OSTIS-2014
(Open Semantic Technologies for Intelligent Systems)

УДК 004.827

ПРОГРАММНЫЙ ИНСТРУМЕНТАРИЙ ДЛЯ РАЗРЕШЕНИЯ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В ТАТАРСКОМ ЯЗЫКЕ

Сулейманов Д.Ш.[*], Гильмуллин Р.А.[*], Гатауллин Р.Р.[*]

[*]*Казанский (приволжский) федеральный университет,*
Научно-исследовательский институт «Прикладная семиотика» АН РТ,
г. Казань, Республика Татарстан

dvd.slt@gmail.com
rinatgilmullin@gmail.com
ramil.gata@gmail.com

В работе описывается программный инструментарий, предоставляющий широкие возможности для создания, редактирования, а также тестирования Базы контекстных правил автоматического разрешения морфологической многозначности в татарском языке.
Ключевые слова: морфема; омоформа; разрешение морфологической многозначности; База контекстных правил.

Введение

Разрешение многозначности в тексте является одной из наиболее сложных и важных задач компьютерной лингвистики. Данная проблема с особой остротой обнаруживается при попытке создания автоматических систем обработки текстов, т.е. полностью автоматизировать такие процессы как поиск информации, перевод текста с одного языка на другой, разметка текстов в электронном корпусе языка, извлечение знаний, контент-анализ и др.

Научные исследования по разрешению многозначности, в частности, лексической многозначности (word sense disambiguation, WSD), исследованию которой для татарского языка наиболее активно занимаются авторы статьи, находятся в поле зрения прикладной и компьютерной лингвистики достаточно давно и имеют многолетнюю историю. Несмотря на то, что с течением времени количество предложенных решений и их эффективность неуклонно растут, а сравнительно-эффективные показатели точности достигли определѐнного, более-менее приемлемого уровня (т.е. достигается ситуация, когда автоматическое разрешение многозначности дает результат, который уже может быть использован в практических задачах, и ручная обработка уже не является предпочтительной в силу ее трудоемкости и дороговизны), полного решения проблема пока не получила, поскольку на пути успешного решения стоит множество задач, напрямую связанных с языковыми особенностями человеческой речи[Зализняк, 2004].

Одной из подзадач WSD является разрешение грамматической омонимии (морфологической многозначности). Исследование этой проблемы весьма важно для задач корпусной лингвистики и компьютерной лексикографии, в особенности при создании электронного корпуса татарского языка, в котором одной из наиболее важных задач является разрешение морфологической многозначности в размеченной коллекции текстов.

Для решения данной задачи необходиморазработатьпринципыклассификациюомоформвтатарскомязыкеипостроитьклассификациюомоформнаосноверазработанныхпринципов, далее – исследоватьконтекстныеограниченияврazличныхтипахомоформиразработатьправиларазрешениямногозначности на основе обнаруженных контекстных ограничений.

В данной работе приводится описание программного инструментария (ПИ), реализованного для решения вышеуказанных задач.

1. Методы разрешения омонимии

В настоящее время в задачах автоматического разрешения омонимии используются следующие методы: контекстный, статистический и гибридный.

Методконтекстногоразрешенияграмматическойомонимиисводитсякразработкедлякаждогофункционального типаомонимиигруппыправил, задающих синтаксический контекст разрешения омонима, и построение управляющей структуры группы, определяющей порядок применения правил. В работе «Разрешение функциональной омонимии в русском языке на основе контекстных правил» [Невзорова и др., 2005] подробно описаны основные достоинства и недостатки данного метода, приведены конкретные структуры обобщенных правил для разрешения функциональной омонимии некоторых типов. Подход, основанный на правилах, является чрезвычайно трудоемким, требуетпроведениятщательнойлингвистическойэкспертизыкаждоготипаомонимии.

Несмотря на указанные недостатки, приходится констатировать, что в настоящее время для татарского языка наиболее предпочтительным является метод основанный на контекстных правилах. Во-первых, в настоящее время не имеется достаточно объемного электронного корпуса татарского языка, позволяющего полноценно задействовать статистические и гибридные модели, во-вторых, регулярность грамматики и строгая подчиненность правилам практически на всех языковых уровнях [Сулейманов Д.Ш., 1994], позволяют рассчитывать на обнаружение и описание четких контекстных ограничений.

