УДК 004.89:004.93

С. А. Большакова, А. В. Ниценко, В. Ю. Шелепов
Государственное учреждение «Институт проблем искусственного интеллекта», г. Донецк
83048, г. Донецк, ул. Артема, 118-б

К ВОПРОСУ ОБ АВТОМАТИЧЕСКОМ СНЯТИИ ОМОНИМИИ РУССКИХ ДЕЕПРИЧАСТИЙ

S. A. Bolshakova, A. V. Nicenko, V. Ju. Shelepov
Public institution «Institute of Problems of Artificial Intelligence», Donetsk
83048, Donetsk, Artema st., 118-b

ON THE QUESTION OF AUTOMATIC DISAMBIGUATION OF RUSSIAN ADVERBIAL PARTICIPLES

С. А. Большакова, А. В. Ниценко, В. Ю. Шелєпов
Державна установа «Інститут проблем штучного інтелекту», м. Донецьк
83048, м. Донецьк, вул. Артема, 118-б

ДО ПИТАННЯ ПРО АВТОМАТИЧНЕ ЗНЯТТЯ ОМОНІМІЇ РОСІЙСЬКИХ ДІЄПРИСЛІВНИКІВ

В статье рассматривается вопрос автоматического снятия омонимии русских деепричастий. Классифицированы виды омонимии деепричастий в русском языке и отобран перечень омонимов каждого вида из обширного словаря русских словоформ. Разработаны правила для снятия частеречной и морфологической омонимии и реализовано экспериментальное программное обеспечение.
Ключевые слова: префиксное дерево, автоматическое снятие омонимии, омонимия деепричастий, правила для автоматического снятия омонимии.

The article deals with the issue of automatic disambiguation of Russian adverbial participles. The types of homonymy of gerunds in the Russian language are classified and a list of homonyms of each type is selected from an extensive dictionary of Russian word forms. Rules for the part-of-speech and morphological disambiguation have been developed and experimental software has been implemented.
Key words: prefix tree, automatic disambiguation, adverbial homonymy, rules for automatic homonymy resolution.

У статті розглядається питання автоматичного зняття омонімії російських дієприслівників. Класифіковані види омонімії дієприслівників у російській мові та відібрано перелік омонімів кожного виду з великого словника російських словоформ. Розроблено правила для зняття частинномовної та морфологічної омонімії та реалізовано експериментальне програмне забезпечення.
Ключові слова: префіксне дерево, автоматичне зняття омонімії, омонімія дієприслівників, правила для автоматичного зняття омонімії.

Введение

С момента создания компьютера люди стремятся к тому, чтобы он понимал человеческую речь (как устную, так и письменную). Автоматическая обработка текста – одно из старейших направлений искусственного интеллекта. Одним из важнейших этапов обработки текста является морфологический анализ, в ходе которого определяются морфологические характеристики слов (часть речи, род, число, падеж, время и так далее) и их начальная форма (лемма).

Задача морфологического анализа осложняется омонимией. В русском языке встречаются различные виды омонимии: частеречная, морфологическая, лексическая. Частеречные и морфологические омонимы можно различить между собой, опираясь на морфологическую и синтаксическую информацию их окружения.

Деепричастие считается глагольной формой. Однако, по мнению некоторых ученых, его следует считать самостоятельной (неизменяемой) частью речи, обозначающей добавочное действие при основном.

Следует отметить, что проблема омонимии деепричастий в русском языке стоит особенно остро. Даже в национальном корпусе русского языка [1] в подавляющем большинстве случаев деепричастия, имеющие омонимы с другими частями речи, размечены неверно.

Целью работы является разработать метод автоматического снятия омонимии русских деепричастий.

Постановка задачи. Для реализации поставленной цели требуется: классифицировать виды омонимии деепричастий в русском языке, отобрать перечень омонимов каждого вида из словаря [2], разработать метод автоматического снятия частеречной и морфологической омонимии русских деепричастий с помощью правил и реализовать его в экспериментальном программном обеспечение.

