

ОПРЕДЕЛЕНИЕ СЕМАНТИЧЕСКОЙ БЛИЗОСТИ ТЕРМОВ С ПОМОЩЬЮ КОНТЕКСТНОГО МНОЖЕСТВА

Бондарчук Д.В.

Уральский государственный университет путей сообщения
ул. Колмогорова, 66, Екатеринбург, Свердловская обл., 620034, Россия
e-mail: dvbondarchuk@gmail.com

Аннотация — В данной работе предлагается способ вычисления семантической близости термов, основанный на предположении, что семантически близкие термы употребляются в одинаковых или схожих контекстах. Для расчета семантической близости вводится понятие "контекстное множество". Контекстное множество показывает множество термов, с которыми целевой терм встречается в одном контексте. С помощью контекстного множества, для каждого слова из словаря можно определить признак, показывающий его встречаемость вместе с целевым термом. Контекстное множество термина удобно строить, используя матрицу корреспонденций термов.

CALCULATING THE SEMANTIC RELATEDNESS OF TERMS WITH THE CONTEXT SET

Bondarchuk D.V.

Ural State University of Railway Transport
ul. Komogorova, 66, Yekaterinburg, Sverdlovsk region, 620034, Russian Federation
e-mail: dvbondarchuk@gmail.com

Abstract — In this paper we propose a method of calculating the semantic relatedness of terms based on the assumption that semantically similar terms are used in the same or similar contexts. We introduce concept of "context set" for semantic relatedness calculation. Context set shows a variety of terms, which occurs with target term in same contexts. The context set, for each word in the dictionary, you can define a sign showing its occurrence along with the target term. The context set of user-friendly building, using the matrix of terms correspondence.

I. Введение

Семантическая близость термов уже давно является неотъемлемой частью теории обработки текстов на естественном языке. Семантическая близость между двумя сущностями с течением времени может изменяться в связи с изменениями корпусов и словарей [1]. Кроме того, семантическая близость двух сущностей может быть различна в различных предметных областях. Например, слово «одноклассники» в российском интернете чаще всего будет связано с фразой «социальная сеть», однако данный смысл вряд ли будет представлен в каком-либо из словарей или корпусов. Человека, ищущего данное слово в интернете скорее всего интересует именно этот смысл данного слова, чем какой-либо другой. Новые слова создаются настолько же часто насколько новые смыслы приписываются старым словам, поэтому поддержание словарей в актуальном состоянии не представляется возможным [2].

Семантическая близость термов может быть вычислена с помощью специализированных баз данных и статистических корпусов. Так же огромное число современных подходов к вычислению семантической близости основано на вычислении расстояний между словами в известной семантической сети WordNet. Российская версия сети WordNet разрабатывается в Петербургском Государственном Университете Путей Сообщения. При расчете семантической близости может использоваться так же связность между словами в контексте, а также ее важность. Чтобы показать разницу между связностью и близостью Резник Ф.А. [3] приводил в пример слова «автомобиль» и «бензин». Данные слова не являются синонимами и их значения далеки друг от друга, однако очевидно, что эти термины все-таки имеют что-то общее. Эти слова могут иметь сильную функциональную

взаимосвязь в контексте, в пример можно привести фразу «автомобили используют бензин в качестве топлива».

Определение. Будем называть множество слов, связанных с заданным термом *контекстным множеством* термина.

Таким образом, для каждого слова из словаря можно определить признак, показывающий его встречаемость вместе с целевым термом.

II. Проблемы существующих подходов

Большинство современных подходов к вычислению семантической близости основано на использовании семантической сети WordNet или Web-источников. В реализации известной семантической сети WordNet, разработанной в Принстонском университете, используются так называемые «синсеты» - синонимические ряды, объединяющие слова со схожим значением. Каждый «синсет» содержит список синонимов или синонимичных словосочетаний и указатели, описывающие отношения между ним и другими «синсетами». Слова, имеющие несколько значений, включаются в несколько «синсетов» и могут быть причислены к различным синтаксическим и лексическим классам.

Возможность свободного использования сети WordNET привела к тому, что появилось множество исследований, которые использовали её в качестве базы для обучения алгоритмов информационного поиска. В ходе этих экспериментов было выявлено множество недостатков этой семантической сети, которые препятствуют его эффективному применению.

