УДК 004.82
DOI

Д.В. Кравцов, Е.А. Леонов

МОДЕЛЬ ЛИНГВИСТИЧЕСКОЙ ОНТОЛОГИИ С НЕЧЕТКИМИ СЕМАНТИЧЕСКИМИ ОТНОШЕНИЯМИ, ГЕНЕРИРУЕМОЙ НА ОСНОВЕ ВИКИПЕДИИ

Обоснована актуальность автоматизации создания лингвистических онтологий и выбор для этой цели Википедии в качестве источника информации. Предложена математическая модель Википедии и лингвистической онтологии, использующая парадигму нечетких семантических отношений между понятиями. Сделан краткий обзор мер семантической близости понятий с учетом их вычислительной сложности, обоснован выбор взвешенной меры Дайса.

Ключевые слова: лингвистическая онтология, лексическая онтология, автоматическое построение онтологий, ontology learning, Википедия, нечеткие семантические отношения, семантическая близость.

D.V. Kravtsov, E.A. Leonov

MODEL OF LINGUISTIC ONTOLOGY WITH FUZZY SEMANTIC RELATIONS GENERATED ON BASIS OF WIKIPEDIA

The application without knowledge of an ontological type allows updating considerably quality of problem solutions in natural language processing. A number of researchers use Wikipedia as a basis for the formation of such resources. This paper reports the formalization method of Wikipedia structures and linguistic ontology used in the developed by the authors system of the linguistic ontology formation a specified subject field from Wikipedia. The papers and references connecting them serve a purpose for formation of a weighted graph of ontology to the graph nodes correspond notions, and to the ribs of graph – fuzzy semantic relations between them. The references obtain different weights depending on entering this or that information unit on a page. By a graph of relations it is possible to estimate numerically the degree of semantic proximity of two arbitrary concepts. For this purpose it is possible to use different measures of semantic proximity. Recursive measures possess considerable computational complexity at insignificant improvement of quality in test problem solution in comparison with non-recursive local measures of the Dice measure type that is unacceptable for the ontology large enough. From these considerations the Dice weighted measure is chosen as a basic one for the system under development.

Key words: linguistic ontology, lexical ontology, automated formation of ontology, ontology learning, Wikipedia, fuzzy semantic relations, semantic proximity.

С развитием технологий Интернета, ростом объема создаваемого контента, наблюдаемым в последние годы, растет и потребность в эффективных и в то же время более интеллектуальных методах обработки и анализа естественного языка. Наряду с традиционными статистическими методами, основанными на машинном обучении, все больше применяются подходы, использующие явные знания о предметных областях, представленные в виде формализованных структур, таких как онтологии. В частности, практика многих зарубежных [1; 2], а также отечественных [3; 6] ученых показывает, что так называемые лингвистические (лексические) онтологии (ЛО) могут весьма успешно применяться в самых разных задачах информационного поиска и обработки естественного языка. Например, в диссертации [1] приведен пример классификации текстов без обучения, чисто на основе онтологических знаний. Неформальное определение ЛО предметной области дается в работе [4]: «…это база знаний онтологического типа о понятийной системе и лексико-терминологическом составе предметной области».

Многие исследователи ищут способы автоматизации построения онтологий, так как создавать их вручную с нуля очень трудозатратно и долго. Этому посвящена отдельная область исследований в онтологическом инжиниринге, именуемая в английском языке «ontology learning». Обычно в области ontology learning в качестве источника для

автоматизированного построения онтологий рассматривают коллекции неструктурированных текстов. В такой постановке построение онтологии является комплексной проблемой, состоящей из ряда весьма нетривиальных подзадач, таких как извлечение терминов, извлечение синонимов, формирование понятий, построение иерархии понятий, выявление произвольных отношений между понятиями и др. (см. «ontology learning layer cake» [5]). В то же время все больше исследователей используют для этой цели Википедию – открытый источник информации, представленной в частично структурированном виде, по широкому перечню предметных областей и общих знаний. Википедия, как показано ниже, относительно легко преобразуется к онтологическому представлению. Помимо этого к преимуществам Википедии стоит отнести постоянную пополняемость и актуализацию большим сообществом волонтеров, мультиязычность (наличие связей между одинаковыми понятиями на разных языках), бесплатность использования (в том числе в коммерческих целях), свободную доступность в виде дампов базы данных.

