Расчет семантической близости концептов на основе кратчайших путей в графе ссылок Википедии

М. И. Варламов, А. В. Коршунов
{varlamov, korshunov}@ispras.ru
Институт системного программирования РАН, Россия, Москва 109004, ул. А. Солженицына, 25

В задачах автоматической обработки текстовой информации часто возникает необходимость определить, насколько сильно та или иная пара концептов (понятий) связана по смыслу, — иначе говоря, оценить степень семантической близости между ними. В данной работе исследуется применимость к вычислению семантической близости пары концептов расстояния между соответствующими им статьями в графе ссылок Википедии. При этом для оценки расстояния между вершинами в графе используется длина кратчайшего пути между ними. Предлагается ряд мер семантической близости, использующих расстояния по различным типам ссылок Википедии; выявляются типы ссылок, наиболее релевантные для данной задачи (внутритекстовые и категорийные). В сравнении с мерой Дайса, используемой в системе анализа текстов Текстерра, показывается, что использование кратчайших путей позволяет как повысить корреляцию получаемых оценок близости с экспертными, так и достичь лучших результатов в задаче разрешения лексической многозначности.

Ключевые слова: семантическая близость; кратчайшие пути; разрешение лексической многозначности; Википедия

Computing semantic similarity of concepts using shortest paths in Wikipedia link graph

M. I. Varlamov, A. V. Korshunov
Institute for System Programming of the Russian Academy of Sciences, 25 Alexander Solzhenitsyn Str., Moscow 109004, Russia

Background: A measure of semantic similarity between concepts characterizes the degree of relatedness between their senses. Texterra system uses Wikipedia-based Dice semantic similarity measure for word sense disambiguation. Since concepts in Texterra are Wikipedia articles, one is interested in precise link-based semantic similarity measures.
Methods: This work presents a global semantic similarity measure based on distances between concepts in Wikipedia link graph. Graph distance is estimated as the shortest path length between a pair of nodes (Wikipedia articles). The difference of the proposed method from existing measures based on shortest paths is in the usage of disparity of different link types. Here, a special data structure is used which allows one to compute the shortest pasts efficiently with acceptable memory costs.
Results: Compared to Dice measure, usage of shortest paths allows both to increase the correlation between computed and expert similarity and to achieve better results in the word sense disambiguation task. Also, it is demonstrated that regular and category links are the most relevant for semantic similarity estimation.
Concluding Remarks: This work shows that distances between articles in Wikipedia link graph can provide an effective basis for computing semantic similarity between corresponding concepts.

Keywords: semantic similarity; shortest paths; word sense disambiguation; Wikipedia

Введение

Современные приложения анализа текстовых документов переходят от рассмотрения отдельных слов и терминов непосредственно к работе с их значениями — понятиями или, как устоялось в современной лингвистической литературе, концептами. Ключевыми элементами методов обработки текстов становятся понимание и правильная интерпретация соотношений между концептами. Стремление воспроизвести способность человека к интуитивному проведению ассоциаций между понятиями привело к появлению задачи оценки семантической близости концептов — т.е. оценки степени их смысловой связанности.

Непосредственным приложением мер семантической близости концептов является задача разрешения лексической многозначности терминов. Многозначность — способность некоторых терминов принимать различные значения в зависимости от контекста — является неотъемлемой частью естественного языка. К приложениям мер семантической близости также относятся классификация и кластеризация текстов, уточнение поисковых запросов и другие.

Система анализа текстов Текстерра [1], разрабатываемая в Институте системного программирования РАН, предоставляет пользователю широкий набор средств для работы с текстовыми документами. Система использует базу знаний, извлекаемую из Википедии — свободной интернет-энциклопедии с широким покрытием различных предметных областей. В модели Текстерра концепты — статьи Википедии, а термины — названия статей и тексты гиперссылок на статьи. Сами гиперссылки образуют граф связей между концептами.

Граф ссылок Википедии представляет собой пример так называемой безмасштабной (scale-free) сети. Доля вершин со степенью k в безмасштабном графе пропорциональна величине $k^{-\gamma}$, где γ — константа, для большинства существующих сетей принадлежащая интервалу $(2, 3)$. В частности, такое распределение означает присутствие как вершин со степенями, значительно превышающими среднее по графу значение, так и длинных «хвостов» из вершин малой степени. Вершины первого рода — хабы — выполняют в сети связующую функцию. Из-за них диаметр графа — максимальное из расстояний между парами его вершин — очень мал и имеет порядок $O(\log \log n)$, где n — число вершин в графе.

Система Текстерра обладает возможностью выделения в тексте терминов и сопоставления им концептов-значений. Для этого осуществляется решение задачи разрешения лексической многозначности терминов с использованием меры семантической близости концептов. В качестве последней используется взвешенная мера Дайса: для пары концептов x и y близость между ними $sim_{\text{Дайс}}(x, y)$ рассчитывается как

$$sim_{\text{Дайс}}(x, y) = \frac{\sum\limits_{z \in N(x) \cap N(y)} [w(x, z) + w(z, y)]}{\sum\limits_{z \in N(x)} w(x, z) + \sum\limits_{z \in N(y)} w(z, y)}, \tag{1}$$

где $N(x)$ — множество соседей концепта x в графе ссылок Википедии (множество статей, связанных с x гиперссылкой), $w(x, z)$ — вес ссылки от x к z, зависящий от того, в какой части статьи встретилась гиперссылка и какие статьи она связывает.

Мера Дайса является локальной в том смысле, что использует для расчета семантической близости только первые окрестности вершин. Несомненным плюсом такой меры является низкая вычислительная сложность. При упорядоченном хранении ссылок для каждой вершины мера Дайса для пары концептов может быть вычислена за время $O(d)$, где d — средняя величина степени вершины в графе ссылок Википедии. Однако на прак-

тике оказывается, что среди всех пар концептов в базе знаний системы Текстерра (5 млн статей англоязычной Википедии) не более 4% имеют ненулевое количество общих соседей и, соответственно, ненулевое значение близости по мере Дайса, что заставляет сомневаться в применимости такой меры для оценки семантической близости между парой произвольных концептов.

Целью данной работы является проверка гипотезы, что семантическую близость между концептами можно эффективно оценивать на основе расстояния между ними в графе ссылок Википедии. Основными результатами работы являются:

— построение меры семантической близости между концептами, использующей значения расстояний между ними в графе Википедии по различным типам ссылок, которая дает лучшие оценки близости концептов по сравнению с мерой Дайса;
— получение знаний о том, какие типы ссылок наиболее информативны для вычисления семантической близости.

Дальнейшее изложение построено следующим образом. Следующий раздел содержит описание существующих методов расчета семантической близости с использованием Википедии. В разделе «Эффективный расчет длин кратчайших путей в графе» описан использованный метод вычисления расстояний между концептами в графе ссылок Википедии. Раздел «Экспертные данные» описывает набор пар концептов с экспертными оценками близости, который мы использовали для вывода и тестирования новой меры. В разделе «Зависимость семантической близости от типа ссылок» описываются исследуемые типы ссылок и приводятся характеристики расстояний между концептами, посчитанных с их использованием. В разделе «Вывод меры семантической близости» описан непосредственно процесс построения меры семантической близости концептов с использованием графовых расстояний между ними. В разделе «Анализ полученных мер близости и дополнительные эксперименты» приводится сравнение полученных мер с другими мерами семантической близости концептов на основе кратчайших путей в графе ссылок Википедии, рассматриваются различные отображения из функции расстояния в функцию близости, а также исследуется применимость разработанных мер к оценке семантической близости для пар концептов, удаленных друг от друга в графе ссылок Википедии. Раздел «Разрешение лексической многозначности» содержит результаты тестирования новой меры для решения задачи разрешения лексической многозначности. Наконец, раздел «Заключение» резюмирует полученные результаты.

