Обнаружение тропов на основе дифференциации смыслов♣

© Д.Н. Богданова, Б.А. Новиков

Санкт-Петербургский государственный университет

dasha.bogdanova@gmail.com, borisnov@acm.org

Аннотация

Задачи, связанные с автоматическим анализом текстов на естественном языке, требуют особого внимания к различным характеристикам текста. Тропы и другие средства выразительности часто бывают полезны для формирования таких характеристик, но их алгоритмическое извлечение является очень сложным. В данной работе мы предлагаем метод, определяющий по данному контексту, является ли слово или выражение употребленным в переносном смысле. Метод основан оценке близости смыслов по расстоянию между наборами документов, содержащих исследуемое выражение и его контекст. В работе экспериментально определяются значения параметров, которые позволяют выделить существенные различия, и затем оценивается качество результатов применения метода.

1 Введение

Задачи, связанные с автоматическим анализом текстов на естественном языке, требуют особого внимания к различным характеристикам текста. Осмысленность алгоритмов, направленных на решение этих задач, во многом зависит от того, какие характеристики принимаются во внимание. Например, в задачах психолингвистики целесообразно уделять внимание характеристикам, отражающим психофизиологические особенности личности автора, а в задаче автоматического определения авторства - отражающим идиостиль. Чаще всего при автоматической обработке текста, например, при классификации или кластеризации документов, используется векторная модель представления текста [14], то есть текст рассматривается как вектор из своих характеристик, в качестве которых нередко выступают частоты различных знаков препинания, средние длины слов и предложений и т. д. Несомненным достоинством таких низкоуровневых характеристик является легкость их извлечения, послужившая причиной их повсеместного использования, например, большинство работ по автоматическим методам определения авторства основано на подобных характеристиках. С другой стороны, в работе [1] показано, что использование низкоуровневых характеристик для решения этой задачи часто ведет к неудовлетворительным результатам. К тому же, лингвистический подход к неавтоматическому определению авторства учитывает такие высокоуровневые характеристики, как особенности использования стилистических фигур, звуковых приёмов (аллитераций, анафор и др.), наличие речевых ошибок (в том числе и в качестве средств выразительности), а также особенности восприятия автора (например, для Гоголя характерен “взгляд сверху” на происходящее) и т. д. Эти характеристики отражают авторский стиль более точно, но их алгоритмическое извлечение является очень сложным. В связи с этим было сделано крайне мало попыток решения задачи автоматического определения авторства на основе высокоуровневых характеристик [17].

Долгосрочной целью нашего исследования является выделение высокоуровневых семантически насыщенных конструкций из текстов на естественном языке. Помимо задачи определения авторства, выделение таких конструкций является важным для большинства задач обработки текстов на естественном языке, поскольку такие высокоуровневые конструкции, как, например, тропы или другие средства выразительности, являются неотъемлемой частью языка. В частности, такими являются задачи, связанные с вопросно-ответными системами, в которых требуется умение правильно интерпретировать выражения, употребленные переносном смысле. Задача выделения тропов на основе дифференциации смыслов, решаемая в данной работе, является одной из задач, направленных на решение обозначенной выше проблемы.

Мы предлагаем решать такую задачу с помощью метода, основанного не следующей гипотезе, названной нами идеей дифференциации смыслов: *существенное различие смысла выражения и смысла его контекста указывает на то, что выражение употреблено в переносном смысле*.

Труды 12[й] Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2010, Казань, Россия, 2010

Целью данной работы является исследование этого метода, включающее в себя проверку гипотезы, на которой он основан, и оценивание качества его работы.

Необходимо подчеркнуть, что в данной работе нас интересует именно различие смыслов, а не понимание смысла текста.

2 Постановка задачи

Задачей, решаемой в данной работе, является автоматическое определение по данному контексту выражения, является ли оно употребленным в переносном смысле. Употребления слов в переносном смысле могут быть представлены, в частности, метафорами (например, "*британский лев* мягко ступает по российскому правовому пространству"), метонимиями ("у нас в семье восемь *ртов*"), метафорическими эпитетами ("хочу я слов *кинжальных*").