1.1. Обобщенный контекстный метод

Обобщенный метод контекстного разрешения функциональной омонимии для татарского языка включает несколько этапов:

1) построение полной классификации типов функциональных омонимов;

2) выделение минимального множества разрешающих контекстов для каждого типа. Минимальность множества означает, что для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу. Затем необходимо построить множество разрешающих контекстов (МРК), имеющих минимальную сложность распознавания. В алгоритмической записи данное требование выражается следующим правилом: если для функционального омонима X, имеющего тип T1 или T2, применено правило из МРК, то тип омонима X определяется примененным правилом, иначе приписывается альтернативный тип;

3) построение управляющей структуры обобщенного правила, обеспечивающего максимальную точность распознавания.

1.1.1. *Пример разрешения функциональной омонимии для татарского языка*

Рассмотрим разрешение функциональной омонимии типа (V+Refl)/(N+3PossSg+Acc),где (V+Refl)–глагол с аффиксом возвратно-страдательного залога, (N+3PossSg+Acc)–существительное с аффиксами принадлежности 3 лица единственного числа, на примере омоформы: *асылын*.

Варианты аффиксальной структуры омоформы:

(1) *ас(V)+Ыл(Refl)+Ын(Refl)* (вешайся) (в предложений «Муенга асылын»)

(2) *асыл(N)+СЫ(3PossSg)+нЫ(Acc)* «истинность, сущность, суть (чего-л., кого-л.)» («Гомернең асылын аңлагыз» «Поймите суть жизни»)

Потенциальные модели, главные компоненты и семантика словосочетаний:

(1) в качестве зависимого компонента не встречается;

(2) N+Acc→V (главный компонент – глагол, семантика прямого объекта).

Данный тип аффиксальной омонимии разрешается следующим правилом:

если в правом контексте находится глагол, возможна потенциальная модель словосочетания N+Acc→V

Соответственно, если реализуется данная модель словосочетания, омонимия разрешается по 2-му варианту морфемной структуры, т.е. *N+3PossSg+Acc: асыл(N)+СЫ(3PossSg)+нЫ(Acc)*.

В формализованном виде обобщенное правило выглядит следующим образом:

$$if\ (\mathrm{X}_{1} \cap_{Acc} V^{*})\ then\ X = N + 3\mathrm{PossSg} + Acc$$
$$else\ X = V + Refl$$

1.2. Ключевые понятия, архитектура контекстного правила

Для решения поставленных задач разработана соответствующая архитектура программного инструментария, включающая следующие базовые объекты и понятия:

- Омоформа (или функциональный омоним) – слова, совпадающие в своем звучании лишь в отдельных формах (той же части речи или разных частей речи;
- База типов омоформ (или База контекстных правил) – иерархически упорядоченный список типов омоформ; для каждого типа определено множество разрешающих контекстов. На основе этих правил происходит разрешение многозначности для отдельно взятого типа омоформ;
- Обобщенное правило разрешения (ОПР) – правило, на основе контекстной информации определяющее актуальный вариант структуры омоформы. Для каждого типа функционального омонима следует оценить сложность распознавания каждой части речи, принадлежащей данному типу

[Сулейманов Д.Ш., 1994];

- Множество разрешающих контекстов (МРК) – совокупность минимальных разрешающих контекстов, достаточных для распознавания функционального омонима как определенного варианта структуры омоформы;

- Управляющаяструктураобобщенногоправил аобеспечиваетиконтролируетпорядокприменениянпр авил (рисунок 1);

- Минимальный разрешающий контекст – неделимое в данном контексте простое условие, имеющее минимальную сложность распознавания.