Методы расчета семантической близости на основе Википедии

Существующие методы расчета семантической близости концептов Википедии можно условно разделить на две категории: контентные и ссылочные.

Контентные меры семантической близости основаны на использовании текстового содержимого Википедии. Как правило, такие меры применяются не к парам концептов, а к текстам соответствующих им статей, представляя их в виде векторов над пространством слов или терминов. Простейшим примером данного класса мер является косинусный коэффициент TF-IDF векторов статей, соответствующих концептам. Среди более сложных методов следует выделить латентный (LSA, [2]) и явный (ESA, [3]) семантический анализ. Первый использует сингулярное разложение матрицы TF-IDF весов терминов в статьях для построения нового факторного пространства и представляет концепты векторами в этом пространстве. Второй же проецирует тексты непосредственно на множество концептов Википедии.

Поскольку база знаний системы Текстерра не сохраняет текстов статей для извлекаемых из Википедии концептов, а внедрение соответствующих модификаций является достаточно трудоемким, контентные меры не представляет такого интереса для наших исследований, как ссылочные. Последние основаны на представлении концептов вершинами в графе гипертекстовых ссылок Википедии и подразделяются обычно на локальные и глобальные.

К локальным ссылочным мерам относится уже упоминавшаяся мера Дайса, а также ряд схожих мер: Жаккара [4], Кульчинского [4], Симпсона [4], Google Distance [5] и другие.

Глобальные меры, в большинстве своем, основываются на модели случайного блуждания в графе. Популярны методы, так или иначе интерпретирующие результаты работы персонализированного алгоритма PageRank [6].

Один из современных методов — WikiWalk [7] — является примером успешной комбинации контентного и глобального ссылочного подходов. Паре концептов сопоставляются ESA-вектора соответствующих им статей. Затем на каждом из них как на начальном распределении запускается PageRank. Результирующие распределения сравниваются с помощью косинусной меры для получения оценки близости.

Как правило, меры на основе случайного блуждания обеспечивают лучшее качество определения семантической близости концептов по сравнению с локальными, однако достигается это за счет существенно большей временной сложности, порядок которой в общем случае не лучше, чем $O(n)$, где n — число вершин в графе (число концептов Википедии). В работе Турдакова и Велихова [4] можно найти сравнение ряда упомянутых методов применительно к системе Текстерра: утверждается, что выбор меры Дайса позволяет достичь сравнимого с блуждающими мерами качества на задаче устранения лексической многозначности.

Известны примеры использования для построения мер близости длин кратчайших путей между концептами. В обзоре Цеша [9] рассматривается мера вида

$$\mathrm{sim}\,(x, y) = D - \mathrm{dist}\,(x, y)\,, \tag{2}$$

где $\mathrm{dist}\,(x, y)$ — расстояние между концептами x и y, D — диаметр графа. Также известна система WikiRelate! [8] и использованная в ней мера Ликока и Чодороу:

$$\mathrm{sim}\,(x, y) = -\log \frac{\mathrm{dist}\,(x, y)}{2D}, \tag{3}$$

рассчитываемая по категорийным ссылкам Википедии. В отличие от меры 2, здесь расстояние между концептами рассчитывается как число *узлов* на кратчайшем пути между ними в иерархии категорий Википедии, а D обозначает максимальную глубину иерархии.

В табл. 1 приводятся оценки корреляции значений семантической близости, полученных некоторыми из рассмотренных мер, с экспертными на стандартном наборе пар *терминов* WordSim-353 [11].

Отметим, что данные оценки были получены на различных по дате версиях Википедии с использованием различных методов сопоставления концептов терминам из набора, потому сравнение методов на их основе не в полной мере корректно. Тем не менее видно, что качество оценок, получаемых на основе кратчайших путей, далеко от оптимального. При этом в классификации выше эти меры попадают в категорию глобальных и имеют тот же порядок сложности. Однако в отличие от блуждающих методов эту сложность можно существенно понизить за счет предварительного вычисления части расстояний. Вкупе

Таблица 1. Существующие методы расчета семантической близости и их корреляция с экспертными оценками

Метод	Корреляция с экспертными оценками
мера Цеша	0,43 [9]
мера Ликока и Чодороу	0,48 [8]
ESA	0,75 [3]
LSA	0,56 [3]
WikiWalk	0,63 [7]

с наблюдением, что не все типы ссылок одинаково релевантны для расчета ссылочной близости (например, мера Ликока и Чодороу показывает лучшие результаты, чем мера Цеша, хотя первая основана только на ссылках в графе категорий), мы можем предположить осуществимость построения эффективной меры семантической близости на основе кратчайших путей в графе ссылок Википедии.

Эффективный расчет длин кратчайших путей в графе

Поскольку граф ссылок Википедии имеет довольно существенный размер — около 5 млн вершин для англоязычного сегмента — предварительный расчет и хранение значений семантической близости для *всех* пар вершин не представляется возможным. Действительно, даже если предположить, что для хранения каждого расстояния достаточно 1 байта, размер хранилища превысит 10 ТБ. Соответственно, для некоторых пар концептов придется осуществлять расчет близости в реальном времени. Данный факт значительно осложняет использование в Текстерре глобальных мер семантической близости.

Однако описанных проблем можно избежать при формировании меры семантической близости на основе расстояния между вершинами в графе. Рассмотрим структуру данных специального вида — *индекс графа с двухшаговым покрытием вершин* (2-hop cover).

Пусть $G =< V, E >$ — некоторый неориентированный граф с множеством вершин V и множеством ребер E. Назовем *меткой* $L(u)$ вершины $u \in V$ множество пар

$$L(u) = \{(w, \mathrm{dist}_G(u, w))\}_{w \in C(u)}, C(u) \subset V, \tag{4}$$

где $\mathrm{dist}_G(u, w)$ — расстояние между вершинами u, v в графе G.

Индекс графа — такое множество меток его вершин, что для любой пары вершин $u, v \in V$ обе их метки содержат расстояние хотя бы до одной вершины w на кратчайшем пути между ними. Если метки $L(u), L(v)$ обладают указанным свойством, расстояние между u и v может быть вычислено как

$$\mathrm{dist}_G(u, v) = \mathrm{dist}_G(u, w) + \mathrm{dist}_G(w, v) = \min_{\substack{(w, \delta_{uw}) \in L(u) \\ (w, \delta_{wv}) \in L(v)}} \{\delta_{uw} + \delta_{wv}\} \tag{5}$$

Для построения индексов Википедии мы воспользовались методом разметки вершин с отсечением (Pruned Landmark Labeling, [10]). Алгоритм хорошо масштабируем, применим к графам с сотнями миллионов ребер. Была использована открытая реализация алгоритма на языке Си++[1]. Необходимо отметить, что эта реализация применима только к

[1] `https://github.com/iwiwi/pruned-landmark-labeling`

неориентированным и невзвешенным графам, однако авторы [10] утверждают, что алгоритм может быть обобщен на эти случаи.

