РАСШИРЕННЫЙ ЭКСПЕРИМЕНТ ПО АВТОМАТИЧЕСКОМУ ОБНАРУЖЕНИЮ И ИСПРАВЛЕНИЮ РУССКИХ МАЛАПРОПИЗМОВ

AN EXTENDED EXPERIMENT ON AUTOMATIC DETECTION AND CORRECTION OF RUSSIAN MALAPROPISMS

Е. И. Большакова *(bolsh@cs.msu.su)*
Московский государственный университет им. М.В. Ломоносова

И. А. Большаков *(igor@cic.ipn.mx)*
Национальный политехнический институт (IPN), Мексика

А. П. Котляров *(koterpillar@gmail.com)*
Московский государственный университет им. М.В. Ломоносова

Малапропизм – это семантическая ошибка, при которой одно знаменательное слово заменяется в тексте другим, близким по звучанию, но с иным смыслом. Обсуждаются результаты расширенного эксперимента, который тестирует предложенный ранее метод обнаружения и исправления малапропизмов, базирующийся на интернет-статистике и числовом показателе смысловой совместимости. Результаты подтверждают высокие качества метода и уточняют введенные ранее пороговые правила.

Введение

Обнаружение в текстах семантических ошибок пока не выполнимо современными автокорректорами и редактирующими системами. К лексико-семантическим ошибкам относятся так называемые ***малапропизмы*** – речевые ошибки, при которых одно знаменательное слово текста заменяется другим, близким по звучанию, но отличным по смыслу и потому обычно не соответствующим контексту, например, ***трафик*** *ввода*, ***сачок*** *цен*, *начать* ***компанию***.

Малапропизмы возникают из-за случайных описок (*сачок* вместо *скачок*) или явных лексических ошибок (*деловой* вместо *деловитый*). Сложность их автоматического выявления связана с тем, что здесь одно существующее (словарное) слово заменяется другим словарным, причем в той же синтаксической роли и с теми же морфологическими характеристиками (число, род, падеж, лицо). Для высокоточного автоматического выявления малапропизмов нужен синтаксический и семантический анализ текста, но в современной компьютерной лингвистике намечены и более доступные пути решения этой задачи [1, 2, 3].

В статьях [2, 3] описывается метод автоматического обнаружения малапропизмов и автоматизированного их исправления, основанный на проверке смысловой совместимости слов. Предполагается, что малапропизмы разрушают семантическую связь внутри пар слов, но сохраняют их синтаксическую связанность. Для обнаружения ошибок рассматриваются все пары синтаксически связанных в предложении знаменательных слов, и проверяется их смысловая совместимость.

В работе [3] для проверки смысловой связи предложено использовать числовой ***показатель смысловой совместимости*** (ПСС), вычисляемый по статистическим данным интернетовского поисковика о встречаемости слов проверяемого словосочетания, измеренной вместе и раздельно. При этом фактически интернет используется как текстовой корпус, огромный, но с большим информационным шумом [4]. Малапропизм считается обнаруженным, если для пары синтаксически связанных знаменательных слов предложения значение ПСС оказывается ниже установленного порога, при этом малапропизмом считается обнаруженная пара слов целиком. Для исправления найденных малапропизмов предложено использовать заранее составленные словари паронимов (слов, сходных буквами, или слов с одинаковыми корнями). Из этих словарей, достаточно подробно описанных в [5], берутся кандидаты на исправление ошибочных слов, и для выявления лучших кандидатов применяются основанные на ПСС эвристические пороговые правила. Первые проведенные эксперименты показали обнадеживающие результаты по обнаружению и исправлению малапропизмов.

Данная статья продолжает работу [3], существенно расширяя эксперименты с российским поисковиком Яндекс с целью дополнительного обоснования предложенного метода. В экспериментах использовался набор из 370 русских малапропизмов всех часто встречающихся в русском языке типов словосочетаний. Поскольку слова-компоненты словосочетаний в текстах могут быть разделены другими словами, при проведении экспериментов были учтены наиболее вероятные расстояния между словами, заранее определенные на основе яндекс-

статистики. Результаты экспериментов подтвердили ранее выявленные закономерности распределений значений ПСС и в то же время позволили уточнить эвристические пороговые правила метода.

