DOI: 10.15514/ISPRAS-2022-34(5)-11

Контекстное разрешение омонимии на основе центроидно-контекстной модели

[1,2,3] Александр А. Хорошилов, ORCID: 0000-0003-4885-3232 <khoroshilov@mail.ru>
[1,4] Ю.В. Никитин, ORCID: 0000-0002-7641-0247 <yuri.v.nikitin@gmail.com>
[2,5] А.В. Кан, ORCID: 0000–0001–9410–406X <kanav@nrczh.ru>
[4] Я.Д. Козловская, ORCID: 0000–0002–1780–5687 <yana_kozlovskaia@mail.ru>
[2] Е.А. Евдокимова, ORCID: 0000-0003-4719-2786 <evdokimovaekan@mail.ru>

[1] Федеральный исследовательский центр «Информатика и управление» РАН,
119333, Россия, г. Москва, Вавилова, д.44, кор.2
[2] Московский авиационный институт (национальный исследовательский университет),
125993, Россия, г. Москва, Волоколамское шоссе, д. 4
[3] 27-й Центральный научно-исследовательский институт МО РФ
123007, г. Москва, 1-й Хорошевский пр-д, д. 5
[4] Научно-промышленная компания «Высокие технологии и стратегические системы»,
107023, Россия, Москва, ул. Электрозаводская, д. 27, стр. 9
[5] Национальный исследовательский центр «Институт имени Н.Е. Жуковского»,
125319, Россия, г. Москва, ул. Викторенко, д.7

Аннотация. В статье описывается новый метод контекстного разрешения омонимии на основе центроидно-контекстной модели (ЦКМ). Предлагаемый метод выявления случаев омонимии в корпусе текстов и ее разрешения с помощью модели ЦКМ базируется на теоретической концепции фразеологического концептуального анализа текстов (ФКАТ) и уникальной машинной грамматике, в основу которой положена система флективных классов русских слов. Заложенное в теоретической концепции флективных классов слов русского языка жесткое соответствие между формой представления слов и их грамматической информацией позволило создать на этой основе новые классы – классы слов, имеющие одинаковые наборы грамматических признаков, соответствующие их формам представления в сходных контекстных окружениях. При разработке этой модели авторы исходили из следующей гипотезы: одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры различных фрагментов текстов. При этом предполагалось, что такая гипотеза верна для любых синтаксических моделей и может быть полезна при решении как глобальных, так и частных задач анализа текста. С помощью этого метода было предложено новое решение задачи разрешение омонимии на основе предлагаемой модели ЦКМ.

Ключевые слова: разрешение омонимии; омонимы; центроидно-контекстная модель; обобщенные синтагмы

Для цитирования: Хорошилов Александр А., Никитин Ю.В., Кан А.В., Козловская Я.Д., Евдокимова Е.А. Контекстное разрешение омонимии на основе центроидно-контекстной модели. Труды ИСП РАН, том 34, вып. 5, 2022 г., стр. 171-182. DOI: 10.15514/ISPRAS-2022-34(5)-11

Context resolution of homonymy based on a centroid-context model

[1,2,3] Alexander A. Khoroshilov, ORCID: 0000-0003-4885-3232 <khoroshilov@mail.ru>
[1,4] Yu.V. Nikitin, ORCID: 0000-0002-7641-0247 <yuri.v.nikitin@gmail.com>
[2,5] A.V. Kan, ORCID: 0000–0001–9410–406X <kanav@nrczh.ru>
[4] Ya.D. Kozlovskaya, ORCID: 0000–0002–1780–5687 <yana_kozlovskaia@mail.ru>
[2] E.A. Evdokimova, ORCID: 0000-0003-4719-2786 <evdokimovaekan@mail.ru>

[1] Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences,
44 building 2, Vavilova st., Moscow, 119333, Russia
[2] Moscow Aviation Institute (National Research University),
4, Volokolamskoye Shosse, Moscow, 125993, Russia
[3] 27th Central Research Institute of the Ministry of Defence of the Russian Federation
5 1st Khoroshevsky Passage, 123007, Moscow, Russia
[4] Scientific and Industrial Company «High Technologies and Strategic Systems»,
27 building 9, Electrozavodskaya st., Moscow, 107023, Russia
[5] National Research Center «Zhukovsky Institute»,
7, Viktorenko st., Moscow, 125319, Russia