Экспертные данные

Несмотря на достаточное количество работ в данной области, нам не удалось найти готовых данных с экспертными оценками близости именно концептов. Исследователи, использовавшие Википедию для оценки семантической близости, ставили эксперименты не на парах концептов, но на парах терминов. Популярен уже упоминавшийся набор данных WordSim-353, содержащий 353 пары терминов с экспертными оценками семантической близости.

Мы вручную сопоставили терминам в наборе WordSim-353 идентификаторы концептов Википедии. Наиболее существенным препятствием к этому стали термины, не имеющие адекватных представлений среди концептов Википедии, в частности, обозначающие слишком отвлеченные понятия. Так, поиск значения термина «announcement» приводит к соответствующей статье в словаре Wiktionary, не использующемся в качестве источника знаний в системе Текстерра. Пары, в которых хотя бы одному из слов не удалось сопоставить статью Википедии, были удалены. Размер результирующего набора данных составил 308 пар концептов с экспертными значениями близости.

Зависимость семантической близости от типа ссылок

Система Текстерра различает пять основных типов ссылок между концептами Википедии (рис. 1).

Обычные (внутритесктовые) ссылки. К данному типу относятся гиперссылки, встречающиеся непосредственно в текстах статей. Наиболее многочисленный тип ссылок.

Категорийные ссылки. Википедия содержит специальные конструкции — категории — обеспечивающие возможность структуризации знаний в энциклопедии. Категории служат для объединения схожих по тематике статей. Каждой статье может быть назначена одна или несколько категорий. Сами категории представляют собой страницы особого вида, содержащие ссылки на статьи, относящиеся к данной категории, а также ссылки на страницы-подкатегории (таким образом, множество категорий Википедии имеет иерархическую структуру). В качестве ссылок категорийного типа в системе Текстерра подразумеваются ссылки

— от статьи к назначенной ей странице-категории;
— от страницы-подкатегории к категории верхнего уровня.

Ссылки в инфобокс-секциях статей. Инфобокс — специальная таблица в правом верхнем углу статьи, резюмирующая некоторые факты или статистические данные о предмете статьи. Обычно ряд статей, связанных общей тематикой, содержит инфобоксы с одинаковыми полями. Так, статья о музыкальной группе содержит в данной секции, как правило, жанр песен, годы активной деятельности, страну и город образования, язык песен, музыкальные компании, издающие альбомы группы, а также ее текущий состав и список бывших участников.

Ссылки из секций «См. также». Секция «См. также» обычно размещается в конце статьи и содержит список статей, тематически связанных с данной.

Ссылки вида «Основная статья». Ссылки данного типа размещаются в соответствующем шаблоне над разделом, содержащим краткое описание некоторого аспекта статьи, и указывают на статью с более детальным описанием этого аспекта.

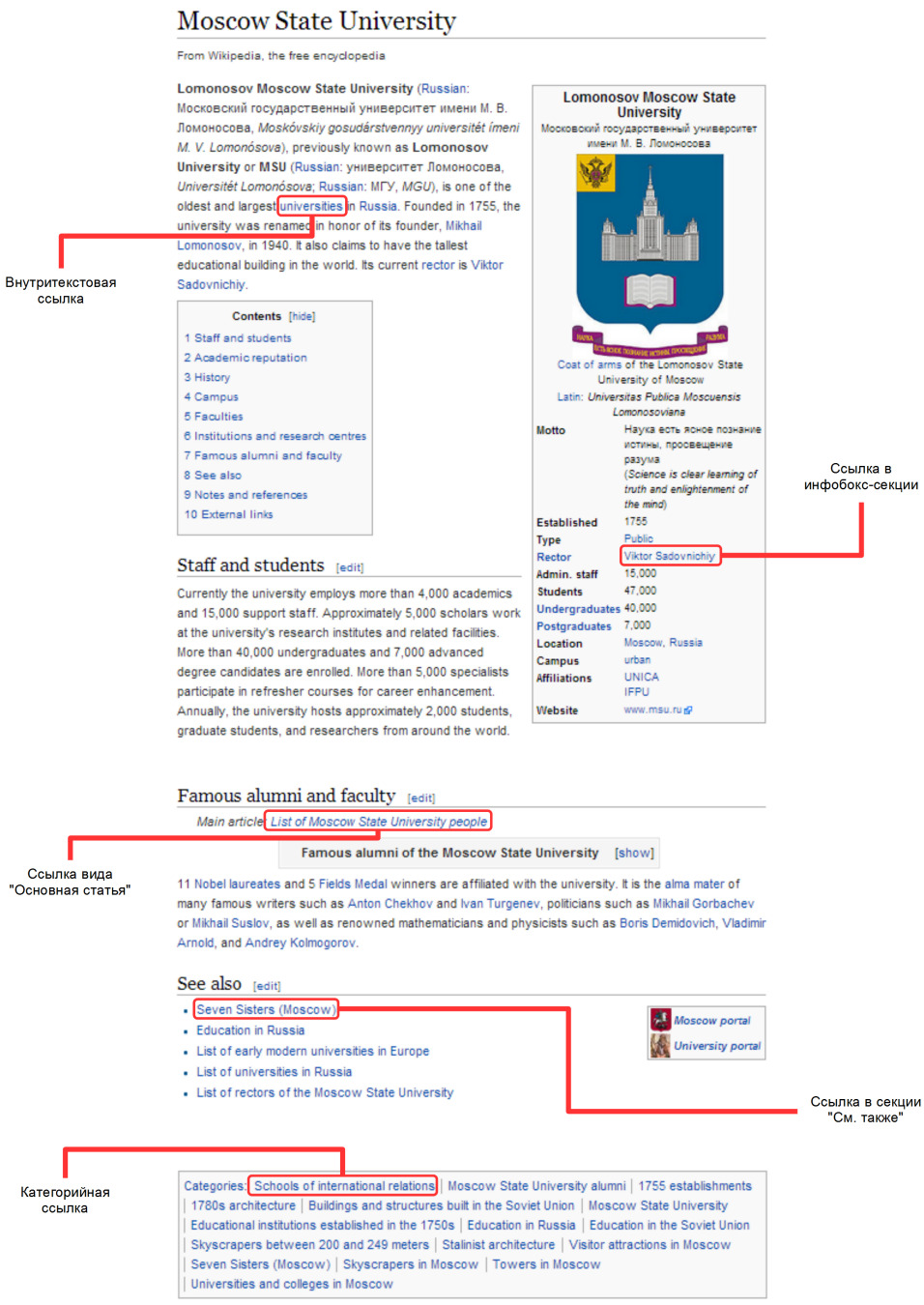


Рис. 1. Фрагмент статьи Википедии с примерами основных типов ссылок

Так как используемый нами метод вычисления расстояний неприменим к ориентированным графам, рассматривались только неориентированные пути между вершинами (иначе говоря, мы считали все ссылки Википедии двунаправленными).

Расстоянием $dist_T(x, y)$ между концептами x и y по ссылкам типа T будем называть длину (число ссылок) кратчайшего пути между ними в графе ссылок Википедии по всем путям, составленным из ссылок типа T. Следует отметить, что такой путь не всегда существует, так как в срезе на определенный тип ссылок граф Википедии (т. е. при удалении из него ссылок других типов) может получиться несвязным. В частности, некоторые статьи могут не иметь секций «инфобокс» или «См. также». В случае отсутствия пути между парой концептов расстояние между ними считалось потенциально бесконечным.

