаточную полноту и точность получаемых результатов.

В исследовании предлагается использовать метод расстояний, с помощью которого определяется мера семантической близости между семантическими эквивалентами. При этом под семантическими эквивалентами мы понимаем текстовые выражения, сопоставленные одному и тому же понятию, а также слова и словосочетания с близким значением, встречающиеся в одном контексте. Метод расстояний, комбинируя статистическую и семантическую составляющие, позволяет дать количественную оценку смысловой близости между терминами толкового словаря. Данная величина формально определяема дефинициями глоссариев как отношение мощностей множеств, образованных теоретико-множественным пересечением и объединением множеств терминов дефиниций [9]:

$$f(t',t'') = \frac{2 \times N(x_1 \bigcap x_2)}{N(x_1 \bigcup x_2)}, \quad (1)$$

где $f(t',t'')$ – величина семантической близости между концептами t’ и t”;

x_1, x_2 – дефиниции лингвистических единиц толкового словаря t’ и t”;

$N(x_1 \bigcap x_2)$ – количество общих слов в определениях концептов t’ и t”;

$N(x_1 \bigcup x_2)$ – количество всех слов в определениях концептов t’ и t”.

Например, для вычисления меры семантической близости между терминами t’ = “computer language” и t” = “interpreted language” определим количество общих слов в соответствующих дефинициях:

x_1 = an artificial language that specifies instructions to be executed on a computer;

x_2 = an artificial language in which instructions are translated into executable form.

Определенная согласно формуле (1) мера семантической близости $f(t',t'')$ концептов t’ = “computer language” и t” = “interpreted language” составляет 67%. Данный показатель находится в диапазоне 30-100%, предварительно определенный экспериментальным путем как допустимый диапазон семантической эквивалентности, т.е. рассматриваемые концепты “computer language” и “interpreted language” будут определены как семантические эквиваленты.

Для формализации таких межконцептуальных отношений, как классификация, гиперонимия, гипонимия и меронимия, применяются шаблоны лексических последовательностей:

$$NN_1 \rightarrow \mathrm{Re}l_z \rightarrow NN_2,$$

где NN_1 и NN_2 – связанные концепты, представленные ключевыми словами и словосочетаниями толкового словаря, $\mathrm{Re}l$ – лексические цепочки, выражающие отношения z:

z=отношение классификации, Rel={“meaning of”, “identification of”, “is a”}.

z=отношение гиперонимии, Rel={“group of”, “class of”, “set of”, “list of”, “collection of”}.

z=отношение гипонимии, Rel={“branch of”, “type of”, “version of”, “study of”}.

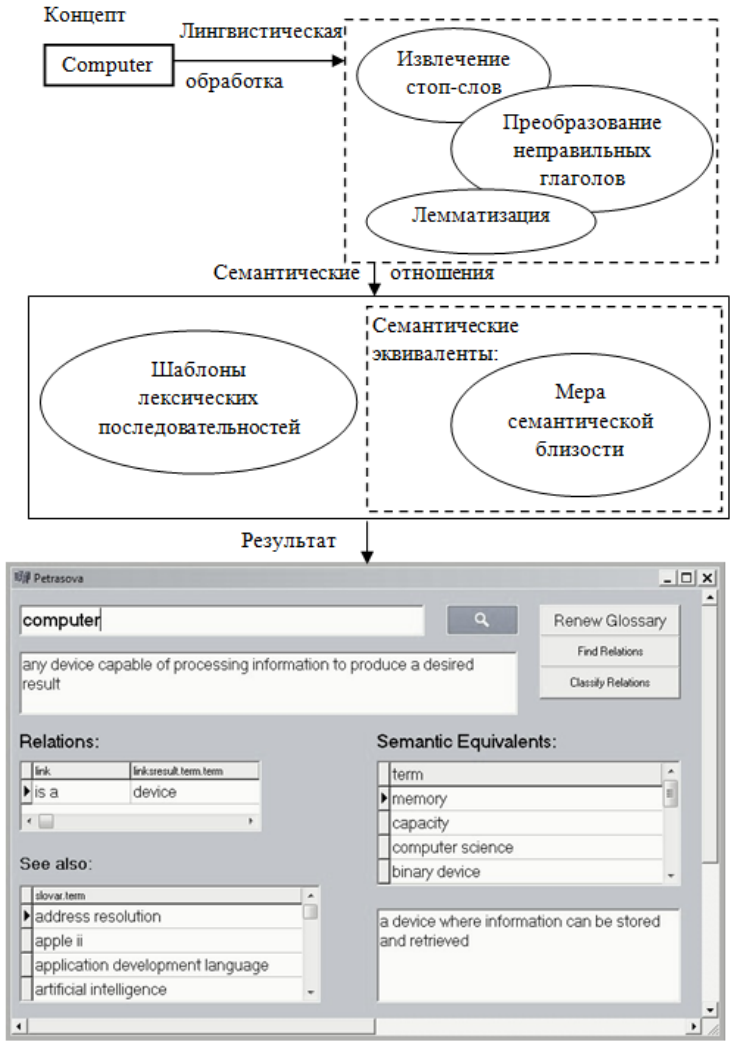
z=отношение меронимии, Rel={“part of”, “component of”, “subset of”, “member of”}.

