Морфологическая разметка текстов на русском языке с использованием формальной грамматики AGFL

Азарова И. В.
(*Кафедра математической лингвистики СПбГУ*)

Система порождения грамматических парсеров AGFL

Система AGFL (Affix Grammar for Finite Lattice) – свободно распространяемое программное обеспечение для решения задач автоматической обработки текстов на естественном языке, использующее формализм AGFL. Система была разработана на отделении Компьютерных исследований Университета Неймеген (Нидерланды) под руководством К. Костера. Формализм AGFL[1] является двухуровневой формальной грамматикой: контекстно-свободная порождающая грамматика дополнена набором признаков (аффиксов) с конечным числом значений, которые позволяют задавать конструкции согласования, координации и управления. Система AGFL позволяет генерировать эффективные парсеры для анализа морфологических и синтаксических структур естественных языков, при этом формат выходной последовательности парсеров можно задавать при помощи трансдукций в правилах формализма. Помимо парсеров, система AGFL дает возможность подключить лексические базы данных большого объема. В конечном итоге, система AGFL позволяет создавать парсеры на основании лингвистических описаний, которые легко создаются и видоизменяются.

Для последней версии системы AGFL 2.0 на сайте представлены парсеры для английского и голландского языков, ориентированные на выдачу данных для информационного поиска и автоматической классификации документов[2]. На завершающей стадии находятся парсеры для испанского, арабского и русского языков. Частичные AGFL-описания были выполнены для греческого и венгерского языков.

AGFL-описание для морфологической разметки текстов на русском языке

Первоначально AGFL-описание морфологии русского языка было выполнено в рамках версии 1.5 в 1997 г., однако значительное число ограничений функций системы AGFL давало возможность использовать это описание как учебную модель для формально-грамматических исследований, проводившихся на кафедре математической лингвистики СПбГУ. Лишь последняя версия предоставила возможности практического применения AGFL-описания для обработки текстов на естественном языке: в частности, создания парсеров, ориентированных на морфологическую разметку текста на русском языке и на представление данных для информационного поиска[3]. Парсер морфологической разметки русского текста будет описан содержательно в данном докладе.

Грамматическое описание в системе AGFL имеет модульный характер, в настоящее время грамматика русского языка включает модули построения именных и предикативных словосочетаний, модули описания словоизменения знаменательных частей речи (имен и глаголов), модули задания флексий словоизменения для знаменательных частей речи. Были проведены эксперименты по описанию конструкций простого предложения, однако описание полной синтаксической структуры сложных предложений с учетом актуального членения находится в настоящее время на этапе исследований.

---

[1] Описание системы AGFL, примеры парсеров, библиография представлены на сайте: http://www.cs.kun.nl/agfl/. См. также сборник статей: Proceedings of the first AGFL Workshop // Eds. C. H. A. Koster, E. Oltmans. Nijmegen, 1996.

[2] Koster C. H. A., Verhoevan T. Head/Modifier Frames for Information Retrieval // Proceedings COLING 2002. 2002.

[3] Азарова И. В. Использование AGFL-разметок текста для автоматической классификации документов // Материалы XXXI Всероссийской научно-методической конференции преподавателей и аспирантов. Вып. 4. Секция прикладной и математической лингвистики. Ч. 2. С. 3–8.

Аффиксы AGFL-описания для русского языка

Аффиксы в формализме AGFL дают возможность признакового описания грамматических объектов. В качестве таких признаков регулярно используются грамматические категории, например, нетерминальный аффикс числа (NUMBER) и его терминальные значения – единственное и множественное (1); нетерминальный аффикс рода (GENDER) и его терминальные значения – мужской, женский, средний (2).

(1) NUMBER :: sing | plur.

(2) GENDER :: masc | fem | neut.

Нетерминальные аффиксы могут определяться через другие нетерминальные аффиксы, например, аффикс-категория падежа (CASE) определяется через терминальный аффикс именительного падежа и нетерминальный аффикс (OCASE) косвенных падежей (3), который в свою очередь определяется через набор терминальных падежных аффиксов, то есть родительного, дательного, винительного, творительного, предложного (4).

(3) CASE :: nom | OCASE.

(4) OCASE :: gen | dat| acc| abl| loc.

Если потребуется различать смысловые варианты падежных форм, например, родительный частичный и собственно родительный или предложный и местный падежи, это можно сделать через систему «вложенных» аффиксов (5–7):

(5) OCASE :: GEN | dat| acc| abl| LOC.

(6) GEN :: gen | gen_pt.

(7) LOC :: prep | loc.

Аффиксы могут также использоваться для деления слов части речи на лексико-грамматические разряды, например, переходные и непереходные глаголы (8).

(8) TRANS :: tr | intr.

Таким образом, некоторые спорные вопросы грамматической теории, например, является ли одушевленность категорией или характерным значением лексико-грамматического разряда существительных, снимаются при использовании специального нетерминального аффикса (ANIM), имеющего два значения: одушевленный и неодушевленный (9).

