ОБНАРУЖЕНИЕ И ИСПРАВЛЕНИЕ МАЛАПРОПИЗМОВ С ПОМОЩЬЮ ИНТЕРНЕТА

И. А. Большаков

Национальный Политехнический институт (IPN), Мехико
igor@cic.ipn.mx

Е. И. Большакова

Московский Государственный Университет им. М.В. Ломоносова
bolsh@cs.msu.su

Малапропизм – это семантическая ошибка, при которой одно знаменательное слово заменяется другим, близким по звучанию, но отличным по смыслу. Описывается метод автоматического обнаружения малапропизмов и автоматизированного их исправления, основанный на интернет-статистике и специально введенном числовом показателе семантической совместимости.

Введение

Современные автокорректоры и редактирующие системы достаточно надежно выявляют орфографические ошибки и некоторые виды синтаксических ошибок. Однако в текстах нередко появляются ошибки другого типа, при которых одно знаменательное слово заменяется иным похожим словом в той же синтаксической роли и с теми же морфологическими характеристиками, например: ***спасательная*** *мысль*, ***неутомимый*** *голод*. Такие ошибки редактирующие программы выявлять не умеют, поскольку для этого автоматический морфосинтаксический анализ недостаточен.

Семантическая ошибка, при которой одно существующее слово заменяется другим, близким по звучанию, но отличным по смыслу и потому не соответствующим контексту, в европейских языках называется малапропизмом (в вольном переводе ‘малоуместное’).

В современной вычислительной лингвистике этому виду ошибок посвящены немногочисленные работы [1, 3].

В статье [1] предложен метод обнаружения малопропизмов, выявляющий знаменательные слова, далекие от прочих по контексту в смысле парадигматических связей лексической базы WordNet [2] (синонимы, гипонимы, гиперонимы и др.). Синтаксические связи слов не учитываются, а требуемый контекст обширен: слова из других предложений и даже абзацев.

В [3] обнаружение малапропизмов основано на синтагматических связях между словами, что позволяет рассматривать много меньший контекст – одно предложение. Предполагается, что малапропизмы разрушают семантическую связь между словами, сохраняя их синтактическую связанность. Для обнаружения ошибок в предложениях рассматриваются пары из двух синтаксически связанных знаменательных слов и проверяется их семантическая совместимость, для чего предлагается использовать один из трех типов лингвистических ресурсов: базу словосочетаний, например, КроссЛексику [4], текстовый корпус, например, Национальный Корпус Русского Языка [5] или интернетовский поисковик, например, Гугл.

Данная статья продолжает работу [3], делая упор на эксперимент с российским поисковиком Яндекс для проверки словосочетаний на семантическую совместимость. Интернет ныне рассматривается как текстовой корпус, огромный, но с большим информационным шумом [6]. Применительно к нему предложенный в [3] метод существенно уточняется и дополняется. Как и ранее, рассматриваются малапропизмы, нарушающие установившуюся в языке сочетаемость слов с сохранением той же синтаксической связи внутри словосочетания. Малапропизм обнаружен, если для пары синтаксически связанных знаменательных слов значение специально введенного числового ***показателя смысловой совместимости*** (ПСС) оказывается ниже установленного порога (далее малапропизмом называется вся обнаруженная пара слов).

Для исправления обнаруженных малапропизмов, в отличие от примененного в [1, 3] слепого подбора кандидатов, предлагается использовать заранее составленные словари паронимов (слов, сходных буквами, или слов с одинаковыми корнями). Из словарей берутся все первичные кандидаты на исправление, и для отсева ненужных предложены основанные на ПСС эвристические пороговые правила, оставляющие человеку-редактору только наилучшие варианты. Пороги найдены при эксперименте над набором из 100 русских малапропизмов. Эксперимент показал обнадеживающие результаты как по обнаружению, так и по исправлению малапропизмов.

Словари паронимов

Для исправления малапропизмов необходимо быстро находить сходные слова. Слова, близкие по звучанию или написанию, обычно называются паронимами [7]. Звуковые паронимы мы не рассматриваем, а среди остальных выделяем ***буквенные*** и ***морфемные***. Первые различаются немногими буквами (*краска* – *каска*), вторые – немногими служебными морфами (*человечный* – *человеческий*).