Основной проблемой, связанной с применением WordNET стала сложность описания многозначных сущностей. В современной версии этой семантической сети содержится самое часто встречающееся значение, что дает возможность

выбирать именно его в случае возникновения проблем с многозначными сущностями. Так же серьезной проблемой практического применения WordNET является так называемая «теннисная проблема»: синсеты, принадлежащие одной и той же предметной области в структуре WordNET часто располагаются очень далеко, что в свою очередь приводит к затруднениям применения в задачах разрешения лексической многозначности.

Таким образом, применение этой семантической сети ограничено, поскольку она оперирует не какими-то определенными предметными областями, а охватывает общую лексику и семантику естественного языка. Кроме того, реальное применение данной базы ограничено, поскольку она содержит только наиболее часто встречающиеся слова. Специализированные слова некоторых предметных областей в ней практически отсутствуют, что приводит к невозможности их обработки. Кроме того, описания синсетов остались на английском языке, что так же накладывает определенные ограничения на область их применения.

Рассмотрим подход основанный на использовании внешних источников, расположенных в сети интернет (web-источники). Существует 2 фактора, которые негативно влияют на вычисление семантической близости, с использованием данного подхода. Этими факторами являются синонимия и полисемия.

Синонимия - случай, когда несколько слов имеют схожий смысл, например, «автомобиль» и «машина».

Полисемия - случай, когда одно слово имеет несколько смыслов, например, «диск».

Проблема синонимии состоит в том, что если документ уже содержит один из синонимов, то вероятность, что он так же содержит другой синоним мала. Это приводит к тому, что связность между синонимами, вычисленная только с помощью подхода, основанного на использовании веб-контента, получается меньше, чем она есть на самом деле. Например, поиск в Google слова «происшествие» даст примерно 440 000 результатов, а поиск его синонима слова «инцидент» - 592 000 результатов и примерно 34 000 результатов, где они встречаются вместе. *NGD* можно вычислить по следующей формуле [4]:

$$NGD(t_1, t_2) = \frac{\max \{ \lg f(t_1), \lg f(t_2), \lg f(t_1, t_2) \}}{\lg N - \min \{ \lg f(t_1), \lg f(t_2) \}} \quad (1)$$

где N - общее количество веб-страниц, обрабатываемых поисковой системой Google, $f(t_1)$ и $f(t_2)$ - количество страниц, на которых термины t_1 и t_2 находятся по отдельности, $f(t_1, t_2)$ - количество страниц, на которых термины t_1 и t_2 находятся вместе.

Вычислим *NGD* (*Normalized Google Distance*) между этими связанными словами $NGD \approx 0.8365$.

Если принять результат за достоверный, то можно сделать вывод, что связи между словами «происшествие» и «инцидент» фактически нет. *NGD* семантически идентичных слов равна 0, а семантически не связанных - 1.

Полисемия дает обратный эффект, состоящий в том, что в документе одно и то же слово может употребляться в нескольких смыслах. Например, слово «диск» может употребляться в различных смыслах: «колесный диск», «компьютерный диск» и т. п. Поиск в Google слова «диск» дает примерно 32

720 000 результатов, однако, допустим, нас интересуют только «колесные диски». По данному запросу мы получим только 238 000 результатов. Наблюдения производились в апреле 2014.

Таким образом, вычисление семантической близости термов, основанное на результатах выдачи глобальной поисковой системы, зачастую не дает ожидаемого результата. Конечно, это утверждение верно, если помимо глобальной поисковой системы никаких других источников данных о связи между словами не используется. Однако извлечь только нужные данные из этих источников и интерпретировать их для вычисления семантической близости достаточно сложно.

Далее рассмотрим подход вычисления семантической близости основанный на семантической сети *WordNet*. Вычислим связь между близкими словами «университет» и «экзамен», используя «синсеты» *WordNet* и метод *Леска* [5] и метод *Резника* [3]. Результаты данного вычисления указаны в следующей таблице:

Таблица 1. Близость между словами «университет» и «экзамен»

Метод	Леска	Резника
Близость	20	0

Далее вычислим те же самые метрики близости для слов, которые имеют совершенно различные значения: «университет» и «растение». Результаты указаны в таблице 2:

Таблица 2. Близость между словами «университет» и «растение»

Метод	Леска	Резника
Близость	28	2.3443

Сравнивая результаты, указанные в таблице 1 и в таблице 2 можно прийти к выводу, что слово «университет» ближе к слову «растение», чем к слову «экзамен», хотя очевидно, что данное утверждение далеко от истины.

