М. А. Грачкова, О. Н. Ляшевская, О. А. Митрофанова,
П. В. Паничева, А. С. Шиморина
M. A. Grachkova, O. N. Lashevskaya, O. A. Mitrofanova,
P. V. Panicheva, A. S. Shimorina

МОДЕЛЬ ДАННЫХ ДЛЯ КАТАЛОГА РУССКИХ ЛЕКСИЧЕСКИХ КОНСТРУКЦИЙ (НА ПРИМЕРЕ ИМЕН РЕЧЕВЫХ ДЕЙСТВИЙ В НКРЯ)

DATA MODEL FOR THE CATALOGUE OF RUSSIAN LEXICAL CONSTRUCTIONS (A CASE STUDY OF NOUNS DENOTING SPEECH ACTS IN RNC)

Аннотация. Цель исследования – разработка технологии автоматического распознавания конструкций, связанных с целевыми словами, и использование этих данных для построения каталога лексических конструкций. Исследование проводится на материале Национального корпуса русского языка (НКРЯ, http://ruscorpora.ru). Выделение конструкций осуществляется с опорой на многоплановую (прежде всего, морфологическую и лексико-семантическую) лингвистическую разметку НКРЯ. В докладе обсуждаются инструменты автоматического выделения и визуализации данных, используемые для построения каталога лексических конструкций.

Abstract. Our research aims at automatic identification of constructions associated with particular lexical items and its subsequent use in building the catalogue of Russian lexical constructions. The study is based on the data extracted from the Russian National Corpus (RNC, http://ruscorpora.ru). The main accent is made on extensive use of morphological and lexico-semantic data drawn from the multi-level corpus annotation. The toolkit that processes corpus samples and learns up the constructions is described. We also discuss the use of visualization module that represents the inner structure of extracted constructions.

1. Введение

Цель обсуждаемого в докладе проекта – предложить основанную на статистических методах технологию автоматического распознавания типичных конструкций, связанных с той или иной лексической единицей.

Данный доклад продолжает цикл работ[1], посвященных автоматическому выделению лексических конструкций в контекстах Национального корпуса русского языка (НКРЯ, **http://ruscorpora.ru**), отличающегося детальностью и многоплановостью лингвистической разметки. В центре исследования находятся именные конструкции – прежде всего, те, которые строятся вокруг имен существительных из разных лексико-семантических групп. В настоящем докладе обсуждаются конструкции имен речевых действий (*дискуссия, комплимент, обращение, обсуждение, ответ* и т.д.).

2. Трактовка понятия «конструкция», принятая в нашем проекте

Современные корпусы текстов дают возможность получать статистические данные о поведении лексической единицы в контексте, составить портрет контекстного окружения слова[2].

---

[1] См., например, *Митрофанова О. А., Ляшевская О. Н., Грачкова М. А., Шиморина А. С., Шурыгина А. С., Романов С. В.* Эксперименты по автоматическому разрешению лексико-семантической неоднозначности и выделению конструкций (на материале Национального корпуса русского языка) // Структурная и прикладная лингвистика. Вып. 9. СПб., 2012; *Ляшевская О. Н., Митрофанова О. А., Грачкова М. А., Шиморина А. С., Шурыгина А. С., Романов С. В.* К построению инвентаря русских именных конструкций // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18). М.: Изд-во РГГУ, 2012.

[2] *Gries St.Th., Divjak D.S.* Behavioral profiles: a corpus-based approach towards cognitive semantic analysis // *Evans V., Pourcel S.S.* (eds.) New directions in cognitive linguistics. John Benjamins, Amsterdam & Philadelphia, 2009.

Эти задачи решаются в ряде лингвистических (в основном, не русскоязычных) проектов, где особое внимание уделяется формализации лексико-синтаксических связей единиц текста, например, PropBank, NomBank, FrameNet, DeepDict, Sketch Engine, StringNet и т.д. Данные ресурсы дают разноплановую информацию о сочетаемости лексических единиц, при этом форма представления результирующих данных, как правило, табличная. Особняком стоят PropBank и NomBank, где основной акцент сделан на семантико-синтаксической разметке контекстов.

Традиционно контекстный профиль представляется в виде наборов *n*-грамм, сгруппированных по определенным признакам. Наш подход состоит в том, что выделяемые последовательности должны интерпретироваться как лингвистически значимые законченные лексические конструкции с ядром – искомым целевым словом. Реализация данного подхода подразумевает кластеризацию фрагментов контекстов употребления той или иной лексемы и выделение повторяющихся в контекстном окружении шаблонов.

