УДК 81'322.2 + 81'322.3

АНАЛИТИЧЕСКИЙ ОБЗОР МЕТОДОВ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ

Р.Р. Гатауллин

Институт вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета
Институт прикладной семиотики Академии наук Республики Татарстан
ramil.gata@gmail.com

Аннотация

Проанализированы основные методы разрешения морфологической многозначности применительно к татарскому языку. Описано текущее состояние работ и приведены основные результаты по данному направлению, сделаны выводы о применимости методов разрешения с оценкой их точности.

***Ключевые слова**: разрешение морфологической многозначности, контекстные методы, статистико-вероятностные методы, татарский язык.*

ВВЕДЕНИЕ

Многозначность языковых форм – одна из природных особенностей естественного языка, способствующая качественному развитию словарного запаса, тем самым «экономящая» словесный материал [2]. Разрешение многозначности (т. н. дизамбигуация) является одной из важнейших задач автоматической обработки естественного языка. Результаты разрешения используются для повышения точности методов классификации и кластеризации текстов, улучшения качества машинного перевода, информационного поиска и других приложений [2].

Исследователи выделяют несколько типов многозначности естественного языка: *морфологическую, синтаксическую и лексико-семантическую многозначности*. Иногда к ним добавляют прагматическую многозначность. *Для* работы с каждым из этих типов существуют собственные методы [2].

Задача разрешения *морфологической многозначности* заключается в определении для слова части речи и грамматических признаков, соответствующих контексту. Морфологическая многозначность, в основном, представлена грамматической омонимией, т. е. совпадением слов в отдельных грамматических формах. Например, слово «стекло» в зависимости от контекста может быть либо существительным, обозначающим материал («смотреть через стекло»), либо глаголом в прошедшем времени 3-го лица единственного числа («масло стекло»).

Задача разрешения синтаксической многозначности (*многозначность синтаксических структур)* заключается в правильном определении функций синтаксических единиц предложения. Примером такой неоднозначности является предложение «мужу изменять нельзя» (словоформа *мужу* – субъект или объект предложения?) [2].

Значения слов могут относиться к одной части речи, но различаться по смыслу, например, «*platform*» – железнодорожная или компьютерная платформа. В этом случае речь идет о *полисемии*, когда у одного слова имеются два или более значения, взаимосвязанных по смыслу и происхождению. Полисемия относится к *лексической многозначности*. Сюда же следует относить и *лексическую омонимию (*слова совпадают в звучании и написании, но имеют разные значения). Такими омонимами являются слова *лук* («оружие») и *лук (*«растение»). Задача разрешения такой неоднозначности состоит в установлении значений слов или составных терминов в соответствии с контекстом, в котором они использовались [2].

Еще один тип неоднозначности возникает в результате употребления местоимений или специальных существительных типа *one, another* (еще один). Так, в предложении «*Она уронила карандаш на стол и сломала его*» невозможно однозначно определить, что именно было сломано – *карандаш* или *стол* (нельзя однозначно разрешить референцию местоимения *его*) [2]. В этом случае говорят о *прагматической неоднозначности*.

Сложность и особенности разрешения многозначности для каждого конкретного языка проявляются по-разному. Например, для английского языка с бедной морфологией и жестким порядком слов в предложении разрешение морфологической многозначности, как правило, сводится к задаче POS-теггинга (от

англ., part of speech – определение части речи слова) и решается применением достаточно простых методов. Для русского языка морфологическая многозначность не столь характерна, как для английского и татарского, но, тем не менее, присуща. Дополнительную сложность добавляет свободный порядок слов в русском языке. В татарском языке, как и в других агглютинативных языках, таких, как турецкий и венгерский, морфемы несут как семантическую, так и синтаксическую информацию. Имея теоретически неограниченное количество присоединяемых к основе морфем, морфологическая многозначность приобретает разнообразные формы, что значительно усложняет задачу разрешения.

КОНТЕКСТНЫЕ МЕТОДЫ

Эти задачи были поставлены еще в 1950–1960-х годах, и теоретические исследования имеют многолетнюю историю. Еще в конце 1950-х годов в работах K.E. Harper [5], A. Caplan [6] основным способом снятия омонимии признавались изучение и описание тех контекстных условий, в которых реализуется то или иное значение слова. При этом под контекстом понималось окружение слова в тексте, т. е. слова, с которыми данное слово употребляется.