(9) ANIM :: anim | inanim.

В качестве аффиксов можно использовать различные формальные показатели, например, тип склонения существительных (DECLTYPE) или тип спряжения глаголов (CONJTYPE), которые определяются через подтипы, например, твердый (ORD) и мягкий (PAL) подтипы склонения; нестандартный (irreg) и нулевой (zero) типы склонения; I (ETYPE) и II (ITYPE) спряжения глаголов; разноспрягаемые (irreg) глаголы (10–11).

(10) DECLTYPE :: ORD | PAL | irreg | zero.

(11) CONJTYPE :: ETYPE | ITYPE | irreg.

Использование аффиксов в правилах AGFL-описания для русского языка

Нетерминальные и терминальные аффиксы используются в правилах формальной грамматики AGFL в качестве признаков нетерминальных символов, при этом состав и порядок перечисления аффиксов должен быть постоянным. Правило (12) задает построение формы существительного (nounform) путем приписывания к основе существительного (nounstem) окончания (nounending). Набор аффиксов формы имени существительного включает спецификацию категорий падежа, числа, рода, и аффикса одушевленности, аффиксы основы существительного задают терминальные значения аффиксов рода, одушевленности и типа склонения, аффиксы окончания – значения падежа, числа, рода, одушевленности, типа склонения. Использование одного и того же нетерминального аффикса в правиле обеспечивает идентичность значений в правой и левой частях, т. е. для терминального значения "gen" не-

терминального аффикса падежа в левой части для нетерминала формы существительного (nounform) это же значение будет задано для нетерминала флексии существительного (nounending). Перечни терминальных основ слов задаются в специальном формате файлов лексикона (13), цепочек терминальных флексий – в модулях грамматики (14).

(12) nounform (CASE, NUMBER, GENDER, ANIM) :
nounstem (GENDER, ANIM, DECLTYPE),
nounending (CASE, NUMBER, GENDER, ANIM, DECLTYPE).

(13) "двер" nounstem (fem, inanim, irreg)

(14) nounending (gen| dat| loc, sing, fem, ANIM, irreg): "-и".

В правиле (14) терминальная последовательность в кавычках содержит специальный символ – дефис, который указывает для генератора парсеров, что флексия приписывается к символьной строке основы справа без разделителя, например пробела. В этом же правиле используется символ объединения (|) терминальных аффиксов, который позволяет компактно записывать несколько однотипных конструкций.

Перечень нетерминальных аффиксов задает прямое произведение их значений, т. е. для формы существительного предполагается 72 комбинации для шестипадежной системы. Часть комбинаций терминальных аффиксов может и не иметь грамматической интерпретации, например, область определения нетерминала *nounform (dat, plur, masc, anim)* тождественна *nounform (dat, plur, fem, inanim)*, поскольку в формах множественного числа существительных, как правило, нет родовых различий, также не важны для формы дательного падежа значения аффикса одушевленности. Нетерминал *nounform (dat, plur, GENDER, ANIM)* более четко показывает реальную комбинаторику значений аффиксов.

Учет комбинаторики терминальных аффиксов необходим в формально-грамматическом описании AGFL. Например, в системе категорий русского глагола значения категорий наклонения и времени естественным образом объединяются в одном аффиксе (MOOD):

(15) MOOD :: TENSE | imper| subjunc.

(16) TENSE :: past | PRESFUT.

(17) PRESFUT :: pres | fut | presfut.

Дополнительная дистрибуция категорий лица (в настоящем и будущем временах) и рода (в прошедшем времени) также естественным образом описывается при помощи одного аффикса:

(18) GENPER :: GENDER | PERSON.

(19) PERSON :: first | second | third.

Формат морфологической разметки в AGFL

В правила AGFL можно включить трансдукцию (transduction), т. е. указание того, в каком виде представлять информацию в выходном файле морфологического парсера. Трансдукционная часть отделяется от самого правила знаком "/", за которым указывается перечень нетерминалов и нетерминальных аффиксов правой части правила, значение которых будет печататься в выходном файле.

Правило (20) указывает, что терминальное существительное (terminal noun) может быть формой существительного, которая и будет помещена в выходную последовательность.

(20) terminal noun (CASE, NUMBER, GENDER, ANIM) :
nounform (CASE, NUMBER, GENDER, ANIM) / nounform.

Если мы планируем xml-формат выходного файла, то можно вставить знаки перевода каретки, отступы табуляции и xml-маркеры:

(21) terminal noun (CASE, NUMBER, GENDER, ANIM) :
nounform (CASE, NUMBER, GENDER, ANIM) /
"\n\t\t\t<word>", nounform, " </word>".