Рисунок 2 иллюстрирует распределения экспертной близости относительно длины кратчайшего пути по каждому из типов ссылок на данных набора WordSim-353.

В случае внутритекстовых и категорийных ссылок прослеживается наиболее четкая тенденция к монотонному убыванию семантической близости с увеличением средней длины пути между концептами. Для остальных типов зависимость такого вида выражена слабее, что может быть объяснено разреженностью полученных подграфов либо дефектами построения набора данных. Отметим также малую величину диаметров подграфов Википедии по рассматриваемым типам ссылок, что может указывать на сохранение этими подграфами свойств безмасштабных сетей.

Вывод меры семантической близости

Меры близости по отдельным типам ссылок. Как показывает рис. 2, семантическая близость между концептами тем больше, чем меньше расстояние между ними. Для каждого типа ссылок T мы ввели меру семантической близости вида

$$\mathrm{sim}_T(x, y) = \frac{1}{1 + (\mathrm{dist}_T(x, y))}, \tag{6}$$

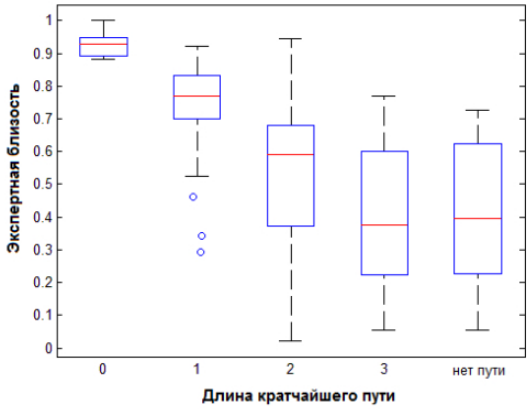
С помощью каждой из мер были вычислены оценки семантической близости пар концептов из набора WordSim-353 и найдена корреляция последних с экспертными оценками. Для сравнения была также рассчитана корреляция с экспертными оценками значений близости, полученных на тех же данных с помощью меры Дайса.

Помимо стандартных для данной задачи коэффициентов корреляции Пирсона и Спирмена, авторы настоящей статьи использовали также коэффициент корреляции по расстоянию (distance correlation [12]). Последний интересен тем, что коэффициенты Пирсона и Спирмена могут быть нулевыми для зависимых величин [12], корреляция же по расстоянию равна нулю тогда и только тогда, когда рассматриваемые величины независимы. Таким образом, корреляция по расстоянию может дать более точную оценку зависимости семантической близости концептов от длин кратчайших путей между ними в графе ссылок Википедии, чем корреляция Пирсона и Спирмена.

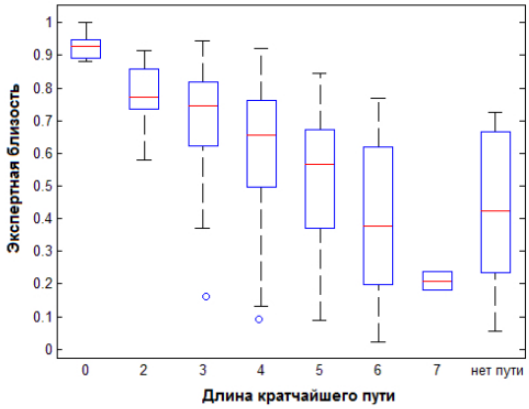
Результаты эксперимента приведены в табл. 2.

Видно, что наилучшую корреляцию среди рассматриваемых типов ссылок показывают внутритекстовые и категорийные. Тем самым подтверждаются результаты, приведенные в предыдущем разделе. При этом для обоих отображений мера близости на основе внутритекстовых ссылок лучше коррелирует с экспертными оценками, чем мера Дайса, по коэффициентам Пирсона и расстояния.

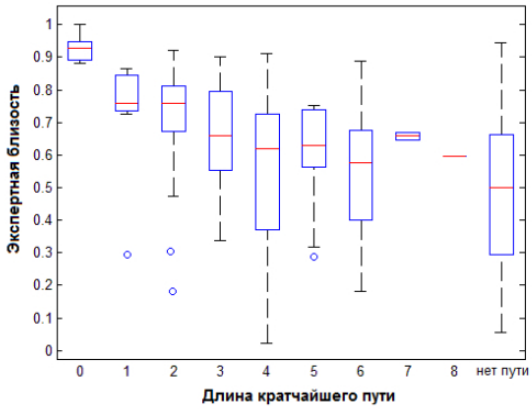
Справедливо будет отметить, что отображение $1/(1+x)$ в формуле (6) не является единственным способом перевода значения расстояния в величину близости. Для построения



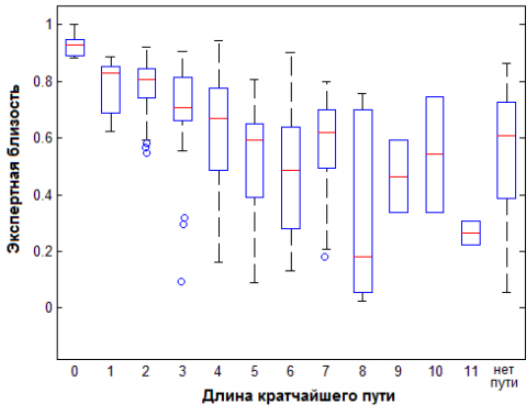
а) Внутритекстовые ссылки



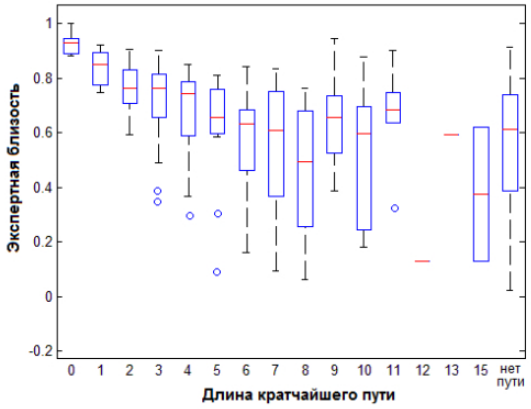
б) Категорийные ссылки



в) Ссылки в инфобокс-секции



г) Ссылки в секции «См. также»



д) Ссылки вида «Основная статья»

Рис. 2. Зависимость экспертной семантической близости от расстояния по различным типам ссылок

Таблица 2. Корреляция мер семантической близости по различным типам ссылок с экспертными оценками на наборе данных WordSim-353

Мера близости		Корреляция с экспертными оценками		
		по Пирсону	по Спирмену	по расстоянию
Типы ссылок	Внутритекстовые	**0,5072**	0,6059	**0,5409**
	Категорийные	0,4252	0,5695	0,4805
	Инфобокс	0,3633	0,3986	0,3684
	См. также	0,3850	0,4593	0,4273
	Основная статья	0,3388	0,3013	0,3209
Мера Дайса		0,3376	**0,6804**	0,4901

меры, сопоставляющей менее удаленным концептам большие значения близости, подойдет любая функция $f(x)$, монотонно не возрастающая на полуинтервале $[0, +\infty)$. Мера Дайса принимает значения из отрезка $[0, 1]$; чтобы меру близости на основе расстояния можно было использовать в Текстерре, логично потребовать от f того же множества значений, причем

— $f(0) = 1$;
— $\lim\limits_{x \to +\infty} f(x) = 0$.