Чтобы решить данную проблему предлагается использовать метод, использующий множество связанных слов, которое, как уже было сказано выше, предлагается назвать контекстным множеством слова. Используя распространенные корпуса и словари, несложно сформировать данное множество для любого слова.

III. Подготовка данных к анализу

Очевидно, что текстовые описания формируются обычными людьми, и как следствие часто имеет место сильная зашумленность данных. В связи с этим, прежде, чем переходить к анализу данных, необходимо произвести ряд действий для освобождения текста от шумов. Для этого предлагается использовать: семантическое ядро и стемминг.

Стемминг - это процесс нахождения основы слова для заданного исходного слова.

Основа слова необязательно совпадает с морфологическим корнем слова. Алгоритм стемматизации представляет собой давнюю проблему в области компьютерных наук. Первый документ по этому вопросу был опубликован в 1968 году. Данный процесс применяется в поисковых системах для обобщения поискового запроса пользователя [8].

Конкретные реализации стемматизации называются алгоритм стемматизации или просто

стеммер. Наиболее удачный алгоритм стемминга — стеммер Портера.

Оригинальная версия стеммера была предназначена для английского языка. Алгоритм не использует баз основ слов, а лишь, применяя последовательно ряд правил, отсекает окончания и суффиксы, основываясь на особенностях языка, в связи с чем работает быстро, но не всегда безошибочно [8].

Семантическое ядро - это подборка понятий, имеющих существенное значение для данной предметной области. Точное определение семантического ядра зависит от области применения. Так, в лингвистике, семантическим ядром называют «не упрощаемое замкнутое подмножество языка», подразумевая при этом скорее смысловую составляющую языка, а не грамматические конструкции.

Для использования в статистическом анализе текста приведем определения нескольких подборок смысловых единиц, сходных с семантическим ядром.

1. Специфичные слова предметной области

Это такие слова, которые встречаются исключительно в текстах предметной области и позволяют установить принадлежность текста этой предметной области.

2. Высокоинформативные слова предметной области

Это такие слова, которые позволяют рубрицировать тексты внутри предметной области. Например, для предметной области «поиск подходящих вакансий» такими словами являются: «няня», «сантехник», «репетитор» и т.д.

Семантическое ядро проще всего сформировать, анализируя большой объем текстов по предметной области. В него попадают слова, которые чаще всего встречаются в анализируемых текстах, исключая так называемые стоп-слова, например, предлоги, союзы и прочие слова, которые не несут смысловой нагрузки. Считается, что каждое из этих общих стоп-слов есть во всех документах выборки.

Кроме того, в некоторых предметных областях имеет смысл удалять имена собственные.

В некоторых предметных областях имеет смысл поработать с так называемыми зависимыми стоп-словами. Идея зависимых стоп-слов состоит в том, чтобы не учитывать наличие некоторых слов в документе без наличия других. Например, разбирая тексты предметной области «поиск вакансий разовой работы», при анализе фразы «гибкий график», имеет смысл рассматривать слово «гибкий» только в сочетании со словом «график».

IV. Построение контекстного множества

Предлагается следующий алгоритм построения контекстного множества терма.

Возьмем *матрицу корреспонденций термов*, описанную в работах [7; 8].

Определение. Матрица корреспонденций термов $G = \{g_{ij}\}$ - это квадратная матрица, элементами которой являются коэффициенты g_{ij} отражающие близость -го и j -го термов, для которых выполняются следующие условия:

1. $g_{ij} = g_{ji}$
2. $-1 < g_{ij} < 1$ для всех i и j .
3. $g_{ij} = 0$ при отсутствии взаимосвязи между термами.

Основное назначение матрицы G – отображение взаимосвязей термов внутри документов,

построенное на основе знаний частоте об их совместных употреблениях. На рисунке 1 изображен случай, когда термы t_1 и t_2 совместно встречаются в документе d_2 , а термы t_2 и t_3 - в документе d_1 . Таким образом, термы t_1 и t_3 так же связаны между собой через терм t_2 .