Из наиболее известных проектов, в которых для построения баз знаний онтологического типа использовалась Википедия, можно выделить проекты DBpedia, YAGO, Texterra [6]. Первые два являются относительно высокоформализованными и, по-видимому, больше ориентированы на использование различными интеллектуальными агентами в рамках концепций Linked open data, Semantic Web, чем на автоматическую обработку текстов. Проект Texterra – разработка российских ученых из Института системного программирования РАН – ориентирован как раз на анализ текстов и основан на численной оценке семантической близости понятий с использованием (как и у многих других исследователей) графа ссылок Википедии. Таким образом, подход к автоматизированному построению лингвистических онтологий на базе информации, извлекаемой из Википедии, является перспективным и востребованным.

Для того чтобы разработать эффективный алгоритм и программную реализацию для построения ЛО на базе Википедии, требуется формализация этих понятий. С точки зрения поставленных целей Википедию можно представить в виде кортежа

$$W = < P, L, R, A>.$$

Здесь P – множество страниц, в котором каждая страница $p_i = <t, k, B>$, где t – заголовок страницы, k – тип страницы ($k \in$ {обычная, страница-перенаправление, страница-категория, страница-дизамбигуация}), $B=\{(b_i, s_i, w_i)\}$ – контент страницы, представленный в виде совокупности информационных блоков b_i , их весов w_i и их типов $s_i \in$ {шапка страницы, раздел «История», инфоблок, основной текст, текст ссылок типа «Основная статья», блок «См. также», навигационный шаблон}; L – множество ссылок между страницами; $R \subset (P \times P) \times L$ – отношение, задающее связь конкретного экземпляра ссылки с парой страниц; $A: L \rightarrow B$ – алгоритм (функция), который выделяет на странице блоки и ставит им в соответствие ссылки, находящиеся в блоке. $\mu (p_i, p_j,) \rightarrow [0, 1]$ – весовая функция, рассчитывающая для пары страниц вес ссылочной связи от i-й страницы к j-й. Этот вес передается на вход модели лингвистической онтологии в качестве степени принадлежности нечеткого отношения $r_{i,j}$.

Вес блока, определяемый его типом, распространяется на его ссылки. Вес $\mu(A, B)$ вычисляется на данный момент как простая сумма весов всех ссылок из A в B. В дальнейшем, после проведения тестов, возможно введение некоторых нормировочных коэффициентов (например, деление на количество ссылок, логарифмирование).

При разработке математической модели ЛО авторы исходили из двух соображений: 1) ее функционала должно быть достаточно для использования в задачах автоматической обработки

текстов; 2) она будет создаваться преимущественно автоматическими методами из информации, которую можно получить из Википедии.

Предлагаемую модель ЛО в общем виде можно представить следующим набором:

$$ЛО = < C, T, L, M, D, R, A >.$$

Ниже дано описание всех составляющих его элементов.

C – множество понятий (концептов), основных единиц онтологии. Каждой статье Википедии (кроме некоторых видов страниц, например страницы-перенаправления и др.) соответствует понятие в ЛО. В предлагаемой модели онтологии не делается различий между понятиями и их экземплярами (instances), которые также рассматриваются как понятия.

T – множество терминов (лексикон) онтологии, которыми понятия могут выражаться в текстах. Терминам соответствуют названия статей (основные и названия страниц-перенаправлений), текст гиперссылок из других статей (с определенными оговорками).

M – отношение, задающее связь терминов с понятиями (значениями терминов): $M \subset T \times C$ или $M: T \times C \rightarrow \{0, 1\}$. Одному понятию могут соответствовать несколько терминов (синонимы, квазисинонимы), в то же время один термин может быть связан с несколькими понятиями (многозначность). При обработке текстов с использованием ЛО необходимо определять нужное значение многозначного термина, для чего разработаны соответствующие методы разрешения лексической многозначности.

D – подмножество дескрипторов $(D \subset T)$, т. е. терминов, которые являются предпочтительными для понятия и однозначно идентифицируют. Каждому понятию сопоставлен один дескриптор, т.е. отношение *M* задает биекцию – $M: D \leftrightarrow C$.