1 Алгоритм поиска омонимов в словаре с использованием дерева

Для поиска омонимов деепричастий будем использовать обширный словарь русских словоформ [2], содержащий более 4 миллионов словоформ и грамматическую информацию относительно каждой словоформы. Этот словарь организован как множество строк-словоформ с сопровождающей грамматической информацией, собранных в парадигмы, каждая из которых начинается леммой (начальной формой слова). Леммы расположены по алфавиту. Парадигмы разделены пустыми строками. Морфологическая информация состоит из цепочки сокращений, разделенных пробелами и отражающих часть речи и грамматические характеристики для существительных, глаголов и так далее (например, число, падеж, лицо). Например,

ехать | гл несов непер инф
едемте | гл несов непер пов мн
ехал | гл несов непер прош ед муж
ехала | гл несов непер прош ед жен
ехало | гл несов непер прош ед ср
ехали | гл несов непер прош мн
едут | гл несов непер наст мн 3-е
еду | гл несов непер наст ед 1-е
едешь | гл несов непер наст ед 2-е
едет | гл несов непер наст ед 3-е
едем | гл несов непер наст мн 1-е
едете | гл несов непер наст мн 2-е
…

Для обеспечения быстрого поиска в данном словаре мы используем представление его в виде префиксного дерева, состоящего из множества узлов. В каждом узле дерева хранится метка – один из символов алфавита $A = \{a_1, a_2, \ldots, a_d\}$. Ключом, который соответствует некоторому узлу, является путь от корня дерева до узла, а точнее строка $c_1c_2\ldots c_m$, составленная из меток узлов, повстречавшихся на этом пути. Если соответствующая строка есть в словаре, то с узлом ассоциируется индекс, являющийся ее порядковыми номером в словаре. В словаре может быть несколько словоформ-омонимов с одинаковым написанием, но отличающихся по смыслу или морфологической информацией. Поэтому вместо одного упомянутого индекса может присутствовать некоторый список таких индексов. Наличие одного или нескольких индексов указывает на то, что узел является конечным, то есть соответствующим концу некоторого слова. В противном случае узел является лишь промежуточным по дороге в какой-либо другой, который является конечным. Корень дерева, очевидно, соответствует пустому ключу. Имеет смысл также для каждого узла использовать множество дочерних узлов следующего нижнего уровня (смежные узлы – потомки). Количество их может меняться от *0* до *d*.

Хранение дерева в памяти осуществляется с помощью списка всех его узлов и списка номеров смежных узлов – потомков для каждого из узлов. Для узлов, не имеющих ни одного потомка, список смежных узлов пуст. Если в дереве имеется *L* узлов, пронумерованных 0,…,*L-1*, то в памяти будет храниться *L* списков потомков, собранных в главный список.

0: {1, 5, 9, 15}
1: {2, 4}
2: {3}
…
L-1: {}

Порядковый номер каждого списка смежных узлов соответствует номеру данного узла в списке.

В списке узлов хранятся данные для каждого узла, состоящие из двух полей – символа алфавита и списка индексов строк (для промежуточных узлов последний список будет пустым).

0: [' ',{}]
1: ['а',{0,1,2}]
2: ['б',{}]
3: ['а',{3}]
4: ['ж',{}]

Алгоритм поиска омонимов с использованием дерева можно описать следующим образом. Пусть есть слово или словосочетание, для которого необходимо найти все возможные совпадающие по написанию формы (омонимы) в словаре, представленном в виде дерева. Будем спускаться из корня дерева на нижние уровни, каждый раз переходя в узел, чей символ совпадает с очередным символом строки. После того как обработаны все символы строки, узел, в котором остановился спуск, и будет искомым узлом. Индекс, ассоциированный с этим узлом, есть номер строки в словаре. Если в процессе спуска не нашлось узла с символом, соответствующим очередному символу строки, или спуск остановился на промежуточной вершине (с пустым списком индексов), то искомый ключ отсутствует в дереве, а строка – в словаре.

Таким образом, для нахождения всех омонимов деепричастий нужно для каждого из деепричастий, входящих в словарь, произвести поиск совпадающих по написанию словоформ по вышеописанному алгоритму. Если список индексов для данной словоформы содержит более одного индекса (т.е. таких словоформ в словаре несколько), значит, деепричастие имеет омонимы.

Ниже приведены некоторые примеры найденных в словаре омонимов деепричастий:

бая | *сущ одуш ед муж род*
бая | *сущ одуш ед муж вин*
бая | *деепр несов пер/не наст*

благодаря | *деепр несов перех наст*
благодаря | *предл дат*

богатея | *сущ одуш ед муж род*
богатея | *сущ одуш ед муж вин*
богатея | *деепр несов непер наст*

буря | *деепр несов перех наст*
буря | *сущ неод ед жен им*
…

Среди них можно выделить следующие виды омонимии деепричастий: омонимия деепричастий и предлогов, деепричастий и существительных, деепричастий и прилагательных, деепричастий и причастий, омонимия деепричастий переходного и непереходного глагола. Рассмотрим эти случаи омонимии и сформулируем некоторые правила для ее автоматического разрешения.