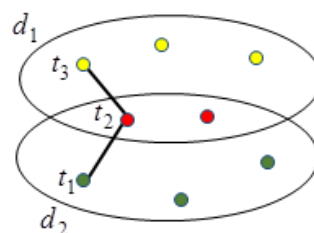


Рисунок 1. Иллюстрация взаимосвязей термов

Строки и столбцы данной матрицы соответствуют термам, пересечения показывают признак встречаемости данных термов совместно.

Предположим, необходимо построить контекстное множество -того терма, в этом случае необходимо рассматривать i -тую строку или i -тый столбец матрицы корреспонденций термов, исключая i -тый элемент, поскольку он показывает встречаемость терма с самим собой.

$$G = \left\{ (x_i, x_j) \right\}_{j=1, i \neq j}^n \quad (2)$$

где x_i, x_j – векторы термов, n – количество термов.

После этого, вычислим среднее арифметическое среди элементов вектора из формулы (2). Обозначим данное среднее \bar{G}_i . Далее, отбросим все компоненты вектора G_i меньше данного среднего значения. Контекстное множество -того терма будет состоять из термов, соответствующих оставшимся компонентам вектора G_i .

V. Метод вычисления семантической близости

Предположим, что w_1 и w_2 - слова, для которых необходимо вычислить семантическую близость. Предлагаемый метод можно условно разделить на несколько шагов:

1) Формирование контекстных множеств слов w_1 и w_2 .

Пусть $C_1 = \{c_{11}, c_{12}, \dots, c_{1n}\}$ и $C_2 = \{c_{21}, c_{22}, \dots, c_{2m}\}$ - контекстные множества слов w_1 и w_2 соответственно.

Данные множества содержат слова, с которыми слова w_1 и w_2 часто употребляются в одном контексте. Затем мы формируем общее контекстное множество слов:

$$C = C_1 \cup C_2 \quad (3)$$

Очевидно, что мощность данного множества будет равна $n + m$.

2) Вычисление нормализованных близостей между общим определителем и каждым из слов w_1 и w_2 :

$$\begin{aligned} \text{близость}(c_i, w_1) &= \frac{\text{частота}(c_i, w_1)}{\text{макс.частота}(w_1)} \\ \text{близость}(c_i, w_2) &= \frac{\text{частота}(c_i, w_2)}{\text{макс.частота}(w_2)} \end{aligned} \quad (4)$$

где $\text{частота}(c_i, w_1)$ - количество документов, где c_i и w_1 встречаются вместе, а $\text{макс.частота}(w_j)$ рассчитывается по формуле как максимум частот по всем словам из объединенного контекстного множества C :

$$\text{макс.частота}(w_j) = \max \{ \text{частота}(c_i, w_j) \} ; c_i \in C \quad (5)$$

3) Расчет семантической близости

Рассмотрим расчёт семантической близости между словами w_1 и w_2 .

Для этого рассчитаем коэффициенты R_i для всех слов из контекстного множества C по формуле:

$$R_i = \frac{\min \{ \text{близость}(c_i, w_1), \text{близость}(c_i, w_2) \}}{\max \{ \text{близость}(c_i, w_1), \text{близость}(c_i, w_2) \}} \quad (6)$$

Обозначим p_i - коэффициент совместной встречаемости w_1 и w_2 во всей выборке, равный 2 в случае, когда оба слова встречаются в одном документе и 1 в противном случае, s коэффициент синонимии, равный 1, если слова w_1 и w_2 являются синонимами и 0, в противном случае. Тогда семантическая близость слов w_1 и w_2 рассчитывается по формуле:

$$\text{сем.близость}(w_1, w_2) = \frac{\sum_{i=1}^k \left(\frac{p_i R_i}{1 + R_i} + s \right)}{1 + s} \quad (7)$$

В результате применения формулы (7) получится число в промежутке $[0, 0.75 \cdot (n + m)]$, чтобы получить семантическую близость в диапазоне $[0, 1]$ необходимо разделить получившийся результат на $0.75 \cdot (n + m)$. Для семантически близких слов, коэффициент близок к 1.