Понятие "употребления в переносном смысле" тесно связано с понятием контекста, поскольку выражение можно назвать употребленным буквально или метафорически только относительно контекста. Например, такое физическое понятие, как импеданс, употребленное в контексте литературы по базам данных (impedance mismatch), безусловно, является метафорой. Как было упомянуто во введении, мы предлагаем решать описанную выше задачу с помощью метода, основанного на идее сравнения смыслов контекста и самого выражения.

3 Обзор существующих подходов

Задача определения тропов рассматривалась многими исследователями.

В работе [16] предлагается отличать буквальные употребления выражений от их употреблений в переносном смысле по наличию семантических связей (cohesive links) между словами рассматриваемого выражения и окружающего контекста. При наличии таких связей алгоритм считает рассматриваемое выражение употребленным буквально, а при отсутствии – употребленным в переносном смысле. Для большинства рассмотренных в работе выражений точность превышает 50% (точность для различных выражений варьируется от 11% до 98%). В работе [9] авторы предлагают улучшение своего алгоритма, добавляя в качестве второго этапа классификацию, основанную на методе опорных векторов: обучающая выборка для алгоритма классификации формируется описанным выше способом. Таким образом, авторам удалось повысить среднюю точность до 90%. Алгоритм, предлагаемый нами, также основан на предположении, что слабая семантическая связь между выражением и его контекстом указывает на то, что выражение употреблено не в буквальном смысле. Но в отличие от вышеописанных работ, рассматривающих фразеологизмы, алгоритм, предлагаемый нами, направлен в большей степени не на выделение устойчивых выражений (*кот в мешке*, *водить за нос* и т. д.), а на выделение авторских тропов (*в горах моё сердце*, *живой костёр из снега и вина* и т. д.). Более того, для работы улучшенного алгоритма [9] требуется обучающая выборка, алгоритм, предлагаемый в данной работе, не требует обучающей выборки.

В [8] представлен алгоритм, который также является методом контролируемого обучения. Рассматриваемая задача формулируется как задача классификации: авторы строят обучающую выборку, а затем применяют алгоритм ближайших соседей (k-Nearest Neighbor Classification, [6]).

В исследовании [4] для решения описанной задачи формируются два набора текстов (seed sets): первый набор состоит из различных буквальных употреблений выражений, а второй – из употреблений в переносном смысле. Авторы вычисляют семантическую близость между рассматриваемым контекстом и двумя описанными наборами и определяют, как употреблено выражение – буквально или в переносном смысле, исходя из того, расстояние до какого из соответствующих наборов меньше. Алгоритм, описанный в [4], так же как и подход, предлагаемый в данной работе, обращается к концепциям из области Разрешения Лексической Неоднозначности (Word Sense Disambiguation). Но этот алгоритм, в отличие от нашего, направлен только на выделение употреблённых метафорически глаголов. Авторы формулируют вышеописанную задачу как задачу разрешения лексической неоднозначности одного слова (one word или targeted disambiguation) и применяют конкретный существующий метод, в то время как мы лишь пользуемся идеями из данной области.

Как уже было упомянуто, наш алгоритм должен уметь определять по контексту, является ли выражение употребленным в переносном смысле. При этом употребления в переносном смысле могут быть выражены, например, метафорами, метонимиями, метафорическими эпитетами и т. д. Стоит заметить, что в литературе встречаются алгоритмы, направленные на выделение конкретных тропов, например, в работе [13] описан метод машинного обучения для обнаружения метонимий. Авторы формулируют задачу выделения метонимий как проблему классификации: употребление выражения нужно либо определить как буквальное, либо отнести к одному из предопределенных типов метонимии (например, "место как человек", "место как событие" и т. д.).

