Формирование лингвистических баз знаний: статистика vs. грамматика

М. Г. Мальковский, С. А. Шевелев
Факультет вычислительной математики и кибернетики МГУ им. М.В. Ломоносова

Рассматривается задача построения лингвистических баз знаний для компьютерных систем обработки текста и звучащей речи. В силу очевидной трудоемкости создания подобных ресурсов вручную, не перестают быть актуальными разработки, связанные с автоматизацией этого процесса. Важным фактором, стимулирующим исследования, является современная информационная среда, динамичность которой требует постоянного обновления лингвистических ресурсов как на основе описаний языка, созданных лингвистами-профессионалами, так и на основе корпусов текстов. Обсуждаются вопросы автоматического и автоматизированного формирования лингвистических баз знаний, участия в этом процессе человека-эксперта. Анализируются современные решения, используемые в данной области, особенности входных текстов (наличие структурной разметки, гиперссылок). Серьезное внимание уделяется сопоставлению статистических и лингвистических подходов, приемам, позволяющим компенсировать недостаточную репрезентативность исследуемого корпуса текстов (обучающей выборки).

Введение

Лингвистические базы знаний (ЛБЗ) широко используются в большинстве современных систем, предназначенных для решения задач в области обработки естественного языка (ЕЯ), таких, как информационный поиск, распознавание речи и др. Следует отметить, что этот компонент является одной из важнейших составных частей системы обработки естественного языка (ЕЯ-системы), поскольку от его качества зачастую существенно зависит качество работы системы в целом.

Состав ЛБЗ может варьироваться в зависимости от предназначения системы, ее функций и характера решаемых задач. В качестве компонентов ЛБЗ выступают различного рода электронные словари, от широко распространенных словарей словоформ и частотных словарей до словарей синтаксических и семантических моделей управления, применяемых в приложениях, требующих глубокого лингвистического анализа (рис. 1).Обычно каждый вид словарей содержит описания характеристик определенного типа языковых явлений, которые традиционно разделяются на несколько уровней сложности.

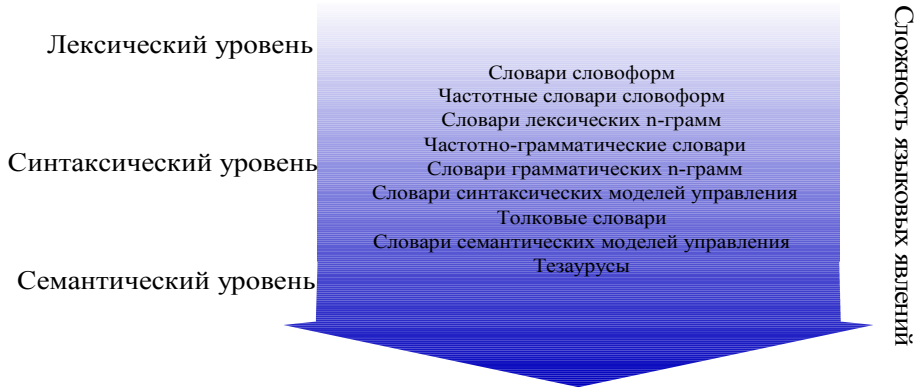

Рисунок 1. Соотношение уровней языковых феноменов и компонентов ЛБЗ

Следует отметить, что приведенное здесь деление на «уровни» весьма условно, т.к. во многих случаях одни и те же словарные компоненты могут быть квалифицированы с различных точек зрения как отражающие разные языковые явления. Например, словари лексических n-грамм содержат информацию о статистических характеристиках сочетаемости лексем, которые на самом деле отражают синтаксические закономерности. В качестве другого примера можно привести работу [2], в которой развивается идея о том, что толковый словарь может рассматриваться как источник частотно-грамматической информации.

Современные тенденции в области обработки естественного языка заключаются в стремлении найти альтернативу использовавшимся ранее простым методам анализа, которые, похоже, достигли своего естественного предела и не могут обеспечить требуемого многими современными приложениями качества ([5]). В связи с этим растет роль лингвистически содержательных методов обработки, а значит, возрастает и важность базы знаний в составе лингвистического компонента.

Процесс построения ЛБЗ включает в себя сбор и формализацию лингвистической информации, формирование представления собранных данных, пригодного для использования в программных системах. Разумеется, наиболее экономичным с точки зрения временных и людских ресурсов является полностью автоматическое построение ЛБЗ (и в настоящее время существует ряд методов, позволяющих избежать или свести к минимуму участие человека в этом процессе), однако во многих случаях сформированные таким образом ЛБЗ не обеспечивают желаемого уровня качества. С другой стороны, процесс создания этих ресурсов вручную требует огромного количества времени и, кроме того, зачастую является источником случайных ошибок.

