УДК 004.89:004.93

С. А. Большакова
ГУ «Институт проблем искусственного интеллекта», г. Донецк
283048, г. Донецк, ул. Артема, 118-б

О СНЯТИИ ОМОНИМИИ «ПРЕДИКАТИВ-ПРЕДЛОЖНАЯ ГРУППА» ДЛЯ НЕКОТОРЫХ РАСПРОСТРАНЕННЫХ СЛОВОСОЧЕТАНИЙ В РУССКОЯЗЫЧНЫХ ТЕКСТАХ

S. A. Bolshakova
PI «Institute of Artificial Intelligence Problems»
283048, Donetsk, str. Artema, 118-b

ON THE «PREDICATIVE-PREPOSITIONAL GROUP» DISAMBIGUATION FOR SOME COMMON WORD COMBINATIONS IN RUSSIAN TEXTS

Статья посвящена проблеме автоматического разрешения неоднозначности для ряда словосочетаний в случае, когда один из вариантов является предикативным словосочетанием, а другой – группой слов с предлогом и существительным. В первом случае словосочетание интерпретируется как единое целое – элемент словаря с характеристикой части речи «предикатив». Во втором случае оно является предложной группой и требует представления в виде отдельных слов словаря.
Ключевые слова: обработка естественного языка, автоматический анализ текста, снятие омонимии, предикативное словосочетание, предложная группа, национальный корпус.

The article is devoted to the problem of automatic disambiguation for a number of phrases in the case when one of the options is a predicative phrase, and the other is a group of words with a preposition and a noun. In the first case, the phrase is interpreted as a single dictionary element with a "predicate" part of speech tag. In the second case, it is a prepositional group and requires presentation as separate dictionary words.
Key words: natural language processing, automatic text analysis, disambiguation, predicative phrase, prepositional group, national language corpus.

Введение

Одной из проблем, возникающей при машинном анализе текстов на естественном языке, являются слова-омонимы. Процедуры снятия омонимии являются необходимым и важным этапом для качественного анализа и обработки текстов. Исследования в этом направлении активно ведутся и в настоящее время. На сегодняшний день проблема является нерешенной, хотя ею занимается в мире довольно большое количество специалистов, работающих с различными естественными языками [1–5].

Статья посвящена проблеме автоматического разрешения неоднозначности для ряда словосочетаний в случае, когда один из вариантов является предикативным словосочетанием, а другой – группой слов с предлогом и существительным. В первом случае словосочетание интерпретируется как единое целое – элемент словаря с характеристикой части речи «предикатив». Во втором случае оно является предложной группой и требует представления в виде отдельных слов словаря.

Алгоритм снятия омонимии словосочетаний «предикатив-предложная группа» работает с файлом *Предл гр.txt* [6-11]. Этот файл состоит из групп слов (иногда с метками справа), относящихся к словосочетаниям (управляющие группы). Программа просматривает отрезок текста между двумя соседними знаками препинания, содержащий словосочетание, и интерпретирует последнее как предложную группу, если вместе с ним отрезок содержит словоформу слова из управляющей группы. В противном случае словосочетание интерпретируется как предикатив.

Ниже приведены управляющие группы для ряда словосочетаний, которые могут быть только омонимами «предикатив-предложная группа». Результаты получены в результате анализа *всех* найденных предложений из Национального корпуса русского языка с рассматриваемыми словосочетаниями. Ради экономии места управляющие группы, в основном, записаны в две колонки. Обсуждаемое словосочетание снабжено в файле восклицательным знаком (здесь также выделено жирным курсивом).

1.

в курсе !(предл гр)

<список дисциплин>
быть (-1) (предик-предл гр)
гений [предик]
глава
заправила (предик-предл гр)
звук
идея (-1)
излагать
изложить
изучать
изучить
использовать
использоваться
к (1)
лекция
монета
некто
обучение
падать
упадать
упасть
подниматься
подход
причина
| род (предик-предл гр)
пройти
проходить
разбираться
разобраться
расхваливать
расхвалить
революция (предл гр-предик)
речь
сообщение
ставить
поставить
уделять
уделить
уделяться
участие
учить
научить
обучать
обучить
обучаться
обучиться
читаться
чтение

2.

в ударе !(предл гр)

быстрота
двигаться
же (1)
заключаться
захватить
иметься
использование
клавиша
кровь
ошибиться
падать
плечо
практиковаться
присутствовать
| тв (1)
развернуться
резкость
смысл
смычок
срастаться
участвовать
участвующий
участие
четкость
по&оргкомитету
по&москве
по&мячу
+, в бою

3.