Abstract. The article describes a new method for contextual resolution of homonymy based on a centroid-context model (CCM). The proposed method of detecting cases of homonymy in the corpus of texts and its resolution using the CCM model is based on the theoretical concept of phraseological conceptual analysis of texts (FCAT) and unique machine grammar, which is based on a system of inflective classes of russian words. The rigid conformity between the form of presentation of words and their grammatical information laid down in the theoretical concept of inflective classes of words of the Russian language made it possible to create on this basis new classes - classes of words that have the same sets of grammatical features, conforming to their forms of representation in similar contextual environments. When developing this model, the authors proceeded from the following hypothesis: the same sequences of generalized characters of word classes (generalized syntagms) should correspond to the same syntactic structures of various fragments of texts. At the same time, it was assumed that such a hypothesis is true for any syntactic models and can be useful in solving both global and particular problems of text analysis. Using this method, a new solution to the problem of resolving homonymy based on the proposed CCM model was proposed.

Ключевые слова: homonymy resolution; homonyms; centroid-context model; generalized syntagms

For citation: Khoroshilov Alexander A., Nikitin Yu.V., Kan A.V., Kozlovskaya Ya.D., Evdokimova E.A. Context resolution of homonymy based on a centroid-context model. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 5, 2022. pp. 171-182 (in Russian). DOI: 10.15514/ISPRAS-2022-34(5)-11

1. Введение

В процессе автоматической обработки текстов нередко возникает ситуация, когда одни и те же словоформы в различных контекстных окружениях принимают разные смысловые значения. Такое явление в лингвистике называется омонимией. Разрешение (снятие) омонимии является одной из важнейших проблем автоматической обработки естественного языка. Решение этой проблемы необходимо для многих приложений компьютерной лингвистики, в частности, для интеллектуальных систем обработки текстов при формализации смысловой структуры текстов и построении их формальной модели метаданных, для поисковых систем для повышения точности обработки некоторых классов запросов и ряда других аналогичных задач.

В настоящее время наметилось три основных подхода к решению проблемы снятия омонимии: а) подход, основанный на правилах, б) подход, основанный на статистике, в) подход, основанный на машинном обучении.

Суть метода, основанного на правилах, сводится к тому, что в некоторых ситуациях анализ контекста помогает понять синтаксическую структуру части предложения, а с ее помощью и

формы слов. Например, в конструкции вида: *ни ..., ни ...* оба слова обычно принадлежат одной и той же части речи и находятся в одной и той же форме. Если одно из слов окажется неомонимичным, определить форму второго несложно. Этот метод требует ручного составления правил, для каждого из правил требуется написать самостоятельный программный модуль.

Подсчёт статистики различных вариантов разбора по корпусу является простейшим способом снятия морфологической омонимии. При этом по размеченному корпусу со снятой омонимией происходит вычисление апостериорных вероятностей каждого из разборов. В современных системах анализа текстов на естественном языке применяются несколько способов подсчёта таких вероятностей, однако большее влияние на точность снятия омонимии оказывает сам корпус: его представительность, объём, точность разметки. Для разрешения омонимии может быть реализован метод простого подсчёта вероятности для каждого из набора тегов и слов корпуса. Более точные результаты даёт учёт контекста слова.

Для решения проблемы снятия омонимии используются такие методы машинного обучения, как скрытые марковские модели [1], условные случайные поля [2], рекуррентные нейронные сети [3] и др. (подробнее см. [4]). Для обучения метода классификации также используются размеченные корпуса, например, НКРЯ (национальный корпус русского языка) с вручную снятой омонимией. Для контекстного снятия омонимии может использоваться, например, метод CRF (Conditional Random Field, условные случайные поля). Этот метод хорош в задачах определения части речи, определения именованных сущностей и др. [5]. CRF является дискриминативной вероятностной моделью. Одним из главных достоинств этой модели является то, что она не требует моделировать вероятностные зависимости между так называемыми наблюдаемыми переменными.

В настоящей работе описан метод, позволяющий предиктивно, на основе принципа лингвистической аналогии выявлять словоформы, принадлежащие к двум или более словоизменительным парадигмам и назначать им соответствующие грамматические характеристики. В словарном комплексе эти словоформы помечаются как омонимы и обе словоформы с грамматической информацией помещаются в соответствующие словари. В процессе автоматического анализа текстов на основе применения центроидно-контекстных моделей (ЦКМ) разрешается конкретная языковая ситуация по ее контекстному окружению. Рассмотрим более подробно этапы выявления омонимичных словоформ и разрешения случаев омонимии на основе метода контекстного разрешения омонимии.