2 Общие правила для автоматического снятия омонимии деепричастия

Правило 1: Пусть предложение содержит омоним деепричастия. Для того чтобы это действительно было деепричастие, необходимо, чтобы оно содержалось в отрезке предложения, выделенном знаком препинания в начале или в конце предложения или с двух сторон в середине предложения.

Правило 2: Если отрезок содержит предикатив, личную форму глагола, краткое прилагательное или краткое причастие, то он не может содержать деепричастие. Исключение «будучи». Пример: «*Мальчик, будучи* (прич) *определен в кадетский корпус, с раннего детства жил вне семьи*». В то же время возможно деепричастие с инфинитивом. Пример: «*Стремясь* (деепр) *помочь, он потянул дверь*».

3 Омонимы деепричастие-существительное

В словаре [2] имеются следующие омонимы деепричастие-существительное:

бая, богатея, буря, воя, гвоздя, горя, гостя, доля, душа, ежа, заезжая, залив, заплыв, запоров, застав, клея, клича, корча, кроя, лая, лишая, мая, меча, моля, моря, моча, мытаря, нагоняя, надоев, нажив, налив, напев, нарыв, неволя, обогрев, отлив, отрыв, отсев, отстоя, пав, паря, переборов, перелив, перерыв, пища, плача, плюща, поборов, подлив, подогрев, пожив, полив, поля, пошив, приезжая, приколов, прилив, пристав, проезжая, проколов, пролив, прорыв, разлив, размыв, разогрев, разрыв, расколов, рея, ржа, родня, роя, руля, свища, сев, селя, сеча, сколов, слив, споров, сторожа, строя, суша, туша, уколов, устав, хворая, хромая, чая.

Данные омонимы различаются в первую очередь по правилам 1 и 2.

Пример: *Свиная туша* (сущ) *у мясника стоит тысячу рублей.*

Задохнулся, туша (деепр) *пожар в своей мастерской.*
От громкого лая (сущ) *собак звенело в ушах.*
Понесся дальше, сбивая снежинки хвостом и лая (деепр) *от счастья.*