Пример: существование в левом контексте слова с набора морфем вида «V+Refl» в пределах трех слов; возвращает «истинно»или «ложно»;

- Текст – совокупность предложений; обладает рядом свойств, которые могут быть полезным и при расширении контекстного метода;

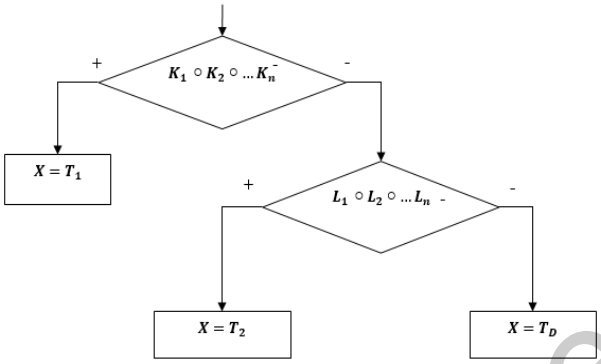

Рисунок 1 – Управляющая структура обобщенного правила $T_1/T_2/T_d$

Здесь, T_1 – морфологический тип, соответствующий распознаванию МРК $K_1 \circ K_2 \circ \ldots K_n$,

T_2 – морфологический, тип соответствующий распознаванию МРК $L_1 \circ L_2 \circ \ldots L_n$,

T_d – морфологический тип по умолчанию; K_i и L_i – минимальные разрешающие контексты для соответствующих типов, $\circ$ – булевы операции (AND, OR ит.д.).

Процесс распознавания омонимии происходит следующим образом:

1) У анализируемого слова определяется тип функциональной омонимии, и в соответствии с этим типом из Базы контекстных правил находится обобщенное правило разрешения;

2) Управляющая структура задает порядок применения правил;

3) При применении каждого правила, проверяется каждый минимальный контекст разрешении этого правила;

4) Если при проверке правила получили подтверждение о его истинности, то функциональная омонимия распознается в соответствии с этим вариантом структуры омоформы;

5) Иначе, если есть другое правило, осуществляется переход к следующему правилу и выполняется то же самое;

6) И если нет другого правила, то в качестве структуры выбирается тип по умолчанию;

7) Если нет такого типа, то многозначность помечается как неразрешенная.

Минимальными разрешающими контекстами могут служить:

- Наличие в левом (в правом, или с обеих сторон) контексте слова определенной формы (или конкретного ожидаемого слова) на определенном расстоянии от проверяемого слова. Проверяется в пределах предложения;

- Наличие определенных характеристик предложения. В программе нужно обеспечить расширяемость этих характеристик (элемент синтаксического анализа);

- Наличие определенных знаков пунктуации на определенном расстоянии слева (или справа);

- Наличие определенной лексемы в предложении.

2. Программный инструментарий для создания и тестирования контекстных правил

2.1. Общие требования к программному комплексу и функционалу

Программный комплекс обеспечивает автоматизацию процесса создания Базы контекстных правил и является «дружественным», т.е. удобным для конечного пользователя.

Программный комплекс содержит следующий функционал:

- создание, редактирование и удаление типов омоформ;
- создание, редактирование и удаление правил и минимальных разрешающих контекстов;
- тестирование: нахождение слов, подходящих выбранному типу омоформ; отображение контекста и морфологической структуры слова.

2.2. Входные и выходные данные

В процессе работы конечный пользователь (в частности, филолог) имеет возможность создавать, редактировать и тестировать Базу контекстных правил. В качестве выходных данных формируется база контекстных правил, которая впоследствии может быть интегрирована в другие программы и использована в целях разрешения морфологической многозначности.

База контекстных правил сохраняется в файл с помощью стандартных методов языка программирования Java.

Для тестирования, в качестве входных данных, программа принимает текст с уже проведенным морфологическим разбором. Текст устроен таким образом, что каждой словоформе отводится две строки: 1) сама словоформа, 2) морфологический разбор данной словоформы (варианты разбора разделяются символом «;»).

2.3. Рабочие области программного инструментария

Программный комплекс содержит два основных модуля: а) модуль для создания, редактирования и тестирования Базы контекстных правил иб) модуль для разрешения морфологической многозначности на базе этих контекстных правил.

2.4. Модуль для создания, редактирования Базы контекстных правил

Основное окно (рисунок 2) состоит из рабочих областей. Каждая область отображает определенную информацию о состоянии базы.

