074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 475

МНОГОЦЕЛЕВАЯ СЛОВАРНАЯ ПОДСИСТЕМА ИЗВЛЕЧЕНИЯ ПРЕДМЕТНОЙ ЛЕКСИКИ

MULTIPURPOSE DICTIONARY SUBSYSTEM FOR EXTRACTION OF SUBJECT LEXICON

Сидорова Е.А. *(lena@iis.nsk.su)*
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск

Рассматривается словарная технология, предназначенная для создания предметно-ориентированных словарей и их использования при решении различных задач анализа текста в информационных системах. Исследуются проблемы одновременного использования нескольких словарей и согласования разнотипной словарной информации.

1. Введение

Современные требования к качеству содержательного анализа текстовых документов приводят разработчиков к необходимости ограничивать как предметную область содержания документа, так и стиль оформления текста. Кроме того, большинство информационных систем, если они не являются специализированным литературным ресурсом, работают с документами, написанными в жанре «деловой прозы» [1]. Однозначный контекст, строгое выражение мысли в таких текстах значительно уменьшает количество способов выражение одной и той же информации. Служебная либо общезначимая лексика, характерная для текста на естественном языке, исполняет роль связывания или замещения значимой лексики; не используются образные, литературные, эмоциональные выражения. Безусловно, это упрощает создание словарей – позволяет создавать специализированные словари небольшого объема и эффективно их использовать.

С другой стороны, разнообразие целей (задач) анализа текста даже в рамках одной информационной системы требует включать в словари самую разнообразную информацию. Например, при создании информационного Интернет-портала [2] для решения различных задач потребовалась следующая информация:

- для оценки релевантности документа тематике портала – статистика, многословные устойчивые термины, сокращения и аббревиатуры предметной области и т.п.;
- для автоматического определения жанра документа – признаки сегментов, лексические конструкции, формальные модели документов;
- для классификации или рубрикации – поддержка иерархии признаков и статистика;
- для простой индексации (словами) – морфология;
- для содержательной индексация – семантические признаки, связи между терминами (синонимы, антонимы и т.п.);
- для обработки незнакомых слов (поиск и исправление ошибок) – хэш-словари словоформ.

На практике такой словарь зачастую представляет собой объединение нескольких разнотипных и разноформатных словарей, к тому же создаваемых разными авторами или группами авторов. Использование набора словарей для реализации конкретной задачи ставит вопрос о согласовании данных из разных словарей (таких как семантические характеристики, связи между элементами словарей и т.п.), а также согласовании получаемых результатов словарной обработки текста с помощью словарных компонентов.

В данной работе описывается система, назначение которой создание ПО-словарей и их использование для анализа текста в различных информационных системах.

2. Информационное наполнение словарей

Одной из задач разрабатываемой технологии было создание гибких механизмов, позволяющих специалисту проводить тонкую настройку структуры словаря. Это выражается в возможности в значительной степени формировать структуру словарной статьи для терминов одного вида.

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 476

2.1. Типизация признаков словарной статьи

Все признаки, хранящиеся в словарной статье термина, мы условно разделили на четыре группы, в зависимости от их функционального назначения.

Терминообразующие признаки служат для того, чтобы с одной стороны выявить термин в тексте (анализ), с другой – послужить основой для построения терминов (синтез). Для разных видов терминов, набор терминообразующих признаков различен. В случае терминообразующих признаков возможность изменять структуру словарной статьи сильно зависит от вида термина.

Семантические признаки приписываются терминам словаря и передаются внешним программам вместе с найденными в тексте терминами. Эта семантическая информация позволит связать элементы словаря с онтологическими классами проблемной и предметной области и в дальнейшем будет использоваться на стадии семантического анализа текста. Набор признаков и их тип определяется пользователем при создании и наполнении словаря.

Семантическая информация может быть выражена в словаре по-разному. Во-первых, это семантический класс (или несколько классов), к которому приписывается термин словаря. Во-вторых, это атрибуты, добавляемые пользователем в словарную статью терминов. Наличие и тип значения атрибута фиксируется в структуре словарной статьи для всех или выделенной группы терминов (класса) словаря. В-третьих, это связи между элементами словаря, такие как отношения синонимии, омонимии, часть-целое и пр.

Статистические признаки накапливают статистическую информацию о появлении термина в обрабатываемых текстах. Часто такие признаки служат для наполнения словарей. Так, при создании информационных систем, как правило, изначально имеется большая выборка ресурсов, размеченная и соотнесенная тематическим разделам, и, используя классические методы обучения, можно сразу получить начальное наполнение словаря, которое, в противном случае, пришлось бы вводить вручную многочисленным специалистам. Помимо этого, наличие в словаре такой информации позволяет использовать статистические методы классификации (рубрикации) для определения общей тематики документа.