в чести !(предл гр)

блюсти
дело&только
завидовать
мой (1)
нарицание
непреклонный
ни (-1)
отказать
отказывать
понижать
понимать
| род (1)
потребность
пребывать [предик]
прибыток
расти
сомневаться
убежденность
уверен&будучи
умаление
участие
явиться
+, которую

4.

на сносе !(предл гр)

делаться
кладбище
настаивать

Запись <список дисциплин> означает, что программа должна обратиться к файлу *Список дисциплин. txt* . Если она найдет в нем слово, словоформа которого содержится в анализируемом отрезке текста между двумя знаками препинания, то словосочетание есть предложная группа. Нижеприведенный список, естественно, может быть пополнен.

Список диципли:
агробиология, агрономия, агрофизика, агрохимия, агроэкология, аквакультура, акупунктура, акустика, акушерство, алгебра, альгология, анатомия, андрология, анестизиология, антропогенез, антропология, архитектура, астробиология, астрономия, астрофизика, аэробиология, аэрофотосъемка, библиология, биогеография, биоинформатика, биология, биотехнология, биофизика, биохимия, ботаника, ветеринария, виноградарство, вирусология, гематология, геммология, генетика, география, геодезия, геология,

геометрия, геоморфология, геофизика, геохимия, гериатрия, герменевтика, герпетология, гидрогеология, гидрография, гидрология, гинекология, гистология, гляциология, граждановедение, диабетология, диетология, дизайн, драматургия, египтология, живопись, журналистика, зоология, зубопротезирование, иммунология, интерлингвистика, информатика, искусствоведение, история, ихтиология, каллиграфия, кардиология, кардиохирургия, картография, кибернетика, киноведение, климатология, кораблестроение, космонавтика, криобиология, криогеника, криптозоология, ксенобиология, культурология, лесоводство, лимнология, лингвистика, литература, литературоведение, литология, логика, логистика, маркетинг, математика, материаловедение, медицина, менеджмент, метеорология, механика, микология, микробиология, минералогия, мифология, морфология, музыка, музыковедение, неврология, нейробиология, нейропсихология, нейрохиругия, нейрохирургия, нефрология, нутрициология, океанография, океанология, онкология, онтология, оология, оптика, орнитология, ортодонтия, ортопедия, отоларингология, офтальмология, палеоантропология, палеобиология, палеогеография, палеоклиматология, палеонтология, паразитология, парапсихология, патология, педагогика, педиатрия, периодонтология, планетология, политология, помология, поциолингвистика, почвоведение, поэтика, право, прагматика, приматология, психиатрия, психоанализ, психология, психометрия, психопатология, психотерапия, психофизика, пчеловодство, радиология, ревматология, религиеведение, рисование, риторика, садоводство, семантика, семиотика, серология, синтаксис, систематика, системотехника, скульптура, социолингвистика, социология, спелеология, статистика, стоматология, строительство, сценография, таксономия, теория, терапия, тератология, термодинамика, технология, травматология, трансфузиология, урология, фармакология, фармация, физика, физиология, физиотерапия, филология, философия, фольклор, фонетика, фонология, фотография, хемоинформатика, химия, хирургия, хореография, хронобиология, цитология, экология, экономика, электромагнетизм, электротехника, эмбриология, эндокринология, энология, энтомология, эпидемиология, эпистемология, эстетика, этика, этноботаника, этнография, этноистория, этнолингвистика, этнология, этология,

Метка (1) означает, что соответствующая словоформа должна находиться сразу за словосочетанием. Аналогично метка (-1) обозначает непосредственное предшествование словосочетанию. Наличие метки [предик] значит, что наличие соответствующей словоформы внутри рассматриваемого отрезка текста превращает словосочетание в предикатив. Метка (предик-предл гр) означает, что для словосочетания возможны оба варианта в зависимости от контекста. Программа по умолчанию выберет предикатив. При метке (предл гр-предик) – наоборот. Запись | род обеспечивает выбор предложной группы при наличии после словосочетания в пределах отрезка родительного падежа существительного или местоимения-существительного. Запись | род(1) то же для непосредственного следования. Запись | тв (1) – аналог для творительного падежа.