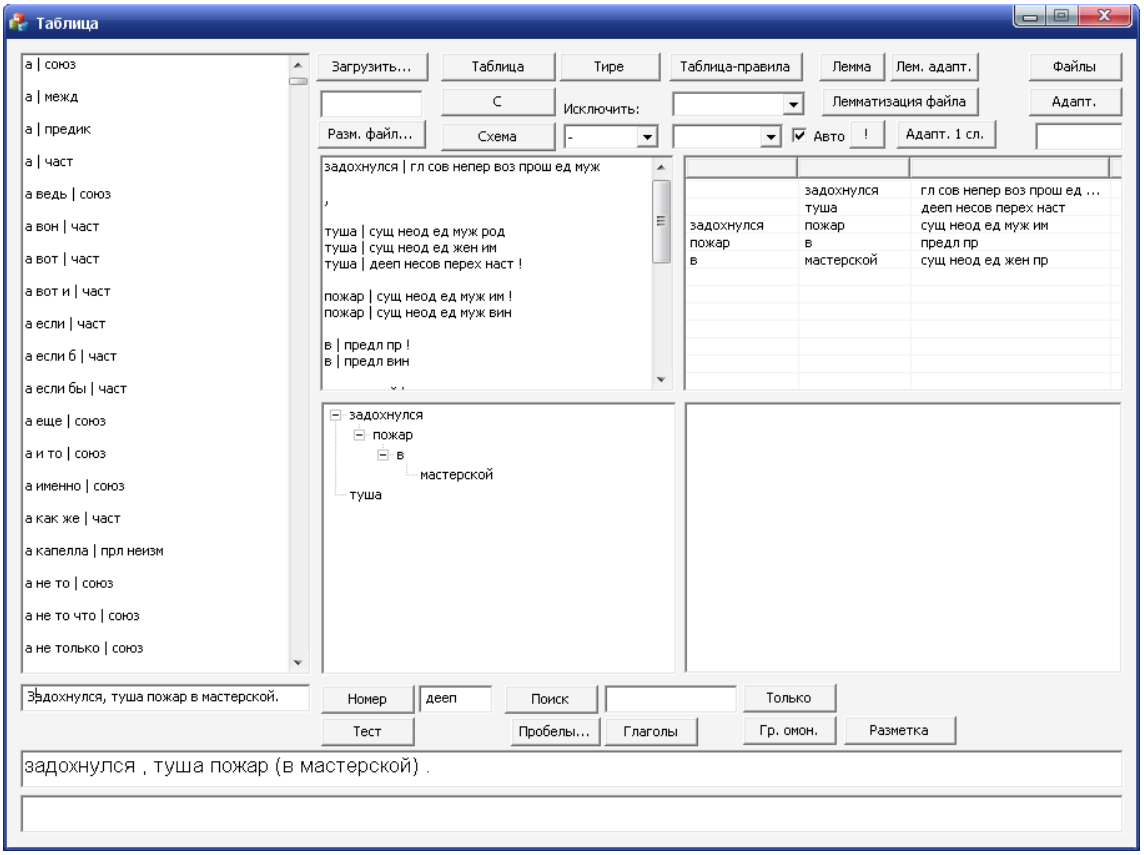

Рисунок 1 – Экранная форма экспериментального программного обеспечения

4 Омонимы деепричастий и предлогов

благодаря, включая, для, исключая

Правило 3: Если после слова есть существительное, которое согласуется по падежу с предлогом, то выбирается предлог. В противном случае – деепричастие.

Омоним *благодаря* определяется как предлог, если далее в пределах отрезка найдется существительное (местоимение-существительное) в дательном падеже. В противном случае это деепричастие.

Омоним *для* определяется как предлог, если далее в пределах отрезка найдется существительное (местоимение-существительное) в родительном падеже. В противном случае это деепричастие.

Омонимы *включая, исключая* как в случае с предлогом, так и в случае с деепричастием употребляются совместно с существительным в родительном падеже, поэтому требуют дополнительных исследований. Примеры: «*Уехали все, исключая* (предл) *древних стариков*». «*Он действовал жестко, исключая* (деепр) *нерадивых учеников*».

5 Омонимы деепричастий и прилагательных

бухая, вещая, витая, вшив, горяча, заезжая, приезжая, проезжая, синя, скупая, строгая, хворая, хромая

Правило 4: Если на отрезке, содержащем омоним деепричастие-прилагательное, есть существительное, согласованное с предполагаемым прилагательным, то это действительно прилагательное.

Пример: *Скупая* (прил) *старуха выжила в самые сложные времена. Он выживал в сложные времена, скупая* (дееп) *ценности.*

6 Омонимы деепричастий и причастий

витая, обитая, питая

Правило 5:

1) если за словом «*обитая*» следует предлог или слова «*там*», «*тут*», «*здесь*», то это деепричастие. В остальных случаях – причастие. Примеры: *Он жил, обитая* (дееп) *здесь уже год. Там дверь, обитая* (прич) *железом.*

2) если за словом «*питая*» следует существительное в винительном падеже, то это деепричастие. В остальных случаях – причастие. Пример: *В стакане вода, питая* (прич) *вчера. Дельта разливается, обильно питая* (дееп) *окрестности водой.*

3) если за словом «*витая*» следует предлог «*из*», или слова «*с помощью*», «*с использованием*», то это причастие. В остальных случаях – деепричастие. Примеры: *Веревка, витая* (прич) *из рогожи. Он жил, витая* (дееп) *в облаках.*

7 Омонимы деепричастий переходного и непереходного глагола

вкалывая, досадив, досадивши, досыпая, запахнув, запахнувши, засыпая, метя, мешая, наказывая, находя, недосыпая, отсыпая, отточив, отточивши, парируя, перетрусив, перетрусивши, пикировав, пикируя, планируя, повалив, поваливши, поводя, помешав, помешавши, поправив, поправивши, провалив, проваливая, проваливши, саботируя, свалив, сваливая, сваливши, снуя, тая, точа, узрев, узревши, целя

Правило 6: Если есть омонимы деепричастий переходного и непереходного глагола, то деепричастие переходного глагола выбирается, если за ним следует существительное в винительном падеже, в противном случае выбирается деепричастие непереходного или пер/не глагола. Примеры: *Он работал с утра, кося* (перех) *траву. Она сидела, кося* (неперех) *в его сторону.*

Омонимы деепричастий переходного и переходного/непереходного глагола:

выхаживая, грабя, дожав, дожавши, коля, кося, лупя, меля, метая, метя, наболтав, наболтавши, наваливая, нажав, нажавши, наказав, наказавши, наметав, наметавши, написав, написавши, оря, отжав, отжавши, отколов, отколовши, перевесив, перевесивши, перевешивая, пережав, пережавши, подкосив, подкосивши, пожав, пожавши, потопив, потопивши, причитая, просыпая, разрывая, сжав, сжавши, скашивая, скосив, скосивши, смыкая, спланировав, спланировавши, стегая, стирая, считая, теша, топя.