Актуальным для исследуемой задачи также являлся вопрос о минимальном разрешающем контексте. В этой связи заслуживают внимания результаты, полученные A. Caplan [6] по исследованию минимального разрешающего контекста. В работе анализировались 140 многозначных употребительных английских слов (в основном, лексических омонимов), находившихся в различных контекстных условиях. Автором выделены следующие виды контекстов:

- сочетание с предшествующим словом – P1;
- сочетание с последующим словом – F1;
- сочетание с предшествующим и последующим *словами* – B1 (both);
- сочетание с двумя предшествующими словами – P2;
- сочетание с двумя последующими словами – F2;
- сочетание с двумя предшествующими и двумя последующими словами – B2;
- все предложение в целом – S (sentence).

Основной вывод заключался в том, что цепочка B1 по эффекту редуцирования многозначности (отношение количества значений слова в конкретном контексте к их количеству в нулевом контексте) более продуктивна, чем контекст, состоящий из двух предшествующих или двух последующих слов (P2 и F2), и приближается к эффекту, даваемому целым предложением (S) [6].

В другом выводе подчеркивается важное значение материального типа контекста, т. е. входят ли в непосредственное окружение знаменательные слова, или слова, называемые автором «particles» (предлоги, союзы, глаголы типа will или do, артикли, местоимения и наречия типа there и др.). Первый тип контекста дает значительно большую редукцию многозначности, чем контекст, содержащий слова без конкретного лексического наполнения [6, 7].

Общие выводы A. Caplan сводятся к тому, что наиболее практичным является контекст, состоящий из одного слова слева и одного слова справа от анализируемой многозначной лексемы. Если же одно из слов окружения – «particle», то следует «усилить» контекст до двух слов с обеих сторон [6, 7].

Исследования такого подхода для русского языка [7] показали, что его применимость в реальных контекстах вряд ли возможна. Реальная ситуация с разрешением омонимии в русском языке значительно сложнее и не может быть разрешена на основе упрощенных схем. В отличие от английского, в русском языке порядок слов свободный, предполагается, что количество возможных контекстов из-за этого увеличивается. Для решения этой проблемы для русского языка была предложена усложненная структура правил, а также предполагается в качестве контекста использовать все предложение [7]. С учетом этого замечания было разработано программное средство разрешения функциональной омонимии, которая для некоторых типов дает точность распознавания, равную 100% при тестировании не менее 100 примеров, в наихудших случаях – точность не менее 95% [7].

При исследовании омонимии в татарском языке в центре внимания были лексические омонимы. Тем не менее, есть несколько работ, посвященных и грамматической омонимии [9–11]. Но до настоящего времени специальные исследования и классификации грамматической омонимии практически не проводились [1].

В работе [12] приведены основные формально-грамматические модели словосочетаний в татарском языке (15 основных, 80 частных типов) с указанием главного и зависимого слов. Актуальной задачей является проверка возможности использования этих моделей в качестве основы для определения разрешающих контекстов. Определенная строгость агглютинативной синтаксической структуры позволяет рассчитывать на обнаружение четких контекстных ограничений [1].

Для разрешения морфологической многозначности на основе контекстных правил в татарском языке в НИИ «Прикладная семиотика» Академии наук Республики Татарстан создан программный инструментарий для разработки и тестирования контекстных правил. Первые результаты экспериментов по построению контекстных правил показали работоспособность метода, однако для окончательных выводов требуются дополнительные исследования [23].

Подход, основанный на правилах, является чрезвычайно трудоемким, требует проведения тщательной лингвистической экспертизы каждого типа омонимии. Полная классификация типов омоформ является прагматически неоправданной задачей, так как татарский язык относится к агглютинативным языкам, для которых количество присоединяемых к основе морфем теоретически не ограничено. Например, в указанном корпусе татарских текстов объемом более 21 млн. словоупотреблений число типов омоформ превышает 7000 [1]. Здесь под типом морфологической многозначности (т. н. типом омоформ) подразумевается комбинация возможных аффиксальных цепочек, соответствующих слову. Например, тип, состоящий из «N» (сущ.) и «V+Neg» (глагол в повелительном наклонении с отрицанием), приписан словам «*алма», «басма», «тартма*» и др.

Чрезмерная трудоемкость этого подхода требует поиска более оптимальных путей решения. Одним из направлений является попытка комбинирования данного подхода со статистико-вероятностными методами и методами машинного обучения.