4 Подход к решению задачи

Мы можем рассматривать переносное значение слова (выраженное, например, метафорой) как дополнительное, чаще всего не включенное ни в один словарь. Хотя нужно заметить, что некоторые метафоры в словарях встречаются, например, *игольное ушко* или *горлышко бутылки*. Такие метафоры-катахрезы появляются, чтобы восполнить пробелы в языке [5], и никак не отражают идиостиль: автор использует их только потому, что упоминаемый

объект нельзя назвать по-другому. Такие метафоры не очень интересны с точки зрения задачи изучения авторского стиля и извлечения семантически насыщенных характеристик текста, поэтому разрабатываемый нами подход не уделяет внимания метафорам такого рода.

Наш подход к выделению употреблений слов в переносном значении основан на идее дифференциации смыслов.

4.1 Представление смысла

Определение смыслов не является нашей целью, т. к. мы не решаем задачу понимания текста. Имея выражение и его контекст, мы пытаемся выяснить не сами смыслы, а то, насколько они близки или далеки друг от друга. Тем не менее нам требуется представление смысла. Чтобы получить такое представление, мы обращаемся к идее, на которой основано большинство исследований в области Разрешения Лексической Неоднозначности (Word Sense Disambiguation). Смысл предлагается рассматривать как совокупность употреблений (sense is a group of contextually similar occurrences of a word) [15]. Таким образом, мы представляем смыслы слова и его ближайшего контекста как наборы документов, связанных со словом и окружающим это слово текстом соответственно. Такое представление, конечно, не может учитывать субъективные оттенки смысла, однако дает возможность объективной оценки того, каким образом слово или выражение используется.

Такие наборы текстов мы получаем с помощью поисковой системы. Для слова в качестве запроса можно использовать его само. Для текста запрос можно формировать как текст целиком или как более значимые его части: в качестве таких частей мы предлагаем использовать выделенные из текста лексические цепи, определенные в работе [12].

4.2 Различие смыслов

Имея представление смысла, мы должны выяснить, как измерять различие между смыслами и какое различие является достаточно сильным для того, чтобы считать рассматриваемое слово или выражение употребленным в переносном смысле. Так как смыслы представлены наборами документов, мы можем измерять различия между ними как расстояние между векторными представлениями [14] этих наборов документов. В нашей работе мы предлагаем использовать косинусную меру близости, широко используемую в задачах автоматического анализа текстов [3]. Так как рассматриваемая нами мера является мерой близости, различие между наборами зависит от нее обратно пропорционально.

Более перспективный способ нахождения расстояния между наборами документов предложен в [11]: авторы предлагают использовать текстовую меру близости (text-to-text similarity), которая основана на вычислении семантической близости между словами (word semantic similarity), входящими в тексты. Обычно подобные меры семантической близости между словами зависят от длины пути между соответствующими синсетами[1] в WordNet'е [19]. К сожалению, использование подобной меры близости недоступно для русского языка из-за отсутствия WordNet-подобных словарей.

5 Эксперименты

Мы разделили эксперименты на два этапа: на первом этапе мы проверяем гипотезу, на которой основан исследуемый нами метод. В случае принятия гипотезы требуется также оценить параметры предлагаемого метода, то есть выяснить, какое различие является достаточно сильным для того, чтобы считать рассматриваемое слово или выражение употребленным в переносном смысле. Второй этап экспериментов направлен на оценку качества полученного алгоритма.

5.1 Обучение алгоритма

Для первого этапа экспериментов мы выбрали несколько выражений (в т. ч. состоящих и из одного слова), для которых возможно как буквальное, так и метафорическое употребление. Для каждого такого выражения мы подобрали по два контекста: в одном из них выражение употреблялось буквально, а в другом – в переносном смысле. Из каждого полученного контекста мы удалили само целевое выражение.

Мы условились считать набор текстов, семантически связанных с данным, за представление смысла данного текста. Для формирования этого набора мы рассмотрели два различных способа.