Программная реализация предложенного метода. Все вышеприведенные этапы построения семантической сети программно реализованы в приложении, позволяющем выявлять парадигматические отношения между концептами толкового словаря Microsoft Computer Dictionary [10].

На первом этапе проводится лингвистическая обработка лексем в дефинициях концептов глоссария, заключающаяся в лемматизации, удалении стоп-слов, преобразовании неправильных глаголов. На следующем этапе происходит поиск и классификация семантических отношений между концептами толкового словаря с помощью шаблонов лексических последовательностей.

После проведенной обработки словарных статей глоссария для каждого запрашиваемого пользователем концепта вычисляется мера семантической близости с остальными концептами, в соответствии с которой определяется отношение семантической эквивалентности между рассматриваемыми концептами (см. рисунок).

Экспериментальные результаты. Для оценки качества созданной подсистемы автоматической экстракции парадигматических отношений между понятиями толкового словаря использовались метрики, характеризующие деятельность пользователей до и после внедрения данного метода.



Информационно-лингвистическое обеспечение автоматической экстракции парадигматических отношений между концептами толкового словаря

Для определения коэффициентов точности – precision и полноты – recall вычисляем отношение правильно определенных системой связей выбранного концепта (термина словаря) – n_{yy} к общему количеству определенных связей – $n_{yy}+n_{yn}$ (2) и количество верно определенных связей – n_{yy} к общему числу интеллектуально определенных связей данного концепта – $n_{yy}+n_{ny}$ (3):

$$precision = \frac{n_{yy}}{n_{yy} + n_{yn}}, \quad (2)$$

$$recall = \frac{n_{yy}}{n_{yy} + n_{ny}}. \quad (3)$$

Для определения качества разработанного программного обеспечения исследовалась выборка из ста терминов-запросов. В результате эксперимента было определено 130 межконцептуальных отношений, агрегируемых в отношения принадлежность к классу, гиперонимии, гипонимии и меронимии, и 706 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=0,9201. Коэффициент шума, определяющийся отношением числа неправильно определенных системой связей к общему числу связей, выданных системой, error=0,0771.

Данные показатели являются начальным результатом и в дальнейшей разработке должны быть учтены для повышения качества работы системы, осуществляющей автоматизированное построение элементов семантической сети.

Выводы. Неоднозначность толкования и представления естественного языка является характерной особенностью текстовых ресурсов, не позволяющей однозначно формализовать извлечение концептов и межконцептуальных отношений из текстов. В работе рассмотрен метод автоматической экстракции парадигматических отношений между понятиями, основанный на использовании толкового словаря как естественно-языкового текста, наиболее полно концентрирующего знания.

Для выявления межконцептуальных отношений принадлежности к классу, гиперонимии, гипонимии и меронимии использовались шаблоны лексических последовательностей; а для определения семантической эквивалентности вычислялась мера смысловой близости лингвистических единиц глоссария.