2. Выявление омонимичных словоформ русского языка

В соответствии с теоретической концепцией фразеологического концептуального анализа текстов (ФКАТ) [6-8] некоторые виды омонимии словоформ русского языка частично разрешаются с помощью тематических концептуальных словарей, в которых значения слов устанавливаются в соответствии с их доминантными значениями, принятыми в конкретной тематической области и зафиксированные в составе терминологических словосочетаний. Поэтому основное внимание необходимо сосредоточить на методах автоматического выявления и разрешения грамматической омонимии для отдельных словоформ.

В основу метода выявления грамматической омонимии словоформ была положена гипотеза, что *в достаточно большом корпусе политематических текстов должна отобразиться значительная часть словоформ, имеющих один и тот же буквенный состав, но принадлежащих к различным грамматическим словоизменительным парадигмам*. Другими словами, если для всех словоформ достаточно большого корпуса текстов автоматически построить их словоизменительные парадигмы и назначить им грамматические признаки исходной словоформы, то все словоформы, содержащие одинаковый буквенный состав, но имеющие различные грамматические характеристики с большой вероятностью будут являться ***грамматическими омонимами***.

В качестве базовых признаков, на основе которых возможно идентифицировать каждую текстовую словоформу и автоматически построить ее грамматическую парадигму были выбраны номер флективного класса словоформы и ее грамматическое окончание. Это сочетание признаков однозначно характеризует смысловое содержание словоформы и ее конкретную форму в контекстном окружении. При этом мы исходили из того, что флективный класс словоформы однозначно соотносится с грамматическим классом словоформы, типом ее словоизменения, а в связке с грамматическим окончанием однозначно устанавливается расширенный состав грамматических характеристик конкретной словоформы. Рассмотрим эту идею в виде модели процесса выявления омонимов в корпусе тестов.

Пусть имеются две упорядоченные последовательности словоформ $\{w_i^1\}_{i=1}^{12}$ и $\{w_i^2\}_{i=1}^{12}$, где w_1^1 и w_1^2 – первая и вторая исходные словоформы соответственно, а $w_2^1, w_3^1, \dots, w_{12}^1$ и $w_2^2, w_3^2, \dots, w_{12}^2$ – их словоизменительные парадигмы. Имеются две упорядоченные последовательности $\{p_i^1\}_{i=1}^{12}$ и $\{p_i^2\}_{i=1}^{12}$, содержащие в себе признаки словоформ из последовательностей $\{w_i^1\}_{i=1}^{12}$ и $\{w_i^2\}_{i=1}^{12}$ соответственно. Признак p_i включает в себя информацию о флективном классе, окончании и грамматической информации (род, число, падеж, лицо) $p_i^j = \{FK_i^j, OK_i^j, GI_i^j\}$. Последовательности $\{p_i^1\}_{i=1}^{12}$ и $\{p_i^2\}_{i=1}^{12}$ обладают свойством:

$$\exists i: p_i^1 \neq p_i^2, i \in \{2,3,\dots,12\}.$$

Проблема разрешения омонимии возникает, когда встречаются $w_i^1 = w_j^2$, $i, j \in \{1,2,\dots,12\}$.

Подтверждением выше приведенной гипотезы служит пример, приведенный в табл.1, в которой приведены словоизменительные парадигмы для слов: «суд» и «судно».

Табл. 1. Автоматически сформированные грамматические словоизменительные парадигмы текстовых словоформ
Table 1. Automatically generated grammatical inflectional paradigms of text word forms

Исходная словоформа:	***Исходная словоформа***:
суд ОК=00, FK=001	судно ОК=01, FK=071
Словоизменительная парадигма №1:	***Словоизменительная парадигма №2:***
Им.п., ед.ч. = суд ОК=00, FK=000	Им.п., ед.ч. = судно ОК=00, FK=071
Род.п., ед.ч. = **суда ОК=01, FK=001**	Род.п., ед.ч. = судна ОК=01, FK=071
Дат.п., ед.ч. = суду ОК=01, FK=001	Дат.п., ед.ч. = судну ОК=01, FK=071
Вин.п., ед.ч. = суд ОК=00, FK=001	Вин.п., ед.ч. = судно ОК=00, FK=071
Тв.п., ед.ч. = судом ОК=02, FK=001	Тв.п., ед.ч. = судном ОК=02, FK=071
Пр.п., ед.ч. = суде ОК=01, FK=001	Пр.п., ед.ч. = судне ОК=01, FK=071
Им.п., мн.ч. = суды ОК=01, FK=001	Им.п., мн.ч. = суда ОК=01, FK=071
Род.п., мн.ч.= **судов ОК=02, FK=001**	Род.п., мн.ч.= судов ОК=02, FK=071
Дат.п., мн.ч.= **судам ОК=02, FK=001**	Дат.п., мн.ч.= **судам ОК=02, FK=071**
Вин.п., мн.ч.= суды ОК=01, FK=001	Вин.п., мн.ч.= **суда ОК=01, FK=071**
Тв.п., мн.ч.= **судами ОК=03, FK=001**	Тв.п., мн.ч.= **судами ОК=03, FK=071**
Пр.п., мн.ч.= **судах ОК=02, FK=001**	Пр.п., мн.ч.= **судах ОК=02, FK=071**

