Литература

1. V. Devedžić. Understanding Ontological Engineering // Communications of the ACM. – Vol. 45, № 4ve. – 2002. – P. 136 – 144.

2. D. King, K. Kimble. Uncovering the epistemological and ontological assumptions of software designers // Proceedings 9e colloque de l'AIM, Evry, France, May 2004. – 9 p.

3. F. Steimann. On the representation of roles in object-oriented and conceptual modeling. // Data &Knowledge Engineering. – Vol. 35, № 1. – 2000. – P. 83 – 106.

4. M. Baldoni, G. Boella, L. van der Torre. powerJava: ontologically founded roles in object-oriented programming language. // Procs. of OOPS Track of ACM SAC'06. – ACM – 2006. – P. 1414 – 1418.

ПОСТРОЕНИЕ СЛОВАРЯ ДЛЯ МОРФОЛОГИЧЕСКОГО АНАЛИЗА РУССКИХ СЛОВ

П.А. Вейник, Н.Д. Походенько
Беларусь, г. Минск

Задача автоматизированного морфологического анализа русских слов становится все более актуальной, так как ее решение будет способствовать развитию технологий обработки текста. Такой анализ может быть применен в системах поиска текстовой информации и перевода, в автоматическом реферировании, для автоматической расстановки ударений. Кроме того, оно служит основой для дальнейших этапов анализа текста — синтаксического и семантического.

На данный момент полностью формализован и реализован метод получения всех форм слова по его начальной форме (задача морфологического синтеза). При этом используется грамматический словарь для синтеза словоформ (далее — словарь синтеза), содержащий начальные формы слов вместе с их характеристиками.

В оперативной памяти словарь представляется в виде 34-арного дерева, отсортированного по начальной форме [1]. В словаре синтеза по начальной форме слова определяются его характеристики и в соответствии с ними могут быть построены все его формы. Если в словаре нет интересующего слова, реализована возможность дополнить словарь недостающей информацией. Очевидно, что описываемый метод нуждается в том, чтобы словарь синтеза был по возможности полным как, например, «Грамматический словарь — словоизменение» А. А.Зализняка, содержащий около 100000 слов. Однако все доступные грамматические словари ориентированы на восприятие человеком и являются непригодными для работы в автоматическом режиме. Поэтому электронная версия грамматического словаря Зализняка в текстовом формате была преобразована в более удобный формат XML.

Существенным достоинством XML как языка представления данных является то, что он базируется на текстовом формате и потому может быть с легкостью понят как человеком, так и прочитан машиной. Более того, XML файлы могут быть без труда прочитаны, созданы или модифицированы при помощи

стандартного текстового редактора. Автоматическая обработка документов в формате XML облегчается существованием большого количества библиотек. Этот формат является общепринятым, потому что позволяет без каких либо затруднений выражать произвольную структуру данных, причем в качестве тегов могут быть использованы осмысленные строки.

Кроме описанной задачи синтеза форм решена также и обратная задача: по словоформе получить ее характеристики и начальную форму (задача анализа). Для ее решения, как и для решения задачи синтеза, необходим словарь. Но словарь синтеза, даже преобразованный в формат XML, для этой цели непригоден. Чтобы убедиться в этом, рассмотрим в общих чертах тот алгоритм, который для решения задачи анализа словоформы использует словарь, предназначенный для синтеза словоформ. Поскольку основа слова неизвестна, из словаря синтеза придется выделить подмножество тех слов, которые частично совпадают с анализируемым, а затем по очереди сравнивать наше слово со всеми формами каждого из слов подмножества. Следует иметь в виду, что существительное, к примеру, имеет 12 форм, а глагол — 185 и, кроме того, среди частей речи глагол в русском языке представлен наиболее широко. Все это делает такой алгоритм неприемлемым с точки зрения времени выполнения. Проблему не решает даже предварительная сортировка слов подмножества по убыванию степени совпадения с анализируемым словом и исключение из числа проверяемых словоформ тех, которые имеют окончания, не укладывающиеся в анализируемое слово справа.

Возможен также и другой подход: искать грамматические характеристики и начальную форму слова в сформированном заранее словаре анализа. Структура словаря анализа была предложена Кулагиной О.С.[4, 5]. Ей также был разработан алгоритм определения грамматических характеристик с его помощью. В словаре слова разбиты на пять морфологических классов [4]. Каждый класс характеризуется своими признаками и своим алгоритмом анализа. К первому классу относятся слова субстантивного склонения, ко второму слова адъективного склонения, к третьему — глаголы, к четвёртому — числительные. Так как невозможно ограничить многообразие морфологических форм русского языка какими-либо правилами, кроме указанных четырех морфологических классов введен еще один класс для неизменяемых слов или таких форм изменяемых слов, которые не вкладываются в общую парадигму своего класса и являются исключениями. Эти формы названы особыми.

