Методы улучшения качества синтаксического анализа фраз естественного языка.

Н.В.Одинцев

МГУ

nickone@nvkvist.ru

В настоящее время на кафедре алгоритмических языков разрабатывается лингвистический процессор текстов на русском языке. Сейчас он включает в себя базу данных с лингвистической информацией и морфологический и синтаксический анализаторы. Семантический и прагматический анализаторы, как и некоторые другие компоненты, планируется разрабатывать в дальнейшем. Данная статья посвящена синтаксическому анализатору и перспективным идеям его развития.

Часть средств, используемых для синтаксического анализатора, разработана ранее и описана в работах [1] и [2]. Это прежде всего способ представления грамматики с помощью расширенной сети переходов, а также синтаксис и семантика описания грамматики и операторов для проверки контекстных условий. Кроме того, широко используются модели управления, описанные в [3], но в более общем виде. Помимо обобщенных моделей управления, были продуманы и применяются новые подходы к улучшению качества синтаксического анализа с помощью ранжирования вариантов анализа и/или выделения наиболее вероятного из них с учетом специфики анализируемого текста. Для этого были введены понятия *контекста* и *вероятности* и адаптированы модели управления. Кроме того, система работает не с экспериментальным, а с реальным словарем, содержащим около 100 тыс. основ и неизменяемых слов.

В целом схема анализа достаточно традиционна. На вход системе поступает текст, далее запускается предсинтаксический анализатор, разбивающий этот текст на отдельные предложения. Для каждой словоформы в предложении морфологический анализатор, использующий базу данных, определяет значения грамматических характеристик. На основе этой информации осуществляется синтаксический анализ.

Обобщенные модели управления.

В работе [3] были описаны *модели управления* глаголов, являющиеся формализованной записью ограничений на зависимые от глаголов слова. Они используются для указания прямых и непрямых дополнений и обстоятельств и возможностей их совместного использования. В этом направлении можно пойти дальше: практически у всех слов могут быть зависимые от них другие слова, на которые могут быть наложены ограничения, связанные с конкретным словом (а не только с классом слов), т. е. правила сочетаемости отдельных слов и синтаксических групп.

С точки зрения синтаксиса описания моделей управления акцент делается не на управляющие слова, а на управляемые. В частности, модели управления глаголов в [3] фактически являются моделями управления именными и предложными группами. Такой способ описания управления можно обобщить и на другие группы. Для этого в модели управления слова надо указать тип "управляемой" группы (или слова) и требуемое значение ее (его) грамматических переменных.

Использование базы данных в качестве хранилища лингвистической информации накладывает отпечаток на представление моделей управления. Само по себе это не сказывается на быстродействии системы: в конечном счете всю информацию в начале работы можно перенести из базы данных в оперативную память и не тратить время на ее извлечение в процессе работы. Но при дополнении и изменении лингвистической информации (а это может быть долгим и итерационным процессом) важно легко получать доступ к разным атрибутам слов и их моделей управления и формировать для них удобное представление.

Предлагается новый способ организации хранения информации о моделях управления в базе данных.

Отдельно хранятся *атомарные* модели управления. Таблица, где они хранятся, содержит идентификатор зависимой лингвистической группы и набор допустимых значений ее грамматических переменных (например, для глаголов это могут быть именные группы и конкретные значения рода, числа и падежа).

Другая таблица представляет описание структуры дерева моделей управления, где каждая вершина представляет собой логическую связку двух поддеревьев. Таким образом, каждая запись содержит логическая связку (И или ИЛИ) и идентификаторы корневых вершин левого и правого поддеревьев.

В основной таблице базы данных, содержащей словарные статьи, находятся ссылки на корневые вершины деревьев моделей управления слов; каждому слову соответствует одна ссылка на корневую вершину его дерева моделей управления.