Если в случае семантических признаков, пользователь имеет полную свободу в формировании словарной статьи, то в случае статистических терминов, такая свобода отсутствует. Пользователь может только задавать иерархию рубрик или тем, согласованных с обучающей выборкой, в соответствии с которой будет автоматически накапливаться статистика, и, при желании, вручную устанавливать веса терминов в той или и иной теме, которые будут при дальнейшей обработке иметь приоритетное значение по сравнению с автоматически получаемыми значениями.

Еще одна важная группа признаков, упрощающая работу со словарями внешним программам, это динамические признаки, которые появляются у терминов после того, как они найдены в тексте документа в результате словарного анализа текста. К ним относятся такие признаки, как статус обновления статистических параметров термина, видимость найденных терминов после словарной обработки текста (см. Рис.1.), позиция найденного термина в тексте и др. К этой же группе можно отнести и те признаки, которые влияют на значения динамических признаков. Например, значимость термина позволяет управлять параметром видимости. Этот атрибут имеет следующие значения:

- самостоятельно незначимый термин всегда «невидим» и используется только в составе других терминов. Например, аббревиатуры в составе других названий (ОАО, СО);

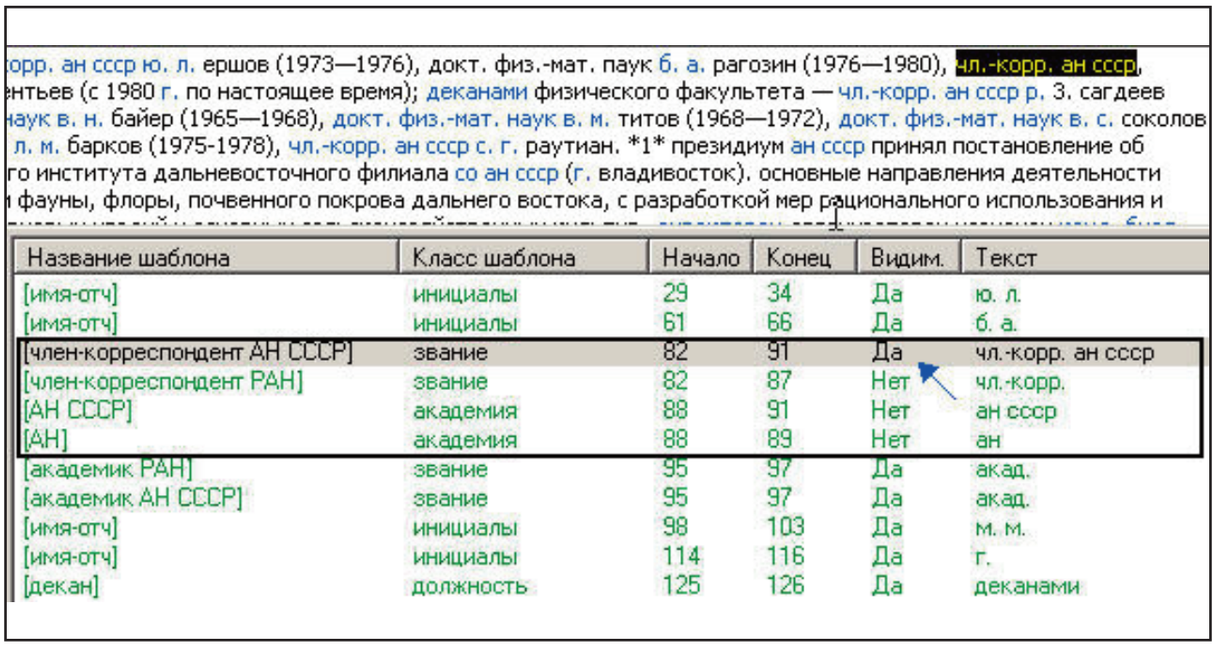
- самостоятельно значимый термин имеет положительную видимость, если не входит в состав другого самостоятельно значимого или абсолютно значимого термина и отрицательную видимость в противном случае (это означает, что строка текста, покрываемая термином, строго вложена в часть текста, покрываемом другим термином). К таким терминам относится большинство предметных терминов;

- универсально значимый термин является самостоятельно значимым, но не влияет на видимость вложенных в него терминов (например, разрывные термины или термины, характеризующие целый текстовый сегмент);

- абсолютно значимый термин всегда имеет положительную видимость независимо от того, входит он в состав другого термина или нет (например, такие жанровые термины как обращение, заключительная реплика и т.п.).

