Синтаксический анализатор для русского и английского языков

А. А. Антонова, А. В. Мисюрев

Описываются возможности синтаксического анализатора для русского и английского языка, ориентированного на универсальную обработку текста независимо от жанра и предметной области. Используемое синтаксическое описание основано на отношениях зависимости между словами (Dependency Theory). Синтаксический анализатор может быть использован при разработке систем различного профиля, таких, как системы распознавания речи, рубрикации, текстовой атрибуции и машинного перевода.

Введение

В статье кратко описываются возможности синтаксического анализатора для русского и английского языка, который может быть использован при разработке систем различного профиля, таких, как системы распознавания речи, рубрикации, текстовой атрибуции и машинного перевода.

Развитие методов автоматической обработки текста невозможно без соответствующего инструментария (в частности, программного обеспечения), разработка которого сама по себе является сложной задачей. Отсутствие или труднодоступность такого инструментария часто вынуждает исследователей тратить большую часть времени на его подготовку и/или состыковку с данными для исследования. Переход от предварительных экспериментов к более содержательным может быть осложнен тем, что старый инструментарий может оказаться непригодным для решения новых задач. Таким образом, для исследований, предполагающих дальнейшее развитие, значение технологической составляющей существенно возрастает.

Стадии обработки текста

На вход синтаксическому анализатору поступает файл с текстом на русском или английском языке. Можно выделить следующие стадии обработки текста:

- разбиение текста на предложения и слова;
- морфологический разбор;
- синтаксический разбор;
- интерпретация результатов синтаксического разбора.

Грамматические атрибуты слов

Каждому слову сопоставляется набор его грамматических атрибутов в соответствии с результатом разбора. Грамматические атрибуты для русского языка: *часть речи*, *падеж*, *число*, *род*, *лицо*, *одушевленность*, *финитность*, *время*, *залог*, *наклонение*, *краткая форма* (*для прилагательных и причастий*), *сравнительная степень* (*для прилагательных и наречий*).

Грамматические атрибуты для английского языка: *часть речи*, *падеж*, *число*, *лицо*, *финитность*, *время*, *степень сравнения.*

Синтаксический разбор

В формальных определениях задачей синтаксического анализа объявляется определение принадлежности входной цепочки слов некоторому языку. Применительно к естественному языку, особенно в практических приложениях, более осмысленной представляется задача построения наилучшей интерпретации *для любой* входной цепочки слов. Эта задача обычно называется частичным разбором предложения.

Синтаксический анализатор ориентирован на универсальную обработку текста независимо от жанра и предметной области. Используемое синтаксическое описание основано на отношениях зависимости между словами (Dependency Theory) [2–21]. Хотя изначально синтаксический анализ разрабатывался для нужд машинного перевода, он зависит только от входного языка и не подстраивается ни под какую языковую пару. Благодаря этому увеличивается область его применения в других приложениях, в том числе не связанных с переводом.

Синтаксические связи, выделяемые синтаксическим анализатором, обладают следующими свойствами.

- В каждой синтаксической связи участвует два слова из разбираемого предложения.
- Слова во входном предложении линейно упорядочены, поэтому в любой связи есть левое слово (расположено ближе к началу предложения) и правое слово (расположено дальше от начала предложения).
- Если тип связи — подчинение, то одно из слов является главным, а другое — зависимым. Стрелка всегда направлена от зависимого слова к главному.

Пример результатов синтаксического разбора для предложения *Тося договорился с Хубертом выкупить свое кольцо, и деньги для этого должен буду дать я.*

Дерево синтаксического разбора:

```
{ Top }
     { hi }
         & { lo }
            ^ договориться
              (subj_es > fin_es) тося
              (head_es < inf_es) выкупить
                 (prepnp_es > inf_es) с
                    (prep_es < noun_es) хуберт
                 (head_es < acc_es) кольцо
                    (adj_es > noun_es) свой
            ^ ,
            ^ дать
              (conj_es > sent_es) и
              (acc_es > fin_es) деньги
                 (noun_es < prepnp_es) для
                    (prep_es < noun_es) этот
              (aux_es > inf_es) должен
                 (head_es < aux_es) быть
              (head_es < subj_es) я
            ^  .
```

Каждое имя связи состоит из трех частей:

elem1 — условный идентификатор для левого слова;

linktype — тип связи (подчинение “>” или “<”);

elem2 — условный идентификатор для правого слова.