Область№1

В этой области отображается выборка слов из загруженного текста с неоднозначным морфологическим разбором. При выборе правила, каждое слово подсвечивается зеленым цветом, если слово по типу подходит выбранному правилу, иначе – красным цветом.

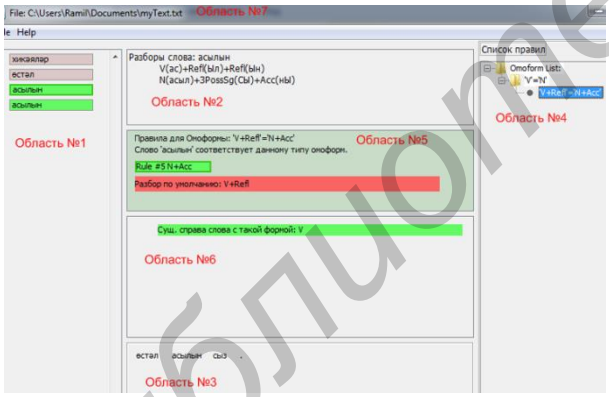

Рисунок 2 – Рабочие области программного инструментария

Область№2

В этой области отображаются разборы выбранного слова (изОбластей№1 или№5)

Область№3

В этой области отображается предложение, в которое входит выбранное слово. Подводя к слову указатель мыши, можно узнать ее морфологический разбор в Области №2.

Область№4

В этой области отображается текущая База правил. Для того, чтобы загрузить сохраненную базу, необходимо нажать в верхнем меню *Файл->Загрузить Базу омоформ* и в появившемся окне выбрать нужный файл с базой правил. Также с помощью *Файл->Загрузить Базу по умолчанию* можно загрузить базу по умолчанию с одним единственным правилом. Для сохранения необходимо открыть *Файл->Сохранить Базу омоформ* и в появившемся окне выбрать папку для сохранения.

Для добавления нового правила, нужно сначала выбрать к какому типу оно будет относиться. Правила привязаны к типам. У каждого типа может быть подтип. База правил достаточно гибкая. После того как определен тип правила, выбирается родительский тип (если это подтип), или корневой элемент (Список омоформ), далее необходимо нажать правую кнопку мыши и в контекстном меню выбрать *Добавить новый тип омоформ*. Затем появится окно с несколькими полями: первый тип разбора, второй тип разбора, тип разбора по умолчанию и ID, которые необходимо заполнить. Для редактирования или удаления правила, необходимо также выбрать нужный тип омоформ и выполнить нужное действие.

Область№5

После выбора типа омоформы, в этой области отображается информация о правилах. Здесь же их можно добавлять, изменять и удалять, нажав на правую кнопку мыши и выбрав соответствующий функционал в контекстном меню (рисунок 3).

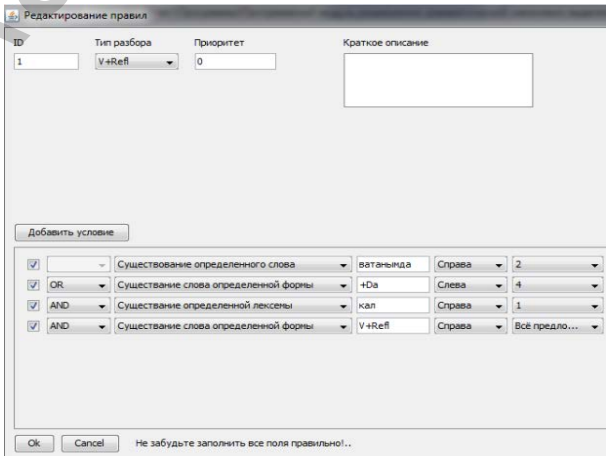

Рисунок 3 – Окно создания и редактирования правила

Кроме правил, в конце списка может располагаться разбор по умолчанию. Выбрав нужный тип омоформ из Области №4, и выбрав нужное слово из Области №1, можем проверить соответствие правила этому слову:

- если область окрашивается в красный цвет (и появляется надпись «Слово не соответствует данному типу»), то это значит, что неоднозначность разбора слова не соответствует данному типу. Иначе неоднозначность соответствует данному типу, т.е. она может быть разрешена с помощью этого правила;

- если одно из правил окрашивается в зеленый цвет, то это означает, что неоднозначность разбора

будет разрешена по данному правилу (также, если нажать на данное правило, то в Области №6 тоже загорятся выполненные атомарные условия)

- если, ни одно из правил не будет окрашено в зеленый цвет, т.е. если ни одно из правил не будет выполнено, то неоднозначность разбора будет разрешена разбором по умолчанию, если разбор по умолчанию отсутствует, то неоднозначность останется неразрешенной.