Метод обнаружения и исправления малапропизмов

Обсуждаемый метод опирается на понятие ***коллокации***. В зарубежной и отечественной компьютерной лингвистике нет единого точного понимания этого термина. Для наших прикладных целей мы считаем коллокацией синтаксически связанную и семантически допустимую пару знаменательных слов предложения (например, *начинает урок*). Если нарушается хотя бы одно из указанных условий, соответствующая пара словоформ коллокацией не считается. К примеру, в паре *он рисует* первое слово не является знаменательным, в *великий художница* знаменательные слова синтаксически не согласованы, в *кипящая цель* слова семантически не совместимы.

Компонентами русских коллокаций могут быть слова, играющие роль одной из четырех главных частей речи: существительных, глаголов, прилагательных (включая причастия), или наречий (включая деепричастия). Синтаксические связи между знаменательными словами могут реализоваться непосредственно или через служебные слова (обычно – через предлоги, например, *идти в лес*). Поэтому мы включаем такие служебные слова в соответствующие коллокации.

Внутри предложений компоненты коллокации могут быть разделены не только предлогами, но и многими другими словами, обычно от них зависящими. Так, во фрагменте *дала ему короткое рекомендательное письмо* компоненты коллокации *дала письмо* разделены тремя другими словами, подчиненными одной из компонент.

Нами рассматриваются малапропизмы, разрушающие коллокации, точнее, нарушающие семантическую связь между словами при сохранении их синтактической связи. Главная идея обнаружения малапропизмов – просмотр всех пар знаменательных слов в анализируемом предложении текста с проверкой их на синтаксическую связанность и семантическую совместимость. Если пара синтаксически связана, но семантически несовместима, сигнализируется малапропизм.

Для исправления ошибочного слова в малапропизме привлекаются ***паронимичекие*** словари – ***буквенный*** и ***морфемный***. Первый содержит группы русских однобуквенных паронимов, т.е. слов, отличающихся одной буквой от заглавного слова группы: к примеру, для заглавного слова *каска* в группу входят *качка*, *кашка*, *краска* и др. Второй словарь содержит группы русских морфемных паронимов – слов, отличающихся немногими служебными морфами: например, *человечный*, *человеческий*, *очеловеченный* и др. Если первый словарь пригоден для исправления однобуквенных ошибок (как правило, случайных), то второй – для исправления обычно неслучайных многобуквенных ошибок, связанных с употреблением слов с одинаковыми корнями.

Поскольку не известно, какое слово в обнаруженном малапропизме ошибочно, то делаются попытки исправить оба компонента по отдельности. Для этого с помощью обоих паронимических словарей составляются всевозможные исправляющие пары с варьированием обоих компонентов. При этом из морфемного словаря берутся слова, отличающиеся не более чем двумя морфами. Полученные пары образуют список ***первичных кандидатов***. Среди первичных кандидатов есть ровно один, соответствующий нарушенной малапропизмом исходной коллокации; этот кандидат назван ***истинным исправлением***.

Каждый первичный кандидат проверяется на смысловую допустимость, и все не выдерживающие теста пары отбрасываются. В итоге остается список ***вторичных*** кандидатов, он упорядочивается, и из него берутся ***лучшие***. Они и предъявляются человеку-редактору текста, за которым остается окончательное решение об исправлении малапропизма.

Как и в [3], для проверки смысловой совместимости пары слов (V, W) мы используем ***Показатель Смысловой Совместимости*** (ПСС), высчитываемый по данным интернетовского поисковика – количеству $N(V)$ и $N(W)$ страниц, на которых слова встретились по отдельности, и количеству страниц $N(V, W)$ их совместной встречаемости:

$$ПСС(V, W) \equiv \begin{cases} 16 + \log_2 N(V, W) - (\log_2 N(V) + \log_2 N(W)) / 2 & \text{если } N(V, W) > 0, \\ -\infty & \text{если } N(V, W) = 0, \end{cases}$$

Здесь $-\infty$ – большая по величине отрицательная константа. По сравнению с [3], аддитивная константа и основание логарифма выбраны в этой формуле так, чтобы большинство коллокаций любого вида, интуитивно принимаемых за устойчивые, не попадали в отрицательную область [6]. Таким образом, устойчивость лексико-семантической связи слов V и W грубо определяется условием $ПСС(V, W) > 0$.