На Рис.1. приведен пример результата словарной обработки текста.

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 477



Название шаблона	Класс шаблона	Начало	Конец	Видим.	Текст
[имя-отч]	инициалы	29	34	Да	ю. л.
[имя-отч]	инициалы	61	66	Да	б. а.
[член-корреспондент АН СССР]	звание	82	91	Да	чл.-корр. ан ссср
[член-корреспондент РАН]	звание	82	87	Нет	чл.-корр.
[АН СССР]	академия	88	91	Нет	ан ссср
[АН]	академия	88	89	Нет	ан
[академик РАН]	звание	95	97	Да	акад.
[академик АН СССР]	звание	95	97	Да	акад.
[имя-отч]	инициалы	98	103	Да	м. м.
[имя-отч]	инициалы	114	116	Да	г.
[декан]	должность	125	126	Да	деканами

Рис. 1. Пример разбора цепочки вложенных терминов.

Найденные термины представляются в виде объектов, у которых, в данном случае, имеются динамические признаки *Начало, Конец, Видимость* и *Текст*. Границы терминов *АН, АН СССР, член-корреспондент РАН* (элемент *РАН* необязательный) включены в интервал, определяемый границами термина *член-корреспондент АН СССР*, поэтому для этих терминов признак видимости отрицателен. Отрицательное значение признака *видимость* означает, что термин не будет принимать участие в дальнейшем анализе.

2.2. Типизация терминов

Нами были выделены несколько видов терминов, для каждого из которых реализован отдельный словарный компонент.

Лексема или однословный термин представляет собой слово во всей совокупности его форм и значений. Словарная статья лексемы содержит следующие терминообразующие признаки:

- нормальная форма,
- основа – неизменяемая при склонении или спряжении часть лексемы (у некоторых лексем основа может быть пустой, тогда лексема определяется парадигмой),
- парадигма – набор псевдофлексий или изменяемых частей слова,
- морфологический класс (класс, в частности, включает часть речи и словообразующие морфологические признаки лексемы),
- тип: слово универсального словаря, предсказание, слово служебного словаря и т.п.

Набор морфологических признаков, включаемых в словарную статью лексемы, может легко настраиваться пользователем, являющимся специалистом в данной области.

Термин-словокомплекс – устойчивое терминологическое сочетание, характерное для выбранной предметной области. Наиболее распространенными структурами здесь являются сочетания существительного с прилагательным, существительного с существительным в косвенном падеже, существительного с другим существительным в качестве приложения. Имеются также многословные термины, иногда состоящие из трех и более слов [3].

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 478

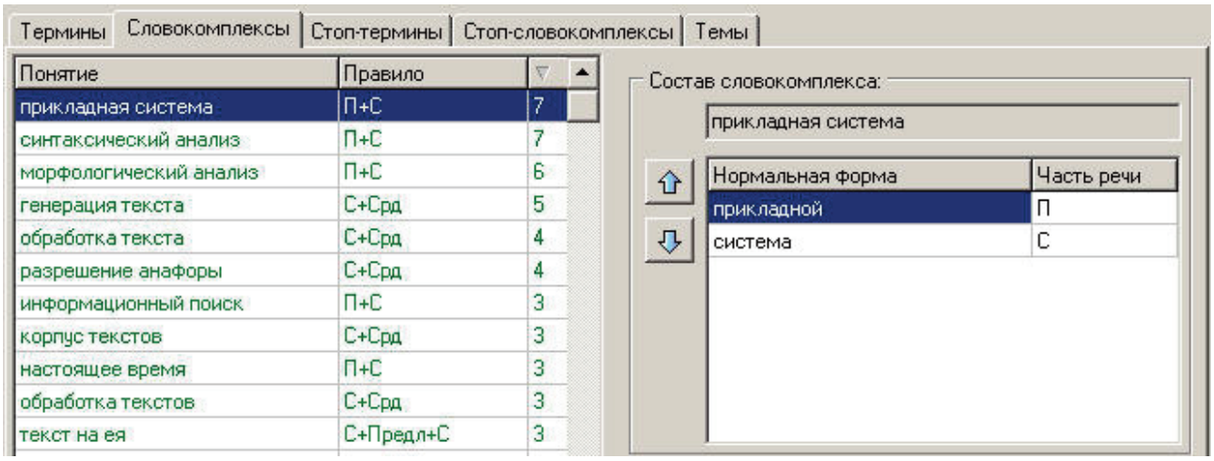


Рис. 2. Фрагмент словаря словокомплексов.