Особенности современных ЛБЗ

Среди важнейших характеристик лингвистических баз знаний, оказывающих непосредственное влияние на сложность процесса их построения, можно выделить следующие:

- *полнота* лингвистической информации;
- *актуальность* данных;
- *точность* описаний.

Рассмотрим эти характеристики, их влияние на особенности формирования ЛБЗ и применяемые методы, более подробно.

Для достаточно сложных предметных областей требование *полноты* обычно приводит к большим объемам хранимой лингвистической информации, необходимой для покрытия соответствующего подъязыка. Это очевидным образом свидетельствует о необходимости

автоматизации процесса построения ЛБЗ, который в противном случае может привести к большому количеству ошибок при ручном вводе информации, требуя при этом значительных усилий и временных затрат.
Требование *актуальности* данных становится особенно важным в условиях современной информационной среды, характеризующейся

- огромными (и, тем не менее, постоянно увеличивающимися) объемами электронной информации,
- высокой динамичностью ее обновления,
- большим количеством разнородных сообществ пользователей, объединенных глобальной компьютерной сетью.

Эти условия не могут не оказывать влияния на естественный язык, его лексический состав, распространенность тех или иных терминов и синтактико-семантических конструкций. (См., например, работу [3], в которой обсуждаются некоторые аспекты этой проблемы, связанные с развитием русскоязычного сегмента сети Интернет.) Таким образом, требование актуальности данных делает необходимой возможность автоматического или автоматизированного пополнения ЛБЗ в процессе функционирования системы.
В то время как для некоторых приложений не требуется высокая *точность* представленных в ЛБЗ данных, для многих ЕЯ-систем вопросы качества лингвистической информации являются гораздо более важными. Требуемая точность получаемых описаний накладывает ограничения как на использование *ручного* кодирования (следствием которого при условии больших объемов данных, как уже отмечалось, могут являться многочисленные случайные ошибки), так и на методы, применяемые при *автоматическом* формировании ЛБЗ.

Некоторые подходы к формированию ЛБЗ

Методы, применяемые при построении компонентов ЛБЗ, можно разделить на несколько групп:

- статистические;
- использующие эвристики, специфичные для конкретных предметных областей;
- использующие явную (созданную автором документа) структуру анализируемой информации (задаваемую, например, при помощи гипертекстовой разметки);
- лингвистические (грамматические).

Разумеется, достаточно часто используются различные комбинации перечисленных выше методов.
Статистические методы обычно подразумевают анализ специально подобранных текстов (например, из некоторой предметной области) с целью выявления статистических характеристик определенных языковых явлений (например, частотности отдельных слов, цепочек слов и их взаимной встречаемости). Эти методы используются в основном при формировании частотных компонентов ЛБЗ, таких как частотные словари словоформ или словари лексических n-грамм, однако существуют также подходы к использованию статистических методов для формирования более сложных компонентов, таких как тезаурус (при серьезном ограничении на подъязык и область применения полученного компонента). Достоинствами статистических моделей являются их простота и эффективность реализации. Среди недостатков наиболее существенным является не слишком высокий уровень качества получаемых словарей. Основные причины указанного недостатка связаны с трудностью подбора входных текстов таким образом, чтобы они покрывали всю исследуемую область и отражали весь спектр языковых явлений. Использование в качестве исходного материала общедоступных корпусов текстов (таких как Penn Treebank, Brown corpus) также не решает проблемы, т.к. многие из существующих корпусов устарели и не являются по современным

меркам достаточно представительными; кроме того, они зачастую не отражают существующих в настоящее время языковых реалий.

Методы, которые используют различные ad hoc *эвристики*, отражающие особенности, специфичные для конкретных предметных областей и источников данных, обычно применяются в совокупности с другими методами в целях улучшения качества обработки. Подобные методы могут применяться, например, в системах фильтрации спама как средство борьбы с «новыми технологиями» замусоривания текста, внедряемыми спамерами с целью снижения эффективности фильтрующих машин ([6]). Приемы аналогичные спамерским часто используются авторами сайтов, содержание которых связано с незаконной деятельностью либо нарушает общественные, религиозные и другие нормы. Естественно, что методы борьбы с ними могут быть аналогичны тем, что используются современными антиспамерскими системами.