Быстрый рост массивов интернета показывает, что практически любое семантически допустимое словосочетание в конце концов появляется в интернете и притом несколько раз, так что мы в праве считать семантически допустимыми все словосочетания, число появлений которых превысило довольно низкий порог, близкий к нулю.

Название типа коллокации	Синтаксическая структура подтипа	Примеры коллокаций	% в наборе
определяемое слово → определитель	Adj ← N Adv ← Adj Adv ← V Adj → Adv V → Adv	полевую форму вполне удачный торопливо шел повернутые налево пройдут нормально	35%
существительное → его дополнение	$N \rightarrow N_{доп}$ $N \rightarrow Prep \rightarrow N_{доп}$	рост возмущения отчетов о доходах	18%
глагол → его дополнение	$V \rightarrow N_{доп}$ $V \rightarrow Prep \rightarrow N_{доп}$ $N_{доп} \leftarrow V$ $V \rightarrow V_{инф}$ V → Adj Adj ← V	заметить разницу прибрала к рукам плешь переели решили продать было забыто главным остается	29%
сказуемое → его подлежащее	$N_{им} \leftarrow V$ $V \rightarrow N_{им}$ $Adj_{крат} \rightarrow N_{им}$ $N_{им} \leftarrow Adj_{крат}$ $Adv \rightarrow V_{инф}$	судно пропало затонула лодка отправлен груз доклад проверен нужно ожидать	7%
прилагательное → его дополнение	Adj → Prep → $N_{доп}$ ↓ ⌐ ⌐ Prep → $N_{доп}$ Adj Adj → $N_{доп}$ $N_{доп} \leftarrow$ Adj Adj → $V_{инф}$	дошедший до ручки по сути неверный нагруженный лесом плащом закрыты умеющих выживать	6%
сочиненная пара	N → *и* →N Adj → *и* → Adj V → *и* → V	ходы и выходы наземный и воздушный грабить и убивать	5%

Таблица 1. Типы и синтаксические структуры коллокаций

Малапропизм и первичные кандидаты	**Значение ПСС**	**Лучшие кандидаты со значениями ПСС**
1L упрощение грехов	-∞	прощение грехов 10,77
1L опрощение грехов	3,17	опрощение грехов 3,17
1L!! прощение грехов	10,77	укрощение грехов –1,64
1L укрощение грехов	-1,64	
1L уплощение грехов	-∞	
1L упрочение грехов	-∞	
2L упрощение греков	-∞	
2L упрощение огрехов	-∞	
2L упрощение орехов	-∞	
2L скверный ветер	2,44	северный ветер 9,72
2L скверный веер	-∞	скверный вечер 1,93
2L!! скверный вечер	1,93	
1L северный ветер	9,72	
1L отчистить дороги	-3,85	очистить дороги 3,75
1L обчистить дороги	-∞	
1L отчестить дороги	-∞	
1L!! очистить дороги	3,75	
2L отчистить вороги	-∞	
2L отчистить пороги	-∞	

Рисунок 1. Примеры образцов малапропизмов и кандидатов на их исправление со значениями ПСС

Пара слов (Vm, Wm) признается малапропизмом, если ПСС (Vm, Wm) < P, где P – положительная константа, подбираемая экспериментально. Первичный кандидат (V, W) на исправление считается вторичным, если срабатывает иное пороговое правило: ПСС (V,W) > Q, где Q – константа (-□ □ Q < 0), тоже подбираемая экспериментально. Вторичные кандидаты упорядочиваются по значению ПСС, и лучшими считаются несколько первых.