Построенный нами словарь русских буквенных паронимов состоит из групп слов, включающих заглавное слово и набор однобуквенных паронимов – слов, получающихся в результате применения к заглавному слова элементарной операции редактирования: опущения одной буквы; вставки буквы в любой позиции; замены одной буквы на другую; перестановки двух соседних букв. К примеру, одну группу образуют заглавное слово *белка* и его паронимы *булка*, *елка*, *телка*, *челка*, *щелка*.

В группы берутся слова одинаковой части речи, причем у существительных по отдельности строятся группы для множественного числа и разных родов единственного числа, а у глаголов рассматриваются отдельно личные формы, причастия и деепричастия. Такое разделение гарантирует синтаксическую правильность предложений после паронимической замены. Пытаясь исправить малапропизм, мы извлекаем подозрительную словоформу из текста (например, *белкой*), восстанавливаем ее словарную форму (*белка*), находим в словаре ее пароним (например, *булка*), морфологически изменяем его в ту же форму, что и у подозрительного слова (*булкой*), а затем подставляем в текст вместо него.

Описываемый словарь пригоден для исправления однобуквенных ошибок, обычно случайных. Для исправления многобуквенных ошибок, связанных с употреблением слов с одинаковыми корнями (например, *массивный* вместо *массовый* или *весовой* вместо *весомый*), и обычно неслучайных, построен словарь русских морфемных паронимов. Единицей словаря является группа паронимов, например, группа слов *бегающий*, *беглый*, *беговой*, *бегущий*. В группу включены слова одной части речи, имеющие одинаковый или омонимичный корень, но различающиеся вспомогательными морфами (суффиксами и/или префиксами).

Построенный словарь русских однобуквенных паронимов содержит ныне 17,4 тыс. паронимических групп среднего объема 2,65 паронимов в группе, а словарь морфемных паронимов – 1310 паронимических групп среднего объема 7,1. Оба словаря подробнее описаны в [8].

Метод обнаружения и исправления малапропизмов

Главная идея обнаружения – просмотр всех пар знаменательных слов в анализируемом предложении текста с проверкой их на синтаксическую связанность и семантическую совместимость. Если пара (V,W) синтаксически связана, но семантически несовместима, сигнализируется малапропизм.

Поскольку не известно, какое слово в обнаруженном малапропизме ошибочно, для его исправления составляются, с помощью обоих паронимических словарей, всевозможные исправляющие пары с варьированием как первого, так и второго слова в паре. При этом из словаря морфемных паронимов берутся слова, отличающиеся не более чем двумя морфами. Полученные пары составляют список первичных кандидатов.

Каждый первичный кандидат проверяется на семантическую совместимость, и пары, не выдерживающие теста, отбрасываются. В итоге остается список вторичных кандидатов, который упорядочивается, и перед показом человеку-редактору в нем оставляются лучшие.

Для тестирования пар (V,W) на семантическую совместимость с помощью интернета, рассматриваемого как текстовый корпус, необходим статистический критерий. В соответствии с одним из них, пара совместима, если относительная частота $N(V,W)/S$ совместного ее выпадения на ограниченном расстоянии в пределах всего текстового корпуса больше произведения относительных частот $N(V)/S$ и $N(W)/S$ выпадений слов V и W, рассматриваемых по отдельности (N – частоты выпадений, S – размер корпуса). Логарифмируя, получаем пороговое правило

$$\text{ПВИ}(V, W) \equiv \ln N(V, W) + \ln S - \ln N(V) - \ln N(W) > 0,$$

где $\text{ПВИ}(V, W)$ – показатель взаимной информации [9].

