АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 001.8:81'322.2-161.1

Г. Г. Белоногов, Р. С. Гиляревский, Ал-др А. Хорошилов, Ал-сей А. Хорошилов

Автоматическое распознавание смыслового тождества и смысловой близости русских слов на основе их словообразовательного и словоизменительного анализа и синтеза*

Описывается метод автоматического распознавания смыслового тождества и смысловой близости русских слов на основе их словообразовательного и словоизменительного анализа и синтеза. Метод позволяет проводить автоматическую нормализацию слов текстов и генерацию их словообразовательных вариантов.

Русский язык имеет богатую систему словоизменения и словообразования, что затрудняет распознавание смыслового тождества и смысловой близости слов при решении многих задач автоматической обработки информации, например, таких, как поиск информации в текстах, их автоматическое реферирование, автоматическое индексирование и автоматическая классификация. Поэтому необходимо разработать такие процедуры, которые позволяли бы автоматически идентифицировать различные формы слов, имеющих один и тот же или примерно один и тот же смысл.

При отождествлении слов можно применять два подхода. Первый из них заключается в том, что различные словообразовательные и словоизменительные варианты слов, имеющие примерно один и тот же смысл, заменяются на одну нормализованную каноническую форму. Тогда отождествление одинаковых по смыслу слов можно свести к отождествлению их канонических форм. Другой подход состоит в том, что при отождествлении двух слов текстовая форма одного слова заменяется на множество эквивалентных ей по смыслу нормализованных словообразовательных вариантов, а текстовая форма другого слова нормализуется только на уровне словоизменения. Затем эта форма второго слова сравнивается со всеми нормализованными словообразовательными вариантами первого и, в случае совпадения с одним из них, она считается эквивалентной по смыслу первому слову.

Обычно под нормализованной (канонической) формой слова понимается такая форма, которая традиционно указывается в словарях. Например, для существительного — это форма именительного падежа единственного или (в случае pluralia tantum) множественного числа, для глагола — форма инфинитива, для прилагательного — форма именительного падежа единственного числа мужского рода. Процедура замены исходной вариантной формы слова на каноническую называется процедурой лемматизации.

Но в системах автоматической обработки текстовой информации для получения канонической формы слова вместо лемматизации можно применять и другую операцию — операцию замены исходной формы слова на сочетание ее словоизменительной основы и номера флективного (словоизменительного) класса. Например, формы слов *стола, столу, столом, столе, столы, столов, столам, столами, столах* могут быть представлены в виде записи "стол001", в которой буквосочетание *стол* является основой слова, а код 001 — номером флективного класса. Этот код обозначает перечень окончаний, которые могут быть у слова *стол* (нулевое окончание "+" и окончания *-а, -у, -ом, -е, -ы, -ов, -ам, -ами, -ах*). Такие же окончания бывают у всех слов, принадлежащих к флективному классу 001 (например, у слов *вездеход, самолет, катер, снаряд, взрыв*). Формы слов *санатория, санаторию, санаторием, санатории, санаториев, санаториям, санаториями, санаториях* могут быть представлены в виде записи "санатори005", в которой буквосочетание *санатори* является основой слова, а код "005" — номером его флективного класса. Такие же окончания бывают у слов *сценарий, глоссарий, критерий, лекторий, гербарий, профилакторий* (подробнее о системе флективных классов русского языка см. в [1, с. 160–163]).

При нормализации слов на словообразовательном уровне каноническая форма слова должна представлять по возможности всю его словообразовательную парадигму. Например, тождественные или близкие по смыслу формы слов *испытают, испытав, испытавший, испытай, испытайте, испытал, испытан, испытание, испытанный, испытать, испытывавший, испытываемый, испытывают, испытывай, испытывайте, испытывать,*

* Работа выполнена при поддержке Российского Фонда Фундаментальных Исследований (РФФИ). Грант N

испытываться, испытывал, испытывая, принадлежащие к различным частям речи, могут быть заменены на одну форму существительного *испытание.* А формы слов *замолчит, замолчав, замолчавши, замолчавший, замолчал, замолчать, замолчи, замолчите* — на одну форму инфинитива *замолчать.*

Выбор канонической формы слова, представляющей множество его словообразовательных вариантов, имеющих примерно одинаковый смысл, должен производиться с учетом системы словообразования русского языка. С целью выявления этой системы в начале 80-х гг. прошлого века авторами были проведены соответствующие исследования. Их результаты опубликованы в 1984 г. в книге Г. Г. Белоногова, Б. А. Кузнецова и А. П. Новоселова "Автоматическая обработка научно-технической информации. Лингвистические аспекты" [1, с. 25–35 и 163–311].