Словарная статья словокомплекса (Рис.2.) содержит следующие терминообразующие признаки:

- нормальная форма,
- список составляющих терминов,
- правило, согласно которому образуется данное сочетание.

Структура словарной статьи словокомплекса фиксирована и не может быть расширена пользователем.

Лексическая конструкция. Под лексической конструкцией понимается несловарная единица, имеющая регулярную структуру, например, номер телефона, дата, инициалы и т.п. Также с помощью лексических конструкций могут описываться специфические термины предметной области, отсутствующие в универсальном словаре русского языка. Для создания словаря лексических конструкций используется технология Alex [4]. Лексическая конструкция в словаре Alex – это множество фрагментов текста произвольной сложности (в общем случае, разрывных), представляющее собой список альтернатив, связываемых с определенной строковой конструкцией.

Отдельной группой стоят формальные сегменты, служащие для сопоставления тексту его жанровой модели, выделения текстовых структур, которые впоследствии могут существенно ограничивать смысл, содержащейся в текстовом фрагменте информации.

2.3. Словари формальных сегментов

Формальный сегмент, по сути, это та же лексическая конструкция, но которая определяет не весь сегмент, а только маркирующие его элементы. Элементы либо задают начало и конец сегмента явным образом, либо неявно через границы другого формального сегмента (в этом случае маркирующий элемент должен быть вложен в сегмент). В качестве маркирующего элемента может выступать либо произвольная строка, либо любой термин, определенный в других словарях, используемых совместно со словарем сегментов.

К терминообразующим признакам сегмента относятся ограничения, на основании которых осуществляется поиск и формирование сегмента:

- single – сегмент не должен пересекаться с сегментами того же типа; частный случай этого ограничения – отсутствие вложенности,
- min – выбираются минимальные из возможных сегменты,
- max – выбираются максимальные из возможных сегменты,
- count – задается максимальное количество сегментов, извлекаемых из текста.

Возможны и другие типы ограничений.

Семантические признаки сегментов находят свое отражение в жанровой модели документа, которая определяется набором сегментов, порядком их следования и вложенностью. Модель документа впоследствии может быть использована для автоматического определения жанра документа.

Примеры жанровых моделей документов можно найти в [4, 5].

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 479

3. Архитектура словарной подсистемы

Для того, чтобы применять несколько словарей и словарных компонент для решения некоторой задачи, необходим менеджер, осуществляющий управление потоками данных и скрывающий от внешнего пользователя (программы) особенности используемых словарей. Для разных задач менеджер загружает различные сценарии, которые определяют порядок применения словарей, набор входных параметров и схему согласования результатов словарного анализа.

Рис. 3. Архитектура словарной подсистемы

Созданная словарная технология включает словарные компоненты и обработчики, которые обеспечивают с одной стороны, автоматизацию создания и наполнения словарей, с другой стороны, словарный анализ и последующую работы со словарной информацией найденных в тексте терминов (означивание классов термина и его индивидуальных признаков).

Для целей создания словарей были разработаны следующие модули.

- Модуль просмотра конкорданса – позволяет в выбранном корпусе текстов просматривать места встречаемости термина словаря и его контекст.
- Модуль тематизации – обеспечивает анализ текста в различных режимах: наполнение словаря (обучение), ведение статистики встречаемости терминов, классификация на основе статистики. Последовательный анализ текста в разных режимах позволяет поддерживать механизм расширения иерархии классов и «дообучения» словаря.
- Модуль выявления стоп-терминов – позволяет отделить шумовую или общеупотребительную лексику от предметно-зависимой.
- Модуль сборки словокомплексов – извлекает из текста словосочетания по фиксированному набору правил. Основной задачей модуля является выявление наиболее важных терминообразующих синтаксических групп, большинство из которых представляют собой именные группы либо строятся на их основе.

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 480

- Модуль морфологического анализа Lemmatizer, созданный компанией Диалинг (www.aot.ru), который, в частности, используется для предсказания морфологических признаков незнакомых слов.