СТАТИСТИКО-ВЕРОЯТНОСТНЫЕ МЕТОДЫ

В тех же 1950–1960-х годах вслед за контекстными методами для задач дизамбигуации стали использовать статистико-вероятностные методы. Отсутствие больших информационных ресурсов и языковых баз данных значительно осложняло эксперименты и их дальнейшее применение.

После появления репрезентативных электронных корпусов эксперименты с вероятностно-статистическими методами показали достаточно хорошие результаты. Например, для английского языка, как было отмечено, задача снятия морфологической омонимии сводится, как правило, к проблеме разрешения многозначности на уровне частей речи (так называемого POS-теггинга). При этом используются алгоритмы, основанные на статистических моделях, таких, как скрытая марковская модель HMM [13] и марковская модель максимальной энтропии MEMM [14], учитывающие вероятность появления тега той или иной части речи в данном контексте. Для английского языка эти алгоритмы дают приемлемый результат с точностью не менее 96% [15].

Среди известных методов, применяемых при снятии морфологической многозначности в текстах английского языка, следует также отметить метод опорных векторов (Support Vector Machines, SVM) и деревья решений (Decision Trees). Например, точность SVM составила 97.2% при тестировании на текстах новостных статей из корпуса The Wall Street Journal, что является достаточно хорошим результатом [16].

Статистические методы для разрешения морфологической омонимии применительно к русскому языку стали использоваться сравнительно недавно. Зеленков и др. [17] предложили алгоритм, предназначенный для разрешения морфологической омонимии слов, которые совпадают лишь в нескольких грамматических формах. Метод основан на использовании автоматически полученного словаря контекстов, выведенного из уже размеченных текстов [16].

При адаптации к русскому языку некоторых методов необходимо учесть некоторые особенности языка. Во-первых, морфологическая омонимия в русском языке в отличие от английского языка не сводится к частеречной омонимии, а охватывает большое количество различных грамматических признаков. Во-вторых, хорошая работа статистических моделей на материале английских текстов объясняется тем, что в английском языке существует фиксированный порядок слов. Это обстоятельство упрощает создание модели, так как позволяет, к примеру, опираться только на локальный контекст слова (соседние слова) без учета дальних зависимостей. Именно поэтому для морфологической дизамбигуации в

английском языке часто успешно используются алгоритмы, основанные на марковских моделях и учитывающие зависимость каждого набора тегов только от одного элемента контекста – непосредственно предшествующего ему набора тегов [15].

В русском языке, напротив, порядок слов свободный, так что предполагается, что количество возможных контекстов из-за этого увеличивается, и эффективность обучения простой модели, основанной на локальных зависимостях, снижается. Поэтому, наряду с марковскими моделями, для снятия морфологической омонимии в русском языке используются более сложные статистические модели или гибридные системы [15], в которых статистика дополняется набором правил (см., например, Transformation-Based Learning [18], а также [17]).

Алгоритм, основанный на использовании скрытой марковской модели (HMM), требует предварительного обучения системы на уже размеченной выборке текстов большого объема. Предварительные результаты экспериментов показали точность работы алгоритма для русского языка не менее 95% [15, 19].

В [15] отмечается, что при сравнительном анализе алгоритмов, основанных на скрытой марковской модели и марковской модели максимальной энтропии, оба алгоритма неплохо (точность не менее 95%) справляются с задачей частеречной дизамбигуации, но значительно хуже снимают омонимию по расширенному набору грамматических тегов. Как правило, алгоритмы ошибаются при разметке имен собственных, местоимений, римских цифр, инициалов и сокращений. Помимо этого, модели не работают со случаями субстантивации прилагательных и выбором некоторых падежных форм: в первую очередь, с разграничением между номинативом и аккузативом, что связано с особенностями порядка слов в русском языке. В заключении этой работы делается вывод, что алгоритм MEMM в целом работает лучше в применении к задаче POS-теггинга, чем HMM [15].

Для агглютинативных языков, таких, как венгерский [20] и финский [21], метод, основанный на HMM, также дает не менее 97% точности. В работе [24] утверждается о достижении 98% точности разрешения морфологической многозначности для турецкого языка при использовании HMM совместно с персептронным алгоритмом (англ, Perceptron Algorithm [25]).

