ИНФОРМАТИКА И ЕЁ ПРИМЕНЕНИЯ, 2016. Т. 10. Вып. 1. С. 119–128

ПРИМЕНЕНИЕ КОНТЕКСТНО-СВОБОДНЫХ ГРАММАТИК ДЛЯ ИЗВЛЕЧЕНИЯ ОНТОЛОГИИ ИЗ ТЕКСТОВ КОРОТКИХ ОПИСАНИЙ СТАТЕЙ БИОЛОГИЧЕСКОЙ ТЕМАТИКИ*

Д. А. Алексеевский[1]

Аннотация: Обработка текстов биологической и медицинской тематики представляет интерес как с точки зрения биологии, для которой она предоставляет ценные результаты, так и в качестве источника более сложных задач для обработки текстов. Одной из важных задач автоматической обработки текстов является построение онтологий. Предложен метод построения онтологий промежуточного уровня по корпусу текстов на ограниченном подмножестве английского языка. Онтологии промежуточного уровня служат одним из инструментов решения задачи установления соответствия между фактами в априорных онтологиях и фрагментами текста. Предложен новый подход, основанный на расширенном определении контекстно-свободных (КС) грамматик, позволяющий порождать онтологии, обладающие указанным свойством. Показаны преимущества использования корпусов на ограниченном подмножестве естественного языка для построения таких онтологий.

Ключевые слова: КС-грамматики; построение онтологий; биомедицинские тексты

DOI: 10.14357/19922264160111

1 Введение

За последние десятилетия биология, а следом за ней и медицина претерпели несколько научных переворотов, каждый из которых приводил к бурному росту числа публикаций, а также и прочих текстов этих тематик. Многие полученные данные были собраны в базы данных, которые играют большую роль в этих науках. В то же время с ростом объема опубликованных текстов обнаруживаются новые виды данных, доступные в текстовом виде, требующие структурирования и верификации. Этим объясняется растущая актуальность темы извлечения фактов из текстов биологической и медицинской тематики. Следует заметить, что эта тема имеет существенные отличия от автоматической обработки текстов в целом, что обусловливает выделение ее в отдельную область.

Задачам автоматической обработки текстов медицинской и биологической тематики посвящено много работ. Среди современных направлений исследований: извлечение и нормализация именованных сущностей [1], извлечение событий и составных отношений [2], анализ дискурса и кореференции [3], построение и пополнение онтологий и баз данных [4]. Среди наиболее широко используемых биологических баз данных встречаются ресурсы, совмещающие структурированные данные (ссылки на другие базы данных, числовые характеристики объектов, номенклатурные названия объектов и т. п.), неструктурированные текстовые данные (текстовые описания, цитаты из статей и энциклопедий) и частично формализованные текстовые данные (описания на ограниченном подмножестве языка с использованием контролируемых словарей) [5, 6].

Наряду с задачей извлечения фактов, соответствующих заранее заданной онтологии, для некоторых областей актуальна задача определения онтологической структуры и извлечения самих онтологических элементов. В настоящей статье предложен метод преобразования частично структурированных текстовых описаний в онтологии, основанный на использовании гетерогенных частотных списков и семантически ориентированных КС (СОКС) грамматик.

Для иллюстрации работы метода выбраны краткие аннотации статей, используемые в одной из баз данных (см. подразд. 2.4). Приведена последовательность действий по преобразованию аннотаций в онтологическое представление, дана оценка применимости метода в выбранном примере. В настоящее время указанные краткие аннотации заполняются кураторами вручную, но затем автоматически посредством простых шаблонов по ним определяется уровень доверия к записи в базе данных. Приведение таких аннотаций к онтологическому представлению является необходимым первым ша-

*Работа выполнена при частичной поддержке РФФИ (проект 15-07-09306).

[1]НИУ Высшая школа экономики, dalexeyevsky@hse.ru

гом для последующего автоматического построения аннотаций по тексту статьи.