Из табл. 1 видно, что словоформы «***суда***», «***судах***» «***судам***» «***судами***» принадлежат к двум различным грамматическим парадигмам (№1 и №2), в которых эти словоформы имеют разные грамматические признаки. Например, для словоформы «суда» в парадигме №1 эта словоформа имеет характеристики: FK=001, OK=01, в парадигме №2 та же словоформа имеет характеристики: FK=071, OK=01.

В качестве тестового корпуса текстов был выбран массив сообщений СМИ объемом 740 тыс. документов и содержащий более 12 млн. слов. Далее обработка этого корпуса текстов производилась по следующей технологической схеме:

Этап 1. По этому корпусу текстов был построен частотный словарь словоформ, включающий 880 тыс. разных словоформ.

Этап 2. Все словоформы были обработаны процедурой морфологического анализа, и им был назначен набор грамматических характеристик, из которых для дальнейшей обработки были выбраны только две характеристики – номер FK и OK.

Этап 3. Для каждой словоформы корпуса была построена ее грамматическая словоизменительная парадигма и всем членам парадигмы были назначены грамматические признаки (FK и OK) той словоформы, на основе которой была построена ее грамматическая словоизменительная парадигма (см. табл. 1).

Этап 4. Все автоматически сгенерированные члены парадигмы всех словоформ корпуса были слиты в один массив и отсортированы. Далее были исключены все словоформы с одинаковыми грамматическими характеристиками (FK и OK) и сохранены только словоформы, имеющие различные грамматические характеристики.

Этап 5. Полученный массив рассортирован по грамматическим характеристикам, исключены дублирующие записи с одинаковыми характеристиками и назначены каждому типу омонимии его трансформационную модель (табл. 2).

3. Создание словарей ЦКМ для разрешения омонимии

Для реализации процесса разрешения омонимии в текстовых документах необходимо создать декларативные средства для решения этой задачи. Как было выше указано разрешение омонимии словоформы возможно только с учетом ее контекстного окружения. Разработанная авторами центроидно-контекстная модель (ЦКМ) содержит грамматическую синтагму словоформы-омонима и ее контекстное окружение в виде синтагм словоформ контекстного окружения, а результатом разрешения омонимии является однозначный выбор конкретной словоформы-омонима. При разработке этой модели авторы исходили из следующей гипотезы: *одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры различных фрагментов текстов*. Кратко рассмотрим идею модели ЦКМ применительно к задаче разрешения омонимии.

3.1 Модель ЦКМ для разрешения омонимии

На вход ЦКМ подается упорядоченное множество объектов $\mathrm{Sem} = \{s_i\}_{i=1}^{n}$. Каждому объекту множества (синтагме) Sem ставится в соответствие элемент упорядоченного множества $\mathrm{Pr} = \{p_i\}_{i=1}^{n}, s_i \leftrightarrow p_i$, где p_i – свойства объекта.

Задается радиус n, выбирается целевой токен (центроид) s_k, составляется позиционная модель, которая задается упорядоченной последовательностью (рис. 1):

$$\mathrm{PM} = \{s_k, s_{k+1}, s_{k-1}, s_{k+2}, s_{k-2}, \dots, s_{k+n}, s_{k-n}\}. \quad (1)$$

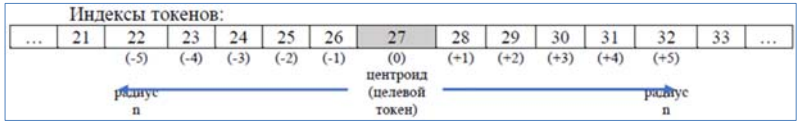

Рис. 1. Позиционная модель
Fig. 1. Positional model

Заполнение позиционной модели показано на рис. 2.