Любой поисковик выдает «сырую» статистику выпадений запрашиваемого слова или комбинации слов, обычно измеряемую количеством содержащих выпадения страниц. После некоторых прикидок, мы ввели в качестве меры смысловой совместимости Показатель Семантической Совместимости (ПСС), сходный с ПВИ:

при $N(V,W) > 0$ $\quad \text{ПСС}(V,W) \equiv \ln N(V,W) + \ln P - (\ln N(V) + \ln N(W)) / 2$

при $N(V,W) = 0$ $\quad \text{ПСС}(V,W) \equiv NEG$

где N – число релевантных страниц; NEG – отрицательная константа; P – положительная константа, подбираемая экспериментально.

Достоинством ПСС по сравнению с ПВИ является то, что для его вычисления не нужно знать размер корпуса, т.е. полное число страниц поисковика. В то же время, как и ПВИ, ПСС практически не зависит от флюктуирующего роста всех статистических данных поисковика во времени: если $N(V,W)$, $N(V)$ и $N(W)$ нарастают примерно с одинаковой скоростью, то деление логарифмов двух последних величин на 2 компенсирует этот рост.

Когда $\text{ПСС}(V_m,W_m) < 0$, пара (V_m,W_m) признается малапропизмом, в то время как первичный кандидат на его исправление (V,W) признается вторичным, если срабатывает иное пороговое правило:

$$(\text{ПСС}(V_m,W_m) = NEG) \textbf{ И } (\text{ПСС}(V,W) > Q)$$
$$\textbf{ИЛИ } (\text{ПСС}(V_m,W_m) > NEG) \textbf{ И } (\text{ПСС}(V,W) > \text{ПСС}(V_m,W_m)),$$

где Q, $NEG < Q < 0$, константа, подбираемая экспериментально. Иными словами, при малапропизме, отсутствующем в поисковике, кандидату достаточно преодолеть довольно низкий фиксированный порог, если же малапропизм из-за интернетовского шума в поисковике имеется, кандидат должен превышать его по значению ПСС.

Оставшиеся вторичные кандидаты упорядочиваются по значению ПСС. Лучшими признаются все кандидаты с положительным ПСС (пусть их будет n), а из кандидатов с отрицательными значениями в лучшие берется при $n = 1$ только один кандидат с максимальным ПСС, а при $n = 0$ только два максимальных.

Экспериментальный набор малапропизмов

20 образцов малапропизмов взяты в экспериментальный набор на вскидку, в том числе – шутки вроде *жизни на Марксе* и ошибки, допускаемые даже образованными русскими, типа *проектироваться на экран*.

Остальные 80 образцов сформированы из потока интернетовских новостей: бралась пара смежных слов, образующая коллокацию (т.е. синтаксически связанное и семантически совместимое словосочетание), и один ее компонент затем фальсифицировался при помощи буквенного или морфемного словаря паронимов, с сохранением исходных морфологических характеристик (например, форма *точками* заменялась на *тучками*).

В итоге получился набор из 100 образцов, куда буквенные ошибки, ожидаемые много чаще морфемных, вошли в пропорции 86 : 14. Набор содержит все пять часто встречающихся в русском языке синтаксических типа коллокаций: “определяемое слово → определяющее слово” (*актовый зал*, *вполне приемлемо*); “глагол → его дополнение в виде существительного” (*уделить внимание, ворвались в здание*); “существительное → его дополнение” (*тайники с оружием*, *жертвы теракта*); “адъектив (прилагательное или причастие) → его дополнение” (*занятый трудом*); и “сказуемое → подлежащее” (*спасатели обнаружили*, *отправлен груз*).

```
1)2L  жизнь на Марксе          274,жизнь:34871341,Марксе:9021
  2L!!жизнь на Марсе           49288, Марсе:440004
2)1L  (совершать) полые акты   0,полые:37385,акты:2357875
  1L  голые акты               1,голые:2729404
  1L!!подлые акты              12,подлые:63984
  1L  полные акты              4,полные:1264157
  1L  пошлые акты              0,пошлые:209498
  2L  полые акры               0,акры:8065
  2L  полые пакты              0,пакты:8676
  2L  полые такты              0,такты:22226
  2L  полые факты              0,факты:3729898
3)1L  истерический центр       0,истерический:46860,центр:33808389
  1M  истеричный центр         0,истеричный:14736
  1L!!исторический центр       46199,исторический:2029436
  1L  стерический центр        0,стерический:461
  2L  истерический цент        0,цент:174231
4)1L  (без) призраков жизни    2,призраков:293225,жизни:43588136
  1L!!признаков жизни          581,признаков:1092302
5)2L  добиваться суждения      0,добиваться:568690,суждения:387344
  2L!!добиваться осуждения     107,осуждения:163625
  2L  добиваться сужения       18,сужения:56756
  1L  добираться суждения      1,добираться:233950
  1L  добываться суждения      0,добываться:7472
  1L  доживаться суждения      0,доживаться:28
  1L  доливаться суждения      0,доливаться:179
  1L  дошиваться суждения      0,дошиваться:3
```