2 Контекст работы

2.1 Специфика обработки биологических текстов

Медицинская и биологическая тематика текстов привносит особенности во многие этапы их обработки. В значительной мере именно это и обусловливает выделение bionlp как отдельной предметной области.

Один из часто используемых шагов обработки текстов — идентификация фрагментов текста, соответствующих известным сущностям в базах данных, так называемое извлечение именованных сущностей.

В биологических текстах этот шаг обработки усложнен несколькими обстоятельствами:

- именованная сущность может являться лишь частью слова, например в предложении «The acid-promoted expression of the PmrD protein was phoPQ-dependent, which is in agreement with the fact that PhoP is the only known direct transcriptional activator of pmrD (Kox *et al.*, 2000)» в слове phoPQ-dependent выделяют двухбелковый комплекс «phoPQ», состоящий из белка «phoP» и белка «Q»;
- некоторые сущности, такие как белки или химические соединения, имеют множество синонимичных названий, при этом в текстах могут использоваться не полные названия, а их сокращения, смысл которых возможно восстановить лишь из контекста. Например, белок[1], имеющий в базе данных названия «Sweet protein mabinlin-2», «Mabinlin II», «MAB II», «Sweet protein mabinlin-2 chain A», «Sweet protein mabinlin-2 chain B», может встречаться в статьях как «heat-stable sweet protein, mabinlin-II», «mabinlin» (в пределах текста одной статьи это название может в разных контекстах обозначать как название класса белков, так и конкретный белок), «Cm-MaIIA» (обозначение одной цепочки модифицированного белка, введенное в статье) [7];
- для некоторых сущностей после определения их названия необходимо точнее идентифицировать сущность, о которой идет речь: например, одно и то же название белка может иметь несколько аллелей в одном организме, белок может различаться или не различаться в зависимости от ткани, для которой проводился эксперимент, одно имя могут иметь схожие, но различные белки из разных организмов. Для каждого белка, имеющего то же имя, имеется отдельная запись в базе данных, и необходимо определить, о какой именно записи идет речь.

2.2 Задача построения онтологий

В литературе встречаются разнообразные определения понятия онтологии в зависимости от темы и специфики выбранной задачи. Встречающиеся определения этого понятия в контексте извлечения фактов описывают способы представления знаний, как правило, состоящие из описаний сущностей, их свойств, классификации, связей между ними и логических правил пополнения их свойств и связей [8].

Выделяют онтологии, построенные априори путем логической классификации и лексические, в которых отражаются семантические связи между языковыми единицами [9]. Они обладают разными свойствами: априорная точнее отражает предметную область и позволяет применять богатые механизмы логического вывода, в то время как сущности лексических онтологий, как правило, проще выделять в тексте. В связи с этим одна из часто возникающих задач состоит в установлении соответствия между сущностями лексической и априорной онтологии [10].

Подход, предлагаемый в настоящей статье, позволяет построить онтологию, занимающую промежуточное положение. Такая онтология строится частично по базам данных как априорная, частично по корпусу текстов как лексическая. Это определяет ее главное достоинство: она содержит как ссылки на конкретные сущности из базы данных, так и их текстовое представление.

2.3 Семантически-ориентированные контекстно-свободные грамматики

Для настоящей работы в качестве формализма для описания синтаксической структуры предложения были выбраны КС-грамматики. Контекстно-свободная грамматика — это способ описания структуры предложения в виде иерархии составляющих частей [11]. Дадим ей формальное определение.

Определение 1. Контекстно-свободной грамматикой называется четверка $G = (V, \Sigma, R, S)$, где V — конечное множество нетерминальных символов; Σ — конечное множество терминальных символов; $R \subset \{V \times (V \cup \Sigma)^*\}$ — множество правил вывода вида $v \to a_1 a_2 \cdots$, где $v \in V$, $a_i \in V \cup \Sigma$; $S \in V$ — начальный символ.

[1] UniProt AC P30233, http://www.uniprot.org/uniprot/P30233.

