Статистические языковые методы. Коллокации и коллигации

Кочеткова Н.А.
МИЭМ НИУ ВШЭ, ФИТиВТ

В последнее время наблюдается накопление массивов специализированных текстовых документов. В ходе своего существования предприятия формируют архивы документации колоссального объема. Эти данные требуют не только хранения, но и соответствующей обработки. А так как значительная часть документов представляет собой текстовое описание то, для выполнения этих задач требуется использование методов автоматической обработки текстов на естественном языке.

Анализ текстов на естественном языке

Этапы анализа текста на естественном языке практически не зависят от выбранного языка.

Обычно различают следующие этапы анализа текста :

1. графематический анализ — выделение структурных единиц из входного текста;
2. морфологический анализ — определение морфологических характеристик каждого слова — часть речи, падеж, склонение, спряжение и т.д.;
3. синтаксический анализ — построение синтаксического представления предложения;
4. семантический анализ — построение аргументно-предикатной структуры высказываний или другого вид семантического представления предложения.

На каждом из этих этапов возникают различного рода неоднозначности, и из-за невозможности кодирования всех явных и неявных знаний о языке, для решения многих задач обработки текста используются статистические языковые методы.

Статистические языковые методы

Статистические языковые методы рассматривают естественный язык как случайный процесс, что позволяет переопределить многие задачи, связанные с обработкой естественного языка, в более строгом математическом смысле.

Для формального описания статистической модели языка используется модель скрытых Марковских цепей. Самая распространенная версия этой модели - модель n-грамм, соответствующая скрытой Марковской цепи n-1 порядка, где вероятность P(w) появления какого-то символа последовательности $w=w_1,w_2,..w_t$ длины T - это первое разложение в соответствии с теоремой Байеса.

Определенная таким образом модель n-грамм позволяет предсказать появление символов в некотором ограниченном наборе, основываясь на контексте из n-1 предшествующего элемента. С увеличением длины контекста возникают серьезные трудности при вычислениях такой модели, и применение ее на практике требует огромных затрат памяти.

Коллокации

Коллокации - это частный случай модели n-грамм. Коллокации – это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости.

Так как основное внимание уделяется частоте совместной встречаемости, то коллокации в корпусной лингвистике могут быть определены как статистически устойчивые словосочетания. За последние годы появилось большое число

исследований и разработок, посвященных коллокациям, затрагивающих как теоретические аспекты статистического подхода к данному понятию, так и практические методы выявления и применения коллокаций.

Извлечение коллокаций — довольно простой процесс – необходимо лишь разбить весь текст на последовательности заданной длины и подсчитать частоту каждой уникальной последовательности. Коллокации нашли широкое применение в системах машинного перевода, а так же часто используются для снятия неоднозначностей и поиска ошибок. И хотя точность полученных таким образом инструментов не велика, они существенно сокращают объем ручной обработки.

Также коллокации достаточно успешно применяются в таких задачах, как выделение терминологии и семантических предпочтений. Однако стоит отметить, что в большинстве работ, помимо самого словаря коллокаций вводятся разного рода дополнения: от фильтрации по словарям до фильтрации по частеречным и синтаксическим шаблонам. Естественно любые привнесенные лингвистические правила существенно улучшают качество получаемых инструментов.

Также существует множество «мер устойчивости», которые позволяют выделять те или иные «интересные» коллокации. К наиболее распространенным относятся MI и T-score. MI — позволяет выделять наиболее редкие и своеобразные коллокации и подходит для выделения терминологии, имен собственных и прочих конструкций, в которых частота составляющих коллокацию слов в тексте вне этой коллокации ничтожно мала. T-score — напротив позволяет найти наиболее распространенные частые обороты. Эти (и другие) меры сводятся к простым преобразованиям частотных характеристик коллокации и каждого из коллокатов, и дают существенно меньшую точность выделения, чем введение фильтров, основанных на лингвистических правилах.

Кроме того, стоит отметить, что хотя речь чаще всего идет о коллокациях вообще, на практике обычно используют биграммы, гораздо реже триграммы, а работ по четырехграммам и более применительно к задачам автоматизации нет. Так для четырехграмм до сих пор нет ни одной общепринятой меры устойчивости.