В результате этих исследований было выявлено около 1200 различных словообразовательных суффиксов и сочетаний суффиксов. Суффиксы и сочетания суффиксов сопровождались номерами совместимых с ними флективных классов. Например, в словах *обновлять, обновить, обновление* входящие в их состав сочетания суффиксов *-лять, -ить* и *-лени* представлялись как *-лять*144, *-ить*144 и *-лени*073.

Авторами книги было введено понятие "словообразовательный класс". Словообразовательный класс слова характеризовался перечнем суффиксов (сочетаний суффиксов), совместимых с его словообразовательной основой. При этом, если у двух слов перечни суффиксов (сочетаний суффиксов) отличались друг от друга хотя бы одним элементом, то они считались принадлежащими различным словообразовательным классам. Количество суффиксов (сочетаний суффиксов) в одном словообразовательном классе варьировало от двух до тридцати восьми и в среднем составляло 11,7, а всего было выявлено около 1400 словообразовательных классов слов.

В табл. 1 представлено четыре примера словообразовательных классов слов.

Таблица 1

Примеры словообразовательных классов слов

	0010 — нагрев	
+001	а116	авш105
аем103	ай143	айте143
ал125	ани073	ател003
ательн103	ать144	аться144
ающ105	ая152	аясь152
	0879 — безопас-ность	
ен126	н103	н126
нее152	но152	ност055
	1015 — экватор	
+001	иальн103	
	1029 — буржуаз-ия	
и061	н103	

В этих примерах для каждого словообразовательного класса указан его четырехзначный номер, слово -представитель, расчлененное дефисом на словообразовательную основу и суффикс или сочетание суффиксов (если они имеются), и перечень словообразовательных суффиксов, сопровождаемых трехзначными номерами совместимых с ними флективных классов.

С помощью табл. 1 могут порождаться словообразовательные варианты слов путем присоединения к их словообразовательным основам соответствующих суффиксов (сочетаний суффиксов) и окончаний, совместимых с этими суффиксами (сочетаниями суффиксов). Например, для слова *нагрев*, принадлежащего словообразовательному классу 0010, могут быть порождены его словообразовательные варианты *нагревает, нагревавший, нагреваемый, нагревал, нагревание, нагреватель, нагревательный* и др., а для слова *экватор*, принадлежащего словообразовательному классу 1015, — его словообразовательный вариант *экваториальный.*

Система словообразовательных классов русских слов, опубликованная в [1], создавалась с целью разработки комплекса программ автоматического обнаружения и исправления орфографических ошибок в текстах. При этом ставилась задача при заданном эталонном орфографическом словаре обеспечить контроль правильности написания как можно большего числа слов, а распознавание смыслового тождества слов, встречающихся в текстах, и слов из эталонного словаря играло подчиненную роль. Поэтому в одну словообразовательную парадигму слова иногда включались его производные формы, обозначающие различные понятия. Это имеет место, например, в словообразовательном классе 0010 (см. табл. 1), где в одной словообразовательной парадигме встречаются такие слова, как *нагрев, нагревание* и *нагреватель.* Если слова *нагрев* и *нагревание* можно считать синонимами, то слово *нагреватель* обозначает другое понятие, хотя оно и связано по смыслу с двумя первыми словами.

По аналогии с канонической формой слова, представляющей множество его словоизменительных вариантов, для представления множества членов словообразовательной парадигмы можно также ввести свою каноническую форму. В качестве такой формы может выступать существительное, если оно является членом парадигмы, или (если нет существительного) инфинитив. В тех случаях, когда в составе парадигмы нет ни существительного, ни инфинитива, в качестве канонической формы может выступать прилагательное. Если и прилагательного нет, то любая другая форма.