Составление расширенного набора малапропизмов

Использование ПСС как меры семантической связности требовало дополнительного экспериментального подтверждения. Расширенный набор русских малапропизмов содержит 370 образцов против 100 в [3]. Как и ранее, большинство образцов получены из текстов интернет-новостей путем фальсификации одной из компонент текстовых коллокаций с помощью словаря паронимов и с сохранением исходных морфологических характеристик. Остальные образцы подбирались так, чтобы обеспечить некоторое представительство в наборе морфемных ошибок (они встречаются в текстах реже), а также ошибок в коллокациях не привлекавшихся ранее типов. Были привлечены и сочиненные пары типа *шум и гам, пахнуть и цвести*.

Итоговый набор на 95% состоит из буквенных и на 5% из морфемных ошибок и содержит образцы шести синтаксических типов с 26 подтипами (они определяются частью речи компонентов, их порядком в тексте и наличием вспомогательного предлога). Типы и подтипы приведены в Таблице 1. Компоненты коллокаций помечаются как существительные *N*, прилагательные *Adj*, глаголы *V* и наречия *Adv*; индекс *им* маркирует существительное-подлежащее в именительном падеже; *доп* означает существительное-дополнение с падежом, зависящем от управляющего предлога *Prep* или непосредственно от управляющего слова; *инф* означает инфинитив; *крат* – краткую предикативную форму прилагательного.

В наборе оказалось достаточно много (42) образца ошибки, похожей на малапропизм и называемой нами ***квазималапропизмом***. Эта ошибка превращает одну существующую коллокацию с другую существующую, но, как правило, более редкую и противоречащую прочему контексту, например, *равные труппы* вместо правильного *равные группы*.

Для каждого образца из набора с помощью обоих паронимических словарей были составлены первичные кандидаты на исправление, всего 2688 кандидатов, т.е. в среднем 7,26 кандидата на малапропизм.

Несколько образцов набора представлены в первом столбце рисунка 1. Сначала идет сам малапропизм, далее идут строки с первичными кандидатами. Все строки содержат номер ошибочного слова (1 или 2) и символ использованного словаря (L – буквенный, M – морфемный). Второй из представленных образцов является квазималапропизмом. Знаком !! помечается истинное исправление.

Результаты экспериментов

В текстах интернета слова-компоненты тестируемого словосочетания могут быть разнесены. Сильно разнесенные компоненты могут оказаться случайной встречей этих слов, и интернет гарантирует правильные результаты только при близко расположенных компонентах. При этом для надежного обнаружения ошибок на основе статистического критерия разумно брать наиболее вероятное (частое) расстояние между компонентами.

Для выяснения наиболее вероятных расстояний нами были исследованы на нескольких примерах коллокаций разных типов частоты их совместной встречаемости в зависимости от расстояния между компонентами. Для получения яндекс-статистики (измеряемой в числе релевантных страниц) использовались запросы вида +"словоформа1" /(1 n) +"словоформа2" (например, +"столб" /(1 2) +"дыма"), выдающие частоту совместной встречаемости заданных словоформ, находящихся в одном предложении текста на расстоянии, не большем n (n=1 соответствует смежным словам). Полученные статистические данные, частично представленные в Таблице 2, показывают, что для всех типов коллокаций наибольшая совместная встречаемость их компонент достигается, когда они стоят либо рядом, либо через одно слово, причем оба эти случая обычно покрывают более 60% совместных выпадений (см. последний столбец таблицы).