В нынешнем описании моделей управления разделяются ограничения на слова в постпозиции и в препозиции (в отличие от [3]); например, они могут отличаться*вероятностями* их использования. Так, в словосочетаниях "абсолютно однозначно" и "однозначно абсолютно" с главным словом "однозначно" используется одна и та же модель управления, но в первый вариант намного меньше "режет слух". Для ранжирования этого ощущения можно использовать вероятности, приписанные атомарным моделям управления, представляющие собой числа. Таким образом, для "абсолютно однозначно" и "однозначно абсолютно" будут использованы одинаковые атомарные модели управления, но им приписаны разные вероятности. Без введения вероятностей приходилось бы заранее определять, допустимо ли словосочетание "однозначно абсолютно" или нет. При его отбрасывании сужается сфера применимости анализатора, при допущении будут неизбежно появляться лишние дополнительные варианты. А при введении вероятностей

модели управления "однозначно абсолютно" можно приписать очень маленькую вероятность, и анализатор выдаст такой способ разбора, если не будет ничего лучше.

Для сопоставления атомарным моделям из дерева моделей управления их вероятностей для конкретного управляющего слова служит специальная таблица. Она содержит начальную форму управляющего слова, признак препозиции или постпозиции, номер атомарной модели управления и собственно вероятность.

Заметим, что одну и ту же модель управления можно указывать для целой группы слов, например, принадлежащих одному классу, в качестве базовой модели управления. Такой подход уменьшит число записей в таблицах и упростит ввод информации.

Подчеркну одну из причин усиления роли моделей управления. Многие из ограничений, которые накладывают модели управления, можно реализовать специальными операторами, и работать такая система будет не хуже и с точки зрения времени, и с точки зрения адекватности результата. Но ее модификация будет сопряжена со значительными трудностями из-за неизбежной разбросанности этих ограничений и из-за весьма вероятного возникновения побочных последствий при изменении операторов. Взаимосвязь операторов может быть очень сложной, и всю их совокупность придется пересматривать при изменении одного оператора. Перенесение же части ограничений на модели управления позволит менять эти ограничения, не слишком заботясь о том, будет ли система функционировать вообще после такой модификации.

Контексты.

Важной частью описываемого лингвистического анализатора является т. н. контексты. Как известно, человек практически всегда выбирает один из возможных смыслов фразы, хотя их может быть много больше. Человек как бы "настраивается" на стиль изложения, на привычные в рамках читаемого текста грамматические конструкции, и очень часто подобная "настройка" помогает ему не размышлять помногу над каждым предложением. Фактически человек выступает в роли узкоспециализированного семантико-синтаксического анализатора, который, наткнувшись, например, на нетипичную грамматическую конструкцию, вынужден будет "подумать". Так, когда среди текста о работе на приусадебном участке будет встречено "огороды русские под холмом седым", то небольшое замешательство будет вызвано, в частности, и постпозицией прилагательных.

Учитывая такие особенности обработки предложений, можно ввести понятие контекста. Под контекстом здесь понимается категория текстов, для которой является специфичным употребление определенных слов и грамматических конструкций. Выше уже приводился пример, где использовалась строка из стихотворения. Но и для обычных прозаических текстов введение категорий может быть оправдано. Например, конструкцию предложения "Моросит", наверно, нельзя встретить в статье из словаря или энциклопедии.

Можно подчеркнуть разницу между предметной областью и контекстом. Возможно, научные статьи по химии ферментов напоминают по типам наиболее часто используемых грамматических конструкций кулинарные рецепты, но в то же время весьма далеки от, например, "Комментария к роману А. С. Пушкина "Евгений Онегин" Набокова, который можно считать статьей по литературоведению. Таким образом, кулинария и химия могут попасть в один контекст, а "Комментарий", будучи научной работой, как и статьи по химии, все-таки в другой. Таким образом, контекст применяется по отношению к синтаксису, а не к семантике.

При разборе предложения можно задавать его контекст, и это позволит отбрасывать маловероятные конструкции и тем самым значительно сокращать время анализа. В случае, если не получится удовлетворительно разобрать предложение в рамках данного контекста, то можно попробовать изменить контекст. Если обрабатывается связный текст, то можно осуществить настройку контекста по первым предложениям. Вероятно, чтобы реализовать эту возможность, потребуется широко использовать различные эвристики.

Понятно, что разделение контекстов является сложной задачей. В то же время ясно, что для их совокупности линейная организация не является самой подходящей: есть более близкие между собой контексты, есть более далекие, есть контексты весьма специальные, вплоть до уровня "язык Д. С. Лихачева в статьях о произведениях древнерусской литературы", и общие, например, "научно-технические тексты". Таким образом, естественной структурой организации контекстов будет дерево. На рисунке показана часть возможного дерева контекстов. В корне этого дерева будет некий обобщенный контекст, в рамках которого можно разобрать любое предложение. Работа системы при указании этого контекста по существу не будет отличаться от любой универсальной системы такого рода. Будет выдано большое количество вариантов, о многих из которых автор предложения даже и не подозревает, при этом время анализа будет значительным.