В том случае если мы хотим вместо текстовой формы слова задавать его исходную форму (24), то мы должны предусмотреть этот параметр в качестве аффикса (LEMMA) в правиле (22 вместо 12) и задать этот параметр в лексиконе (23 вместо 13):

```
(22)    nounform (CASE, NUMBER, GENDER, ANIM, LEMMA) :
            nounstem (GENDER, ANIM, DECLTYPE, LEMMA),
                nounending (CASE, NUMBER, GENDER, ANIM, DECLTYPE).
(23)    "двер"          nounstem (fem, inanim, irreg, "дверь")
(24)    terminal noun (CASE, NUMBER, GENDER, ANIM) :
          nounform (CASE, NUMBER, GENDER, ANIM, LEMMA) /
            "\n\t\t\t<word>", LEMMA, " </word>".
```

Аналогичным образом можно задавать принадлежность к части речи и значения грамматических категорий, т. е. подробную морфологическую информацию:

```
(25)    terminal noun (CASE, NUMBER, GENDER, ANIM) :
          nounform (CASE, NUMBER, GENDER, ANIM, LEMMA) /
            "\n\t<word>", nounform,
            "\n\t\t<noun>", "LEMMA=", LEMMA, "CASE=", CASE,
                            "NUMBER=", NUMBER, "GENDER=", GENDER, "</noun>"
            "</word>".
```

Особенности морфологических правил русского языка AGFL-описания

Морфологическими модулями являются модули задания флексий и модули правил генерирования форм. Основную проблему для морфологических модулей представляет изменение основ в парадигмах и использование вариативных или нетиповых флексий.

В русской морфологии регулярно наблюдается варьирование основ в частных парадигмах, например, в парадигмах единственного и множественного числа существительных. В таком случае вводится несколько нетерминалов для основ (nounstem_sg для единственного числа, nounstem_pl для множественного числа), усложняются правила построения форм (26–27), в лексикон добавляется описание основ (28), у которых обычно меняется тип склонения, возможно, и другие признаки.

```
(26)    nounform (CASE, sg, GENDER, ANIM) :
            nounstem_sg (GENDER, ANIM, DECLTYPE ),
                nounending (CASE, sg, GENDER, ANIM, DECLTYPE).
(27)    nounform (CASE, pl, GENDER, ANIM) :
            nounstem_pl (GENDER, ANIM, DECLTYPE ),
                nounending (CASE, pl, GENDER, ANIM, DECLTYPE).
(28)    "ух"   nounstem_sg (neut, inanim, ordk, "ухо")
        "уш"  nounstem_pl (plmasc, inanim, ordsh, "ухо")
```

По такой же схеме вводятся вариантные основы частных парадигм для слов с неполной парадигмой, например, singularia tantum (*гнев, листва*) и pluralia tantum (*дрожжи, брюки*), хотя отсутствие ограничений на порождение форм привело бы к так называемому избыточному порождению (overgeneration), что не является серьезной ошибкой парсера, поскольку таких форм просто не будет во входных текстах.

Сложнее описывать нетиповые флексии, например, для существительного среднего рода "ухо" (28) в именительном падеже множественного числа используется флексия "-и", характерная для существительных мужского и женского рода. В таком случае, используется специальное родовое значение для существительных во множественном числе plmasc, которое также используется для описания родовой принадлежности существительных pluralia tantum, нечто типа реконструированного родового показателя у А. А. Зализняка[1].

[1] Зализняк А. А. Грамматический словарь русского языка: Словоизменение. М., 1977.

Элементы поверхностного синтаксического описания на уровне синтаксиса словосочетания были введены в парсер для снятия неоднозначности морфологических форм. Например, у существительных 3-го склонения типа "дверь" не различаются 5 форм: ед. ч. родительный, дательный, предложный падеж и мн. ч. именительный и винительный падеж. Если такая форма встречается в комбинации с прилагательным (*тяжелой двери, тяжелые двери*) или в предложной конструкции (*к тяжелой двери, на тяжелые двери*) (29–30), то неоднозначность морфологической разметки существенно снижается.

(29) nounphrase (CASE, NUMBER, GENDER, ANIM) :
terminal adjective (CASE, NUMBER, GENDER, ANIM),
terminal noun (CASE, NUMBER, GENDER, ANIM).

(30) prepphrase (CASE) : terminal preposition (CASE),
nounphrase (CASE, NUMBER, GENDER, ANIM).

Заключение

В докладе были описаны основные принципы построения системы морфологической разметки русских текстов с использованием формализма AGFL. В настоящее время завершено ядро морфологического парсера. После включения комментариев в модули грамматического описания они будут помещены на сайте AGFL http://www.cs.kun.nl/agfl/ для свободного использования. В ближайшие задачи входит апробирование парсера морфологической разметки на русских газетных текстах для выявления отрицательного материала (неоднозначных или неправильных интерпретаций текста) и расширение лексикона системы за счет включения в него новых основ и обобщенных описаний с использованием типовых словообразовательных элементов (суффиксов, префиксов).

Грамматическое описание русского языка в формализме AGFL не ограничено каким-либо образом. Предлагаемый вариант является результатом поиска оптимального решения, в котором большая часть может быть изменена под влиянием предпочтений исследователя. Самое ценное в описании состоит в том, что оно «открыто», может быть «прочитано», переосмыслено и изменено.