Отображение $1/(1 + x)$ показалось наиболее простым примером описанного вида преобразований и потому было выбрано для дальнейших экспериментов. Результаты исследования некоторых других отображений приводятся ниже в разделе «Анализ полученных мер близости и дополнительные эксперименты».

Взвешивание мер. После исследования мер близости на основе отдельных типов ссылок возник вопрос, можно ли улучшить корреляцию с экспертными оценками с использованием комбинации рассмотренных мер расстояния или близости.

Было предложено построить комбинированную меру близости в виде линейной комбинации мер по отдельным типам ссылок и меры Дайса:

$$\text{sim}\,(x, y) = \left(\sum_{p \in \{\text{типы ссылок}\}} w_p \text{sim}_p(x, y) \right) + w_{\text{Дайс}} \text{sim}_{\text{Дайс}}(x, y) \tag{7}$$

Здесь w_p, $p \in \{\text{типы ссылок}\}$, и $w_{\text{Дайс}}$ — неотрицательные вещественные коэффициенты.

Для поиска весовых коэффициентов w логично использовать обучение на экспертных данных. Распространенным [13] в задачах вывода мер расстояния и близости подходом является использование набора относительных ограничений вида «концепт x должен быть ближе к концепту y, нежели к концепту z»:

$$R = \{(x, y, z) : \text{sim}_{\text{эксперт}}(x, y) > sim_{\text{эксперт}}(x, z)\} \tag{8}$$

Мы смогли построить 606 таких триплетов на данных набора WordSim-353.

Значения весов w далее выводились с помощью простого итеративного подхода из семейства пассивно-агрессивных алгоритмов [14].

На каждом шаге t алгоритм получает на вход случайный триплет $(x, y, z) \in R$ и строит весовой вектор w^t, решая следующую задачу оптимизации:

$$w^t = \arg\min_w \frac{1}{2} \|w - w^{t-1}\|^2 + C\varepsilon \tag{9}$$

при условиях

$$\text{loss}\,(x, y, z) \leqslant \varepsilon,\ \varepsilon > 0 \tag{10}$$

Выше $\text{loss}\,(x, y, z) = \max\{0, 1 - (\text{sim}\,(x, y) - \text{sim}\,(x, z))\}$ — функция потерь алгоритма на триплете (x, y, z), а C — параметр агрессивности алгоритма, определяющий компромисс между минимизацией функции потерь и сохранением вектора весов w^{t-1}, полученного на предыдущем шаге. Решение задачи 9, 10 может быть записано в виде:

$$w^t = w^{t-1} + \text{loss}\,(x, y, z) \cdot \min\left\{C, \frac{\text{loss}\,(x, y, z)}{\|\Delta_{\text{sim}}(x, y, z)\|^2}\right\}, \tag{11}$$

где

$$\|\Delta_{\text{sim}}(x, y, z)\|^2 = \sum_{p \in \{\text{типы ссылок}\} \cup \{\text{Дайс}\}} (\text{sim}_p(x, y) - sim_p(x, z))^2 \tag{12}$$

На начальном этапе веса всех мер инициализировались нулями. Проводилось 10^6 итераций алгоритма, после чего полученные веса нормировались, чтобы свести множество значений комбинации мер к отрезку $[0, 1]$. В экспериментах использовалось значение параметра агрессивности C, равное 0.01.

Идея использования именно пассивно-агрессивного алгоритма над относительными ограничениями для нахождения параметров меры семантической близости была заимствована нами у метода OASIS [13]. Необходимо отметить, однако, что вид и природа меры близости, выводимой его авторами, существенно отличаются от рассматриваемых в данной работе. Тем не менее далее для сокращения записи мы будем называть линейную комбинацию мер 7 с коэффициентами, полученными пассивно-агрессивным алгоритмом 9-12, OASIS-мерой. При этом непосредственно как «OASIS» будет обозначаться линейная комбинация мер, построенная без участия меры Дайса ($w_{\text{Дайс}} = 0$); включение меры Дайса в комбинацию будет указываться явно.

В табл. 3 приводятся оценки корреляции меры 7 с экспертной близостью, полученные в результате скользящего контроля с 10 блоками на наборе данных WordSim-353.

Таблица 3. Корреляция комбинированных мер семантической близости с экспертными оценками на наборе данных WordSim-353

Мера близости	Корреляция с экспертными оценками		
	по Пирсону	по Спирмену	по расстоянию
Мера Дайса	0,3376	0,6804	0,4901
OASIS	**0,5388**	0,6390	**0,6183**
OASIS+мера Дайса	0,5193	**0,6853**	0,5856

OASIS-мера, построенная без меры Дайса, имеет лучшие коэффициенты корреляции по Пирсону и по расстоянию, а подключение к ней меры Дайса показывает лучшую корреляцию по Спирмену. В целом, использование кратчайших путей позволило улучшить показатели меры Дайса относительно всех рассмотренных коэффициентов корреляции.

Чтобы понять, какие типы ссылок вносят наибольший вклад при вычислении меры OASIS, мы провели 10^6 итераций его обучения на относительных ограничениях, построенных по всему набору WordSim-353. Полученные веса показаны в табл. 4. Подтверждаются

Таблица 4. Веса мер близости по различным типам ссылок, полученные при обучении OASIS-меры (без нормировки)

Тип ссылок	Внутритекстовые	Категорийные	См. также	Инфобокс	Основная статья
Вес	**0,3307**	**0,5843**	0,0317	0,0406	0,0126

результаты предыдущего раздела о том, что наиболее ценными типами ссылок для расчета семантической близости концептов являются внутритекстовые и категорийные.

Анализ полученных мер близости и дополнительные эксперименты

Сравнение с другими мерами на основе кратчайших путей. Мы реализовали меры Цеша и Ликока–Чодороу в соответствии с описаниями их расчета, приведенными в статьях [9] и [8]. Для вычисления расстояния в мере Цеша поиск осуществлялся поиск кратчайшего пути между концептами по произвольным типам ссылок. Глубина поиска кратчайшего пути ограничивалась 5 узлами. При расчете меры Ликока–Чодороу, в отличие от рассмотренных ранее мер по категорийным ссылкам, расстояние между концептами рассчитывалось как сумма их расстояний до страницы-категории, являющейся наименьшим общим предком данных концептов в иерархии категорий Википедии. Глубина поиска наименьшего общего предка ограничивалась 4 узлами.

В табл. 5 представлены значения корреляции мер Цеша и Ликока–Чодороу с экспертными оценками на данных WordSim-353 в сравнении с мерой Дайса и мерами по внутритекстовым и категорийным ссылкам. Мера Ликока–Чодороу показывает лучшую корреляцию Пирсона среди рассматриваемых мер; коэффициенты же Спирмена и расстояния у мер Цеша и Ликока–Чодороу сравнимы с мерой близости по внутритекстовым ссылкам. По всей видимости, мера Ликока и Чодороу эффективнее использует иерархию категорий Википедии, чем введенная нами мера по категорийным ссылкам; тем не менее для других типов ссылок подход Ликока–Чодороу теряет смысл.