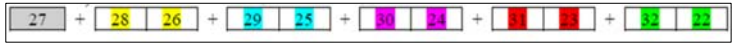

Рис. 2. Заполнение позиционной модели
Fig. 2. Filling in the positional model

К упорядоченной последовательности PM применяется вектор-функция fPR, которая каждому токену из PM ставит в соответствие признаки из множества Pr:

$$\mathrm{fPR}(s_i) = \begin{cases} p_i, & s_i \in \mathrm{Sem}, \\ p_0, & s_i \notin \mathrm{Sem}. \end{cases}$$

Здесь p_0 – фиктивный признак фиктивного элемента. В результате применения функции получается упорядоченная последовательность:

$$\mathrm{fPR}(\mathrm{PM}) = \mathrm{PMp} = \{p_k, p_{k+1}, p_{k-1}, p_{k+2}, p_{k-2}, \dots, p_{k+n}, p_{k-n}\}.$$

Пусть существует функция fSim (similar function), которая проверяет, на сколько два вектора схожи между собой:

$$\mathrm{fSim}(a, b) = i, \text{если } \forall k \in \{1, \dots, i\}\ a_k = b_k \text{ и } a_{i+1} \neq b_{i+1}.$$

Задается порог совпадения начальной части модели bv и во множестве MOD шаблонов синтагм ЦКМ из словаря шаблонов ищется наиболее схожая структура, наиболее схожая в смысле максимума процента покрытия CovPer:

$$\mathrm{PMpMod} = \underset{m \in \mathrm{MOD}}{\operatorname{argmax}}\, \mathrm{CovPer}(\mathrm{PMp}, m) = \begin{cases} 0\%, & \mathrm{fSim}(\mathrm{PMp}, m) < \mathrm{bv} \\ \frac{l - \mathrm{fHam}(\mathrm{PMp}, m)}{l} \cdot 100\%, & \mathrm{fSim}(\mathrm{PMp}, m) \geq \mathrm{bv} \end{cases},$$

где $\mathrm{fHam}(\cdot,\cdot)$ – функция, которая считает расстояние Хемминга (количество не совпавших элементов), $l = 2n + 1$ – количество элементов синтагмы.

Множество MOD содержит в себе шаблоны синтагм ЦКМ из словаря шаблонов. Каждый элемент $m \in \mathrm{MOD}$ имеет признак, который принимает значение 0 или 1.

Пусть функция ftakePR – функция взятия признака элемента $m \in \mathrm{MOD}$.

В результате ЦКМ содержится ответ на вопрос: *Является ли проверяемая структура той структурой, на соответствие которой мы ее проверяем или нет?* Ответ дается следующим образом:

$$\mathrm{answer} = \begin{cases} \text{да}, & \mathrm{ftakePR}(\mathrm{PMpMod}) = 1, \\ \text{нет}, & \mathrm{ftakePR}(\mathrm{PMpMod}) = 0. \end{cases}$$

3.2 Автоматизированное формирование словарей ЦКМ для разрешения омонимии

Необходимыми данными для формирования словарей ЦКМ по корпусу текстов служит массив омонимичных словоформ. В процессе его формирования выявляются предложения, в которых содержатся омонимичные словоформы и для каждой такой словоформы строится ЦКМ. В процессе обработки корпуса текстов устанавливаются наиболее частые формы омонимов (доминантные словоформы-омонимы). При этом пользователю предоставляется возможность корректировать грамматическую информацию словоформы-омонима, соответствующую конкретному контексту. Таким образом, создаются два словаря ЦКМ: словарь доминантных ЦКМ и словарь ЦКМ для редких (альтернативных) словоформ-омонимов. Но, поскольку требуется установить один из двух вариантов омонимов, необходимо определить: *является ли словоформа доминантной или не является,* то для этого достаточно только одного словаря доминантных словоформ. В табл.2 показан макет интерфейса для формирования и коррекции словаря ЦКМ для разрешения омонимии.

Табл. 2. Представление визуального интерфейса для ручного разрешения омонимии в ЦКМ
Tab. 2. Presentation of the visual interface for manual resolution of homonymy in the CCM

Перенумерованное исходное предложение №1:
00 Океанские || 01 суда || 02 стояли || 03 на || 04 рейде || 05 . ||

Исходное предложение, обработанное процедурой МА:
00 Океанские #OK=2#FK=106#GI=*0210*0240#GK=A#OS=Pg#TD=E
01 суда #OK=1#FK=070#GI=*1120#GK=N#OS=ГВ#CK=A#TD=S
02 стояли #OK=1#FK=125#GI=*0200#GK=L#OS=Яf#PG=t#CK=R#TD=S
03 на #OK=0#FK=164#GI=*0040*0060#GK=F#OS=ыA#CK=e#TD=S
04 рейде #OK=1#FK=001#GI=*1160#GK=N#OS=AK#TD=K
05 . #OK=0#FK=0#TW=0#GI=*0000#OS=..#GK=.#TD=Z