Формальным определением для описания онтологии в настоящей работе было выбрано следующее: онтология — это ориентированный граф, в котором каждая вершина и каждое ребро сопровождаются пометой. С помощью пометы множество вершин делится на вершины-классы и вершины-экземпляры. Пометы на ребрах устанавливают тип отношений, в которых находятся две выбранные вершины. Приведем более формализованное определение.

Определение 2. Онтологией называется пара $O = (G_O, L_O)$ из ориентированного графа и меток к нему. В свою очередь, граф $G_O = (E_O, R_O)$ состоит из множества вершин E_O, называемого множеством сущностей, и множества ребер $R_O \subset E_O \times E_O$, называемого множеством отношений; метки $L_O = (T_E, T_R, L_E, L_R)$ задаются алфавитом возможных меток для вершин T_E, алфавитом возможных меток для ребер T_R, отображением $L_E : E_O \to T_E$ вершины на ее метку и отображением $L_R : R_O \to T_R$ ребра на его метку.

Одно из свойств предлагаемого в настоящей работе алгоритма состоит в простоте выделения онтологических фактов из деревьев синтаксического разбора. Такой алгоритм требует введения нового понятия: семантически ориентированной КС-грамматики. Контекстно-свободная грамматика является семантически ориентированной для данной онтологии, если часть ее правил описывает сущности и отношения в онтологии. Предлагается следующее определение.

Определение 3. Семантически ориентированной КС-грамматикой называется тройка $S = (G, O, M)$ из КС-грамматики $G = (V, \Sigma, \ldots)$, онтологии $O = ((E - O, R_O), (T_E, T_R, L_E, L_R))$ и отображения M между ними. Отображение $M = (M_E, M_R)$ состоит из отображения $M_E \subset (\Sigma \cup V, E_O)$, где $\forall (v, e), (v', e') \in M_E : v = v' \Leftrightarrow e = e'$; символов грамматики на вершины онтологии и отображения $M_R \subset (V, L_R)$, где $\forall (v, r), (v', r') \in M_R : v = v' \Leftrightarrow r = r'$.

Терминальный символ грамматики может быть отображен на вершину-класс или вершину-экземпляр либо не использоваться в онтологии. В последнем случае терминальный символ будем называть синтаксическим по последнему этапу обработки текста, в котором он используется. Нетерминальный символ может быть отображен на вершину-класс, метку ребра (тип отношения), в том числе одновременно, либо не использоваться в онтологии. Как и в случае с терминальными вершинами, в последнем случае такой нетерминал будет называться синтаксическим.

2.4 База данных UniProt

Материалом для разработки и тестирования предлагаемой процедуры построения онтологий послужила свободно распространяемая база UniProt [6].

UniProt является хранилищем аминокислотных последовательностей белков наряду с их краткими описаниями. База содержит ссылки на другие базы данных, посвященные исследованиям белков специфическими методами. Кроме того, частью описания белка в базе является список литературы, описывающей белок.

Для каждого белка база содержит:

- описание его аминокислотной последовательности (поле « » — два пробела);
- обозначения белка согласно различным номенклатурам (поля «DE» и «GN»);
- идентификаторы в различных биологических базах данных самого белка (поля «ID», «AC» и «DR») и его носителя («OC» и «OX»);
- биологический контекст белка (поля «OS», «OG» и «OH»);
- библиографическую информацию (поля «RN», «RP», «RC», «RX», «RG», «RA», «RT» и «RL»);
- описания известных свойств белка: текстовые (поле «CC»), на ограниченном подмножестве английского языка (поля «RP» и «KW»), формализованные (поле «FT»);
- уровень доверия данной записи (поле «PE»);
- прочую служебную информацию (поля «DT» и «SQ»).