Переход от вариантной словообразовательной формы слова к его канонической можно представить себе как замену суффикса или сочетания суффиксов вариантной формы на суффикс (сочетание суффиксов) канонической формы. Для этого необходимо уметь выделять в слове его словообразовательную основу и суффиксы, иметь ассоциативный словарь суффиксов, в котором для каждого суффикса (сочетания суффиксов) будет указан один или несколько вариантов его замены на суффикс или на сочетание суффиксов соответствующей канонической формы, и иметь процедуру проверки правильности такой замены (проверки совместимости словообразовательной основы слова и присоединенных к ней суффикса или сочетания суффиксов).

Несколько фрагментов из ассоциативного словаря суффиксов представлены в табл. 2. Этот словарь был составлен на основе словаря словообразовательных классов слов.

Таблица 2

Фрагменты ассоциативного словаря суффиксов

изировав152 — изаци061/изировать144
изировал125 — изаци061/изировать144
изирован125 — изаци061/изировать144
изированн103 — изаци061/изировать144
изировать144 — изаци061/изировать144
изироваться144 — изаци061/изировать144
изиру116 — изаци061/изировать144
изируем103 — изаци061/изировать144
изируй143 — изаци061/изировать
изируйте143 — изаци061/изировать144
изируйте144 — изаци061/изировать144
изирующ105 — изаци061/изировать144
изируя152 — изаци061//изировать144
изируясь152 — изаци061/изировать144

.....

онировавш105 — онировани073/онировать144
онировал125 — онировани073/онировать144
онировани073 — онировани073/онировать144
онированн103 — онировани073/онировать144
онировать144 — онировани073/онировать144
онироваться144 — онировани073/онировать144
ониру116 — онировани073/онировать144
онируйте143 — онировани073/онировать144
онирующ105 — онировани073/онировать144
онируя152 — онировани073/онировать144

.....

ыва116 — ывани073/ывать144
ывавш105 — ывани073/ывать144
ываем126 — ывани073/ывать144
ываемост055 — ывани073/ывать144
ывай143 — ывани073/ывать144
ывайте143 — ывани073/ывать144
ывал125 — ывани073/ывать144
ывал126 — ывани073/ывать144
ывани073 — ывани073/ывать144
ывать144 — ывани073
ывающ105 — ывани073/ывать144
ывая152 — ывани073/ывать144

.....

Проверку совместимости словообразовательной основы слова и присоединяемых к ней новых суффиксов можно было бы проводить с помощью системы словообразовательных классов типа, приведенных в табл. 1. Но для этого необходимо иметь мощный словарь словообразовательных основ слов, в котором для каждой основы должен быть указан номер ее словообразовательного класса или (в случае омонимии основ) сочетание номеров классов. Составление такого словаря — довольно трудная задача, если учесть, что для хорошего покрытия текстов нужно иметь словарь объемом, как минимум, в несколько сотен тысяч основ слов.

Более реальным является другой путь — проверка совместимости по словарю словоизменительных основ слов, сопровождаемых номерами их флективных классов. Такой словарь был составлен авторами статьи по политематическим текстам общим объемом около двух гигобайтов. При этом сначала был составлен частотный словарь словоформ, а затем, с помощью процедуры морфологического анализа [2], был построен словарь словоизменительных основ слов. Словарь основ слов получился объемом около 1.100 тыс. лексических единиц. Фрагменты этого словаря приведены в табл. 3.

Таблица 3

Фрагменты словаря словоизменительных основ слов

хлеб010
погреб010
зоб010
рукав010

.....

берег010
рог010
стог010
луг010

.....

луговск110
воровск110
складск110

.....

окаж120
докаж120
покаж120
скаж120

.....

учител030
мучител030
поручител030

.....

лингвистик060
металингвистик060
экзолингвистик060
биолингвистик060
социолингвистик060
этнолингвистик060
нейролингвистик060
макролингвистик060
психолингвистик060
интерлингвистик060
компаративистик060
когнитивистик060
архивистик060

.....

свиноводств070
зерноводств070
животноводств070
льноводств070
звероводств070
осетроводств070
лесоводств070

.....