Коллокация	Число промежуточных слов:					Процент в случаях 0 и 1
	0	1	2	3	4	
затонувшее судно	**10250**	496	189	642	128	92
сбор информации	**141395**	32342	54354	31326	13566	64
уделить внимание	52248	**72433**	9111	3537	1335	90
приведем пример	**30665**	13106	6343	1376	580	84
спасатели обнаружили	**18534**	2440	929	524	740	91
ходит слух	**15397**	1879	385	201	43	96
порт открыт	4926	**7916**	1802	1261	576	78

Таблица 2. Статистика совместной встречаемости компонентов коллокаций

Коллокация	Число промежуточных слов									
	0, 0	1, 0	0, 1	2, 0	1, 1	0, 2	3, 0	2, 1	1, 2	0, 3
тайники с оружием	**4775**	1	43	10	1	29	3	0	0	38
ворвались в здание	**9869**	123	156	69	0	74	24	1	9	2
справиться с управлением	**5744**	16	177	2	0	11	2	0	0	12
умер от ран	**19186**	379	1269	448	67	223	355	164	10	121
прописан в законе	**2120**	78	795	14	146	142	17	35	25	17

Таблица 3. Статистика совместной встречаемости компонентов коллокаций с предлогами

Просмотрев первые два десятка найденных при этом страниц, мы убедились, что подавляющее большинство найденных близкорасположенных слов действительно являются коллокациями, а не словами, случайно оказавшимися рядом. Тем самым в дальнейших экспериментах мы могли полагаться на то, что наиболее вероятным расположением компонент коллокаций в текстах является их нахождение рядом либо через одно слово.

Что же касается коллокаций, компоненты которых связаны через предлог, то собранная подобным же образом яндекс-статистика показала, что для них наиболее вероятным является смежное расположение всех трех слов. В Таблице 3 представлено распределение частоты встречаемости нескольких типичных коллокаций для разных комбинаций двух расстояний – между первым словом и предлогом и предлогом и вторым словом. Например, комбинация 0, 1 соответствует случаю, когда первое слово стоит рядом с предлогом, а второе отстоит от него на одно слово, например, *ворвались в горящее здание*.

Для проведения основного эксперимента была создана компьютерная программа, которая собирала яндекс-статистику встречаемости словоформ по раздельности и совместно, причем в последнем случае коллокации с предлогами запрашивались как смежные слова, а для коллокаций без предлогов запрашивались случаи соседства или расположения через одно слово.

В ходе эксперимента сначала для каждого образца набора и его первичных исправлений была собрана статистика встречаемости их компонент раздельно и совместно. В 193 случаях (58%) малапропизмы оказались вообще отсутствующими в массивах поисковика, а квазималапропизмы – в 7 случаях (18%).

По собранной статистике были вычислены значения ПСС, закономерность их распределений в целом совпала с указанной в [3]. Основные характеристики распределений для малапропизмов, квазималапропизмов и истинных исправлений представлены в таблице 4.

Простейшим решающим правилом было бы установление порога для малапропизмов, близким к максимальному встреченному для них значению ПСС, а именно 5,6. Тогда гарантировалась бы полнота обнаружения малапропизмов, но пострадала бы точность, поскольку многие редко встречающиеся коллокации были бы сочтены малапропизмами. Более разумным кажется установление порогового значения близким к величине $M–D$ для истинных исправлений (эта величина чуть больше 3). Конкретно, в качестве порога для малапропизмов было взято число 4. При этом только шесть малапропизмов набора остаются необнаруженными: *порыв трубы* (грубая, но очень частотная ошибка), *жизнь на Марксе*, *оральную поддержку* (оба образца часто используются на интернет-форумах), *начать компанию*, *виноват в преступлении* (опять очень частотные ошибки), *оправлено в отставку* (высокочастотная описка).

Несмотря на то, что при выборе порога квазималапропизмы не учитывались, наш метод обнаружил большинство из них: 29 из 42 (69%). Необнаруженные квазималапропизмы относятся к частотным словосочетаниям, таким как *халатный врач* (вместо *палатный врач*).

Первичные кандидаты на исправления отсутствовали в интернете в 1838 случаях, т.е. для дальнейшего анализа оставалось 850 кандидатов (в среднем 2,3 кандидата на ошибку). Столь существенный отсев (68%) позволил взять порог для вторичных кандидатов $Q = -\infty$ (т.е. условием принятия вторичного кандидата является его хотя бы единичное присутствие в массивах Яндекса).