Примеры:

Пары слов	Имя связи	Исходный текст
“red” “flower” “a” “beer” “you” “leave”	adj_es > noun_es; det_es > noun_es subj_es > fin_es	“red flower” “a beer” “you leave”
«весь» «срок» «срок» «выполнение» «выполнение» «проект»	adj_es > noun_es noun_es < gen_es noun_es < gen_es	«весь срок выполнения проекта»

В большинстве случаев существенная информация о связи содержится в теге зависимого слова. Например, если нужно извлечь из текста

случаи, когда глагол управляет прямым дополнением, нужно извлечь связи, где тег зависимого слова — acc_es (для русского языка), sm_es (для английского языка).

Тег главного слова содержит вспомогательную информацию о синтаксической категории слова. Часто ту же самую информацию можно получить из грамматических атрибутов главного слова. Следует отметить, что тег главного слова может зависеть от конструкции, в которой участвует данная связь, и его использование не всегда прозрачно (в частности, встречаются связи с неопределенным тегом главного слова head_es).

Построение дерева разбора

В большинстве случаев число слов в результирующем дереве равняется реальному числу слов во входной цепочке. Однако в разбор заложено некоторое количество конструкций, которые формально состоят из нескольких слов, но функционируют как словарная единица. Такие конструкции в результатах программы разбора объединяются в один узел. Примеры составных слов: союзы («*как только*», «*по мере того как*»), предлоги («*вкупе с* товарным знаком», "*in addition to* showing images"), частицы («*чуть ли не* у самой стены») наречия («*на редкость* замечательный», "*of course*") и др.

В дереве разбора создаются дополнительные узлы для конструкций, между словами которых не нужно устанавливать подчинительную связь. Слова в подобных конструкциях объявляются зависимыми от дополнительного узла.

- Для слов, входящих в перечисление, создается дополнительный узел типа '+'.
- Для сложных лексических единиц с изменяемой частью (например, фразовых глаголов) создается дополнительный узел типа '#'.
- Для частей предложения, между которыми нет других синтаксических связей, создается дополнительный узел типа '&'.

Кроме того, в дереве разбора вводится дополнительный самый верхний узел (Top).

Разрешение неоднозначностей во время синтаксического разбора

Неоднозначность в синтаксическом разборе проявляется в следующем.

- Одному и тому же слову исходного текста может быть сопоставлено несколько возможных наборов морфологических атрибутов.

- Для одной и той же цепочки слов могут быть установлены различные наборы синтаксических связей.
- Для длинных предложений возможны «комбинаторные взрывы», т. е. ситуации, когда программа разбора не способна перебрать все варианты. В таких случаях в процессе разбора возникает проблема выбора наиболее перспективных частичных вариантов.

В отличие от достаточно известной системы aot.ru [1], в которой разрешение морфологической неоднозначности производится до начала синтаксического разбора, в описываемом синтаксическом анализаторе эти задачи решаются согласованным образом, т. е. при выборе морфологических вариантов учитываются синтаксические связи между словами.

Существует взаимосвязь между сложностью грамматических правил и эффективностью методов разрешения неоднозначностей. Более сложный набор грамматических правил может описывать больше различных грамматических конструкций. При этом в процессе разбора порождается больше различных вариантов, соответственно, требуются более точные методы разрешения неоднозначностей. В значительной степени разработка грамматики языка, предназначенной для практических применений, сводится к поиску компромисса между ошибками, возникающими из-за неполного набора грамматических правил, и ошибками, возникающими из-за неспособности программы разбора выбрать правильный вариант из многих возможных. Соответственно, разработка эффективных методов разрешения неоднозначностей позволит использовать более полные наборы грамматических правил.

Применение пакета

Синтаксический анализатор предоставляет возможность разметки текстовых данных с учетом морфологии и синтаксиса. Он может выполнять функцию одного из базовых инструментов при работе, например, с корпусами текстов на одном языке, с корпусами параллельных текстов.

Автоматическая разметка позволяет получить необходимый объем материала для работы по следующим направлениям.

- Изучение статистических закономерностей встречаемости слов и конструкций в языке. Такого рода закономерности могут быть использованы, в частности, при разрешении неоднозначностей, возникающих во время синтаксического разбора и во время выбора вариантов перевода при машинном переводе, а также при построении статистических моделей языка для распознавания слитной речи.
- Сравнение текстов на основе встречаемости в них различных слов и конструкций. Областями применения могут быть классификация документов, определение авторства, определение стиля документа.