5.2 Использованиелексических цепей

Первый способ основан на выделении лексических цепей. Лексическая цепь, согласно [12], представляет собой набор семантически связанных слов в тексте. Рассмотрим, например, следующий отрывок из рассказа Рэя Брэдбэри: "*Огоньки останавливающегося по требованию пассажиров полуночного экспресса плясали на рельсах. Только когда поезд отошел от станции, я выглянул в дверь пульмановского вагона и посмотрел назад*". В этом отрывке существительные "*экспресс*", "*рельсы*", "*поезд*", "*вагон*", "*пассажиры*", "*станция*" образуют лексическую цепь. Лексические цепи широко применяются для разрешения лексической неоднозначности [7] и реферирования текстов [2].

Мы производили процесс извлечения таких цепей из текстов вручную, но он может быть автоматизирован с помощью, например, методов, представленных в [7, 10]. Далее мы сформировали набор семантически близких к данному текстов из результатов, возвращенных поисковой системой по запросу, совпадающему с самой длинной из найденных лексических цепей. При отсутствии в контексте

[1] элементарная единица тезауруса WordNet, представляющая собой набор синонимических лексем

очевидно семантически связанных слов, мы формировали запрос из двух существительных, ближайших к рассматриваемому выражению. Для наших экспериментов мы использовали Google, хотя выбор поисковой системы не является принципиальным. Возможно, использование другой поисковой системы могло бы дать лучшие результаты, однако контрольные запросы показывают, что различия в наборах найденных документов не очень значительны.

Таблица 1. Значения косинусной меры близости при использовании лексических цепей

	5 текстов		**10 текстов**		**15 текстов**	
	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)
вьюга	0,05	0,17	0,09	0,21	0,12	0,27
дыхание	0,11	0,16	0,13	0,20	0,15	0,21
кинжальный	0,08	0,10	0,08	0,12	0,10	0,13
плясать	0,09	0,16	0,08	0,17	0,12	0,17
стебель гибкий	0,09	0,11	0,10	0,21	0,14	0,21
утонул	0,05	0,17	0,08	0,20	0,12	0,21
хрустальными	0,07	0,21	0,06	0,21	0,09	0,21
шотландской волынкой	0,09	0,29	0,10	0,27	0,12	0,36
мед	0,04	0,15	0,06	0,21	0,09	0,25
лекарство	0,08	0,16	0,10	0,20	0,11	0,22

Таблица 2. Значения косинусной меры близости при использовании грамматических основ

	5 текстов		**10 текстов**		**15 текстов**	
	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)	Косинусная мера (переносный смысл)	Косинусная мера (буквальный смысл)
вьюга	0,11	0,12	0,15	0,18	0,19	0,20
дыхание	0,10	0,13	0,09	0,14	0,09	0,14
кинжальный	0,06	0,09	0,08	0,12	0,09	0,12
плясать	0,09	0,17	0,15	0,20	0,17	0,23
стебель гибкий	0,11	0,11	0,15	0,20	0,15	0,19
утонул	0,10	0,14	0,16	0,20	0,16	0,20
хрустальными	0,08	0,10	0,08	0,12	0,09	0,14
шотландской волынкой	0,06	0,20	0,09	0,23	0,10	0,25
мед	0,08	0,09	0,09	0,14	0,09	0,14
лекарство	0,11	0,14	0,14	0,18	0,15	0,18

5.3 Представление смысла контекста с помощью грамматических основ

Второй способ получения набора основан на выделении грамматических основ из предложений. Так как во многих случаях при запросе, совпадающем с выделенными основами, несколько первых десятков результатов поиска содержали только сам рассматриваемый текст, мы добавили к запросу требование об отсутствии в результатах целевого выражения.

5.4 Сравнения представлений смыслов. Оценка параметров алгоритма

В качестве представления смысла самого выражения мы также взяли результаты, возвращенные поисковой системой по запросу, совпадающему с этим выражением. Поиск проводился по точной фразе.

Как было упомянуто, для сравнения текстов мы выбрали косинусную меру близости. Предварительно все тексты прошли процедуру стемминга с помощью стеммера Snowball [18] для русского языка.