Заключение

В статье приведено описание алгоритмов автоматического снятия омонимии словосочетаний, которые могут выступать в роли предикативов, разработанных с использованием данных Национального корпуса русского языка. Алгоритмы были реализованы на языке программирования C++ в экспериментальном программном обеспечении для снятия омонимии. Полученные результаты могут быть использованы для повышения точности морфологической разметки при автоматической обработке текстов на русском языке.

Список литературы

1. Порохнин А.А. Анализ статистических методов снятия омонимии в текстах на русском языке [Текст] // Вестник астраханского государственного технического университета. – 2013. – № 2. – С. 168-174.
2. Лесько О. Н. Использование онтологии предметной области для снятия омонимии в естественно-языковых текстах [Текст] / О. Н. Лесько, Ю. В. Рогушина // Проблеми програмування. – 2017. – № 2. – С. 61-71.
3. Рысаков С.В. Статистические методы снятия омонимии [Текст] / С.В.Рысаков, Э.С.Клышинский // Новые информационные технологии в автоматизированных системах. – 2015. – № 18. – С. 555-563.
4. Епифанов М. Е. Итеративное применение алгоритмов снятия частеречной омонимии в русском тексте [Текст] / М.Е. Епифанов, А.Ю. Антонова, А.М. Баталина, Т.Ю. Кобзарева, Д.Г. Лахути // Компьютерная лингвистика и интеллектуальные технологии: материалы Международной конференции «Диалог-2010» (Бекасово, 26–30 мая 2010 г.). – 2010. – С.119–123.
5. Хаген, М. А. Полная парадигма. Морфология [Электронный ресурс]. – Режим доступа : URL: http://www.speakrus.ru/ dict/#morph-paradigm (дата обращения: 10.12.2021).
6. Ниценко, А. В. О снятии омонимии словосочетаний, которые могут быть предикативами [Текст] / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Проблемы искусственного интеллекта. – 2021. – № 1(20). – С. 53–63.
7. Ниценко, А. В. К вопросу об автоматическом снятии омонимии русских предикативов [Текст] / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Сборник трудов VIII Международной конференции «Знания-Онтологии-Теории» (г. Новосибирск, 8-12 ноября 2021г.) – 2021. – С. 218–225.
8. Ниценко, А. В. Об автоматическом снятии омонимии предикативных словосочетаний. Результаты работы с национальным корпусом русского языка [Текст] / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Проблемы искусственного интеллекта. – 2021. – № 3(22). – С. 46–56.
9. Ниценко, А. В. Исследование омонимии предикативных словосочетаний на основе национального корпуса русского языка [Электронный ресурс] / А. В. Ниценко, В. Ю. Шелепов, С. А. Большакова // Сборник трудов VII Международной научно-технической конференции «Современные информационные технологии в образовании и научных исследованиях» (г. Донецк, 23 ноября 2021г.). 2021. – Режим доступа: http://pm.conf.donntu.org/index.php (дата обращения: 13.12.2021).
10. Большакова, С. А. К вопросу о снятии омонимии в некоторых группах омонимов, включающих предикатив [Текст] / С. А. Большакова, А. В. Ниценко, В. Ю. Шелепов // Искусственный интеллект: теоретические аспекты и практическое применение: материалы Донецкого международного научного круглого стола. – Донецк : ГУ «Институт проблем искусственного интеллекта» (ГУ «ИПИИ»), 2022. – 216 с. – С. 152–158.
11. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа : URL: http://www.ruscorpora.ru/ (дата обращения: 10.12.2021).

References

1. Porokhnin A.A. Analiz statisticheskikh metodov snyatiya omonimii v tekstakh na russkom yazyke [Analysis of statistical methods for removing homonymy in Russian texts] *Vestnik astrakhanskogogo sudarstvennogo tekhnicheskogo universiteta* [Bulletin of the Astrakhan State Technical University] 2013, No 2, pp. 168-174.
2. Les'ko O. N., Rogushina Yu. V. Ispol'zovaniye ontologii predmetnoy oblasti dlya snyatiya omonimii v yestestvenno-yazykovykh tekstakh [Using the domain ontology for removing homonymy in naturallanguage texts]. *Problemi programuvannya* [Problems of the program] 2017, No 2, pp. 61-71.
3 Rysakov S.V., Klyshinskiy E.S. Statisticheskiye metody snyatiya omonimii [Statistical methods for removing homonymy]. *Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh* [New information technologies in automated systems], 2015, No 18, pp. 555-563.
4. Epifanov M. E. Iterative application of algorithms for removing partial homonymy in the Russian text [Text] / M.E. Epifanov, A.Y. Antonova, A.M. Batalina, T.Y. Kobzareva, D.G. Lahuti // Computational linguistics and intellectual technologies: materials of the International Conference "Dialog-2010" (Bekasovo, May 26-30, 2010). – 2010. – pp.119–123.