VI. Экспериментальная часть

Проверим эффективность метода на словах «автомобиль» и «поезд», обучив на основе новостей, представленных на сайте одного федерального СМИ. Примем, что w_1 - «машина», w_2 - «поезд». Составим так же контекстные множества для данных слов:

$$C_1 = \{ \text{автомобиль, мотор, колесо, пассажир, двигатель} \}$$

$$C_2 = \{ \text{рельсы, транспорт, двигатель, груз, пассажир} \}$$

Поскольку слова «машина» и «автомобиль» - синонимы, то будем считать, что если документ содержит слово «автомобиль», то он содержит и слово «машина». В таблице 3 представлены нормализованные близости между общим контекстным множеством и словами, вычисленные по формуле 4.

Таблица 4. Нормализованные близости между общим контекстным множеством и словами «машина» и «поезд»

	машина	поезд
рельсы	0.19	0.13
транспорт	1.00	0.89
двигатель	0.15	0.63
груз	0.07	0.06
пассажир	0.11	0.10
автомобиль	0.63	1.00
мотор	0.89	0.68
колесо	0.11	0.30

Далее рассчитаем коэффициенты R_i , результаты указаны в таблице 5. Расчёт был произведен для каждого слова из общего контекстного множества.

Таблица 5. Коэффициенты R_i

рельсы	0.71
транспорт	0.89
двигатель	0.24

груз	0.84
пассажир	0.93
автомобиль	0.63
мотор	0.77
колесо	0.37

После вычисления данных коэффициентов можно переходить к непосредственному вычислению семантической близости. Рассчитаем близость между словами «машина» и «поезд» по формуле (7), получим 3.107553949 (0.51792719 после нормализации). Рассчитаем так же семантическую близость с помощью расстояния Джакарда [6], получим 0.55321. Без проведения дополнительных расчетов очевидно, что результаты достаточно близки друг к другу.

Проведем еще несколько экспериментов. В таблице 6 представлен результат сравнения эффективности предложенного метода в сравнении с расстоянием Джакарда. Правая колонка иллюстрирует среднее арифметическое оценок людей. Группе из 50 человек было предложено оценить близость между двумя словами по столбальной шкале. В столбце представлен средний и нормализованный результат. Строка «корреляция» показывает коэффициент корреляции между результатами, полученными в результате применения каждого метода и оценкой реальных людей.

Таблица 6. Проверка результатов

Пара слов	Расстояние Джакарда	Представленный метод	Средняя оценка группы
Веревка - Улыбка	0.102	0.137	0.16
Побережье - Лес	0.016	0.649	0.41
Бухта - Холм	0.444	0.559	0.87
Машина - Путешествие	0.071	0.443	0.33
Фрукт - Еда	0.753	0.685	0.55
Автомобиль - Машина	0.654	0.939	1
Полдень - Обед	0.106	0.876	0.97
Джем - Варенье	0.295	0.836	0.84
Корреляция	0.45	0.851	

Довольно высокий коэффициент корреляции показывает, что результаты предложенного метода, ближе к объективным, чем метод основанный на вычислении расстояния Джакарда [7]. Исходя из полученных результатов, можно судить, что представленный метод является эффективным. При этом решены проблемы, возникшие в главе 3, связанные с необходимостью хранения и построения словарей гиперонимов и словарей определений.

К сожалению, при применении предложенного подхода возникает новая проблема, состоящая в сложности подбора определителей термов. Для этого, например, могут быть использованы любые внешние источники данных (вебсайты, журналы, книги, словари, корпусы). Наибольшую эффективность метод показывает при использовании в качестве источников данных веб-сайтов с ясной структурой (например, веб-энциклопедии).

VII. Выводы

В работе предложен способ вычисления семантической близости между словами, который предлагается в качестве альтернативы использованию известной семантической сети WordNET. Метод основан на сборе контекстных множеств термов в форме набора слов. Экспериментальные результаты показывают, что метод является эффективным. Однако задача подбора контекстных множеств термов является достаточно сложной. Метод может быть использован в системах интеллектуального поиска, системах формирования персональных рекомендаций, например, для решения проблемы лексической многозначности.