Результаты, представленные в таблицах 1, 2, показывают, что наборы, представляющие смыслы

буквальных употреблений, действительно ближе к смыслам своих контекстов, чем смыслы употреблений метафорических, что подтверждает гипотезу, на которой основан исследуемый метод. Результаты показывают, что различие значений более заметно при использовании лексических цепей (табл. 1), чем при использовании грамматических основ (табл. 2). При этом различие лучше прослеживается на наборе из 10 первых результатов. В этом случае для 9 из 10 контекстов выражений, употребленных метафорически, значение косинусной меры составляет 0,1 и менее. А для 8 из 10 буквальных контекстов – 0,2 и более. Таким образом, мы принимаем значение косинусной меры, равное 0,1 и менее, за признак существенного различия наборов текстов, а значение, равное 0,2 или более, предлагаем считать признаком буквального употребления.

В результате первого этапа экспериментов мы подтвердили гипотезу, на которой основан исследуемый метод, и выяснили, что построение набора семантически близких текстов, являющегося представлением смысла контекста, на основании лексических цепей показывает лучшие результаты, чем построение с использованием грамматических основ. При этом различие смыслов наиболее четко прослеживается при рассмотрении 10 первых результатов поиска. В качестве различия, достаточного для того, чтобы считать рассматриваемое в контексте выражение употребленным в переносном значении, мы принимаем достаточно малое значение косинусной меры близости между наборами, представляющими смысл самого выражения и смысл контекста. За достаточно малое значение косинусной меры мы принимаем 0,1 и менее.

5.5 Оценка качества алгоритма

На втором этапе экспериментов мы выбрали несколько контекстов различных выражений и применили к ним уточненный алгоритм, полученный в результате первого этапа. Как показано на рисунке, 67 % метафорических употреблений было классифицировано правильно, 8 % – неправильно, оставшиеся 25 % попали в *область неопределенности* (значение косинусной меры от 0,1 до 0,2). В случае буквальных употреблений правильно было классифицировано 62 %, 23 % – неправильно, 15% попало в область неопределенности.

6 Заключение

В данной работе мы предложили и исследовали метод обнаружения выражений, употребленных в переносном смысле. Метод основан на идее дифференциации смыслов, заключающейся в том, что существенное различие между смыслом выражения и смыслом окружающего его контекста указывает на то, что выражение в этом контексте употреблено в переносном смысле. Мы рассмотрели два способа представления смысла контекста – с помощью лексических цепей и грамматических основ – и выяснили, что представление смысла на основе лексических цепей показывает лучшие результаты в данной задаче. Мы провели два этапа экспериментов: на первом этапе подтвердили гипотезу, на которой основан метод, и оценили параметры алгоритма. На втором этапе мы рассмотрели работу алгоритма с учетом параметров, полученных на первом этапе, и выяснили, что алгоритм показывает хорошую точность.

Одним из направлений дальнейшей работы является более глубокое изучение проблемы автоматического выделения метафорических употреблений, в частности, разработка алгоритма *извлечения* подобных употреблений (нахождения в произвольном тексте выражений, употребленных метафорически) с учетом опыта данного исследования.

Мы также планируем рассмотреть задачи выделения других высокоуровневых конструкций, в частности, конструкций, интересных с точки зрения определения индивидуального авторского стиля.

