Е.А. Сидорова

ПОДХОД К ПОСТРОЕНИЮ ПРЕДМЕТНЫХ СЛОВАРЕЙ ПО КОРПУСУ ТЕКСТОВ[1]

Введение

До последнего времени огромные базы накопленной текстовой информации недостаточно использовались в задачах автоматической обработки текстов, связанных с извлечением содержательной информации на основе экспертных знаний. Хотя наличие даже неразмеченного корпуса текстов определенной тематики (например, деловая документация в системах документооборота) могло бы значительно упростить создание лингвистических ресурсов.

Используя классические методы обучения и статистического анализа встречаемости терминов в текстах, можно автоматизировать создание морфологических и синтаксических анализаторов [2] и предметных словарей [3]. Эти методы используют вручную размеченной корпус текстов. В свою очередь, существует и обратная связь – словари могут использоваться для

[1] Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149).

[2] *Андреев А.М., Березкин Д.В., Симаков К.В.* Обучение морфологического анализатора на большой электронной коллекции текстовых документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды седьмой всероссийской научной конференции (RCDL–2005). Ярославль, 2005. С. 173–181.

[3] *Лукашевич Н.В., Добров Б.В., Чуйко Д.С.* Отбор словосочетаний для словаря системы автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». М., 2008. С. 339–344.

разметки корпуса (морфологической, синтаксической, тематической)[1].

В данной работе описывается технология, предназначенная для автоматизированного создания специализированных словарей и параллельной тематической разметки используемого корпуса текстов.

1. Информационное наполнение предметного словаря

Большинство информационных систем ориентированы на небольшой класс задач и ограниченную предметную область. По этой причине при разработке рассматриваемой технологии были выдвинуты следующие требования:

1) поддержка автоматической наполняемости словарей на базе корпусов текстов;
2) возможность настраивать и приписывать предметные характеристики элементам словаря;
3) выполнение лексического анализа текста – сегментация и извлечение из текста заданных в словаре терминов и их характеристик;
4) накопление данных о статистико-комбинаторных свойствах лингвистических явлений;
5) наличие конкорданса.

В соответствии с этими требованиями была разработана следующая структура словаря. Лексическое наполнение разрабатываемых словарей включает наборы терминов следующего вида: лексемы, словокомплексы и лексические конструкции, описываемые шаблонами. Структура словарной статьи терминов содержит набор терминообразующих, статистических и семантических признаков. Также были разработаны механизмы, позволяющие

[1] *Сичинава Д.В.* К задаче создания корпусов русского языка. 2002. // URL: **http://rscorpora.narod.ru/article.html**

специалисту в значительной степени формировать структуру словарной статьи для терминов одного вида[1].

Для поддержки обучения и автоматической классификации текстов в словаре хранятся признаки, которые накапливают статистическую информацию о появлении терминов в обрабатываемых текстах. Для ведения статистики необходимо, чтобы пользователь задал систему связанных между собой тем (если система тем не задана, тогда статистика ведется по одной теме, отражающей всю выборку), а также наличие тематически-размеченного корпуса текстов.

Таким образом, в словаре для каждого термина хранится встречаемость в обучающей выборке, количество текстов выборки, в которых хотя бы один раз встретился термин, а также, встречаемость и количество текстов по каждой теме.

Ряд статистических параметров могут вычисляться динамически: частота встречаемости в выборке, а также частота и вес по каждой теме с учетом или без учета иерархии.

Темы связываются отношением наследования (включая множественное наследование) и образовывать иерархию, которая имеет следующие конструктивные ограничения:

- отсутствие циклов – не должно быть ситуации, когда тема является сама себе родителем;
- для любой темы в списке родителей не должно быть двух тем, которые были бы связаны между собой отношением наследования; наличие такой ситуации повлекло бы за собой семантическую непрозрачность системы описания мира.