В [16] проанализирована применимость метода опорных векторов для задач снятия многозначности. Основная идея SVM-метода заключается в поиске разделяющей гиперплоскости с максимальным зазором между векторами двух различных классов. Для нахождения разделяющей гиперплоскости потребуется уже размеченный набор текстов. Механизм метода опорных векторов довольно прост, и как показывает практика, эффективен. Гибкость алгоритма позволяет успешно сочетать его с уже существующими методами определения частей речи и снятия омонимии [16].

В работе [22] описан интересный подход с генерацией правил разрешения из размеченного корпуса со снятой многозначностью. Метод применялся для турецкого языка, эксперименты показали точность не менее 96%. Отличительной особенностью подхода является выявление контекстных ограничений не для всей аффиксальной цепочки в целом, а для каждой морфемы отдельно. Турецкому языку, как и татарскому, свойственна возможность теоретически неограниченно присоединять морфемы к основе, что приводит к многообразию форм слов, а это с свою очередь, – к разреженности данных при обучении. Данный подход в определенной мере способствует решению проблемы с разреженностью данных.

Проблема разреженности данных стоит и для татарского языка. На данный момент языковой корпус татарского языка находится на стадии разработки. Морфологическая разметка осуществляется автоматически адаптированным морфологическим анализатором на базе двухуровневой модели морфологии татарского языка [26]. Снятие морфологической многозначности выполняется экспертами вручную, поэтому объем корпуса со снятой многозначности незначителен. Поэтому экспериментальная проверка применимости всех описанных статистико-вероятностных методов для татарского языка в настоящее время не представляется возможной ввиду отсутствия размеченного корпуса. Тем не менее, типологическая и генетическая близость турецкого и татарского языка дает основание полагать, что статистические методы способны показать хорошие результаты для татарского языка.

Таким образом, текущими задачами являются подготовка татарского размеченного корпуса и применение описанных методов для решения морфологической многозначности. В первую очередь предполагается применять те методы,

которые показали хорошие результаты для близкородственных языков.

ЗАКЛЮЧЕНИЕ

В настоящей работе представлен аналитический обзор основных методов разрешения морфологической многозначности. Точность работы описанных методов составляет не ниже 95%. В основном методы являются языконезависимыми, но точность разрешения варьируется в зависимости от конкретного языка (см. табл. 1).

Таблица 1

Класс метода	**Методы**	**Язык**	**Точность**
Контекстные методы	-	английский	99,5% [19]
	-	русский	95% [7]
Статистико-вероятностные методы	HMM	английский	96% [15]
		русский	95% [15, 19]
		финский	97% [21]
		венгерский	97% [20]
		турецкий	98% [25]
	MEMM	английский	96% [14, 15]
		русский	95% [15]
	SVM	английский	97,2% [16]
		русский	95,7% [16]
	GPA	турецкий	96% [22]

Для английского языка, имеющего *бедную морфологию*, проблема разрешения морфологической многозначности, как правило, сводится к разрешению многозначности на уровне частей речи (POS-теггинг), что, в свою очередь, заметно

облегчает задачу. В агглютинативных языках, таких, как турецкий, венгерский и татарский, к основе слова присоединяются морфемы, которые, кроме семантики, определяют и синтаксические связи. Морфологическая многозначность в этих языках проявляется разнообразными формами. В некоторых случаях для разрешения морфологической многозначности могут потребоваться как синтаксический, так и семантический анализ.

С другой стороны, жесткий порядок слов в предложениях на английском языке позволяет использовать минимальный размер контекста, тогда как для русского языка иногда требуется в качестве контекста использовать все предложение [7], тем самым усложняя задачу поиска разрешающего контекста. Размер минимального контекста для татарского языка еще предстоит исследовать. Тем не менее, есть основания полагать, что определенная строгость синтаксическая структуры позволит рассчитывать на обнаружение четких контекстных ограничений в ближайшем контексте [1].

Применение статистических алгоритмов для снятия многозначности позволило сместить акценты разработки на подготовку размеченных корпусов для обучения статистико-вероятностных моделей.

Несмотря на указанные сложности, можно констатировать, что для английского, русского и турецкого языков проблема разрешения морфологической многозначности, в основном, решена. Используя различные надстройки над алгоритмами (либо увеличивая обучающую выборку для статистических методов), точность методов разрешения можно довести до уровня не ниже 97%.

Типологическая и генетическая близость турецкого и татарского языка дает основание полагать, что данные методы способны дать приемлемые результаты и для татарского языка.