Область№6

В этой области отображаются атомарные условия правила, выбранного в Области №5. В настоящее время в программе реализованы две разновидности: 1) Проверка существования определенного слова, 2) Проверка существования определенной формы слова

Каждому типу можно добавить дополнительные условия: 1) Контекст проверки (слева, справа или с обеих сторон), 2) Расстояние проверки.

Атомарные условия определяются логическими операциями (AND, OR, AND NOT, OR NOT).

Область№7

Здесь отображается адрес выбранного текста.

2.5. Модуль разрешения морфологической многозначности с использованием Базы контекстных правил

Модуль разрешения морфологической многозначности представляет собой программное приложение, использующее ту же модель, что и модуль для создания, редактирования и тестирования Базы контекстных правил. В отличие от предыдущего модуля, данный модуль предназначен, в первую очередь, для разрешения многозначности в тексте, используя Базу контекстных правил. При отсутствии определенного правила имеется возможность пометить текст вручную. Выходной файл текста сохраняется в том же виде (с той же структурой), что и входной файл.

Модуль состоит из следующих основных разделов (рисунок 4):

1. Область№1 – Основной раздел текста,
2. Область№2 – Раздел Базы контекстных правил,
3. Область№3 – Раздел разбора слова и редактирования форм,
4. Область№4 – Консоль,
5. Область№5 – Титул окна с адресом файла текста,
6. Область№6 – Кнопки базовых действий.

Область№1 – Основной раздел текста

Здесь отображается текст из загруженного файла. Как и в предыдущем модуле структура файла такова, что каждой словоформе соответствуют две строки: 1) сама словоформа, например *өстәл* (стол), 2) Варианты морфологического разбора для данной словоформы: V(өстә (добавь))+REFL(Hl):V(өстәл)–добавься/N(өстәл) - *стол*;

Загружая такого рода текст, для каждого слова создается объект. Слова собираются в предложения, предложения в абзацы. Таким образом, массив слов отображается в этой области. В отличие от предыдущего модуля здесь присутствуют все слова, независимо от количества разборов и их типов. Это объясняется тем, что, в первую очередь, с текстом будет работать человек, которому для разрешения неоднозначности необходим весь контекст. Для снятия неразрешенных модулем неоднозначностей, пользователю предоставляется возможность проверять правильность и самой Базы контекстных правил. Для удобства работы «слова»разделяются на несколько типов (состояний) и отличаются цветовым оформлением (каждому типу соответствует свой цвет фона).

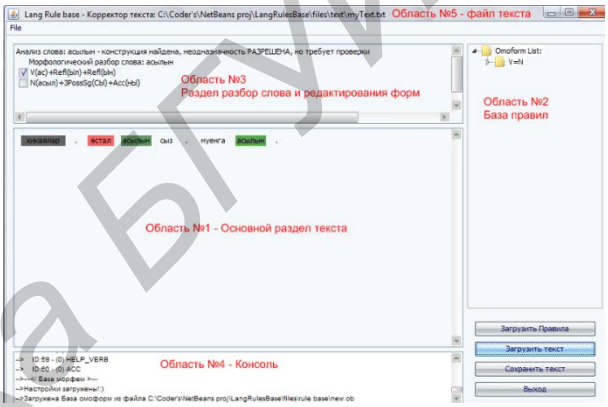

Рисунок 4 – Окно модуля разрешения морфологической многозначности.