Для 364 обнаруженных малапропизмов и квазималапропизмов только в 18 случаях истинные исправления не стояли первыми в списке упорядоченных по значению ПСС вторичных кандидатов (5%). Удивительным было то, что только в одном случае истинное исправление не вошло в список из двух первых лучших кандидатов

Виды пар словоформ	Миним. значение	Максим. значение	Среднее значение M	Стандартное отклонение D	Значений в ($M–D$, $M+D$)
Истинные исправления	-2,33	15,08	6,74	3,38	67,02%
Малапропизмы	-6,36	5,60	-0,81	2,50	63,90%
Квазималапропизмы	-2,18	14,50	3,39	2,99	74,28%
Малапропизмы и квазималапропизмы	-6,36	14,50	-0,01	3,13	67,85%

Таблица 4. Значения ПСС

(исключением явилось словосочетание *часто находит* для исправления квазималапропизма *часто сходит*, и в этом случае ему предшествовали более частотные *часто заходит/приходит/переходит*). Таким образом, редактору текста целесообразно выдавать только первые три-четыре наилучших по значению ПСС кандидата, а их, как правило, всего-то остается один-два (см. рисунок 1, третий столбец).

Заметим, что в списки наилучших попали несколько кандидатов, не являющихся коллокациями, например, *получил гран, раздел тосты*. Оказалось, что они являлись частями строк *получил гран-при*, *раздел «Тосты»* (даже при самых жестких формах запроса Яндекс допускает между словоформами знаки препинания, а дефис всегда делит слово на два). Отмеченные недостатки могут быть преодолены либо путем совершенствования языка запросов к Яндексу, либо переходом от текстов интернета к очень крупным текстовым корпусам. Но ряд ошибок в интернетовских текстах могут статистически «потонуть» и сами – при дальнейшем пополнении массивов интернета правильными текстами.

Заключение

Проведен расширенный эксперимент для подтверждения действенности метода автоматического обнаружения малапропизмов и автоматизированного их исправления, основанного на вычислении эвристически введенного числового показателя смысловой совместимости слов. Уточнение формулы подсчета ПСС и способов сбора интернет-статистики, а также расширение набора образцов малапропизмов, позволили оптимизировать пороговые правила метода и несколько улучшить по сравнению с [3] показатели по полноте и точности обнаружения и исправления малапропизмов. В пределах принятых ограничений эксперимент дал в целом весьма обнадеживающие результаты. Представляется актуальным полная автоматизация метода обнаружения малапропизмов с последующей его экспериментальной проверкой.

Список литературы

1. *Hirst G., St-Onge D. Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms // C. Fellbaum (Ed.) WordNet: An Electronic Lexical Database. MIT Press, 1998, p. 305-332.*
2. *Bolshakov I.A., Gelbukh A. On Detection of Malapropisms by Multistage Collocation Testing. // A. Düsterhöft, B. Talheim (Eds.) Proc. 8th Int. Conf. Applications of Natural Language to Information Systems NLDB´2003, June 2003, Burg, Germany, GI-Edition, LNI V. P-29, Bonn, 2003, p. 28-41.*
3. *Большаков И.А., Большакова Е.И. Обнаружение и исправление малапропизмов с помощью интернета // Компьютерная лингвистика и интеллектуальные технологии: Труды Междунар. Конф. Диалог '2005 / Под ред. И.М. Кобозевой, А.С. Нариньяни, В.П.Селегея – М.: Наука, 2005, С. 59-64.*
4. *Kilgarriff A., Grefenstette G. Introduction to the Special Issue on the Web as Corpus // Computational linguistics, V. 29, No. 3, 2003, p. 333-347.*
5. *Bolshakov I.A., Gelbukh A. Paronyms for Accelerated Correction of Semantic Errors. // International Journal on Information Theories & Applications. V. 10, 2003, p. 198-204.*
6. *Bolshakov I.A., Bolshakova E.I. Measurements of Lexico-Syntactic Cohesion by means of Internet. // MICAI 2005: Advances in Artificial Intelligence. A. Gelbukh, A. Albornoz, H. Terashima-Marin (Eds.). LNAI 3789, Springer-Verlag, 2005, p. 790-799.*