- Построение корпуса параллельных текстов (русский и английский языки).

  Другие применения пакета.

- Разработка инструментария для исследований в области автоматической обработки текста. Состыковка с другими системами (языками программирования, статистическими пакетами, базами данных).
- Разработка методов машинного понимания текстов, извлечения данных из текстов, другие приложения, в которых может потребоваться синтаксический разбор.

Заключение

Описанный синтаксический анализатор используется при написании курсовых и дипломных работ студентами механико-математического факультета МГУ, ОТиПЛ филологического факультета МГУ, МФТИ. Тематика работ: построение модели языка для распознавания речи, определение авторства документов, понимание компьютером команд на естественном языке. Программа распространяется свободно для некоммерческого использования, на момент написания статьи адрес сайта http://cs.isa.ru:10000/dwarf.

Большое значение для дальнейшего развития инструментария имеет состыковка с другими системами (языками программирования, статистическими пакетами, базами данных). Дальнейшее развитие синтаксического анализа предполагает включение в него дополнительных возможностей для извлечения данных из текста, таких, как разметка имен, названий, дат, и пр.

Приложение. Описание демонстрационного примера

Описывается демонстрационный пример использования результатов работы пакета. Пример представляет собой программу на языке Си (Visual C++), читающую из текстовых файлов результаты разбора нескольких произведений различных авторов и оценивающую близость текстов. Сравнивалось 30 текстов 10-ти авторов, по три текста каждого автора. Для каждого текста искался наиболее похожий. Сравнение проводилось двумя способами, различающимися функцией оценки близости.

1-*й способ*. Сравнение проводилось при помощи 19-ти признаков, вычисляемых на основе морфологических атрибутов слов. Признаки были заимствованы с изменениями из [Галяшина Е. И. «Основы судебного речеведения», СТЭНСИ, 2003].

Значения лексико-грамматических параметров для книг нормировались (нулевое среднее значение, единичная дисперсия). В качестве меры близости использовалось евклидово расстояние с весами, равными отношению среднеквадратичного значения различий признаков у текстов разных авторов к среднеквадратичному значению различий у текстов одного автора. (Эти веса вычислялись на небольшом дополнительном наборе текстов других авторов). В результате сравнения текстов на основе атрибутов слов для 20-ти текстов авторы самого текста и ближайшего текста совпали. В 10-ти случаях ближайшим к произведению одного автора оказалось произведение другого автора.

2-*й способ*. Оценивалась близость частот синтаксических связей. Мера близости основана на предположении, что вероятности использования автором синтаксических связей определенного типа являются постоянными величинами и не зависят от контекста. В результате сравнения текстов на основе частот синтаксических связей для 24-х текстов автор самого текста и автор ближайшего текста совпали. В 6-ти случаях ближайшим к произведению одного автора оказалось произведение другого автора.

Второй способ оказался значительно более простым. Помимо функций, которые можно считать стандартными (чтение результатов разбора из текстового файла и подсчет частот синтаксических связей в тексте), он содержит функцию оценки близости книг, к которой, собственно, и сводится содержательная часть программы:

```
// Параметры
// pfr, pto - вероятности синтаксических связей
//            сравниваемых книг
// nFeature - количество типов связей
// Возвращает оценку близости книг

double bookCmp(double pfr[], double pto[], int nFeature)
{
 double sum= 0.0;
 int pi;

 for (pi= 0; pi<nFeature; pi++)
  sum+= pfr[pi]*log(pto[pi])+(1.0-pfr[pi])*log(1.0-pto[pi]);
 return sum;
}
```

Целью данного примера была демонстрация возможностей применения программного пакета для извлечения различных количественных

характеристик текста (с учетом морфологии и синтаксиса). Из результатов синтаксического анализа текста можно извлечь как его лексико-грамматические параметры, так и распределение синтаксических связей, которое, как показывают результаты эксперимента, само по себе является важной характеристикой при оценке стиля текста.