Примеры: *Выхаживая* (перех) *больного брата, она сильно уставала. Он каждый день вставал пораньше, выхаживая* (**пер**/не) *свою милю. Он каждый день вставал пораньше, выхаживая* (пер/**не**) *по берегу реки.*

Омонимы деепричастий непереходного и переходного/непереходного глагола:

бухая, бухнув, бухнувши, валя, вспылив, вспыливши, высыпая, затрусив, затрусивши, метя, пересыпая, перши, поболтав, поболтавши, пописав, пописавши, потрусив, потрусивши, следя, труся, хватая, чертя

Примеры: *Он все-таки пошел, сильно при этом труся* (непер). *Труся* (**пер**/не) *сумочку, она выронила зеркальце. Телега ехала, труся (*пер/**не***) соломой*.

Эти примеры иллюстрируют также омонимию (на письме) деепричастий, отличающихся ударением. Вот их список:

блок\ируясь, блокир\уясь, бр\едя, бред\я, заб\ухав, забух\ав, заб\ухавши, забух\авши, оп\исавши, опис\авши, оп\исав, опис\ав, оп\исавшись, опис\авшись, перед\охнувши, передохн\увши, перед\охнув, передохн\ув, р\оясь, ро\ясь, ш\икнув, шикн\ув.

Существуют также деепричастия-омонимы с чисто семантическим отличием (лексическая омонимия). Пример: *«взрывая»* (одно значение от «рыть», другое – от «устраивать взрыв»). Здесь классификация определяется контекстом и на сегодняшний день не может быть выполнена автоматически с помощью морфологической и синтаксической информации без использования семантики.

Заключение

В статье приведено описание алгоритмов автоматического снятия омонимии в группах омонимов, содержащих деепричастия. Разработанные алгоритмы были реализованы с использованием языка программирования C++ в экспериментальном программном обеспечении для снятия омонимии. Полученные результаты могут быть использованы для автоматизации морфологической разметки естественно-языковых текстов со снятием неоднозначности, что позволит повысить точность классификации и кластеризации текстов, улучшить качество машинного перевода, информационного поиска и других приложений.

Список литературы

1. Национальный корпус русского языка. [Электронный ресурс] – URL: http://ruscorpora.ru/new/index.html. (дата обращения: 10.05.2021).
2. Хаген М. Полная парадигма. Морфология [Электронный ресурс] / М. Хаген // Форум «Говорим по-русски» [сайт]. – 2018. – URL: http://www.speakrus.ru/dict/#morph-paradigm (дата обращения: 19.11.2018)
3. Александрова З. Е. Словарь синонимов русского языка: Практический справочник: Ок. 11 000 синоним. рядов. – 11 изд., перераб. и доп. [Текст] / З. Е. Александрова. – М.: Рус. яз., 2001. – 568 с.
4. Ляшевская О. Н. Частотный словарь современного русского языка (на материалах Национального корпуса русского языка) [Текст] / О. Н. Ляшевская, С. А. Шаров. – М. : Азбуковник, 2009.
5. Ниценко А. В. О подчинительном дереве для простого распространенного русского предложения [Текст] / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Международный рецензируемый научно-теоретический журнал «Проблемы искусственного интеллекта» . – 2019. – № 2 (13). – С. 63–73. – ISSN 2413-7383.
6. Большакова С. А. К вопросу об автоматическом снятии омонимии русских деепричастий [Текст] / С. А. Большакова // Материалы донецкого международного круглого стола «Искусственный интеллект: теоретические аспекты и практическое применение». – Донецк : ГУ «ИПИИ», 2021.
7. Полякова И. Н. Автоматизация проблемы разрешения функциональной омонимии в русском языке [Электронный ресурс] / И. Н. Полякова, В. А. Крутов // Материалы по международной научно-практической Интернет-конференции «Перспективные инновации в науке, образовании, производстве и транспорте ‘2012» URL: https://sworld.com.ua/index.php/technical-sciences-212/informatics-computer-science-and-automation-212/13178-212-446 (дата обращения: 23.09.2021).