Для каждой темы хранится следующая статистическая информация: количество текстов данной темы в обучающем корпусе, количество терминов во всех текстах данной темы.

---

[1] *Сидорова Е.А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008». М., 2008. С. 475–481.

2. Модули автоматизированной настройки словаря

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 работу со словарной информацией найденных в тексте терминов (означивание классов термина и его индивидуальных признаков).

Для создания словарей используются следующие модули.

- Модуль морфологического анализа компании Диалинг[1], который, в частности, используется для предсказания морфологических признаков незнакомых слов.
- Модуль сборки словокомплексов (СК) – 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.
- 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.

[1] URL: **www.aot.ru**

2.1. Общая схема обучения

Под обучением понимается процесс формирования словаря со статистическими показателями, т.е. словаря, элементам которого сопоставляется статистическое распределение по классам (темам). Обучение происходит на основе обучающего корпуса – массива текстов с исходной тематической разметкой – определенной экспертом темы (тем) каждого текста.

Можно выделить следующие этапы обучения.

1. Морфологический анализ текста и сборка СК.

2. Для каждого термина *x*, обнаруженного в тексте, отнесенного к теме *t*, и самой темы *t* корректируются их статистические показатели в словаре. (Если термина *x* в словаре не было, то он туда добавляется.)

3. В результате обработки всех текстов обучающей выборки для «значимого» словаря строится матрица, столбцы которой соответствуют классам, а строки – лексемам и СК. Ячейки этой матрицы на пересечении термина *x* и класса *t* будут отражать коэффициент вероятности отнесения текста, включающего термин *x*, к классу *t*.

4. Для каждого текста корпуса можно при помощи модуля классификации определить, к какой (каким) теме он относится.

5. Полученное достаточно грубое распределение весов анализируется лингвистом для уточнения и исключения случаев, ухудшающих распознавание. При этом используется конкорданс и выбираются случаи, получившие неправильную или недостаточно ясную классификацию.

2.2. Тематизация

Одним из недостатков автоматического обучения является то, что пользователь сразу должен задать иерархию тем, по которой размечается обучающая выборка. Однако на практике типичной является ситуация, когда требуется расширять и углублять существующую иерархию.

Поочередно используя механизмы классификации и «дообучения», можно дать пользователю возможность расширять иерархию тем. Средства, реализующие этот механизм, получили название *тематизация*.

На рис. 1. показана общая схема данного процесса.

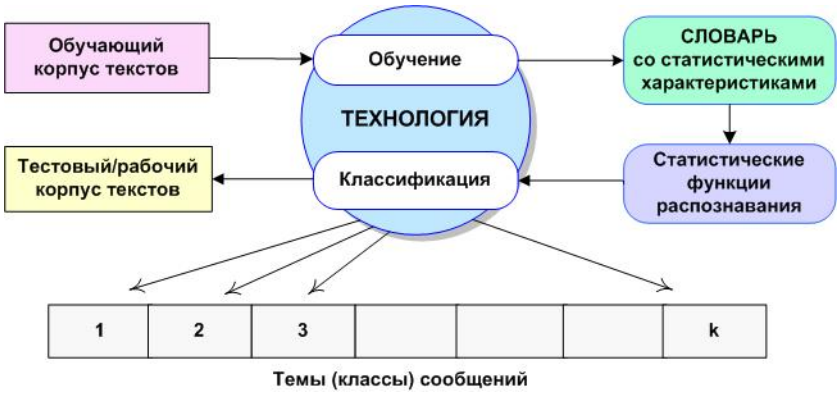

Рис. 1. Схема тематизации

Схема включает следующие компоненты:

- Обучающий корпус текстов – поток сообщений с исходной размеченной классификацией.
- Обучение – формирование статистических функций распознавания тем, соответствующих тексту сообщения.
- Словарь со статистическими показателями – словарь, терминам которого сопоставляется в режиме обучения статистическое распределение по темам.
- Тестовый/рабочий корпус текстов – поток сообщений без разметки, тексты которого классифицируются.
- Статистические функции распознавания – функции, позволяющие определить тему сообщения по распределению его лексики.