Результаты данного исследования так же были опробованы в алгоритме персональных рекомендаций в области поиска работы [11].

В качестве мер оценок результатов использовались *F-measure* [9] и *purity* [10]:

$$F - measure = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (8)$$

где *precision* - количество правильных результатов, *recall* - общее количество результатов.

$$purity(W, C) = \sum_k \max_j |W_k \cap C_j| \quad (9)$$

где *W* - множество документов, *C* - множество категорий - множество документов, отнесенных классификатором к категории *k*, - множество документов, отнесенных к категории *j* экспертом.

Меры, описанные формулами (8) и (9) показывают, насколько результаты работы разработанного классификатора соответствуют представлениям эксперта в предметной области. В таблице 7 представлены результаты данных оценок. Взяты средние значения оценок 1500 текстов:

Таблица 7. Оценка результатов работы

Предложенная мера близости		Семантическая близость методом Леска с помощью WordNET	
F-measure	Purity	F-measure	Purity
0.65	0.66	0.31	0.33

Результаты экспериментов показывают, что использование предложенной меры семантической близости помогает улучшить результаты работы классификатора по сравнению с классическим способом вычисления семантической близости на основе данных из WordNET. В среднем подобный подход к вычислению семантической близости дает на 8-10% более точный результат. Это связано с тем, что данная модель менее чувствительна к «шумам» за счет настройки весовых коэффициентов с помощью вычисления семантической близости. Новые весовые коэффициенты векторов документов учитывают контекст появления термов. Высокие веса связаны с терминами, которые семантически связаны между собой.

Кроме того, эксперименты были проведены над выборками разного рода и объема, на всех из них метод отработал достаточно эффективно. Так же одна из выборок была распределена неравномерно, метод и на ней показал хороший результат в то

время, как результаты векторной модели без учета семантической близости термов оказались неудовлетворительными.

В качестве недостатков метода можно выделить:

- сложность вычисления контекстных множеств;
- большой объем промежуточных вычислений.

Эффективность метода состоит в том, что он:

- требует малый объем вычисленных данных, которые необходимо хранить;
- имеет высокую скорость получения результата;
- метод оптимизирован для реляционных хранилищ.

VIII. Литература

- [1] Forsythe G. E., Malcolm M. A., Moler C. B. Computer Methods for Mathematical Computations // Prentice-Hall. | 1977.
- [2] Зубов А., Зубова И. Основы искусственного интеллекта для лингвистов. - Москва : Логос, 2006.
- [3] Resnik P. Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language // Journal of Artificial Intelligence Research. | 1999. | No. 11. | Pp. 95-130.
- [4] Cilibrasi R. L., Vitanyi P. M. The Google Similarity Distance, ArXiv.org or Clustering by Compression // IEEE Trans. Information Theory. [2004. | No. 51. | Pp. 1523-1545.
- [5] Lesk M. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone // SIGDOC 86. Proceedings of the 5th Annual International Conference on Systems Documentation. | 1986. | Pp. 24{26. [DOI:10.1145/318723.318728.Lesk:1986:ASD:318723.318728.
- [6] Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. [2005. | ISBN 0-321-32136-7.
- [7] Бондарчук Д. В., Тимофеева Г. А. Выделение семантического ядра на основе матрицы корреспонденций термов // Системы управления и информационные технологии. — 2015. — Т. 61, № 3.1. — С. 134—139.
- [8] Willett P. The Porter stemming algorithm: then and now // Program: Electronic Library and Information Systems. | 2006. | Vol. 40, no. 3. | Pp. 219-223.
- [9] Powers D. M. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation // Journal of Machine Learning Technologies. | 2011. | 2(1). | Pp. 37{63.
- [10] Xiong H., Wu J., Chen J. K-means clustering versus validation measures: A data distribution perspective // In KDD. | 2006.
- [11] Бондарчук Д. В., Тимофеева Г. А. Применение машинного обучения для формирования персональных рекомендаций в сфере трудоустройства // Экономика и менеджмент систем управления. — 2015. — Т. 18, № 4.2. — С. 215—221.
- [12] Бондарчук Д. В. Алгоритм построения семантического ядра для текстового классификатора // В мире научных открытий. — 2015. — Т. 68, № 8.2. — С. 713—724.