Таблица 5. Корреляция мер Цеша и Ликока–Чодороу с экспертными оценками на данных набора WordSim-353 в сравнении с мерой Дайса и мерами по отдельным типам ссылок

Мера близости	Корреляция с экспертными оценками		
	по Пирсону	по Спирмену	по расстоянию
Мера Цеша	0,4897	0,6076	0,5350
Мера Ликока–Чодороу	**0,5433**	0,5770	0,5301
Мера Дайса	0,3376	**0,6804**	0,4901
Внутритекстовые ссылки	0,5072	0,6059	**0,5409**
Категорийные ссылки	0,4252	0,5695	0,4805

Для сравнения с OASIS-мерой мы пересчитали оценки мер Цеша и Ликока–Чодороу посредством скользящего контроля на наборе WordSim-353 с 10-блоками. Результаты вычислений приведены в табл. 6. Комбинированная мера показывает лучшую корреляцию

с экспертными оценками, чем меры Цеша и Ликока–Чодороу, в смысле коэффициентов Спирмена и расстояния, и сравнима с ними в смысле коэффициента Пирсона.

Таблица 6. Корреляция мер Цеша и Ликока–Чодороу с экспертными оценками на данных набора WordSim-353 в сравнении с OASIS-мерой

Мера близости	Корреляция с экспертными оценками		
	по Пирсону	по Спирмену	по расстоянию
Мера Цеша	0,5192	0,5893	0,5406
Мера Ликока–Чодороу	**0,5437**	0,5655	0,5580
OASIS	0,5388	0,6390	**0,6183**
OASIS+мера Дайса	0,5193	**0,6853**	0,5856

Влияние вида отображений расстояние-близость. Мы рассмотрели три вида преобразований расстояние-близость:

1. $f(x) = max(1 - ax^b, 0)$
2. $f(x) = 1/(1 + ax^b)$
3. $f(x) = e^{-ax^b}$

Везде выше $a > 0, b > 0$ — некоторые константы.

Путем варьирования значений параметров a, b мы пытались достичь наилучшей корреляции мер по отдельным типам ссылок с экспертами на наборе данных WordSim-353.

Множества значений параметров, по которым осуществлялся поиск, приведены ниже:
— $a \in \{0.01, 0.1, 0.25, 0.5, 0.75, 1, 10\}$;
— $b \in \{0.5, 1, 2, 3, 4\}$

В результате эксперимента выяснилось, что для всех видов отображений оптимальные комбинации параметров позволяют улучшить показатели корреляции, полученные с помощью отображения $1/(1 + x)$, не более чем на $1,5\%$. Тем самым выбор первоначального отображения, сделанный в разделе «Вывод меры семантической близости», можно считать удачным и близким к оптимальному.

Семантическая близость удаленных пар концептов. Большинство пар концептов в наборе WordSim-353 слабо удалены друг от друга в графе Википедии (в смысле путей по произвольным типам ссылок). Нулевое значение меры Дайса имеют всего лишь 30 пар концептов в наборе. Исследование корреляции с экспертами на данном наборе, таким образом, не в полной мере учитывает глобальную природу мер на основе кратчайших путей и их применимость к произвольным парам концептов.

Решено было самостоятельно составить и разметить набор данных, с помощью которого можно было бы оценить, насколько точно меры на основе путей позволяют вычислять близость для *произвольных* пар концептов Википедии. Для облегчения процедуры разметки, а также для устранения шума, который неизбежно возникает при сопоставлении числовых значений близости парам концептов, было предложено разметить набор троек концептов относительной близостью (концепт x ближе к y, чем к z).

Мы сгенерировали всего 1000 троек концептов Википедии. Каждый концепт в каждой тройке генерировался независимо от двух других из равномерного распределения над всеми концептами Википедии, присутствующими в базе знаний системы Текстерра.

В подавляющем большинстве случаев тематика статей в рамках одной тройки разнилась настолько, что не представлялось возможным хоть как-то соотнести их между собой. Процесс разметки руководствовался попытками выделить общий домен для пар концептов, входящих в тройку: так, для тройки 51503 : [Progressive rock]; 415847 : [Rewriting]; 147403 : [The Moody Blues] первый концепт ближе к третьему, так как для них можно выделить общих домен (музыка). Таким образом нам удалось разметить 40 троек концептов.

Применение меры семантической близости к тройке концептов (x, y, z) можно интерпретировать как задачу бинарной классификации с двумя классами:
— x ближе к y, чем к z;
— x ближе к z, чем к y.
Тогда качество решения данной задачи можно оценивать с помощью точности (accuracy), определяемой как отношение числа троек, отнесенных заданной мерой близости в правильный класс, к общему числу троек набора.

Таблица 7 содержит оценки точности определения относительной близости концептов на размеченном нами наборе для меры Дайса и различных мер на основе кратчайших путей в графе ссылок Википедии.

Таблица 7. Точность(accuracy) расчета относительной близости для троек концептов

Мера близости	Точность
Мера Дайса	0,3250
Внутритекстовые	0,3250
Категорийные	0,5250
Инфобокс	0,5250
См. также	0,5250
Основная статья	0,3250
OASIS	**0,7000**

Как и ожидалось, использование кратчайших путей позволяет повысить точность расчета семантической близости для произвольных пар концептов по сравнению с использованием меры Дайса. Неожиданно низкий результат внутритекстовых ссылок связан, скорее всего, с наименьшим среди всех типов ссылок диаметром соответствующего подграфа Википедии: второй и третий концепты часто оказываются равноудалены от первого.

Разрешение лексической многозначности

Мы исследовали применимость полученных мер семантической близости к решению прикладных задач с помощью двух методов разрешения лексической многозначности.

Первый метод используется в системе Текстерра. Для многозначного термина сначала с помощью базы знаний системы строится набор концептов, являющихся предполагаемыми его значениями. Для каждого значения формируется вектор признаков, одним из которых является семантическая близость концепта к однозначному контексту термина. Признаковые описания концептов подаются на вход классификатору на основе метода максимальной энтропии с двумя классами: является ли данный концепт подходящим значением многозначного термина или нет. Концепт с наибольшей вероятностью первого класса сопоставляется термину.