Фоновый цвет слов с однозначным разбором совпадает с фоном самого текста. Слова с неразрешенными неоднозначностями обозначаются красным цветом. Разрешенные слова, но не проверенные человеком, светло-зеленым; проверенные (такого же цвета будут и слова, разрешенные вручную) - темно-зеленым. Слова, для которых нет правил, выделяются серым цветом. При нажатии на слово из этой области в Области№3 отображаются формы разбора, а также информация о правиле, по которой его можно разрешить.

Область№2 – Раздел База контекстных правил

В этом разделе отображается список правил, загруженных с файла База контекстных правил, разработанная с помощью предыдущего программного модуля. Однако, в данном случае, правила нельзя редактировать, так как здесь они представлены только в качестве информационного раздела.

Область№3 – Раздел разбора слова и редактирования форм

Как показано выше, в этом разделе отображается форма разбора выбранного слова, а также информация о правиле, по которой его можно разрешить. Кроме того, здесь же представлена и кнопка «Сохранить» (при нажатии на правую

кнопку мыши в этой области), которая сохраняет в качестве правильной ту форму, которая отмечена галочкой (слева в этой же области).

Область№4 – Консоль

Этот раздел носит вспомогательный характер, отражает информацию об истории и стабильности работы, такого типа, как: «Настройки загружены!», или: «Загружена База омоформ из файла C:\Coder's\NetBeans proj\LangRulesBase\files\rule base\new.ob»

Область№5 – Титул окна с адресом файла текста

Здесь отображается название загруженного файла текста.

Область№6 – Кнопки базовых действий

Базовые действия, такие как «Загрузить правила», «Загрузить текст», функционируют также, как и в предыдущем модуле.

Заключение

В данной работе описан программный инструментарий, предоставляющий широкие возможности для создания, редактирования, а также тестирования Базы контекстных правил автоматического разрешения морфологической многозначности. Выходной файл Базы правил в последующем может быть применен также в других программных приложениях для промежуточногоразрешенияморфологическоймногозначностивсамыхразныхлингвистическихзадачах.

Программный инструментарий также фиксирует проанализированные слова, их типы и результаты разрешения. В итоге собирается статистика, способная помочь в дальнейших разработках и в исправлениях разного рода ошибок в самих контекстных правилах.

В дальнейшем нами планируется расширение возможностей описанного программного инструментария как за счет добавления новых правил для выявленных и описанных неоднозначностей, так и за счет добавления правил разрешения новых типов неоднозначностей. Потенциал повышения эффективности представленного в данной статье программного инструментария видится также в совмещении различных подходов к разрешению многозначности, когда наряду с контекстными методами используются вероятностные модели и статистические методы.

Библиографический список

[Закиев и др., 1993] Татарская грамматика. Т.II. Морфология / Ред. М.З.Закиев, Ф.А.Ганиев, К.З.Зиннатуллина. – Казань: Татарское кн. изд-во, 1993. – 397 с.

[Сулейманов, 1994] Сулейманов Д.Ш. Регулярность морфологии татарского языка и типы нарушений в языке / Д.Ш.Сулейманов // Когнитивная и компьютерная лингвистика / Науч. ред. Р.Г.Бухараев, В.Д.Соловьев, Д.Ш.Сулейманов. – Казань: КГУ, 1994. – С.77-106.

[Зализняк, 2004]Анна А. Зализняк. ФЕНОМЕН МНОГОЗНАЧНОСТИ И СПОСОБЫ ЕГО ОПИСАНИЯ. Вопросы языкознания. — М., 2004. — № 2. — С. 20-45.

[Невзорова, 2005]Разрешение функциональной омонимии в русском языке на основе контекстных правил // Труды межд. конф. Диалог'2005.– М.: Наука, 2005. С. 198-202.

SOFTWARETOOLFORMORPHOLOGICAL DISAMBIGUATION IN THE TATAR LANGUAGE

SuleymanovD.Sh.[*], GilmullinR.A.[*], Gataullin R.R.[*]

[*] *Kazan Federal University, Research Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Russia*

dvd.slt@gmail.com

rinatgilmullin@gmail.com

ramil.gata@gmail.com

The article describes a software tool for creating, editing, and testing base contextual rules for the automatically resolve morphological ambiguity in the Tatar language.