Словари коллокаций существенно зависят от корпуса, из которого они извлечены, и потому трудно сравнимы между собой. Еще одна трудность – пороговые значения, которые зависят от размера корпуса. Чем выше значение порога, тем больше редких коллокаций не входят в словарь, и главным становится вопрос: «как не потерять верные, но редкие события?» Коллокации бывают как контактными — когда слова идут в тексте одно за другим, а бывают и неконтактными — когда из последовательности выбираются скажем 1 и 3 слово, то есть некоторое количество слов последовательности может пропускаться. Однако извлечение неконтактных коллокаций требует больше затрат, в связи с чем неконтактные коллокации используются гораздо реже.

Коллигации

Еще одной статистической моделью являются коллигации. Если коллокации дают представление о поведении слов и словоформ, то коллигации — о поведении грамматических групп. В отличие от чисто лексического характера коллокаций, коллигации имеют смешанный лексико-грамматический характер. Коллигация - это комбинация лексических и грамматических характеристик, имеющих тенденцию к совместной встречаемости. Зачастую в качестве таких характеристик выбирают части речи или более фразовые элементы. Обязательным условием коллигации является то,

что в качестве одного из коллигатов должна выступать словоформа или лемма. Коллигации являются очень удобным промежуточным звеном между фразовыми шаблонами и коллокациями, представляя собой смесь лексических единиц, меток частей речи и фразовых категорий.

Очевидно, что извлечение коллигаций гораздо сложнее извлечения коллокаций. Сначала необходимо разметить части речи, на основе этого выделить фразовые элементы, выбрать из всех вариантов разбора корректные, и установить предпочтительность лексемы или параметра. Никакого систематического описания и массового применения коллигации еще не получили. Наиболее распространены словари коллигаций, в которых первый коллигат представлен словоформой, а второй коллигат - всеми возможными грамматическими параметрами слова, следующего за первым, со статистикой.

Основными недостатками статической модели коллокаций/коллигаций можно считать:

а) заведомо неверное предположение о независимости вероятности появления очередного слова/параметра от более длинной истории

и

б) колоссальные объёмы требующихся обучающих данных.

Однако, как словари коллокаций, так и словари коллигаций, а также словари частот могут быть крайне полезны в сочетании с другими методами, так как они дают общее представление о поведении слов в тексте.

Рассмотрим подробнее применение словарей коллокаций на разных этапах анализа текстов.

Статистический синтаксический анализ

В работе [5] описан синтаксический анализ на основе словаря n-грамм с использованием грамматики основанной на зависимостях. Вместо обобщения синтаксических правил для сегментации последовательности слов на две клаузы используется информация о частях речи, получаемая из n-грамм. Вводится специальная часть речи - сегментационное слово, которое соответствует начальному или конечному символу клаузы. Это позволяет выделять структуру клаузы, и впоследствии сделать процесс разделения автоматическим. Сегментационные слова для каждой клаузы извлекались из небольшого размеченного вручную корпуса. Экспериментальные результаты для предложений на японском языке показали, что синтаксический анализатор использующий описанные выше дополнения достигает 72,2% полноты, что является примерно тем же уровнем производительности, что и у синтаксического анализатора на основе вероятностных контекстно-свободных грамматик с введенными человеком лингвистическими правилами.

Статистические методы в семантическом анализе

Неоднозначность смысла слова распространена во всех естественных языках. Правильный смысл многозначного слова может быть выбран на основе контекста, в котором оно встречается, и, соответственно, проблема разрешения семантической неоднозначности слова может быть определена как задача автоматического назначения многозначному слову наиболее подходящего смысла исходя из контекста. Одним из наиболее частых статистических способов выделения семантической связности слов являются коллокации.

Так в работе [2] используются две идеи:

- слова имеют характерные семантические профили - просодии
- сила связи между словами может быть измерена количественно.