СПИСОК ЛИТЕРАТУРЫ

1. *Хакимов Б.Э., Гильмуллин Р.А., Гатауллин Р.Р.* Разрешение грамматической многозначности в корпусе татарского языка // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. 2014. Т. 156, кн. 5. С. 236–244.

2. *Турдаков Д.Ю.* Методы и программные средства разрешения лексической многозначности терминов на основе сетей документов: автореф. дис. … канд. тех. наук: 05.13.11. Москва, 2010. 20 с.

3. *Бобичев В.Л.* Автоматическое снятие морфологической многозначности при разметке корпуса // Тр. междунар. конф. «Корпусная лингвистика–2008». СПб.: СПбГУ, 2008. С. 45–49.

4. *Tufiş D., Popescu O.A.* Knowledge-based approach to morpho-lexical processing of natural language // in Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991. P. 405–408.

5. *Harper K.E.* Contextual analysis // Mech. Translation. 1956. V. 4, No 3. P. 70–75.

6. *Caplan A.* An experimental study of ambiguity and context // Mech. Translation. 1955. V. 2, No 2. P. 39–46.

7. *Зинькина Ю.В., Пяткин Н.В., Невзорова О.А.* Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005. М.: Наука, 2005. С. 198–202.

8. *Кобзарева Т.Ю., Афанасьев Р.Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Труды междунар. конференции Диалог'2002. М.: 2002. С. 258–268.

9. *Курбатов Х.Р.* Грамматические омонимы в татарском языке // Татар теле һәм әдәбияты. Казан: Татар. кит. нәшр., 1959. Б. 307–311.

10. *Салахова Р.Р.* Омонимичные суффиксы татарского языка. Казань: Gumanitarya, 2007. 204 с.

11. *Салимгараева Б.С.* Омонимы в современном татарском языке. Автореф. канд. дис. Уфа, 1971. 82 с.

12. Татарская грамматика. Казань: Татар. книж. изд-во, 1993. Т. II. Морфология. 397 с.

13. *Weischedel Ralph M.* Coping with ambiguity and unknown words through probabilistic models // Computational Linguistics. Cambridge, MA, USA: MIT Press, 1993. V. 19, Issue 2. P. 361–382.

14. *Ratnaparkhi A.* Maximum entropy model for part-of-speech tagging // Proceedings of the Empirical Methods in Natural Language Processing. Philadelphia, PA, USA, 1996. P. 133–142.

15. *Лакомкин Е.Д., Пузыревский И.В., Рыжова Д.А.* Анализ статистических алгоритмов снятия морфологической омонимии в русском языке. URL: http://aist-conf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf.

16. *Ткаченко М.В.* Модель и алгоритм улучшения распознавания частей речи в текстах, содержащих ошибки. СпбГУ, 2010. 20 с. URL: http://se.math.spbu.ru/SE/YearlyProjects/2010/list.

17. *Зеленков Ю.Г., Сегалович И.В., Титов В.А.* Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог'2005. М.: Наука, 2005. С. 616.

18. *Brill E.* A simple rule-based part of speech tagger // Proceedings of the third conference on Applied natural language processing (ANLC’92). Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. P. 152–155.

19. *Сокирко А.В., Толдова С.Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка (скрытая модель Маркова и синтаксический анализатор именных групп). URL: http://www.aot.ru/docs/RusCorporaHMM.htm.

20. *Orosz G., Novak A.* PurePos 2.0: a hybrid tool for morphological disambiguation // In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2013), Hissar, Bulgaria, 2013. INCOMA Ltd. Shoumen, Bulgaria. P. 539–545.

21. *Kristen Linden, Tommi Pirinen.* Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC // In Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. Editors: Kristiina Jokinen and Eckhard Bick. NEALT Proceedings Series, 2009. V. 4. P. 89–95.

22. *Deniz Yuret, Ferhan Ture.* Learning morphological disambiguation rules for turkish // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. New York, 2006. P. 328–334.

23. *Гатауллин Р. Р., Гильмуллин Р.А.* Контекстные правила для разрешения морфологической многозначности в корпусе татарского языка // Открытые семантические технологии проектирования интеллектуальных систем OSTIS-2016 (OpenSemantic Technologies for Intelligent Systems). Материалы V международной научно-технической конференции (Минск, 18–20 февраля 2016 года). Минск: БГУИР, 2016. С. 389–392.

24. *Hasim Sak, Tunga Gongur, Murat Saraclar*. Morphological disambiguation of turkish text with perceptron algorithm // Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007. P. 107–118.