R – набор отношений нескольких типов между понятиями. Отношения планируется строить в автоматическом режиме на основе различных типов ссылок Википедии. Таким способом мы можем выделить два типа отношений: иерархическое, которое строится на основе ссылок на иерархическую систему категорий Википедии, и ассоциативное, которое строится на основе всех остальных ссылок. Необходимо заметить, что система категорий Википедии является не формальной таксономией, построенной строго на отношении «род - вид», а смешением различных отношений, в том числе «часть-целое». Отношения, построенные таким образом, не обладают формальной строгостью: можно говорить лишь о вероятности наличия отношения, степени его выполнимости на паре понятий или силе семантической связи. Такую характеристику можно выразить величиной в интервале [0, 1]. Таким образом, разумным представляется использование математического аппарата теории нечетких отношений для построения модели нечеткой лингвистической онтологии.

A – набор аксиом онтологии, т. е. правил нечеткого логического вывода, позволяющих распространять нечеткие отношения на понятия, для которых они не заданы явно. В качестве аксиом используются свойства транзитивности и наследования отношений.

Определим нечеткое отношение между понятиями c_i и c_j, принадлежащими C, как функцию, ставящую в соответствие каждой паре понятий степень их принадлежности этому отношению, т. е. $R: C \times C \rightarrow [0, 1]$ или $R(c_i, c_j) \in [0, 1]$, что кратко можно записать как $r_{i,j}$. Такому определению можно поставить в соответствие взвешенный ориентированный граф, вершинам которого соответствуют понятия, ребрам – отношения, весам ребер – значения функции принадлежности. Назовем его *графом отношений* онтологии. Для логического вывода используется свойство транзитивности отношений, которое для нечетких бинарных отношений обычно определяется следующим образом (сильная транзитивность) [9]:

$$R(x, z) \geq \min (R(x, y), R(y, z)) \quad \forall x, y, z \in X .$$

Но интуитивно понятно, что по мере удаления от заданного понятия по графу отношений семантическая связь понятий, т.е. степень принадлежности, должна уменьшаться. Поэтому мы будем далее использовать слабую транзитивность, условие которой для нашей модели можно определить так:

$$R(c_i, c_j) \wedge R(c_j, c_k) \Rightarrow R(c_i, c_k) > 0.$$

Правило нечеткого логического вывода для транзитивных отношений (аксиома транзитивности):

$$A_{tr} \in A: r_{i,j} \wedge r_{j,k} \Rightarrow r_{i,k} = t(r_{i,j}, r_{j,k}),$$

где t – функция транзитивности (в самом простом варианте это произведение степеней принадлежности).

Тогда правило нечеткого логического вывода для отношений, обладающих свойством наследования (аксиома наследования), можно записать как

$$A_{in} \in A:$$
$$r_{i,j} \wedge r`_{j,k} \Rightarrow r`_{i,k} = i(r_{i,j}, r`_{j,k}) = r_{i,j}\, r`_{j,k},.$$

Нечеткие иерархические отношения антирефлексивны, асимметричны и транзитивны. Нечеткие отношения ассоциации мы рассматриваем как несимметричные и транзитивные. Зная эти свойства отношений и применяя к ним правила вывода, можно извлекать заданные подмножества понятий, например полное поддерево иерархии вниз для некоторого понятия (частные понятия) или отранжированный список ассоциативно связанных понятий со степенью принадлежности не менее заданной.

Весьма полезным свойством графа отношений ЛО является возможность численно оценить степень смысловой связанности двух произвольных понятий. Для этого введем понятие *нечеткого отношения (функции) семантической близости* (СБ, semantic relatedness) понятий *rel*:

$$R(c_i, c_j) \vee R(c_j, c_i) \Rightarrow rel(c_i, c_j), \quad \forall c_i, c_j \in C.$$

Функцию принадлежности нечеткого отношения СБ определим как

$$rel(c_i, c_j), = max(R(c_i, c_j), R(c_j, c_i)).$$

Отношение семантической близости рефлексивно (причем семантическая близость понятия с самим собой равна 1), симметрично и транзитивно, т.е. является отношением нечеткой эквивалентности. Взятое отдельно от других отношений, отношение семантической близости преобразует исходный ориентированный граф отношений в неориентированный, и если у пары вершин было более одного ребра, то оставляется только ребро с наибольшим весом.