Для обеспечения словарной обработки текста дополнительно могут использоваться следующие обработчики:

- Подсистема Alex осуществляет идентификацию и сборку лексических конструкций (текстовых фрагментов), на основе системы правил-шаблонов.
- Модуль исправления ошибок может осуществлять обработку незнакомых слов в режиме словарного анализа. Если слово не прошло словарный контроль, то запускается неточный словарный контроль, реализованный на основе метода хеширования по сигнатуре, предложенного в работе [6]. Модуль использует расширенные словари, содержащие универсальную лексику. Исправление ошибок особенно актуально при обработке неформальных документов: деловой переписке, коротких сообщений, транслитерированного текста.
- Модуль сегментации выполняет первичную и жанровую сегментацию. В процессе первичной сегментации осуществляется разбиение линейного представления текста на строковые объекты, оформленные как сегменты и упорядоченные в соответствии с порядком их встречаемости в тексте. Жанровая сегментация осуществляется после лексического анализа на основе полученных словарных объектов, маркирующих тот или иной жанровый сегмент.

4. Организация доступа к словарной информации

Стандартным способом организации доступа к словарной информации, в частности, к результату словарного анализа текста, является разработка соответствующего программного интерфейса (АПИ). Сам словарный компонент реализуется в виде подключаемой библиотеки, обеспечивающей все операции по работе со словарем, а словари хранятся во внешнем источнике и подгружаются библиотекой во время обработки текста. Отдельно реализуется пользовательский интерфейс (рабочее место) для создания словаря.

Указанный способ, несмотря на его эффективность, может оказаться неудобными по различным причинам, например, использование плохо совместимых сред или языков программирования.

Другим способом доступа является использование универсальных промежуточных форматов представления данных. Одним из самых популярных на сегодняшний день форматов, используемых в разных областях информатики, являются xml-схемы и xml-подобные форматы (rdf, rss, xhtml) [7].

В предлагаемой технологии применяется xml-представление данных, которое позволяет осуществлять следующее.

- Согласование данных. Можно указать семантическую эквивалентность словарных классов или атрибутов из разных словарей (которые могут формироваться с помощью разных программных компонент).
- Расширение содержательного наполнения словарей. Можно объединить элементы разных словарей в одну синонимичную группу (в дальнейшем планируется устанавливать произвольные связи между терминами).
- Более точно специфицировать используемые в словаре признаки для внешней программы или внутреннего обработчика, если необходимая возможность не поддерживается словарным компонентом. Например, в одном из компонентов система семантических классов и рубрик для классификации задаются идентичным образом. Для того, чтобы отличить рубрику от класса во внешнем файле, просто перечисляются все индексы классов.

Таким образом, используемое частичное xml-представление необходимо только для согласования словарей и использования результатов словарного анализа, что позволяет достаточно эффективно использовать данный механизм.

Заключение

На текущий момент одной из самых перспективных задач в данной области является создание комплексного инструмента, позволяющего комплектовать в единый программный продукт различные алгоритмы, реализующие требуемые разработчикам информационных систем этапы анализа текста. Важнейшим компонентом такого комплекса является словарная подсистема, предоставляющая средства для создания, наполнения и использования предметно-ориентированных словарей.

Словари, создаваемы с помощью предложенной словарной технологии, могут поддерживать основные этапы анализа текста: морфологический, синтаксический и семантический, а также классификацию на основе статистики.

074-SidorovaE:_Layout 1 13.05.2008 22:54 Page 481

Планируется проводить дальнейшие исследования проблем и потребностей, возникающих при многоцелевом использовании предметно-ориентированных словарей в практических информационных системах — системах документооборота и специализированных интернет-сервисах.

Список литературы

1. Ершов А.П. К методологии построения диалоговых систем: феномен деловой прозы // Избранные труды. Новосибирск: ВО “Наука”, 1994. С.314-330.

2. Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний // Труды международного семинара Диалог’2005 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2005. С. 65-70.

3. Большаков И.А. Какие словосочетания следует хранить в словарях? // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. Протвино: 2002. Т.2. С.61–69.

4. Жигалов В.А., Жигалов Д.В., Жуков А.А., Кононенко И.С., Соколова Е.Г., Толдова С.Ю.. Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог’2002 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2002. Т.2, С.192-208.

5. Кононенко И.С., Сидорова Е.А. Обработка делового письма в системе документооборота // Труды международного семинара Диалог’2002 по компьютерной лингвистике и ее приложениям. М.: Наука, 2002. Т.2. С.299–310.

6. Бойцов Л.М. Использование хеширования по сигнатуре для поиска по сходству // Прикладная математика и информатика. М.: ВМиК МГУ, 2001. № 8. С.135-154.

7. XML // Википедия. http://ru.wikipedia.org/wiki/XML