С точки зрения структурной организации, конструкция – это комбинация целевого слова и слотов, заполняемых регулярными контекстными соседями, среди которых могут быть леммы, грамматические, лексико-семантические и т. п. признаки. По своей природе, конструкция – это абстрактный шаблон, предполагающий лексикализацию, т.е. различные реализации в виде комбинаций лемм/словоформ, ср. V|*дать, найти, предложить…* ОТВЕТ + PR|*на* + speech r:abstr|*вопрос*, r:qual|*простой, неоднозначный…* + ОТВЕТ, ОТВЕТ + t:hum r:concr|*академикам, мудрецам, отцу...*. Тем самым основная функция конструкции – фиксировать регулярную сочетаемость целевого слова в определенном его лексическом значении. Важно заметить, что в конструкциях отражаются контекстные признаки, разграничивающие отдельные значения целевого слова. Структуру многозначности целевого слова можно описать как семейства конструкций. Значение самой же конструкции характеризуется большей или

меньшей устойчивостью, варьирующей от регулярной свободной сочетаемости до высокой идиоматичности.

Наше определение конструкции не противоречит традиционному, однако несколько выходит за его рамки. Предлагаемое нами понимание конструкции позволяет, в отличие от метода *n*-грамм, относиться избирательно к сочетаемостным возможностям целевых слов, учитывать тенденции в сочетаемости целевого слова и его соседей в контексте, описывать как лексическую сочетаемость, так и сочетаемость на уровне классов, не только устойчивые, но и свободные сочетания, важным образом отражающие типовое употребление слова в тексте.

3. Анализ результатов работы инструмента автоматического выделения конструкций

Для представления данных о конструкциях в рамках нашего проекта был создан специализированный модуль CxI на языке Perl, где используются некоторые стандартные средства для обработки контекстных выборок с многоярусной лингвистической разметкой и для эффективной выдачи данных (в частности, XML::LibXML, YAML, Log::Log4perl). Важнейший компонент нашего модуля – пакет Algorithm::Combinatorics, с помощью которого производится выявление частотных комбинаций тегов в контекстах для целевых слов.

На вход программы подается файл с выборкой контекстов с целевым словом, для которого требуется выявить конструкции. Затем пользователь определяет такие параметры обработки данных, как типы тегов, учитываемых при выделении конструкций (*lex, sem, gr*), ширина контекстного окна, в пределах которого ведется поиск частотных комбинаций тегов (от –5 до +5), число конструкций, попадающих в выдачу (от 1 до 50).

Файл с результатами работы инструмента автоматического выделения конструкций содержит наиболее частотные сочетания целевого слова и различных тегов контекстного окружения (*lex, sem, gr*). В зависимости от назначенной ширины контекстного

окна в выдачу попадают комбинации тегов в виде пар, троек, четверок, пятерок и т.д.

В настоящий момент мы можем получать конструкции с двухслойной структурой, т.е. компоненты конструкции могут одновременно характеризоваться не более чем двумя признаками: морфологическими тегами и тегами лемм, или лексико-семантическими тегами и тегами лемм. Например,

(1) ОТВЕТ + PR|на + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*

(2) V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ОТВЕТ + PR|*на* + S m inan sg acc|*вопрос*

Заметим, что большой интерес вызывают конструкции с компонентами, в состав которых входят лексико-семантические теги, поскольку чаще всего с ними ассоциируются группы лемм, выражающих общее значение и характеризующихся близкими дистрибутивными свойствами. Например:

(3) r:rel|*риторический, мировой, процедурный, спорный, шекспировский, практический, методический* + ВОПРОС

(4) ОБСУЖДЕНИЕ + t:ment r:abstr|*проект, концепция* + r:abstr|*благоустройство, реформирование, реформа*

(5) ОТВЕТ + FW + t:speech r:abstr|*запрос, призыв, вопрос, приветствие, просьба, высказывание, похвала, рапорт, реплика*

Наши данные позволяют проследить развертку простейшей структуры в сложную многокомпонентную конструкцию и исследовать видоизменение состава конструкции по пути движения от простого к сложному. Например,

(6) *найти, дать* + *простой, однозначный* + ОТВЕТ + *на* + *вопрос*

t:poss|*дать, получить, давать* + ОТВЕТ

r:qual|*простой, неточный, точный, вероятный, логичный, нужный, вразумительный, ясный, приличный* + ОТВЕТ

r:rel|*готовый, однозначный, стандартный, истинный, числовой, заданный, релевантный, эмоциональный, содержательный, необязывающий, отрицательный, утвердительный, хлесткий,*