Рис. 1. *Фрагмент экспериментального набора с Яндекс-статистикой*

Кроме собственно малапропизмов, набор включает ошибки, названные нами *квазималапропизмами* (8 образцов). В них одна коллокация заменяется на другую, но более редкую и часто противоречащую прочему контексту, например, (*смыло*) *сетевым потоком* вместо *селевым потоком*.

Затем для каждого образца с помощью обоих паронимических словарей были составлены первичные кандидаты на исправление, в итоге получилось 645 кандидатов (в среднем, 6,45 кандидатов на один малапропизм). Для каждого образца среди кандидатов возникала и исходная коллокация, называемая далее истинным исправлением.

Начальный фрагмент экспериментального набора представлен левым столбцом рис. 1. Каждая пронумерованная секция набора начинается строкой с малапропизмом (иногда с поясняющим контекстом в скобках), далее идут строки с первичными кандидатами на исправление. Все строки содержат номер ошибочного слова (1 или 2) и символ использованного словаря (**L** – буквенный, **M** - морфемный). Символ **!!** помечает истинное исправление.

Эксперимент с Яндекс-статистикой

Правый столбец рис. 1 содержит полученную в ходе эксперимента «сырую» статистику: число страниц с совместным выпадением и с выпадением обоих компонентов словосочетания по отдельности (данные для словоформ, уже содержащихся в малапропизме, опущены как повторные). Для получения Яндекс-статистики использовался вариант запроса в кавычках, исключающий поиск разнесенных компонентов словосочетаний.

Малапропизмы оказались в 65 случаях из 100 вообще отсутствующими в массивах поисковика, а первичные их исправления – в 492 случаях из 645. Тем самым сырая статистика Яндекса уже есть некое средство и обнаружения малапропизмов, и отсева их абсурдных исправлений.

При подборе констант в приведенных выше пороговых правилах квазималапропизмы не рассматривались. Чтобы получить для всех малапропизмов отрицательные значения ПСС, была взята константа $P = 200$. Константа $Q = -7.5$ подобрана так, чтобы все кандидаты с ненулевой встречаемостью оказались по значению ПСС выше этого порога. Выбор константы $NEG = -9$ довольно произволен, от нее требуется лишь, чтобы все ненулевые случаи совместной встречаемости давали значение ПСС, большее *NEG*.

Результаты вычисления ПСС и применения пороговых правил для начального фрагмента экспериментального набора показаны на рис. 2.

Несмотря на то, что при выборе констант квазималапропизмы не учитывались, наш метод обнаружил их всех, кроме одного: (*мечтал*) *до последнего вдоха* (вместо *вздоха*).

При отсеве из 645 первичных кандидатов на исправление осталось 150 вторичных, а после выбора из них лучших – 130 кандидата (т.е. почти пятикратный отсев). В 98 из 99 обнаруженных малапропизмов лучшие кандидаты включают истинное исправление. Исключение составил кандидат *введенный в кожу* при отсутствии намеренного *введенный в кому* для исправления малапропизма *введенный в кору*.

Списки лучших кандидатов содержат от 1 до 4 кандидатов (средняя длина списка – 1,3), а обычно кандидат один – истинное исправление.

Только шесть лучших кандидатов оказались, на наш взгляд, не коллокациями, а шумом интернета , вроде *беда мысли* или *оправлен груз*.