Методы, использующие *структурную разметку* обрабатываемого текста, можно в свою очередь разделить на две подкатегории. Во-первых, это методы, которые следует, вероятно, отнести к одной из разновидностей эвристических подходов, позволяющих, например, учитывать различное расположение терминов в анализируемом тексте (в заголовке, в первом абзаце и т.д.), приписывая им затем различную значимость при обработке. Во-вторых, это довольно распространенные в последнее время методы, использующие обработку структурной разметки как основное средство, применяемое при анализе. Примером такого подхода может служить описанный в [4] способ автоматического формирования информационно-поискового тезауруса на основе анализа разметки и гиперссылочной структуры веб-сайтов. В основе предлагаемого в этой работе подхода лежит идея о том, что гипертекстовые ссылки можно рассматривать как дуги семантической сети, формируемой вебмастером при создании сайта. Эта семантическая сеть и служит затем основой для формирования тезауруса. Основные проблемы, с которыми сталкиваются подобные методы, – это отсутствие унифицированной структуры документов, необходимость отличать навигационные и рекламные гиперссылки (не имеющие значения при формировании тезауруса) от ссылок, несущих семантическую нагрузку.

Лингвистические (грамматические) методы являются, пожалуй, наиболее универсальными и могут использоваться при формировании словарных компонентов практически любого уровня. Однако качество существующих методов обработки естественного языка не позволяет использовать их в полностью автоматических процессах (по крайней мере, для случая предметных областей реального масштаба сложности), делая участие человека-эксперта непременным условием. Кроме того, применение лингвистических методов обычно требует наличия вспомогательных компонентов ЛБЗ (например, тезауруса, - см. ниже), необходимых для их успешного функционирования.

Одним из довольно эффективных приемов при формировании ЛБЗ является лингвистическое пополнение, или лингвистическое «обогащение». В основе этого метода лежат следующие типы преобразований исходных данных:

1. Поверхностные грамматические трансформации.

Этот тип преобразования позволяет, например, по имеющимся формам слов сгенерировать другие формы, допустимые в данном контексте с точки зрения синтаксиса. Также это преобразование может учитывать различные (правильные) варианты написания слов и порядка слов в словосочетаниях. Примеры:

web service ⇔ web services
online casino ⇔ casino online
естественноязыковой ⇔ естественно-языковой
e-mail ⇔ email

Применение данного преобразования в случае высокофлективных языков, каким является, в частности, русский язык, может привести к порождению слишком большого числа допустимых слов и словосочетаний. Тем не менее, в ряде случаев подобные трансформации являются оправданными (например, они могут избавить от необходимости морфологического анализа в реальном времени, заменяя его поиском в ЛБЗ, что в некоторых ситуациях может оказаться более эффективным).

2. Глубинные трансформации (на основе связей, задаваемых лексическими функциями [7]).

Набор используемых преобразований может зависеть от конкретных приложений и предметных областей. Среди относительно универсальных преобразований можно указать, например, следующие:

$X \Leftrightarrow Gener(X) \xrightarrow{2} S_0(X)$, например, *графика – искусство графики*

$X \Leftrightarrow S_0(X)$, например, *gambling – gambler*.

3. Использование простейших тезаурусов[1].

Этот метод подразумевает преобразования на основе отношений, представленных в общих или специфичных для конкретной предметной области тезаурусах. В простейших случаях тезаурусы могут описывать синонимические и гипонимические отношения. Примеры:

лингвистика → наука

почтовая рассылка → e-mail рассылка

Указанные методы позволяют по существующей в ЛБЗ информации генерировать новые (отсутствующие в ЛБЗ, но допустимые, по крайней мере, с некоторой вероятностью) данные. В частности, подобные подходы дают ощутимые результаты в совокупности с применением статистических методов. Так, в работе [1] описывается способ коррекции словаря биграмм (традиционно формируемого при помощи статистического анализа) с использованием некоторых из указанных методов «грамматического обогащения» словаря. В работе упоминается опыт успешного использования этого приема для повышения качества обработки в системе распознавания устной речи (точность распознавания была увеличена на 1-2% при исходной точности ≈95%), однако этот и подобные приемы были с успехом использованы авторами и в других задачах, для которых качество, обеспечиваемое чисто вероятностными методами являлось неудовлетворительным. Примером подобной задачи является задача формирования характеристического атрибутного множества категории документов для поддержки работы системы автоматической фильтрации текстовой информации. Подобные ресурсы очень часто формируются при помощи статистических методов, для успешной работы которых требуется непомерно большой объем обучающей выборки (см., например, [8], где для настройки системы автоматической фильтрации требовалось около 10000 заранее отобранных сайтов).