Проведенный эксперимент показал приемлемость значений коэффициентов качества работы программной реализации данного метода. Разработанная программа может быть применена в качестве составляющей системы автоматизированного построения элементов семантической сети.

Список литературы: **1.** *Маннинг К. Д.* Введение в информационный поиск / *К. Д. Маннинг, П. Рагхаван, Х. Шютце;* пер. с англ. – М. : ООО "И.Д.Вильямс", 2011. – 528 с. **2.** *Саломатина Н. В.* О возможностях автоматизации выявления связей между терминами предметной области (на примере катализа) / *Н. В. Саломатина, В. Д. Гусев, Л. Ю. Ильина [и др.]* // Труды междунар. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог-2010). – М. : Наука, 2010. – С. 430–436. **3.** *Hasegawa T.* Discovering relations among named

entities from large corpora / *T. Hasegawa, S. Sekine, R. Grishman* // In Proc. of ACL, 2004. **4.** *Bunescu R.* Learning to extract relations from the web using minimal supervision / *R. Bunescu, R. Mooney* // In Proc. of ACL, 2007. **5.** *Culotta A.* Integrating probabilistic extraction models and data mining to discover relations and patterns in text / *A. Culotta, A. McCallum, J. Betz* // In Procs of HLT/NAACL, 2006. – p. 296–303. **6.** *Мисуно И.С.* Векторные и распределенные представления, отражающие меру семантической связи слов / *И. С. Мисуно, Д. А. Рачковский, С. В. Слипченко* // Математические машины и системы. – 2005. – № 3. – С. 50–66. **7.** *Митрофанова О. А.* Семантические расстояния: проблемы и перспективы / О. А. Митрофанова // Материалы XXXIV междунар. филолог. конф. Вып. 21. Секция прикладной и математической лингвистики. – СПбГУ, 2005. **8.** *Kleinberg J.* Authoritative sources in a hyperlinked environment / J. Kleinberg // Proc. 9th ACM-SIAM Symposium on Discrete Algorithms, 1999. **9.** *Федорченко Л. А.* Метод автоматизированного построения семантической сети терминов учебной дисциплины / *Л. А. Федорченко, Н. Ф. Хайрова, А. И. Довнарь [и др.]* // Радіоелектронні і комп'ютерні системи. – 2011. – № 4 (52). – С. 1–7. **10.** Microsoft Computer Dictionary / *[edited by Alex Blanton]*. – 5-th Ed. – Microsoft Press, 2012. – 656 p.

Надійшла до редколегії 11.11.2013

УДК 004.822

Метод автоматической экстракции парадигматических отношений между понятиями толкового словаря / С. В. Петрасова, З. А. Кочуева, Н. Ф. Хайрова // Вісник НТУ «ХПІ». Серія: Системний аналіз, управління та інформаційні технології. – Х. : НТУ «ХПІ», 2013. – № 62 (1035). – С. 118. – 124. – Бібліогр.: 10 назв.

У статті розглядається метод автоматичного виявлення семантичних парадигматичних відношень між концептами тлумачного словника. Для виявлення міжконцептуальних відношень приналежності до класу, гіперонімії, гіпонімії та меронімії використані шаблони лексичних послідовностей; для визначення семантичної еквівалентності обчислюється міра смислової близькості. Описана розроблена програмна реалізація методу і наводяться експериментально визначені показники якості роботи.

Ключові слова: семантичні відношення, міжконцептуальні відношення, смислова близькість, тлумачний словник.

In this article, the method of the automatic identification of semantic paradigmatic relations between concepts of the glossary has been considered. Patterns of lexical sets have been used for the identification of concepts' relations of belonging to the class, hypernymy, hyponymy and myronymy. A measure of the semantic proximity has been determined for the identification of the semantic equivalence. The developed software implementation has been described and the quality indexes experimentally determined have been produced.

Keywords: semantic relations, concepts' relations, semantic proximity, glossary.