На рис. 3 даны распределения округленных до ближайшего целого значений ПСС для малапропизмов и их истинных исправлений. У малапропизмов два пика распределения. Мощный сосредоточенный пик у значения –9 соответствует ошибкам, не замаскированным шумом (нулевая совместная встречаемость). Другой всплеск, чисто шумовой и рассредоточенный, наблюдается возле значения -2. Пик истинных исправлений располагается между величинами +2 и +3.

```
1)2L  жизнь на Марксе                -0.02 ОБНАРУЖЕН
  2L!!жизнь на Марсе                  3.22 1-Й КАНДИДАТ
2)1L (совершить) полые акты          -9 ОБНАРУЖЕН
  1L!!подлые акты                    -2.78 1-Й КАНДИДАТ
  1L  полные акты                    -5.37 2-Й КАНДИДАТ
3)1L  истерический центр             -9 ОБНАРУЖЕН
  1L!!исторический центр             2.41 1-Й КАНДИДАТ
4)1L  (без) призраков жизни          -3.75 ОБНАРУЖЕН
  1L  признаков жизни                 2.87 1-Й КАНДИДАТ
5)2L  добиваться суждения (вора)     -9 ОБНАРУЖЕН
  2L!!добиваться осуждения           -0.35 1-Й КАНДИДАТ
  2L  добиваться сужения             1.61 2-Й КАНДИДАТ
```

Рис. 2. *Фрагмент набора малапропизмов и лучших кандидатов на их исправление со значениями ПСС*

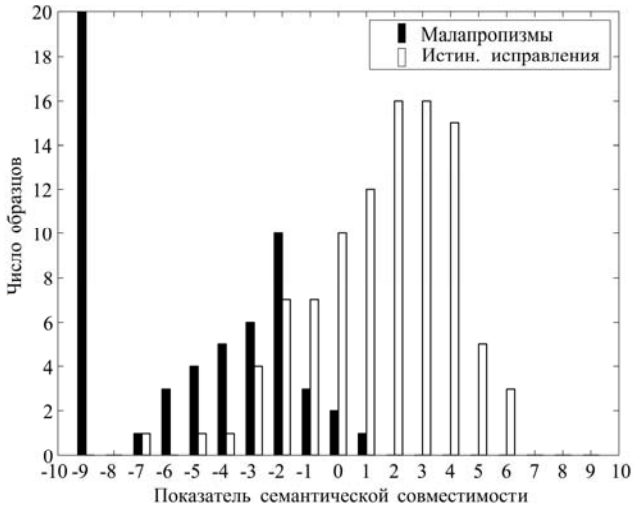

Рис. 3. *Распределение значений ПСС для малапропизмов и их истинных исправлений*

Заключение

Предложен метод автоматического обнаружения малапропизмов и автоматизированного их исправления, основанный на вычислении эвристически введенного числового показателя семантической совместимости пары слов, образующих синтаксически связанное словосочетание. Для проверки его действенности проведен эксперимент, показавший, в пределах принятых ограничений, весьма обнадеживающие результаты как по обнаружению, так и по исправлению малапропизмов.

Представляется актуальным проверить предложенные пороговые правила на иных наборах малапропизмов, в том числе с разнесенными компонентами. Полезно также повторить исследования для других поисковиков.

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) Fellbaum, Ch. (Ed.) *WordNet*: *An Electronic Lexical Database*. MIT Press, 1998.
3) 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.
4) Большаков И.А. Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ, сер 2, № 1, 1994, С. 11-23.
5) *Национальный Корпус Русского Языка*. http://ruscorpora.ru
6) Kilgarriff, A., Grefenstette G. Introduction to the Special Issue on the Web as Corpus // *Computational linguistics*, V. 29, No. 3, 2003, p. 333-347.
7) Бельчиков Ю.А., Панюшева М.С. *Словарь паронимов современного русского языка.* М.: Русский язык, 1994.
8) Bolshakov, I.A., Gelbukh A. Paronyms for Accelerated Correction of Semantic Errors // *International Journal on Information Theories & Applications.* V. 10, 2003, p. 198-204.
9) Manning, Ch. D., Schütze H. *Foundations of Statistical Natural Language Processing*. MIT Press,1999.