ГВЯfPgыAZZAKZZ..ZZZZZZ – доминантная ЦКМ для слова, имеющего ФК=001 и ОК=а)

Перенумерованное исходное предложение №2:
00 Заседание || 01 суда || 02 состоится || 03 в || 04 11 || 05 часов || 06 . ||

00 Заседание#OK=1#FK=073#GI=*3110*3140#GK=N#OS=ЁK#SU=t#CK=A#TD=S
01 суда #OK=1#FK=001#GI=*1120#GK=N#OS=AB#CK=A#TD=S
02 состоится#OK=4#FK=117#GI=*0103#GK=V#OS=Щs#PG=t#PV=t#CK=Q#TD=S
03 в #OK=0#FK=164#GI=*0040*0060#GK=F#OS=ыA#CK=e#TD=S
04 11 #OK=0#FK=313#GI=*0000#GK=0#OS=9A#SU=t#P0=t#TD=0
05 часов #OK=2#FK=001#GI=*1220#GK=N#OS=Az#CK=A#TD=S
06 . #OK=0#FK=0#TW=0#GI=*0000#OS=..#GK=.#TD=Z

ГВЯfPgыAZZAKZZ..ZZZZZZ – доминантная ЦКМ для слова, имеющего ФК=001 и ОК=а)

4. Разрешение омонимии на основе словарей ЦКМ

Под разрешением омонимии понимается установление конкретного набора грамматических и семантических характеристик омонимичной словоформы, соответствующее ее смысловому значению в конкретном контекстном окружении. При реализации данного метода в основном морфологическом словаре (SYS_MA_DCCM) были представлены все выявленные омонимичные словоформы с грамматическими признаками доминантной формы и с указанием признака омонима (РО=t). Альтернативные формы омонимов были представлены в словаре (SYS_MA_ACCM). Разрешение омонимии выполнялось по словарю ЦКМ (SYS_Dom_Hom). В случае полного или частичного совпадения текстовой ситуации и словарного элемента словаря SYS_Dom_Hom выбиралась доминантная форма омонима из словаря SYS_MA_DCCM, если такого совпадения не было, выбиралась альтернативна форма из словаря SYS_MA_ACCM.

Для определения критерия частичного совпадения эталонной модели ЦКМ и текстовой модели ЦКМ был проведен лингвистический анализ различных омонимичных ситуаций в корпусе текстов объемом 12 млн. словоформ. На основе этого анализа было выявлено и проанализировано 43 типа омонимичных трансформаций. Для каждого типа трансформаций был разработан механизм разрешения омонимии (МРО) на основе применения модели ЦКМ.

Табл. 3. Фрагмент таблицы типов омонимичных трансформаций омонимичных словоформ
Table 3. Fragment of the table of types of homonymous transformations of homonymous word forms

№ класса и типа омонимичной трансформации	**Грамматические трансформационные классы омонимов**	**Слово-представитель**	**Обобщенная синтагма № 1**	**Обобщенная синтагма № 2**	**Минимальное совпадение элементов ЦКМ (тест. и слов.)**
1. Местоименные существительные – местоименные прилагательные					
1.1	y-s	ее	0A	pA	3
1.2	y-s	его	яA	pA	3
1.3	y-s	их	4A	pA	3
2. Субстантивированные прилагательные-существительные					
2.1	A-N	легкое	РБ	РБ	6
2.2	A-N	сборная	HI	HI	6
3. Омонимичные существительные - прилагательные					
3.1	N-A	долгом	FГ	FГ	7
3.2	N-A	теплом	ГГ	ГГ	7
3.3	N-A	правом	ГГ	ГГ	7
…					
4. Существительные фамильно-именной группы (мужской род, им. падеж) - существительные					
4.1	N-N	быков	Yz	hA	6
4.2	N-N	волков	Yz	hA	6
…					
5. Существительные фамильно-именной группы (женский род, им. падеж) – существительные фамильно-именной группы (мужской род, косв. падеж)					
5.1	N-N	иванова	jB	hA	6
5.2	N-N	петрова	jB	hA	6

В табл. 4 приведены фрагменты типов омонимии и их доля в текстах, также вероятности разрешения этих типов на основе применения ЦКМ.