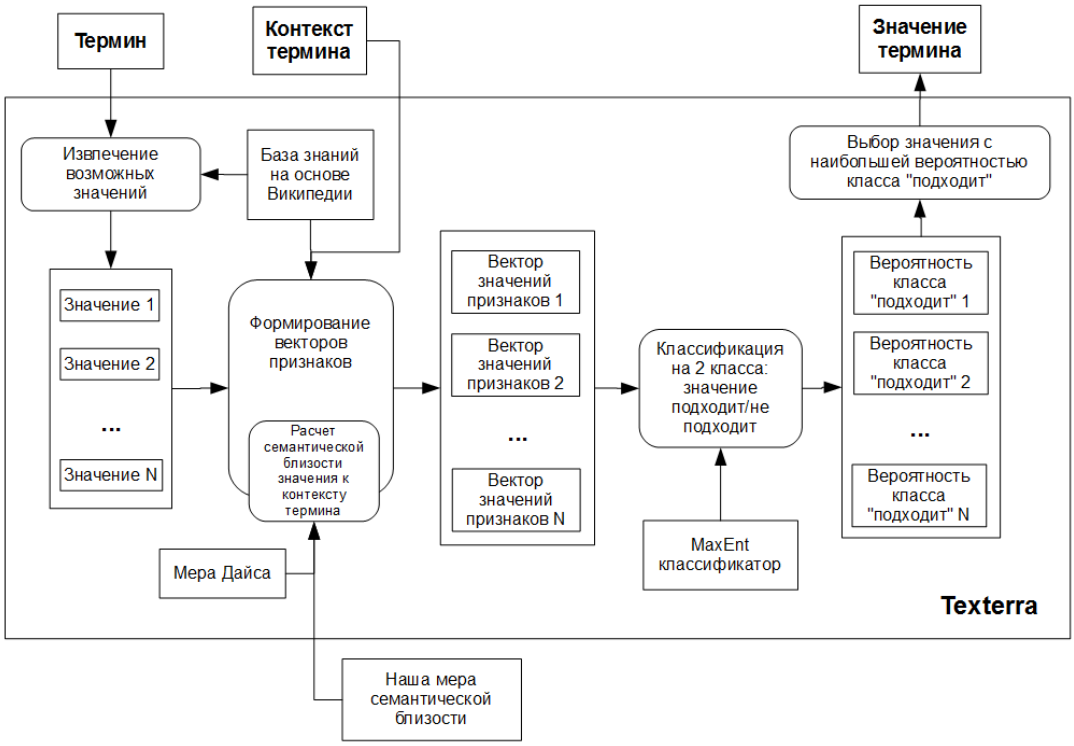


Рис. 3. Схема работы алгоритма разрешения лексической многозначности системы Текстерра

Однозначный контекст термина определяется как набор *всех* терминов текста, для которых база знаний системы Текстерра содержит ровно одно значение. Близость концепта к контексту вычисляется как сумма значений близости к каждому его термину, причем оценка близости для конкретного термина столько раз включается в сумму, сколько раз данный термин встречается в тексте.

Схема работы метода системы Текстерра приведена на рис. 3.

Поскольку другие признаки классификатора могут ослаблять влияние признака на основе семантической близости, для более точной оценки различных мер мы ввели второй, «наивный» метод разрешения лексической многозначности. Схема его работы приведена на рис. 4. Данный метод сопоставляет многозначному термину значение, наиболее близкое к его однозначному контексту в смысле заданной меры семантической близости. Вычисление семантической близости к контексту осуществляется тем же образом, что и в методе системы Текстерра. Никакие другие признаки не используются. При этом метод не требует обучения, что делает его более простым в использовании, чем первый.

В тестировании принимали участие следующие меры семантической близости:

— мера Дайса системы Текстерра;
— мера близости на основе расстояния по внутритекстовым ссылкам;
— мера близости на основе расстояния по категорийным ссылкам;
— комбинация мер близости по всем типам ссылок, построенная по методу OASIS, без участия меры Дайса;

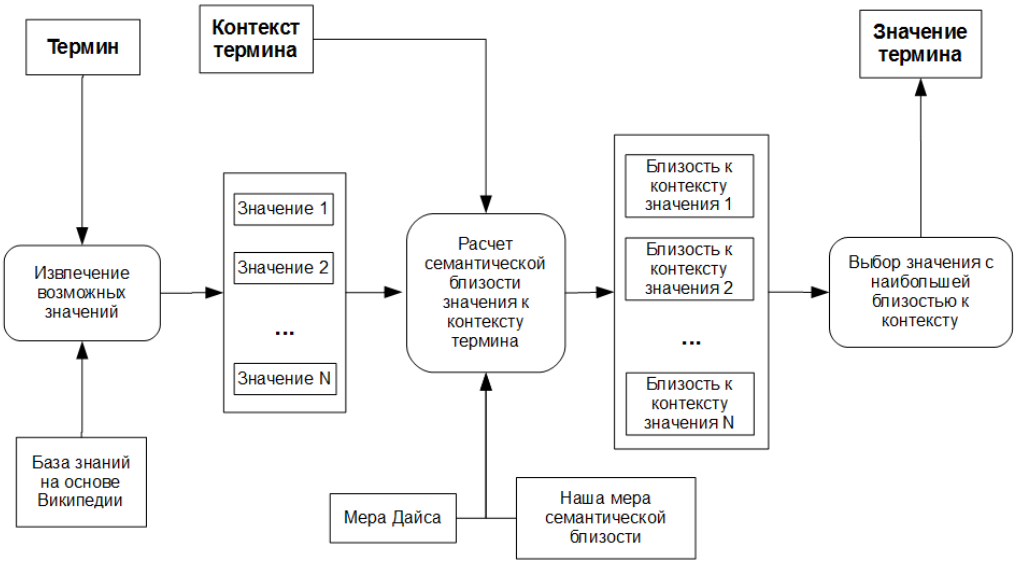


Рис. 4. Схема работы «наивного» алгоритма разрешения лексической многозначности

— комбинация мер близости по всем типам ссылок и меры Дайса, построенная по методу OASIS.

Тестирование проводилось на четырех стандартных наборах данных, используемых в системе Текстерра для оценки качества разрешения лексической многозначности терминов [1]. Их характеристики приводятся в табл. 8.

Таблица 8. Характеристики наборов данных для тестирования разрешения лексической многозначности

Название	Число текстов	Число концептов	Тематика
Board games	35	13410	Статьи о настольных играх
AQUAINT	50	13974	Новостные статьи
MODIS-texts	131	25061	Технические статьи
Wiki	100	104401	Статьи Википедии

Поскольку метод разрешения лексической многозначности, используемый в системе Текстерра, для работы требует обучения классификатора, для его оценки использовался метод скользящего контроля с 10 блоками. Важно отметить, что при обучении алгоритма всегда использовалась та же мера семантической близости, что и при тестировании.

Наивный алгоритм разрешения лексической многозначности не требует обучения, поэтому его тестирование проводилось на полных наборах данных. Результаты тестирования приведены в табл. 9 и 10.

Таблица 9. Разрешение лексической многозначности по методу системы Текстерра

Набор данных	Критерий оценки	Мера Дайса	Внутритекстовые ссылки	Категор. ссылки	OASIS	OASIS + мера Дайса
Board games	Точность	**0,8305**	0,5624	0,6265	0,6191	0,7278
	Полнота	**0,3678**	0,2491	0,2773	0,2740	0,3221
	F-мера	**0,5095**	0,3450	0,3842	0,3796	0,4463
AQUAINT	Точность	**0,8784**	0,8581	0,8626	0,8626	0,8649
	Полнота	**0,5991**	0,5853	0,5883	0,5883	0,5899
	F-мера	**0,7123**	0,6959	0,6995	0,6995	0,7014
MODIS-texts	Точность	**0,8203**	0,6166	0,6997	0,6746	0,7577
	Полнота	**0,4407**	0,3312	0,3759	0,3624	0,4071
	F-мера	**0,5734**	0,4309	0,4890	0,4715	0,5296
Wiki	Точность	**0,9199**	0,5390	0,6541	0,5805	0,7425
	Полнота	**0,6017**	0,3525	0,4278	0,3796	0,4856
	F-мера	**0,7274**	0,4262	0,5173	0,4590	0,5872