5. 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)
6. A. V. Nicenko, V. Ju. Shelepov, S. A. Bolshakova, K. S. Ivashko [On the disambiguation of word combinations that may be predicatives] *Problemy iskusstvennogo intellekta - Problems of artificial intelligence* 2020. No. 1(20). (In Russ.) – pp. 53–63.
7. A.V. Nitsenko, V. Yu. Shelepov, S. A. Bolshakova [On the question of automatic removal of homonymy of Russian predicatives] Proceedings of the VIII International Conference "Knowledge-Ontology-Theory" (Novosibirsk, November 8-12, 2021) – 2021. – pp. 218-225
8. A. V. Nicenko, V. Ju. Shelepov, S. A. Bolshakova [On automatic disambiguation of predicative word combinations. Results of work with the national corpus of the russian language] *Problemy iskusstvennogo intellekta - Problems of artificial intelligence* 2021. No. 3(22). (In Russ.) – pp. 46–56.
9. A.V. Nitsenko, V. Yu. Shelepov, S. A. Bolshakova [Research of homonymy of predicative phrases based on the national corpus of the Russian language [Electronic resource] / // Proceedings of the VII International Scientific and Technical Conference "Modern information technologies in education and scientific research" (Donetsk, November 23, 2021.). 2021. – Access mode: http://pm.conf.donntu.org/index.php (accessed: 13.12.2021).
10. S. A. Bolshakova, A.V. Nitsenko, V. Yu. Shelepov [On the issue of removing homonymy in some groups of homonyms including a predicative] Artificial intelligence: theoretical aspects and practical application: materials of the Donetsk International Scientific Round Table. – Donetsk: GU "Institute of Artificial Intelligence Problems", 2022. – 216 p. – pp. 152-158
11. Natsional'nyy korpus russkogo yazyka [The National Corpus of the Russian language]. URL: http://ruscorpora.ru/new/index.html. (accessed: 10.05.2021).

RESUME

S. A. Bolshakova

On the «predicative-prepositional group» disambiguation for some common word combinations in russian texts

The problem of word disambiguation is one of the most important in the tasks of automatic natural language processing. For the Russian language, this problem is especially relevant, since the number of homonyms in it is very large. Disambiguation are a necessary and important step for qualitative analysis and processing of texts.

The article is devoted to the problem of automatic disambiguation for a number of phrases in the case when one of the options is a predicative phrase, and the other is a group of words with a preposition and a noun. In the first case, the phrase is interpreted as a single dictionary element with a "predicate" part of speech tag. In the second case, it is a prepositional group and requires presentation as separate dictionary words.

Algorithms for "predicative-prepositional group" disambiguation removal are developed on the basis of data from the Russian National Corpus. The developed algorithms were implemented in experimental disambiguation software using the C++ programming language.

The results obtained can be used to improve the accuracy of morphological tagging in automatic Russian texts processing.

РЕЗЮМЕ

С. А. Большакова

О снятии омонимии «предикатив-предложная группа» для некоторых распространенных словосочетаний в русскоязычных текстах

Проблема разрешения неоднозначности слов является одной из важнейших в задачах автоматической обработки естественного языка. Для русского языка эта проблема особенно актуальна, поскольку количество омонимов в нем очень велико. Процедуры снятия омонимии являются необходимым и важным этапом для качественного анализа и обработки текстов.

Статья посвящена проблеме автоматического разрешения неоднозначности для ряда словосочетаний в случае, когда один из вариантов является предикативным словосочетанием, а другой – группой слов с предлогом и существительным. В первом случае словосочетание интерпретируется как единое целое – элемент словаря с характеристикой части речи «предикатив». Во втором случае оно является предложной группой и требует представления в виде отдельных слов словаря.

Разработаны алгоритмы автоматического снятия омонимии «предикатив-предложная группа» на основе данных Национального корпуса русского языка. Алгоритмы были реализованы с использованием языка программирования C++ в экспериментальном программном обеспечении для снятия омонимии.

Полученные результаты могут быть использованы для повышения точности морфологической разметки при автоматической обработке текстов на русском языке.

Статья поступила в редакцию 02.02.2023.