Табл. 4. Типы омонимии и их доля в текстах и вероятности разрешения этих типов на основе применения ЦКМ
Table 4. Types of homonymy and their share in texts and the probabilities of resolving these types based on the use of CCM

	Класс словоформ-омонимов	**Вероятность метода**	**Доля класса в массиве (тыс.)**	**Относительная доля класса в массиве (%)**
1	Местоименные существительные – местоименные прилагательные	**98%**	**727**	**49,0**
2	Субстантивированные прилагательные-существительные	**95%**	**17**	**1,14**

3	Омонимичные существительные - прилагательные	**95%**	**11**	**0,75**
4	Существительные фамильно-именной группы (мужской род, им. падеж) – существительные	**95%**	**16**	**1,07**
5	Существительные одушевленные (мужской род, им. падеж) - существительные неодушевленные	**95%**	**9**	**0,66**
6	Существительные неодушевленные (женский род, им. падеж)	**95%**	**1**	**0,09**
7	Существительные одушевленные муж. рода – существительные неодушевленные жен. рода	75%	16	1,09
8	Существительные неодушевленные жен. рода – существительные неодушевленные муж. рода	75%	57	4,05
9	Существительные одушевленные (женский род, им. падеж) - существительные неодушевленные	75%	6	0,03
10	Существительные одушевленные муж. рода – существительные неодушевленные муж. рода	75%	43	1,45
11	Существительные неодушевленные	75%	155	7,61
12	Глагол личной формы ед. числа 1-го лица	75%	19	1,32
13	Глагол личной формы мн. числа 2-го лица	75%	0,132	0,132
14	Глаголы повелительного наклонения – существительные	75%	0,67	0,67
15	Существительные - краткие причастия	75%	13	1,23

В этом фрагменте таблице (табл. 4) приводится доля классов омонимов от их общего числа и их абсолютное число в массиве, объемом 740 тыс. документов (12 млн. слов). В этом массиве обнаружено 1.5 млн. омонимичных случаев.

4.1 Алгоритм разрешения омонимии словоформ в текстах

Алгоритм разрешения омонимии выполняется на этапе семантико-синтаксического анализа на основе результатов обработки текста процедурой МА и состоит из следующих этапов.

Этап 1. В результатах анализа МА слов текста определяются слова по словарю SYS_MA_DCCM, в ГИ которых имеется признак омонимии (PO=t).

Этап 2. На основе грамматической информации омонимичной словоформы и грамматической информации слов, окружающих слева и справа от нее, строится ЦКМ.

Этап3. Если сформированная текстовая ЦКМ совпадает с одним из элементов словаря SYS_Dom_Hom, а число совпавших синтагм текстовой ЦКМ и словарной ЦКМ больше или равно минимальному числу, указанному в словарной статье этой ЦКМ, то считается, что разрешение данного случая омонимии выполнено и эта словоформа является доминантной.

Этап 4. В случае, если сформированная ЦКМ не совпадает с одним из элементов словаря SYS_Dom_Hom, или число совпавших синтагм текстовой ЦКМ и словарной ЦКМ меньше или равно минимальному числу, указанному в словарной статье этой ЦКМ, то считается, что

разрешение данного случая омонимии выполнено и эта словоформа является альтернативной, а грамматическая информация назначается ей по словарю SYS_MA_ACCM.

5. Заключение

Предлагаемый метод выявления случаев омонимии в корпусе текстов и ее разрешения с помощью модели ЦКМ базируется на теоретической концепции фразеологического концептуального анализа текстов (ФКАТ) и уникальной машинной грамматике, в основу которой положена система флективных классов русских слов. Заложенное в теоретической концепции флективных классов слов русского языка жесткое соответствие между формой представления слов и их грамматической информацией позволило создать на этой основе новые классы – классы слов, имеющие одинаковые наборы грамматических признаков, соответствующие их формам представления в сходных контекстных окружениях. При разработке этой модели авторы исходили из следующей гипотезы: *одинаковым последовательностям обобщенных символов классов слов (обобщенным синтагмам) должны соответствовать одинаковые синтаксические структуры различных фрагментов текстов.*

В процессе выявления омонимичных словоформ применялись методы предиктивного назначения грамматических характеристик словоформам и формирования для всех словоформ их словоизменительных парадигм. Установление расхождений грамматических характеристик одинаковых словоформ позволило автоматически выявить омонимичные словоформы. Установление местоположений омонимичных словоформ в различных контекстных окружениях позволяет автоматически сформировать модели ЦКМ для разрешения омонимии. С помощью визуального интерфейса пользователь способен с минимальными трудозатратами откорректировать словари для разрешения омонимии. Основным достоинством метода является полная автоматизация процесса выявления омонимичных словоформ и высокая степень автоматизации создания декларативных средств для разрешения омонимии.

Разработанный авторами данной статьи новый метод автоматического выявления омонимичных словоформ в корпусе текстов и автоматизированное создание декларативных средств их разрешения позволяет быстро решать проблему разрешения омонимии для текстов любого лексического состава в различных тематических областях.