References

1. *Natsional'nyy korpus russkogo yazyka* [The National Corpus of the Russian language]. URL: http://ruscorpora.ru/new/index.html. (accessed: 10.05.2021).
2. Hagen M. Polnaya paradigma. Morfologiya [The complete paradigm. Morphology] [Electronic resource]. *Forum «Govorim po-russki»* [Forum "We speak in Russian"] [website], 2018, URL: http://www.speakrus.ru/dict/#morph-paradigm (accessed: 19.11.2018)
3. Alexandrova Z. E. *Slovar' sinonimov russkogo yazyka: Prakticheskiy spravochnik: Ok. 11 000 sinonim. ryadov* [Dictionary of synonyms of the Russian language: A practical reference book: About 11,000 synonyms. rows.], 11 ed., reprint. and add, M., Rus. yaz., 2001, 568 p.
4. Lyashevskaya O. N., Sharov S. A. *Chastotnyy slovar' sovremennogo russkogo yazyka (na materialakh Natsional'nogo korpusa russkogo yazyka)* [Frequency dictionary of the modern Russian language (based on the materials of the National Corpus of the Russian Language)], Moscow, Azbukovnik, 2009.
5. Nitsenko A.V., Shelepov V. Yu., Bolshakova S. A. O podchinitel'nom dereve dlya prostogo rasprostranennogo russkogo predlozheniya [On the subordinate tree for a simple common Russian sentence]. *Mezhdunarodnyy retsenziruyemyy nauchno-teoreticheskiy zhurnal «Problemy iskusstvennogo intellekta»* International peer-reviewed scientific and theoretical journal "Problems of artificial Intelligence". 2019. № 2 (13), pp. 63-73, ISSN 2413-7383.
6. Bolshakova S. A. K voprosu ob avtomaticheskom snyatii omonimii russkikh deyeprichastiy [On the issue of automatic removal of homonymy of Russian adverbs]. *Materialy donetskogo mezhdunarodnogo kruglogo stola «Iskusstvennyy intellekt: teoreticheskiye aspekty i prakticheskoye primeneniye»* [Materials of the Donetsk international round table "Artificial intelligence: theoretical aspects and practical application"], Donetsk, SI "IPAI", 2021.
7. Polyakova I. N., Krutov V. A. Avtomatizatsiya problemy razresheniya funktsional'noy omonimii v russkom yazyke [Automation of the problem of resolving functional homonymy in the Russian language]. *Materialy po mezhdunarodnoy nauchno-prakticheskoy Internet-konferentsii «Perspektivnyye innovatsii v nauke, obrazovanii, proiz-vodstve i transporte '2012*» [Materials on the international scientific and practical Internet conference "Promising innovations in science, education, production and transport '2012"] URL: https://sworld.com.ua/index.php/technical-sciences-212/informatics-computer-science-and-automation-212/13178-212-446 (accessed: 23.09.2021).

RESUME

S. A. Bolshakova, A. V. Nitsenko, V. Yu. Shelepov

On the question of automatic disambiguation of Russian adverbial participles

Automatic text processing is one of the oldest areas of artificial intelligence. One of the most important stages of text processing is morphological analysis during which morphological features and the initial form of words are determined. The task of morphological tagging is complicated by homonymy.

The article describes rule based approach for automatic disambiguation in homonymous groups containing adverbial participles.

The developed algorithms were implemented using the C++ programming language in experimental disambiguation software.

The results obtained can be used to automate the morphological tagging of natural language texts with disambiguation, which will improve the accuracy of text classification and clustering, improve the quality of machine translation, information search and other applications.

РЕЗЮМЕ

С. А. Большакова, А. В. Ниценко, В. Ю. Шелепов

К вопросу об автоматическом снятии омонимии русских деепричастий

Автоматическая обработка текста – одно из старейших направлений искусственного интеллекта. Одним из важнейших этапов обработки текста является морфологический анализ, в процессе которого определяются морфологические характеристики и начальная форма слов. Задача морфологической разметки осложняется омонимией.

В статье приведено описание подхода, основанного на правилах, к автоматическому снятию омонимии в группах омонимов, содержащих деепричастия.

Разработанные алгоритмы были реализованы с использованием языка программирования C++ в экспериментальном программном обеспечении для снятия омонимии.

Полученные результаты могут быть использованы для автоматизации морфологической разметки естественно-языковых текстов со снятием неоднозначности, что позволит повысить точность классификации и кластеризации текстов, улучшить качество машинного перевода, информационного поиска и других приложений.

Статья поступила в редакцию 27.09.2021.