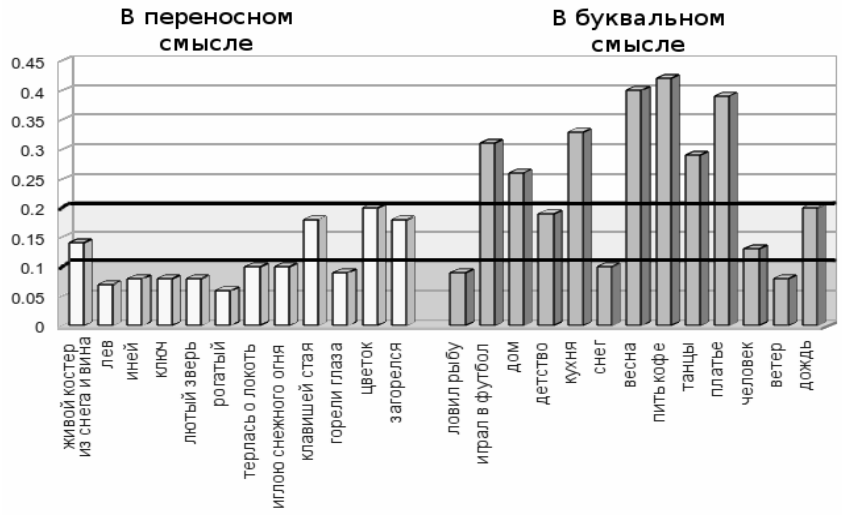

Рис. Значения косинусной меры близости для различных выражений

Литература

[1] Батов В.И., Сорокин Ю.А. Атрибуция текста на основе объективных характеристик// Серия языка и литературы. – 1975. – Т. 34, № 1.

[2] Barzilay R., Elhadad M. Using lexical chains for text summarization// Proc. of the Intelligent Scalable Text Summarization Workshop, 1997.

[3] Berry M.W. Survey of text mining: clustering, classification, and retrieval. – Springer, 2003.

[4] Birke J., Sarkar A. A clustering approach for the nearly unsupervised recognition of nonliteral language// Proc. of EACL-06, 2006.

[5] Black M. Metaphor// Proc. of the Aristotelian Society. – 1954. – P. 273-294.

[6] Cover T.M., Hart P.E. Nearest neighbor pattern classification// IEEE Transactions on Information Theory. 1967. – V. 13, No 1. – P. 21-27.

[7] Galley M., McKeown K. Improving word sense disambiguation in lexical chaining// Proc. of IJCAI-2003.

[8] Katz G., Giesbrecht E. Automatic identification of non-compositional multiword expressions using latent semantic analysis// Proc. of the ACL/COLING-06 Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, 2006.

[9] Li L., Sporleder C. Classifier combination for contextual idiom detection without labelled data// Proc. of the 2009 Conf. on Empirical Methods in Natural Language Processing, 2009. – P. 315-323.

[10] Medelyan O. Computing lexical chains with graph clustering//ACL 2007, 2007.

[11] Mihalcea R., Corley C., Strapparava C. Corpus-based and knowledge-based measures of text semantic similarity// Proc. of AAAI-06, 2006.

[12] Morris J., Hirst G.. Lexical cohesion computed by thesaural relations as an indicator of the structure of text// Computational Linguistics. – 1991. – V. 17, No 1. – P.21-43.

[13] Nissim M., Markert K. Syntactic features and word similarity for supervised metonymy resolution// Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03), 2003. – P. 56-63.

[14] Salton G, Wong A, Yang C.-S. A vector space model for automatic indexing// Communications of the ACM. – 1975. – V. 18. – P. 613-620.

[15] Schutze H. Automatic word sense discrimination// Computational Linguistics. – 1998. – V. 24, No 1. – P. 97-123.

[16] Sporleder C., Li L. Unsupervised recognition of literal and non-literal use of idiomatic expressions// Proc. of EACL-09, 2009.

[17] Stamatatos E. A survey of modern authorship attribution methods// J. of the American Society for Information Science and Technology. – 2009. – V. 60, No 3. – P. 538-556.

[18] The Snowball Home Page. – http://snowball.tartarus.org/.

[19] The WordNet Home Page. – http://wordnet.princeton.edu/.

Figurative language detection techniques based on the sense differentiating

Daria Bogdanova, Boris Novikov

In this paper, we propose an idea of sense differentiation and state a figurative language detection techniques based on this idea. We provide a study of the proposed algorithm: at first, we test the idea of sense differentiation and estimate the parameters of the algorithm, then we evaluate the obtained algorithm. Our experiments show that the proposed techniques provide acceptable precision.

♣Работа выполнена при частичной финансовой поддержке РФФИ (проект 10-07-00156) и гранта компании Google