нитроангидрид001
хлорангидрид001
дихлорангидрид001
трихлорангидрид001
монохлорангидрид001

.....

некогда152
никогда152
иногда152
тогда152
куда152
откуда152

.....

сломив152
разгромив152
сохранив152
заменив152
изменив152
применив152
оценив152

.....

Продолжение табл. 3

электрозондов103
трехзондов103
четырехзондов103
двухзондов103
фондов103
бесфондов103

....

озаглавливани073
обезглавливани073
обуславливани073
улавливани073

.....

стародавн104
древн104
задн104
последн104
предпоследн104

.....

отвезти144
ползти144
зайти144
найти144

.....

убивать144
выбивать144
завивать144
навивать144
развивать144
прививать144

.....

затупивш105
наступивш105
вступивш105
приступивш105
поступивш105
подвыпивш105

.....

разлива116
излива116
заилива116
увилива116
распилива116

.....

Проверку совместимости словообразовательных основ слов и присоединяемых к ним суффиксов можно проводить путем поиска вновь образованных цепочек букв в словаре словоизменительных основ слов. Если сформированная цепочка букв содержится в словаре словоизменительных основ, то она правильная, и совместимость словообразовательной основы слова и присоединенных к ней суффикса или сочетания суффиксов имеет место; если не содержится, то она, вероятно, неправильная, и следует повторить попытку формирования канонической словоизменительной основы слов, используя другие суффиксы или сочетания суффиксов.

Процедуру замены исходной формы слова на каноническую (процедуру нормализации) следует начинать с морфологического анализа исходной формы слова. Далее по результатам морфологического анализа формируется сочетание буквенного кода словоизменительной основы слова и номера ее флективного класса. Если при этом окажется, что сформированная цепочка символов представляет существительное, то на этом процедура нормализации и закапчивается. Если не существительное, то можно попытаться расчленить словоизменительную основу слова на словообразовательную основу и суффикс или сочетание суффиксов.

При расчленении словоизменительной основы исходной формы слова на словообразовательную основу и суффикс или сочетание суффиксов следует сначала найти в ассоциативном словаре суффиксов и сочетаний суффиксов (см. табл. 2) элемент максимальной длины, входящий в расчленяемую основу, и заменить этот элемент на суффикс (сочетание суффиксов) канонической формы слова (существительного или инфинитива). Далее путем поиска в словаре словоизменительных основ слов следует убедиться в том, что сформированная цепочка символов содержится в этом словаре. Если она там содержится, то процесс нормализации закапчивается. Если не содержится, то нужно повторить членение исходной словоизменительной основы с помощью суффикса (сочетания суффиксов) меньшей длины и заменить этот суффикс (сочетание суффиксов) на суффикс (сочетание суффиксов) канонической формы. Эту операцию надо повторять до тех пор, пока не будет найдена необходимая каноническая форма слова. Если она, тем не менее, не будет найдена, то за каноническую форму следует принять исходную словоизменительную основу слова.

Описанная процедура замены вариантных словообразовательных форм слов на канонические довольно сложна. Поэтому авторы применили ее сначала для автоматического формирования канонических форм упомянутого выше большого словаря. Эксперимент оказался успешным. После исправления немногочисленных ошибок было принято решение использовать полученный словарь соответствий между вариантными словообразовательными формами слов и их каноническими формами для автоматической нормализации слов в тестах.

* * *

Следует заметить, что этот словарь может быть использован и для формирования вариантных словообразовательных форм слов, если его представить в двух видах — в прямом и в инверсном. С помощью первого словаря можно заменять текстовые вариантные формы слов на канонические, а с помощью второго — приписывать этим каноническим формам наборы эквивалентных по смыслу вариантных форм.

СПИСОК ЛИТЕРАТУРЫ

1. Белоногов Г. Г., Кузнецов Б А., Новоселов А П. Автоматическая обработка научно-технической информации. Лингвистические аспекты // Итоги науки и техники. Сер. "Информатика". - М.: ВИНИТИ, 1984. – Т. 8.

2. Белоногов Г. Г., Зеленков Ю. Г. Еще раз о принципе аналогии в морфологии // НТИ. Сер. 2.– 1995. № 3.

Материал поступил в редакцию 22.11.02.