Рис. Абстрактный пример части дерева контекстов.

При указании контекста более низкого уровня повышается точность и снижается время, но этому препятствуют две проблемы. Во-первых, это уже упомянутая необходимость предварительно его указывать (и чем дальше от корня в дереве он находится, тем труднее его определить); особенно это трудно при автоматическом определении контекста. Во-вторых, это необходимая кропотливая предварительная работа по описанию контекста. Так, если опуститься до уровня отдельных людей, то, как минимум надо хорошо разбираться в характерных синтаксических особенностях их текстов. Так что уровень детализации контекстов должен быть умеренным. Заметим, что древовидная структура позволяет легко повысить уровень детализации (например, если в ходе работы определится, что фантастика - это слишком грубая категория, к ней можно добавить дочерние вершины - научная фантастика и фэнтэзи, где будет сгруппированы типы грамматических конструкций, которые используются специфичным для этого контекста образом). Разбор предложения в контексте, находящемся в середине дерева, подразумевает использование всех дочерних контекстов.

Если при использовании дочерних контекстов не удастся добиться приемлемых результатов, можно перейти к родительской вершине. Если какая-то конструкция выбивается из контекста, то она все равно будет разобрана, только на это потребуется дополнительное время и могут появиться дополнительные варианты.

Вероятности.

Как уже упоминалось выше, вероятность сопоставляется атомарной модели управления, связанной со словом, и определяется частотой использования этой модели. Здесь понятие вероятности связано с аналогичным понятием в теории вероятностей скорее по смыслу, нежели по использованию, определению и свойствам. Например, удобней для наших целей представлять вероятность в качестве целого числа. При этом истинная вероятность P_j использования j-той модели управления некоторого слова равняется , где p_i – это наша

целочисленная вероятность, а суммирование происходит по всем атомарным моделям управления данного слова. Таким образом, для вычисления целочисленной вероятности не требуется знать обо всех моделях управления слова, что позволит избежать необходимости ее пересчета при изменениях и дополнениях набора моделей управления.

При разборе предложения можно все результирующие структуры предложения ранжировать по их вероятностям, высчитываемым как функции вероятностей примененных моделей управления. При правильно выбранных вероятностях моделей управления структура предложения с максимальной вероятностью будет наиболее адекватной.

Здесь возникает вопрос: а как узнать, какие вероятности приписывать атомарным моделям управления? Лучшим решением было бы сидеть, читая разные тексты, для каждого слова определять примененные модели управления, увеличивать для них вероятность на единицу в базе данных, и т. д. По понятным причинам, это трудноосуществимо; по-видимому, система будет обучаться не быстрее человека, впервые изучающего совершенно незнакомый язык. Очень заманчивой представляется идея автоматического анализа текстов для самообучения и накапливания статистики по использованию моделей управления. Главной проблемой на этом пути является следующее обстоятельство: для правильного выбора используемой атомарной модели управления в конкретном предложении нужно его сначала адекватно разобрать, а гарантированно это можно сделать, лишь накопив достаточную статистику по употреблению моделей управления, что невозможно сделать без правильного выбора атомарных моделей управления. Получается замкнутый круг. Вероятно, в систему необходимо предварительно ввести достаточно большой объем информации, чтобы в процессе самообучения система улучшала качество разбора.

Литература

И.А.Волкова, И.Г.Головин. Об одном подходе к построению синтаксического модуля в системе распознавания устной речи. ДИАЛОГ’97, Труды межд. семинара. М., 1997.

И.А.Волкова, И.Г.Головин. Синтаксический анализ фраз естественного языка на основе сетевой грамматики. ДИАЛОГ’98, Труды межд. семинара. М., 1998.

И.А.Волкова, И.Г.Головин, О.Ф.Кривнова. Компьютерный словарь моделей управления русских глаголов (экспериментальный вариант). ДИАЛОГ’98, Труды межд. семинара. М., 1998.