25. *Collins M.* Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms // Proceedings of EMNLP, 2002. P. 1–8.

26. *Сулейманов Д.Ш., Гильмуллин Р.А.* Двухуровневое описание морфологии татарского языка // Тезисы докладов Международной научной конференции «Языковая семантика и образ мира». Казань: Изд-во Казан. гос. ун-та, 1997. Книга 2. С. 65–67.

REVIEW OF MORPHOLOGICAL DISAMBIGUATION METHODS

R.R. Gataullin

Institute of Computational Mathematics and Information Technologies.
Kazan Federal University
Institute of Applied Semiotics of the Tatarstan Academy of Sciences
ramil.gata@gmail.com

Abstract

This paper describes the morphological disambiguation methods and their application for the Tatar language. The state-of-the art technology is discussed. We analyze the contextual and statistical methods and their evaluations for different languages.

***Keywords**: morphological disambiguation, contextual method, statistical method, Tatar language*

REFERENCES

1. *Khakimov B.E., Gilmullin R.A., Gataullin R.R.* Razresheniye grammaticheskoy mnogoznachnosti v korpuse tatarskogo yazyka // Uchen. zap. Kazan. un-ta. Ser. Gumanit. nauki. 2014. T. 156. kn. 5. S. 236–244.

2. *Turdakov D.Yu.* Metody i programmnyye sredstva razresheniya leksicheskoy mnogoznachnosti terminov na osnove setey dokumentov: avtoref. dis. … kand. tekh. nauk. 05.13.11. Moskva, 2010. 20 s.

3. *Bobichev V.L.* Avtomaticheskoye snyatiye morfologicheskoy mnogoznachnosti pri razmetke korpusa // Tr. mezhdunar. konf. «Korpusnaya lingvistika–2008». SPb.: SPbGU, 2008. S. 45–49.

4. *Tufiş D., Popescu O.A.* Knowledge-based approach to morpho-lexical processing of natural language // In Proceedings of the International Conference for Young Computer Scientists, Beijing, 1991. P. 405–408.

5. *Harper K.E.* Contextual analysis // Mech. Translation. 1956. V. 4, No 3. P. 70–75.

6. *Caplan A.* An experimental study of ambiguity and context // Mech. Translation. 1955. V. 2, No 2. P. 39–46.

7. *Zinkina Yu.V., Pyatkin N.V., Nevzorova O.A.* Razresheniye funktsionalnoy omonimii v russkom yazyke na osnove kontekstnykh pravil // Trudy mezhd. konf. Dialog'2005. M.: Nauka. 2005. S. 198–202.

8. *Kobzareva T.Yu., Afanasyev R.N.* Universalnyy modul predsintaksicheskogo analiza omonimii chastey rechi v RYa na osnove slovarya diagnosticheskikh situatsiy // Trudy mezhdunar. konferentsii Dialog'2002. M., 2002. S. 258–268.

9. *Kurbatov Kh.R.* Grammaticheskiye omonimy v tatarskom yazyke // Tatar tele hem edebiyaty. Kazan: Tatar. kit. neshr. 1959. B. 307–311.

10. *Salakhova R.R.* Omonimichnyye suffiksy tatarskogo yazyka. Kazan: Gumanitarya, 2007. 204 s.

11. *Salimgarayeva B.S.* Omonimy v sovremennom tatarskom yazyke. Avtoref. kand. dis. Ufa, 1971. 82 s.

12. Tatarskaya grammatika. Kazan: Tatar. knizh. izd-vo, 1993. T. II. Morfologiya. 397 s.

13. *Weischedel Ralph M.* Coping with ambiguity and unknown words through probabilistic models // Computational Linguistics. Cambridge, MA, USA: MIT Press, 1993. V. 19, Issue 2. P. 361–382.

14. *Ratnaparkhi A.* Maximum entropy model for part-of-speech tagging // Proceedings of the Empirical Methods in Natural Language Processing. Philadelphia, PA, USA, 1996. P. 133–142.

15. *Lakomkin E.D., Puzyrevskiy I.V., Ryzhova D.A.* Analiz statisticheskikh algoritmov snyatiya morfologicheskoy omonimii v russkom yazyke. URL: http://aistconf.org/stuff/aist2013/submissions/aist2013_submission_33.pdf.

16. *Tkachenko M.V.* Model i algoritm uluchsheniya raspoznavaniya chastey rechi v tekstakh soderzhashchikh oshibki. SpbGU, 2010. 20 s. URL: http://se.math.spbu.ru/SE/YearlyProjects/2010/list.