Объединив эти две идеи можно получить сравнительные семантические профили слов, которые показывают частоту и характерные коллокаты заданного слова, и, следовательно, выявляют семантические отношения между коллокатами.

На данных корпуса размером в 120 млн. слов было показано, что лемма ПРИЧИНА (CAUSE) встречается преимущественно в «неприятных» коллокациях, таких как *причина проблемы (cause trouble)* и *причина смерти (cause death).* Подробное изучение этой леммы используется для пояснения количественных методов исследования коллокаций. Было проведено краткое сравнение семантических особенностей для связанных лемм, например, ПРИЧИНА (REASON) и СЛЕДСТВИЕ (CONSEQUENCE). И результаты показали такие отношения между леммами и семантическими категориями, которые в настоящее время не захвачены ни одним из словарей или грамматик. Использование извлеченных из корпуса коллокаций гарантирует надежность получаемых семантических профилей, из-за повторяемости их на миллионах слов.

Аналогичный метод использования коллокаций для семантических целей представлен в статье [1].

Принцип семантического предпочтения зависит от отношений между набором часто встречающихся коллокатов и некоторыми общими семантическими особенностями.

Эта статья представляет описание семантических профилей, созданных на основе этого принципа с использованием метода, основанного на корпусной лингвистике, включающего в себя четыре этапа: вычисление, сжатие данных, семантический анализ и интерпретация.

Полученные в результате профили и формат их представления делают их легко поддающимися использованию переводчиками. Во-первых, они практически обоснованы и являются представителями того корпуса, из которого они получены. Во-вторых, характер значения, который можно найти в таких профилях, не ограничен лексическим или словарным значением, но включает в себя прагматическую или энциклопедическую информацию, а также ту, которая не могла бы быть предсказана из словарного определения.

Кроме того, можно показать, что частота, с которой некоторые значения многозначных слов, найденных в корпусе, предлагает способ, позволяющий сделать выводы о типе текста или жанре. Проверка числа этих профилей подтверждает предположение Стаббса (2001) о том, что такие профили могут быть использованы в качестве доказательств семантических полей.

Семантические профили, таким образом, могут предложить аккуратное и доступное средство для переводчиков, или, если они применяются в параллельных корпусах то, как часть базы данных для автоматического перевода.

Заключение

До сих пор не существует методов, позволяющих полностью снимать разного рода неоднозначности анализа. Так на уровне слов существует омонимия - ситуация, когда одной словоформе можно приписать несколько нормальных форм. При синтаксическом разборе оказывается, что связи между словами можно назначить более чем одним способом. Может оказаться слишком много вариантов соединения слов, перебор которых займет катастрофически много времени и сложность задачи

будет расти экспоненциально с ростом количества слов. Проблема с семантической неоднозначностью стоит еще острее. Большинство методов опираются лишь на словарные статьи. И хотя за последние десять лет, известность статистических приложений по обработке естественного языка, которые используют статистические, а не только основанные на правилах анализаторы, увеличилась очень значительно поиски более лучших решений продолжаются. Корпусный подход и мощности современной вычислительной техники дают уверенность, что лексические описания в будущем будут предоставлять более точную и исчерпывающую информацию о словах, и дадут доступ к неявным, скрытым моделям языка.

Список литературы

1. Allan Lauder Collocation, Semantic Preference and Translation: Semantic Preference as a Reference Source for Translation, in Proceedings of the The Regional Conference of the International Association of Forensic Linguists 2010, 2010.

2. Michael Stubbs Collocations and semantic profiles: on the cause of the trouble with quantitative studies, in Functions of Language, 2, 1, 1995.

3. Serge Sharoff, Joakim Nivre, The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge. in Proceedings of the Dialog 2011, 2011.

4. Yue Zhang and Joakim Nivre. Transition-based Dependency Parsing with Rich Non-local Features. Proceedings of ACL 2011, 2011.

5. Nobuo Inui; Yoshiyuki Kotani. Robust N-gram Based Syntactic Analysis Using Segmentation Words. Proceedings of the 15th Pacific Asia Conference on Language, Information and Computation, 2001.

6. Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие, М.: МИЭМ, 2011.