Список литературы / References

[1] Baum L.E., Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. The annals of mathematical statistics, vol. 37, issue 6, 1966, pp. 1554-1563.

[2] Lafferty J., McCallum A., Pereira F.C.N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proc. of the Eighteenth International Conference on Machine Learning, 2001, pp. 282-289.

[3] Elman J.L. Finding structure in time. Cognitive science, vol. 14, issue 2, 1990, pp. 179-211.

[4] Manning C.D., Raghavan P., Schütze H. Introduction to Information Retrieval, Cambridge University Press, 2008, 506 p.

[5] Sha F., Pereira F. Shallow parsing with conditional random fields. In Proc. of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, vol. 1, 2003, pp. 134-141.

[6] Хорошилов Александр А., Мусабаев Р.Р. и др. Автоматическое выявление и классификация информационных событий в текстах СМИ. Научно-техническая информация. Серия 2: Информационные процессы и системы, вып. 7, 2020 г., стр. 27-38 / Khoroshilov Alexandr A., Musabaev R.R. et al. Automatic Detection and Classification of Information Events in Media Texts. Automatic Documentation and Mathematical Linguistics, vol. 54, issue 4, 2020, pp. 202–214.

[7] Аблов И.В., Козичев В.Н и др. Средства машинной грамматики русского языка (по Г.Г. Белоногову). Научно-техническая информация. Серия 2: Информационные процессы и системы, вып. 6, 2018 г., стр. 32-46 / Ablov I.V., Kozichev V.N. The Tools of a Machine Grammar of the Russian

Language (based on G.G. Belonogov). Events in Media Texts. Automatic Documentation and Mathematical Linguistics, vol. 52, issue 3, 2020, pp. 142-156.

[8] Калинин Ю.П., Хорошилов Александр А., Хорошилов Алексей А. Современные технологии автоматизированной обработки текстовой информации. Системы высокой доступности, том 11, вып. 2, 2015 г. / Kalinin Yu.P., Khoroshilov Alexander A., Khoroshilov Alexey A. Modern technologies for automated text processing. High Availability Systems, vol. 11, issue 2, 2015 (in Russian).

Информация об авторах / Information about authors

Александр Алексеевич ХОРОШИЛОВ – доктор технических наук, профессор МАИ, ведущий научный сотрудник ФИЦ ИУ РАН, старший научный сотрудник 27 ЦНИИ Минобороны России. Область научных интересов: теоретические основы информатики; вычислительные машины электронные цифровые; машинный перевод; прикладное языкознание; автоматическая обработка речевой информации

Alexander Alexeevich KHOROSHILOV – Doctor of Science, Professor of the Moscow Aviation Institute (National Research University), Lead Researcher of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Senior Researcher of the 27 Central Research Institute of the Ministry of Defense of the Russian Federation. Research interests: theoretical foundations of informatics; electronic digital computers; Machine translate; applied linguistics; automatic processing of speech information

Юрий Викторович НИКИТИН – научный сотрудник ФИЦ ИУ РАН, руководитель группы разработки АО «НПК «ВТ и СС». Научные интересы: методы и технологии семантического анализа текстов, фразеологического машинного перевода и создания декларативных средств для лингвистического программного обеспечения.

Yuri Viktorovich NIKITIN – Researcher of the Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Development Team Leader of the Scientific and Industrial Company "High Technologies and Strategic Systems". Research interests: methods and technologies of semantic text analysis, phraseological machine translation and creation of declarative tools for linguistic software.

Анна Владимировна КАН – кандидат технических наук, доцент МАИ, начальник аналитического отдела ФГБУ «НИЦ «Институт имени Н.Е. Жуковского. Основные научные интересы в области системного анализа, имитационного моделирования и искусственного интеллекта.

Anna Vladimirovna KAN – Candidate of Technical Sciences, Associate Professor of the Moscow Aviation Institute, Head of the Analytical Department of the National Research Center "Zhukovsky Institute". Her main research interests are in the field of systems analysis, simulation and artificial intelligence.

Яна Дмитриевна КОЗЛОВСКАЯ – участник группы разработки АО «НПК «ВТ и СС». Научные интересы: компьютерная лингвистика, анализ текстов.

Yana Dmitrievna KOZLOVSKAYA – development team member of the Scientific and Industrial Company "High Technologies and Strategic Systems". Research interests: computational linguistics, text analysis.

Екатерина Андреевна ЕВДОКИМОВА – студентка. Научные интересы: анализ и обработка текстовых данных.

Ekaterina Andreevna EVDOKIMOVA – student of the Moscow Aviation Institute. Research interests: analysis and processing of text data.