Внутри каждого морфологического класса, кроме последнего, слова разбиты по признаку формального сходства словоизменения на словоизменительные типы, соответствующие типам склонения и спряжения, а в классе глаголов, как в самом сложном, выделено несколько типов окончаний и суффиксов. Каждый тип приписанный к слову, является ссылкой на соответствующую таблицу, содержащую набор флексий, а при каждой флексии хранится набор морфологических признаков, при которых эта флексия может встречаться в словоформах данного слова.

Словарь состоит из словарных основ, при которых хранятся признаки — словоизменительные типы, типы окончаний и суффиксов. Во время анализа в

словаре ищется основа, максимально вкладывающаяся слева в словоформу, и считываются признаки, хранящиеся при этой основе. После этого от словоформы отсекается основа, а в оставшейся части среди наборов флексий, определяемых признаками, ищутся флексии, входящие в словоформу. Морфологические характеристики, хранящиеся при флексиях в наборе, являются результатом морфологического анализа словоформы.

Этот алгоритм с некоторыми дополнительными признаками позволяет очень эффективно и с большой точностью (с точностью до морфологической омонимии) проводить анализ словоформы.

Для практической реализации этого подхода необходимо располагать словарем, имеющим описанную структуру. Однако необходимого словаря нигде обнаружено не было. В связи с этим был предложен алгоритм, с помощью которого словарь анализа строился на основании словаря синтеза [2]. При этом подходе все вычисления выполняются единожды, в процессе формирования словаря анализа. Однако для некоторых статей словаря синтеза алгоритм не может построить статьи словаря анализа. Это объясняется тем, что русский язык богат исключениями. Алгоритм преобразования словарей в том виде, в котором он используется на данный момент является достаточно сложным: в нем использованы идеи алгоритмов синтеза и анализа словоформ. Дальнейшее развитие алгоритма представляется нецелесообразным ввиду неоправданно высоких временных затрат на его разработку и тестирование.

Кроме того, в словаре синтеза присутствуют ошибки, так как он был получен путем преобразования в XML его текстовой электронной версии. При создании текстовой версии словаря также допускались ошибки. Оригинальный словарь является справочным изданием, ориентированным на восприятие человеком грамматической информации Этим обусловлено достаточно свободное описание языковых феноменов: многие отклонения от стандартных правил описываются неформально. Отсюда следует, что в процессе формирования словаря, столь необходимого для морфологического анализа, требуется участие человека.

Для увеличения производительности и устранения ошибок при формировании словаря анализа была написана программа с графическим пользовательским интерфейсом, позволяющая просматривать, проверять и редактировать словарные статьи обоих словарей. Язык программирования Java был выбран как один из современных языков, позволяющий создавать программы, которые могут исполняться на любой платформе. Вторым по важности свойством языка Java является его объектная ориентированность. Этот язык позволяет создавать высоконадежные и легко сопровождаемые приложения. При разработке пользовательского интерфейса была использована распространенная библиотека Swing. Благодаря продуманной декомпозиции визуальных компонентов, разработчики Swing предоставили пользователю возможность заменять часть стандартной реализации компонента своей собственной. Это позволяет настраивать внешний вид и поведение компонента необходимым образом.

При работе с программой пользователь может выбрать файл, содержащий словарь синтеза. После этого программа считывает из файла словарь и форми-

рует из его статей дерево. Такое представление делает время поиска записи в словаре зависящим лишь от длинны требуемого слова и не зависящим от размера словаря. Пользователь может ввести интересующее его слово в поле редактирования, или воспользоваться кнопкой «Дальше» для перехода к следующему в алфавитном порядке слову. Как только слово выбрано, визуальные компоненты интерфейса, управляющие статьей словаря синтеза переходят в состояния, соответствующие содержимому связанной со словом статьи, сазу же строятся все грамматические формы слова и вместе с соответствующими характеристиками отображаются в таблице. При этом для каждой формы указывается ударение. Если пользователь не обнаруживает ошибок, он может подтвердить правильность текущей словарной статьи. В противном случае пользователь может отредактировать статью при помощи элементов управления.