Рассмотрим ситуацию, когда создание словаря, иерархии тем и разметка выборки осуществляются «с нуля». Выделяются следующие этапы.

1. Анализ множества неразмеченных текстов и лексическое наполнение словаря.

2. Пользователь вручную просматривает словарь и выделяет набор ключевых терминов «маркирующих» новые темы.

3. Автоматическая «грубая» классификация неразмеченных текстов и анализ пользователем полученных результатов с целью доуточнения разметки текстов.

4. Обучение оставшейся части словаря.

Если результат шага 3 не устраивает пользователя, он может еще раз перейти ко 2-му этапу, дополнить словарь и заново осуществить разметку текстов, либо вернуться ко 2-му шагу после 4-го. Таким образом эксперт может использовать этот механизм до тех пор, пока результат его не удовлетворит.

Отметим, что данный механизм обеспечивается модулями обучения и классификации, но дополнительно должен отслеживать изменения статистики (чтобы не было ее дублирования при повторных обработках текстов), а также управлять режимами обучения (с/без накопления статистики, с/без добавления новых терминов).

2.3. Конкорданс

Конкорданс – традиционный способ изучения корпуса текстов. Он дает полный индекс терминов в ближайших и расширенных контекстах. Таким образом конкорданс осуществляет обратную связь словаря, словарных терминов с корпусом и обеспечивает своего рода лингвистическую разметку на морфологическом и поверхностно-синтаксическом уровне.

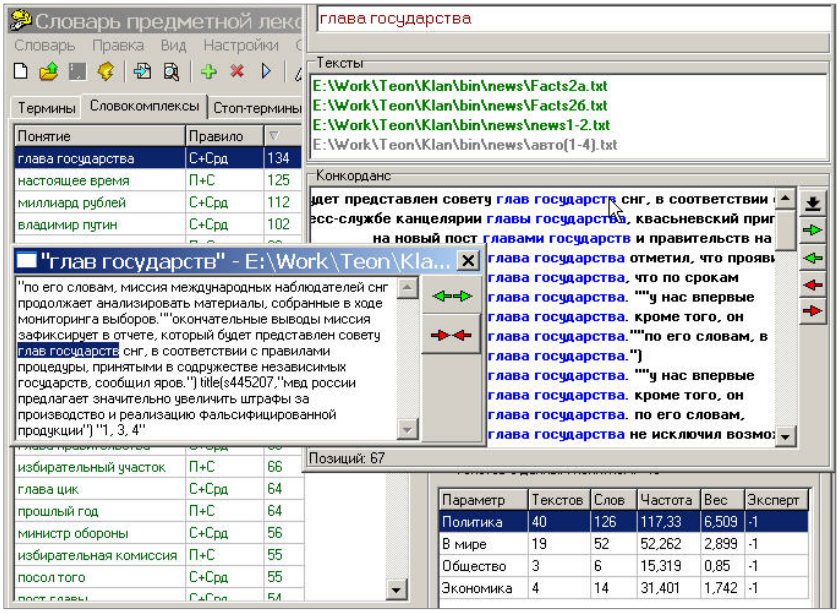

Рис. 2. Модуль конкорданса

Реализованный в системе модуль конкорданса (рис. 2.) работает с текстовыми файлами, каждый из которых может содержать несколько размеченных сообщений. При просмотре контекста вхождения термина пользователь может самостоятельно определять длину просматриваемого фрагмента текста (поддерживается пословное расширение контекста, просмотр абзаца или всего сообщения).

Заключение

Планируется проводить дальнейшие исследования проблем и потребностей, возникающих при многоцелевом использовании корпусов в практических информационных системах – системах документооборота и специализированных интернет-сервисах.