Вычисление функции семантической близости по графу отношений является нетривиальной задачей. Если онтология ограниченной предметной области достаточно маленькая, то может оказаться возможным предварительный расчет СБ для каждой пары понятий. Но для достаточно больших онтологий (сотни тысяч понятий) такой подход потребует слишком много памяти и времени. В то же время если вычислять семантическую близость на лету, то расчет должен выполняться за минимальное время, так как многие задачи, например поиск по запросу, критичны ко времени отклика. В этих условиях актуален вопрос о выборе рациональной меры семантической близости, рассчитываемой на основе графа (в частности взвешенного). Хороший обзор и классификация таких мер сделаны в работе [10]. Их можно разделить на три основные группы:

- меры парного случайного блуждания (SimRank, мера близости Ньюмана);
- меры случайного блуждания (мера Грина, локальный PageRank, PageSim и др.);
- нерекурсивные меры (косинус, меры Дайса, Жаккара, Кульчинского и др.).

Популярная рекурсивная мера парного случайного блуждания SimRank вычисляется по следующей итерационной формуле:

$$S_{ij} = \frac{C}{k_i k_j} \sum_{uv} A_{iu} A_{vj} S_{uv}$$

где S_{ij} – элемент матрицы подобия вершин; A_{ij} – элемент матрицы смежности; k_i – степень i-й вершины; C – коэффициент затухания.

Вычислительная сложность этой меры очень высока – $O(n^3)$, где n – количество ребер графа. Из-за очень маленького диаметра графа Википедии обе меры, SimRank и мера Ньюмана, вычисляют полную матрицу семантической близости, а потому практически невычислимы [10]. Меры случайного блуждания в плане вычислительной сложности существенно превосходят меры парного случайного блуждания (сложность $O(n)$), но все же в больших онтологиях могут оказаться недостаточно эффективными.

Среди традиционных нерекурсивных мер интерес вызывает мера Дайса:

$$Dice(a,b) = \frac{|N(a) \cap N(b)|}{|N(a)|+|N(b)|}$$

где $N(a)$ – множество вершин, соседних с вершиной a.

Несмотря на простую интерпретацию (отношение количества общих соседей к сумме количеств соседей каждой из вершин), согласно экспериментальным данным [7], мера Дайса показывает очень хорошие результаты. Так, при решении задачи разрешения лексической многозначности по методу системы Texterra мера Дайса показывает самые лучшие результаты на всех четырех использовавшихся тестовых наборах данных по сравнению с различными вариациями мер на основе поиска кратчайших путей [8].

В графе отношений нечеткой ЛО семантическая близость пары понятий a и b рассчитывается как взвешенная мера Дайса:

$$Dice(a,b) = \frac{\sum_{i \in N(a) \cap N(b)} (w_{a,i} + w_{b,i})}{(\sum_{j \in N(a)} w_{a,j} + \sum_{k \in N(b)} w_{b,k})}$$

где $w_{a,i}$ – вес ребра между вершинами a и i.

Итак, для расчета семантической близости пары понятий в том случае, если у них есть общие вершины в графе отношений, предполагается использовать взвешенную меру Дайса; если общих вершин нет – одну из мер случайного блуждания, например модификацию меры Грина, предложенную в работе [10], которая показала наилучшие результаты на данных, полученных ручным ранжированием статей Википедии.

Были рассмотрены подходы к построению базы знаний онтологического типа на основе Википедии для применения в автоматической обработке текстов. Дано формализованное представление основных структурных элементов Википедии, используемых в этом процессе. Разработана математическая модель создаваемой лингвистической онтологии, использующая концепцию нечетких семантических отношений между понятиями. Проведен обзор мер (алгоритмов), применяемых для вычисления семантической близости понятий по графу отношений онтологии, в качестве основной выбрана взвешенная мера Дайса.

На основе описанных формализаций ведется разработка программных модулей парсинга дампов Википедии и хранилища онтологии. Предполагается, что лингвистический ресурс, построенный на базе предложенной модели, будет достаточно универсален и сможет использоваться в массе разнообразных задач, таких как смысловое расширение (сужение, дополнение) поисковых запросов, фасетная навигация при поиске, выделение терминов предметной области из текста, разрешение лексической многозначности, построение семантической структуры текстов для улучшения качества их автоматической обработки: поиска, классификации, аннотирования и т. д.

Работа выполнена при поддержке РФФИ (проект № 14-07-31261 мол_а).