окончательный, известный, конкретный, официальный, адекватный, отечественный, обстоятельный, определенный, реактивный, обоснованный, очевидный, зачаточный, энергичный, соответствующий, стойкий + ОТВЕТ

t:move t:poss|*найти* + r:qual|*простой, точный, приличный* + ОТВЕТ + FW + t:speech r:abstr|*вопрос*

t:poss|давать, дать + r:rel|*конкретный, однозначный, окончательный* + ОТВЕТ + FW + *вопрос*

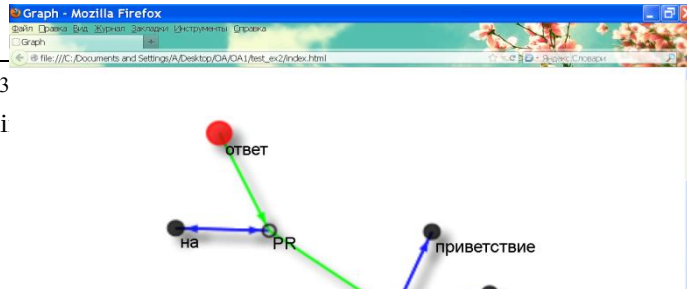
ОТВЕТ + PR|на + t:speech r:abstr|*приветствие, вопрос, высказывание, рапорт, реплика*

V pf tran inf act|*найти, дать* + A m sg acc inan plen|*простой, однозначный* + ОТВЕТ + PR|*на* ++S m inan sg acc|*вопрос*

4. Визуализация структуры и наполнения конструкций

Наша нынешняя задача – из многообразия используемых в компьютерной лингвистике техник визуализации выбрать метод графического представления данных, отличающийся простотой и широкими иллюстративными возможностями. Для получения графических представлений, отражающих структуру и наполнение конструкций, был задействован модуль pattern.graph (**http://www.clips.ua.ac.be/pages/pattern-graph**)[3], разработанный на языке Python и предназначенный для визуализации различных типов связей в тексте. На входе он принимает строку – конструкцию. На выходе создается граф, иллюстрирующий соответствующую конструкцию. Визуализация производится в два этапа: (1) производится парсинг строки конструкции и выявление ее главных и второстепенных элементов с сохранением порядка следования; (2) из них создается граф, отражающий данные структурные соответствия между элементами.

Пример визуализации данных о конструкциях средствами модуля pattern.graph приведен на рис. 1.



3 achine Learni

Рис. 1. Графическое представление конструкции
ОТВЕТ + PR|*на* + t:speech r:abstr| *приветствие, вопрос, высказывание, рапорт, реплика*

Структура конструкции в графах отражается следующим образом: красным цветом помечен узел, содержащий целевое слово, зеленым цветом выделены ребра графа, связывающие между собой элементы разметки конструкции (лексико-семантические и морфологические теги), синим – ребра графа, связывающие теги лемм с лексико-семантическими и морфологическими тегами.

5. Заключение

Проведенные эксперименты дают основания утверждать, что

1) инструмент автоматического выделения конструкций приспособлен для обработки контекстных выборок из НКРЯ, его применение позволило получить списки конструкций для целевых существительных;

2) полученные конструкции различаются по числу компонентов (это пары, тройки, четверки, пятерки, состоящие из тегов контекстного окружения) и по наполнению (это двухслойные

структуры, в состав которых входят либо морфологические теги и теги лемм, либо лексико-семантические теги и теги лемм);

3) задача визуализации данных о выделенных конструкциях успешно решается с помощью модуля pattern.graph, позволяющего наглядно представлять организацию конструкций, иерархию и различные типы их компонентов.

Перспективы развития исследования связаны с решением следующих задач:

1) отражение в конструкции трех слоев разметки (леммы, грамматические теги, лексико-семантические теги) одновременно;

2) учет статуса факультативных элементов конструкции – в нынешней версии такой функционал не предусмотрен;

3) переход к динамической организации модуля визуализации – особенно в тех случаях, когда конструкции содержат много элементов и много лексических вариантов реализации;

4) визуальное представление нескольких конструкций в контексте, когда конструкции с разными лексическими центрами «наслаиваются» друг на друга.

5) сопоставление выделенных наборов лексических конструкций с наборами, который мог бы выделить лексикограф на тех же данных.