Кроме того, программа конвертирует текущую статью словаря синтеза в несколько статей словаря анализа, каждая из которых может описывать несколько словоформ. При этом каждой словоформе соответствует только одна из статей словаря анализа. Все полученные статьи отображаются визуально, и пользователь может редактировать их, удалять или добавлять новые

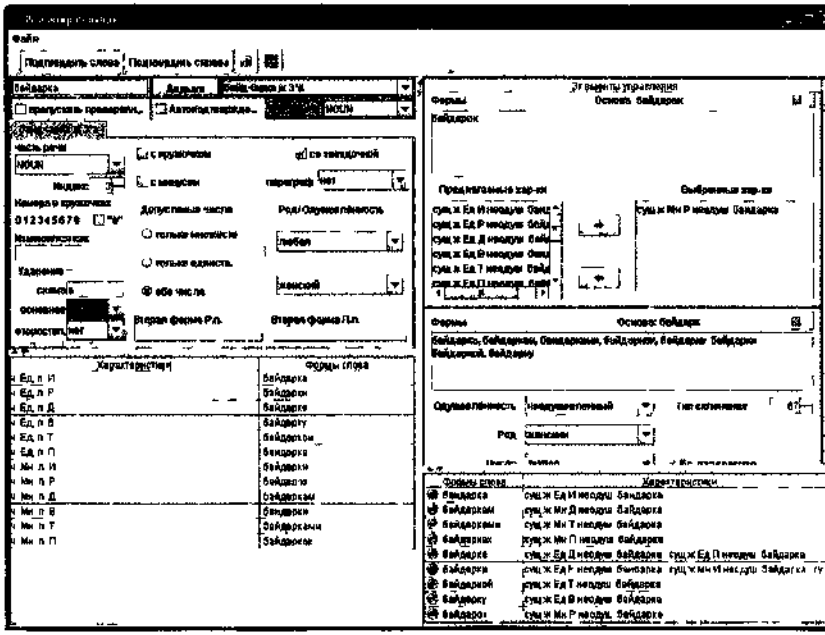

Рис 1 Пользовательский интерфейс программы редактирования словарей

После каждого действия пользователя, модифицирующего статью словаря синтеза или хотя бы одну из статей словаря анализа, полученных путем конвертирования первой, а так же при добавлении или удалении статьи словаря анализа программа производит автоматическую проверку. Проверка заключается в следующем: каждую форму выбранного слова программа пытается проанали-

зировать с помощью статей словаря анализа. Если в результате анализа какой либо из словоформ получены грамматические характеристики, не совпадающие с характеристиками, которые диктует для данной словоформы статья словаря синтеза, то считается, что конвертирование было проведено неправильно либо статья словаря синтеза неправильная, и данная словоформа выделяется красным цветом. Пользователь может подтвердить правильность статей словаря анализа и тем самым добавить их в дерево-образ формируемого словаря анализа. Для удобства использования программы при проверке большого числа статей пользователь может активировать опцию автоматического подтверждения. При условии, что автоматическая проверка не выявила ошибок, эта опция приводит к автоматическому подтверждению правильности всех статей всякий раз, когда пользователь щелкает по кнопке «Дальше». Если активировать опцию «Пропускать проверенное», то при каждом щелчке по кнопке «Дальше» будет выбрана следующая не проверенная статья словаря синтеза.

Результатом работы является удобное и надежное средство получения словарей, используемых для морфологического анализа и синтеза. С помощью этих словарей можно эффективно решать проблемы автоматического перевода, извлечения информации из текста, автоматического индексирования баз данных в информационно поисковых системах, сжатия текстовых данных, проверки грамматической правильности текста, создания электронных словарей и обучающих систем, и на дальнейших этапах исследований.

Литература

1. Вейник П. А., Волосевич А. А. Структура грамматического словаря для синтеза русских словоформ, Инженерный вестник, 2006, 1(21)/3, с.174 - 177.

2. Вейник П. Проблема расстановки ударений в русском тексте, Материалы конференции «Информационные системы и технологии», Часть 2, Мн., 2006.

3. Зализняк А. А. Грамматический словарь русского языка, М., 1983.

4. Кулагина О.С. Морфологический анализ русских именных словоформ, М., 1986.

5. Кулагина О. С Морфологический анализ русских глаголов, М., 1986.

РОЛЕВЫЕ МОДЕЛИ В СОЦИАЛЬНЫХ СЕТЯХ

Заяц Ю.Э.
Беларусь, Минск

Введение

Социальная сеть представляет собой множество социальных объектов, связанных между собой некоторой формой социальных отношений. В связи с высокой популярностью таких сервисов, как MySpace [1], LiveJournal [2], YouTube [3], Flickr [4] социальные сети стали представлять интерес для исследования. В настоящее время данные сервисы широко используются для блоггинга, подкаст-вещания, публикации данных, поиска сотрудников и работода-