СПИСОК ЛИТЕРАТУРЫ

1. Janik, M. Training-less ontology-based text categorization : PhD diss. / Maciej Janik. – University of Georgia, 2008. – 150 p.
2. Syed, Z. S. Wikipedia as an Ontology for Describing Documents / Z. S. Syed, T. Finin, A. Joshi // Proceedings of the Second International

Conference on Weblogs and Social Media. – 2008. – P. 136-144.
3. Добров, Б. В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска / Б. В. Добров, Н. В. Лукашевич //10-я Нац. конф. по искусств. интеллекту с междунар. участием. – 2006. – С. 489-497.
4. Лукашевич, Н. В. Модели и методы автоматической обработки неструктурированной информации на основе базы знаний онтологического типа : дис…. д-ра техн. наук / Н. В. Лукашевич. – М., 2014. – 312 с.
5. Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications / Philipp Cimiano. – Springer US, 2006.
6. Турдаков, Д. Ю. Texterra: инфраструктура для анализа текстов / Д. Ю. Турдаков [и др.] // Труды Института системного программирования РАН. – 2014. – Т. 26. – № 1. – С. 421-440.
7. Turdakov, D. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation / D. Turdakov, P. Velikhov // In proceedings of the SYRCoDIS'2008. – 2008.
8. Варламов, М. И. Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии / М. И. Варламов, А. В. Коршунов // Труды конференции ИОИ-2014: Интеллектуализация обработки информации (5-10 окт. 2014 г., Греция). – 2014. – С. 1107-1125.
9. Нечеткие множества в моделях управления и искусственного интеллекта / под ред. Д. А. Поспелова. – М.: Наука, Гл. ред. физ.-мат. лит., 1986. – 312 с.
10. Велихов, П. Е. Меры семантической близости статей Википедии и их применение к обработке текстов / П. Е. Велихов // Информационные технологии и вычислительные системы. – 2009. – №. 1. – С. 23-37.

1. Janik, M. Training-less ontology-based text categorization : PhD diss. / Maciej Janik. – University of Georgia, 2008. – 150 p.
2. Syed, Z. S. Wikipedia as an Ontology for Describing Documents / Z. S. Syed, T. Finin, A. Joshi // Proceedings of the Second International Conference on Weblogs and Social Media. – 2008. – P. 136-144.
3. Dobrov, B.V. Linguistic ontology on natural sciences and techniques for applications in the field of information retrieval / B.V. Dobrov, N.V. Lukashevich // *The 10-th National Conf. on Artificial Intelligence with International Participation*. – 2006. – pp. 489-497.
4. Lukashevich, N.V. *Models and Methods for Automated Processing Non-Structured Information Based on Knowledge of Ontological Type*: Thesis for D.Eng. Degree / N.V. Lukashevich. – M., 2014. – pp. 312.
5. Cimiano, P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications / Philipp Cimiano. – Springer US, 2006.
6. Turdakov, D.Yu. Texterra: Infrastructure for TYext Analysis / D.Yu. Turdakov [et al.] // *Proceedings of the Institute of System Programming RA*S. – 2014. – Vol. 26. – № 1. – pp. 421-440.
7. Turdakov, D. Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation / D. Turdakov, P. Velikhov // In proceedings of the SYRCoDIS'2008. – 2008.
8. Varlamov, M.I. Computation of concept semantic proximity on basis of the shortest ways in references graph of Wikipedia / M.I. Varlamov, A.V. Korshunov // *Proceedings of the Conf. IOI-2014: Intellectualization of Information Processing* (October 5-10, 2014, Greece). – 2014. – pp. 1107-1125.
9. *Fuzzy Multitudes in Models of Control and Artificial Intelligence* / under the editorship of D.A. Pospelov. – M.: Science, General Editorship of Phys.-Math. Lit., 1986. – pp. 312.
10. Velikhov, P.E. Measures of semantic proximity of Wikipedia entries and their application at text processing / P.E. Velikhov // *Information Technologies and Computer Systems*. – 2009. – №. 1. – pp. 23-37.

Статья поступила в редакцию 13.11.2015.
Рецензент: д.т.н., профессор
Брянского государственного технического университета
Аверченков В. И.

Сведения об авторах:

Кравцов Дмитрий Викторович, программист Брянского государственного технического университета, e-mail: dkrbox@gmail.com.

Леонов Евгений Алексеевич, к.т.н., доцент кафедры «Компьютерные технологии и системы» Брянского государственного технического университета, e-mail: johnleonov@gmail.com.

Kravtsov Dmitry Victorovich, Programmer of Bryansk State Technical University, e-mail: dkrbox@gmail.com.

Leonov Evgeny Alexeyevich, Can.Eng., Assistant Prof. of the Dep. “Computer Techniques of System” Bryansk State Technical University, e-mail: johnleonov@gmail.com.