17. *Zelenkov Yu.G., Segalovich I.V., Titov V.A.* Veroyatnostnaya model snyatiya morfologicheskoy omonimii na osnove normalizuyushchikh podstanovok i pozitsiy sosednikh slov // Kompyuternaya lingvistika i intellektualnyye tekhnologii. Trudy mezhdunarodnogo seminara Dialog'2005. Kazan, 2005. C. 616.

18. *Brill E.* A simple rule-based part of speech tagger // Proceedings of the third conference on Applied natural language processing (ANLC’92). Stroudsburg, PA, USA: Association for Computational Linguistics, 1992. P. 152–155.

19. *Sokirko A.V., Toldova S.Yu.* Sravneniye effektivnosti dvukh metodik snyatiya leksicheskoy i morfologicheskoy neodnoznachnosti dlya russkogo yazyka (skrytaya model Markova i sintaksicheskiy analizator imennykh grupp). URL: http://www.aot.ru/docs/RusCorporaHMM.htm.

20. *Orosz G., Novak A.* PurePos 2.0: a hybrid tool for morphological disambiguation // In Proceedings of the International Conference on Recent Advances in Natural

Language Processing (RANLP 2013), Hissar, Bulgaria, 2013. INCOMA Ltd. Shoumen, Bulgaria. P. 539–545.

21. *Kristen Linden, Tommi Pirinen.* Weighted finite-state morphological analysis of finnish compounding with HFST-LEXC // In Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. Editors: Kristiina Jokinen and Eckhard Bick. NEALT Proceedings Series, 2009. V. 4. P. 89–95.

22. *Deniz Yuret, Ferhan Ture*. Learning morphological disambiguation rules for turkish // Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL. New York, 2006. P. 328–334.

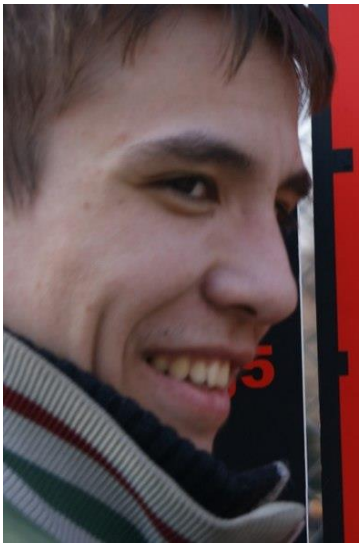
23. *Gataullin R.R., Gilmullin R.A.* Kontekstnyye pravila dlya razresheniya morfologicheskoy mnogoznachnosti v korpuse tatarskogo yazyka // Otkrytyye semanticheskiye tekhnologii proyektirovaniya intellektualnykh sistem OSTIS-2016 (OpenSemantic Technologies for Intelligent Systems). Materialy V mezhdunarodnoy nauchno-tekhnicheskoy konferentsii (Minsk, 18–20 fevralya 2016 goda). Minsk: BGUIR, 2016. S. 389–392.

24. *Hasim Sak, Tunga Gongur, Murat Saraclar*. Morphological disambiguation of turkish text with perceptron algorithm // Computational Linguistics and Intelligent Text Processing, 8th International Conference CICLing, Mexico City, Mexico, February 2007. P. 107–118.

25. *Collins M.* Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms // Proceedings of EMNLP, 2002. P. 1–8.

26. *Suleymanov D.Sh., Gilmullin R.A.* Dvukhurovnevoye opisaniye morfologii tatarskogo yazyka // Tezisy dokl. Mezhdunarodnoy nauchnoy konferentsii "Yazykovaya semantika i obraz mira". Kazan: Izd-vo Kazan. gos. un-ta, 1997. Kniga 2. S. 65–67.

СВЕДЕНИЯ ОБ АВТОРЕ



ГАТАУЛЛИН Рамиль Раисович – аспирант Института вычислительной математики и информационных технологий Казанского (Приволжского) федерального университета.

Ramil Raisovich GATAULLIN, received MS degree in Mathematics from Kazan Federal University (2012). Currently is a graduate student at the Institute of Computational Mathematics and Information Technologies of Kazan Federal University. Current scientific interests: natural language processing, data mining, knowledge extraction technologies.

email: ramil.gata@gmail.com

Материал поступил в редакцию 18 марта 2016 года