Литература

1. *Гершензон Л. М.*, *Ножов И. М.*, *Панкратов Д. В.*, *Сокирко А. В.* Синтаксический анализ в системе РМЛ. http://www.aot.ru/docs/synan.html
2. *Ермаков А. Е.* Неполный синтаксический анализ текста в информационно-поисковых системах: В 2 т. Т. 2. «Прикладные проблемы» // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2002. М.: Наука, 2002.
3. *Chelba C.*, *Engle D.*, *Jelinek F.*, *Jimenez V.*, *Khudanpur S.*, *Mangu L.*, *Printz H.*, *Ristad E.*, *Rosenfeld R.*, *Stolcke A.*, *& Wu D.* 1997. Structure and Performance of a Dependency Language Model // Eurospeech'97.
4. *Collins* 1996 Collins, Michael John. 1996. A new statistical parser based on bigram lexical dependencies. In Proceedings of the 34^{th} Annual Meeting of the Association for Computational Linguistics, Santa Cruz, USA, June. ACL.
5. *Covington* 1990 Covington, Michael. 1990. Parsing discontinuous constituents in dependency grammar. Computational Linguistics, 16:234–236.
6. *Eisner* 1996 Eisner, Jason M. 1996. Three new probabilistic models for dependency parsing: An exploration. In COLING-96. The 16^{th} International Conference on Computational Linguistics, volume 1, pages 340–345, Copenhagen, Denmark, August.
7. *Engel* 1972 Engel, Ulrich. 1972. Bemerkunger der Dependenzgrammatik. Neue Grammatiktheorien und ihre Anwendung auf das heutige Deutsch, 20.
8. *Engel* 1996 Engel, Ulrich. 1996. Tesnière mißverstanden. In Gertrud Gréciano and Helmut Schumacher, editors, Lucien Tesnière — Syntaxe structurale et opérations mentales, volume 348 of Linguistische Arbeiten. Max Niemeyer Verlag, Tübingen, pages 53–61.
9. *Fraser* 1989 Fraser, Norman M. 1989. Parsing and dependency grammar. UCL working papers in Linguistics, 1:296–338.
10. *Hawkins* 1993 Hawkins, John, 1993. Heads, parsing and word-order universals, chapter 11, pages 231–265. CUP.
11. *Hays* 1964 Hays, David G. 1964. Dependency theory: A formalism and some observations. Language, 40:511–525.
12. *Heringer* 1993 Heringer, Hans Jürgen. 1993. Dependency syntax-basic ideas and the classical model. In Joachim Jacobs, Arnim von Stechow, Wolfgang Sternefeld, and Theo Venneman, editors, Syntax — An International Handbook of Contemporary Research, volume 1. Walter de Gruyter, Berlin—New York, chapter 12, pages 298–316.

13. *Hudson* 1991 Hudson, Richard. 1991. English Word Grammar. Basil Blackwell, Cambridge, MA.
14. *Järvinen* 1994 Järvinen, Timo. 1994. Annotating 200 million words: The Bank of English Project. In COLING 94. The 15th International Conference on Computational Linguistics Proceedings, volume I, pages 565–568, Kyoto, Japan, August.
15. *Järvinen and Tapanainen* 1997 Järvinen, Timo and Tapanainen, Pasi. 1997. A Dependency Parser for English. http://www.ling.helsinki.fi/~tapanain/dg/index.html
16. *Marcus P.*, *Satornini B.*, *Marcinkiewicz M.* 1993. Building a large annotated corpus of English: The Penn treebank. Computational Linguistics, 19(2).
17. *Mel'čuk and Pertsov* 1987 Mel'čuk, Igor A. and Nikolaj V. Pertsov. 1987. Surface Syntax of English. A formal model within the meaning-text framework . John Benjamins, Amsterdam.
18. *Mel'čuk* 1987 Mel'čuk, Igor A. 1987. Dependency Syntax: Theory and Practice. State University of New York Press, Albany.
19. *Petkevič* 1987 Petkevič, Vladimir. 1987. A new dependency based specification. Theoretical Linguistics, 14:143–172.
20. *Sleator and Temperley* 1993 Sleator, Daniel and Davy Temperley. 1993. Parsing English with a link grammar. Third International Workshop on Parsing Technologies, August.
21. *Tapanainen and Järvinen* 1997 Tapanainen, Pasi and Timo Järvinen. 1997. A nonprojective dependency parser. In Proceedings of the 5th Conference on Applied Natural Language Processing, Washington, D. C., April. Association for Computational Linguistics.
22. *Tesniére* 1959 Tesniére, Lucien. 1959. Éléments de syntaxe structurale. Editions Klincksieck, Paris.
23. *Voutilainen* 1994 Voutilainen, Atro. 1994. Designing a Parsing Grammar. Publications no. 22, Department of General Linguistics, University of Helsinki, Finland.