Таблица 10. Разрешение лексической многозначности наивным методом

Набор данных	Критерий оценки	Мера Дайса	Внутритекстовые ссылки	Категор. ссылки	OASIS	OASIS + мера Дайса
Board games	Точность	0,7657	**0,7676**	0,6167	0,7235	0,7282
	Полнота	0,3391	**0,3400**	0,2731	0,3205	0,3225
	F-мера	0,4701	**0,4712**	0,3786	0,4442	0,4471
AQUAINT	Точность	0,6689	**0,8086**	0,5766	0,7523	0,7072
	Полнота	0,4562	**0,5515**	0,3932	0,5131	0,4823
	F-мера	0,5425	**0,6557**	0,4676	0,6100	0,5735
MODIS-texts	Точность	0,7103	**0,7775**	0,6189	0,7054	0,7011
	Полнота	0,3816	**0,4177**	0,3325	0,3790	0,3767
	F-мера	0,4965	**0,5435**	0,4326	0,4931	0,4901
Wiki	Точность	0,8978	**0,9241**	0,7601	0,8791	0,9004
	Полнота	0,5880	**0,6053**	0,4978	0,5758	0,5898
	F-мера	0,7106	**0,7314**	0,6016	0,6958	0,7127

Результаты тестов показывают, что меры семантической близости концептов на основе длин кратчайших путей в графе ссылок Википедии позволяют достичь лучших результатов в решении задачи разрешения лексической многозначности по сравнению с мерой Дайса при использовании наивного алгоритма. На всех наборах данных наивысших результатов достигает мера близости по внутритекстовым ссылкам. Использование кратчай-

ших путей здесь позволило поднять точность разрешения лексической многозначности в среднем на 6%, полноту — на 4%, F-меру — на 5%. Внедрение же других признаков делает метод разрешения лексической многозначности менее чувствительным к способу вычисления семантической близости между концептами. Использование меры Дайса дает существенный выигрыш на всех наборах данных: по сравнению с лучшей из мер на основе путей, точность больше в среднем на 13%, полнота — на 6%, F-мера — на 8%.

Заключение

В рамках данной работы исследовалась применимость знания о длине кратчайшего пути между концептами в графе ссылок Википедии к расчету семантической близости между ними. Было показано, что использование длин кратчайших путей между концептами в графе ссылок Википедии позволяет точнее оценивать семантическую близость по сравнению с мерой Дайса, используемой в системе Текстерра. Также были выявлены типы ссылок, предоставляющие наиболее ценную информацию для расчета семантической близости (внутритекстовые и категорийные ссылки).

В рамках продолжения исследований по данной теме планируется исследовать пути в ориентированном/взвешенном графе ссылок Википедии, а также добавить к сравнению меры близости на основе случайного блуждания.

Литература

[1] *Турдаков Д. Ю. и др.* Texterra: инфраструктура для анализа текстов // *Тр. Института системного программирования РАН*, 2014. Т. 26. № 1. С. 421–438.

[2] *Deerwester S. C. et al.* Indexing by latent semantic analysis // *JASIS*, 1990. Vol. 41. No. 6. Pp. 391–407.

[3] *Gabrilovich E., Markovitch S.* Computing semantic relatedness using Wikipedia-based explicit semantic analysis // *IJCAI*, 2007. Vol. 7. Pp. 1606–1611.

[4] *Turdakov D., Velikhov P.* Semantic relatedness metric for Wikipedia concepts based on link analysis and its application to word sense disambiguation // *5th Spring Young Researchers Colloquium on Databases and Information Systems, SYRCoDIS'2008, Proceedings*, 2008.

[5] *Witten I., Milne D.* An effective, low-cost measure of semantic relatedness obtained from Wikipedia links // *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy Proceedings.* — Chicago, USA: AAAI Press, 2008. Pp. 25–30.

[6] *Agirre E., Soroa A.* Personalizing pagerank for word sense disambiguation // *12th Conference of the European Chapter of the Association for Computational Linguistics Proceedings.* Association for Computational Linguistics. 2009. Pp. 33–41.

[7] *Yeh E. et al.* WikiWalk: Random walks on Wikipedia for semantic relatedness // *2009 Workshop on Graph-based Methods for Natural Language Processing Proceedings* . Association for Computational Linguistics, 2009. Pp. 41–49.

[8] *Zesch T., Muller C., Gurevych I.* Using Wiktionary for computing semantic relatedness // *AAAI*, 2008. Vol. 8. Pp. 861–866.

[9] *Strube M., Ponzetto S. P.* WikiRelate! Computing semantic relatedness using Wikipedia // *AAAI*, 2006. Vol. 6. Pp. 1419–1424.

[10] *Finkelstein L. et al.* Placing search in context: The concept revisited // *10th Conference (International) on World Wide Web Proceedings* . ACM, 2001. Pp. 406–414.

[11] *Akiba T., Iwata Y., Yoshida Y.* Fast exact shortest-path distance queries on large net-works by pruned landmark labeling // *2013 Conference (International) on Management of Data Proceedings*. ACM, 2013. Pp. 349–360.

[12] *Szekely G. J. et al.* Brownian distance covariance // *Annals Appl. Stat.*, 2009. Vol. 3. No. 4. Pp. 1236–1265.

[13] *Bellet A., Habrard A., Sebban M.* A survey on metric learning for feature vectors and structured data. 2013. `http://arxiv.org/abs/1306.6709`.

[14] *Crammer K. et al.* Online passive-aggressive algorithms // *J. Machine Learning Research*, 2006. Vol. 7. Pp. 551–585.

References

[1] *Turdakov D., et al.* 2014. Texterra: A framework for text analysis. *Proc. Institute for System Programming of RAS*. 26(1):421–438.

[2] *Deerwester S. C. et al.* 1990. Indexing by latent semantic analysis. *JASIS* 41(6):391–407.

[3] *Gabrilovich E., Markovitch S.* 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. *IJCAI* 7:1606–1611.

[4] *Turdakov D., Velikhov P.* 2008. Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. *5th Spring Young Researchers Colloquium on Databases and Information Systems, SYRCoDIS'2008, Proceedings*.

[5] *Witten I., Milne D.* 2008. An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. *AAAI Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy Proceeding*. AAAI Press, Chicago, USA. 25–30.

[6] *Agirre E., Soroa A.* 2009. Personalizing pagerank for word sense disambiguation. *12th Conference of the European Chapter of the Association for Computational Linguistics Proceedings*. Association for Computational Linguistics. 33–41.

[7] *Yeh E. et al.* 2009. WikiWalk: random walks on Wikipedia for semantic relatedness. *2009 Workshop on Graph-based Methods for Natural Language Processing Proceedings*. Association for Computational Linguistics. 41–49.

[8] *Strube M., Ponzetto S. P.* 2006. WikiRelate! Computing semantic relatedness using Wikipedia. *AAAI* 6:1419–1424.

[9] *Zesch T., Muller C., Gurevych I.* 2008. Using Wiktionary for computing semantic relatedness. *AAAI* 8:861–866.

[10] *Akiba T., Iwata Y., Yoshida Y.* 2013. Fast exact shortest-path distance queries on large networks by pruned landmark labeling. *2013 Conference (International) on Management of Data Proceedings*. ACM. 349–360.

[11] *Finkelstein L. et al.* 2001. Placing search in context: The concept revisited. *10th Conference (International) on World Wide Web Proceedings*. ACM. 406–414.

[12] *Szekely G. J. et al.* 2009. Brownian distance covariance. *Annals Appl. Stat.* 3(4):1236–1265.

[13] *Bellet A., Habrard A., Sebban M.* 2013. A survey on metric learning for feature vectors and structured data. `http://arxiv.org/abs/1306.6709`.

[14] *Crammer K. et al.* 2006. Online passive-aggressive algorithms. *J. Machine Learning Research* 7:551–585.