Применение лингвистических методов совместно со статистическими позволяет компенсировать недостаточную репрезентативность исследуемого корпуса текста или обучающей выборки и тем самым устранить или, по крайней мере, уменьшить влияние одного из основных недостатков, присущих статистическим методам, сохранив при этом их преимущества (эффективность и простоту реализации), являющиеся весьма важными, в частности, для приложений, выполняющих обработку в реальном времени («на лету»). Опыт работы нашей исследовательской группы показывает, что при формировании аналогичных

---

[1] Следует заметить, что преобразования на основе «тезаурусных» отношений фактически являются частным случаем глубинных преобразований (например, задаваемых лексическими функциями Syn и Gener). Тем не менее, этот частный случай заслуживает особого внимания в силу его относительной простоты (по сравнению с общим методом), а также популярности использования тезаурусов для решения других задач.

ресурсов сочетание статистических подходов с лингвистическим обогащением (под контролем эксперта) позволяет существенно сократить объем обучающей выборки.

Заключение

Выбор одного из представленных здесь методов или их комбинации зависит от множества взаимосвязанных факторов, таких как

- Тип формируемого лингвистического ресурса.

Для задач формирования ЛБЗ, отражающих *сложные* языковые явления, *простые* методы могут приводить к удовлетворительным результатам только при наличии существенных ограничений на ПО и способы использования формируемого ресурса.

- Требуемая точность обработки.

Требования, предъявляемые к качеству и точности, влияют не только на выбор методов обработки, но и на необходимость участия в процессе человека.

- Возможность привлечения экспертов.

Участие человека позволяет в значительной степени повысить качество формируемых лингвистических ресурсов. Вместе с тем желательно избавить эксперта от необходимости выполнения рутинных операций, являющихся источником случайных ошибок. Наиболее адекватным в данной ситуации представляется использование специальной инструментальной среды для поддержки работы эксперта, целями которой являются автоматизация рутинных задач, представление результатов обработки в удобной форме, обеспечение возможности контроля и корректировки полученной информации.

- Ограничения на предметную область и подъязык.

Наличие подобных ограничений позволяет успешно применять как лингвистические (грамматические), так и специальные эвристические подходы, а также влияет на качество работы статистических методов.

- Временные ограничения.

- Вопросы эффективности и простоты реализации.

Несмотря на постоянно растущие мощности современных вычислительных систем, эти вопросы очень важны для некоторых классов приложений, выполняющих обработку «на лету».

Наличие большого количества конкурирующих требований означает, что ни один из описанных методов не имеет безусловных преимуществ перед другими, и для достижения разумного компромисса необходимо комбинировать различные подходы. Кроме того, нельзя пренебрегать участием в проекте человека-эксперта.

Литература

1. Мальковский М.Г., Абрамов В.Г., Субботин А.В. Об автоматизированном формировании лингвистических баз знаний // Труды Международного семинара по компьютерной лингвистике и ее приложениям. Т.2 – Казань,1998. – с. 831-836
2. Coughlin, D. A. Deriving Part-of-Speech Probabilities from a Machine-Readable Dictionary. // In Proceedings of the Second International Conference on New Methods in Natural Language Processing, Ankara, Turkey, 1996. ftp://ftp.research.microsoft.com/pub/tr/tr-96-14.ps
3. Беликов В.И. Интернет и орфография. // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог-2003 (Протвино, 11-16 июня 2003). – М.: Наука, 2003.

4. Zheng Chen, Shengping Liu, Liu Wenyin, GeGuang Pu, Wei-Ying Ma. Building a Web Thesaurus from Web Link Structure. // SIGIR2003: pp. 48-55. http://research.microsoft.com/research/pubs/view.aspx?msr_tr_id=MSR-TR-2003-10
5. Kennet W. Church, Lisa F. Rau. Commercial Applications of Natural Language Processing // Communications of the ACM. November 1995/Vol. 38, No.11: pp. 71-79 .
6. Ашманов И., Власова А., Зоркий К., Калинин А., Кошкин С., Тутубалин А. Спам: итоги 2003 года. Аналитический отчет. ЗАО «Ашманов и партнеры», Москва, 2004.
7. Апресян Ю.Д. Избранные труды, т. I. Лексическая семантика: 2-е изд. // М.: Школа «Языки русской культуры», 1995.
8. Russell-Falla, et al. Method for scanning, analyzing and rating digital information content. // United States Patent #6,266,664, 2001.