Значение «PE» уровня достоверности записи базы данных определяется тем, какими экспериментальными средствами установлен факт существования белка и его соответствия представленным данным. Описания того, какие экспериментальные средства применялись к белку, хранятся в базе в полях «CC», «RP» и «KW», для некоторых методов факт их применения можно опознать по свойствам в поле «FT». В базе заданы формальные правила выставления значения уровня доверия («PE») в зависимости от наличия некоторых шаблонных выражений в этих полях [12].

База данных UniProt состоит из двух частей: UniProt/TrEMBL, пополняемой полностью автоматически, и UniProt/Swiss-Prot, пополняемой кураторами вручную на основе материалов UniProt/TrEMBL, существующих публикаций и материалов других баз данных. Поля «KW» и «FT» получают начальные значения автоматически в базе

данных UniProt/TrEMBL, хотя затем могут быть изменены в процессе курирования. Поля «CC» и «RP» заполняются только кураторами вручную.

Этими обстоятельствами обусловлено то, что в качестве материала для настоящей работы был собран корпус предложений в поле «RP» из базы данных UniProt/Swiss-Prot.

3 Материалы и методы

Материалом исследований послужили данные из базы UniProt/Swiss-Prot версии 2015_01. Собранный корпус уникальных атомарных причин цитирования в поле «RP» имеет размер 173 212 предложений.

Всего база UniProt/Swiss-Prot 2015_1 содержит:

- 547 357 записей (одна запись описывает один белок);
- 1 092 817 ссылок на литературу и, соответственно, всего предложений в поле «RP», включая повторяющиеся, среди них;
- 179 616 уникальных предложений в поле «RP», состоящих из одного или нескольких атомарных описаний (в свою очередь также включающих повторения);
- 173 212 уникальных атомарных описаний.

Для дальнейшей работы использовался описанный корпус уникальных атомарных описаний, с тем чтобы наиболее полно покрыть максимально возможное количество особых случаев в языке.

3.1 Особенности предложений в поле «RP»

Предложения в поле «RP» являются полуструктурированными, так как несут признаки как структурированных, так и естественных языковых данных. Предложения порождаются кураторами. Для них не существует формализованного описания структуры или инструмента для валидации. Существует находящаяся на данный момент в стадии разработки инициатива по унификации представления названий различных классов сущностей в таких предложениях с помощью внедрения контролируемых словарей [13]. Наряду с этим для кураторов существует инструкция по заполнению, включающая в себя примеры представления большого числа типов фактов [14].

Каждое атомарное описание является именной группой. Важно заметить, что для краткости описания не содержат упоминаний описываемого объекта. Объект описания устанавливается из факта принадлежности описания записи в базе данных (рис. 1).

3.2 Словники

Для извлечения именованных сущностей и насыщения списка примитивных фактов были использованы словники.

Словник имен белков был построен по значениям в поле «DE», подразделам RecName и AltName

```
X-RAY CRYSTALLOGRAPHY (1.80 ANGSTROMS) OF 44-480 OF WILD-TYPE AND MUTANTS TYR-118;
ARG-168 AND ALA-309 IN ACTIVE AND RESTING STATES AND IN COMPLEX WITH PEPTIDE SUBSTRATE,
FUNCTION, CATALYTIC ACTIVITY, ENZYME REGULATION, SUBSTRATE SPECIFICITY, SUBUNIT, DOMAIN,
PROTEOLYTIC AUTO-CLEAVAGE, ACTIVE SITES, SITES, DISRUPTION PHENOTYPE, MUTAGENESIS
OF VAL-118; ARG-168; SER-309 AND GLN-338, AND PDZ DOMAIN DELETION MUTANT.
```

(*а*)

```
> X-RAY CRYSTALLOGRAPHY (1.80 ANGSTROMS) OF 44-480 OF WILD-TYPE AND MUTANTS TYR-118;
ARG-168 AND ALA-309 IN ACTIVE AND RESTING STATES AND IN COMPLEX WITH PEPTIDE SUBSTRATE
> FUNCTION
> CATALYTIC ACTIVITY
> ENZYME REGULATION
> SUBSTRATE SPECIFICITY
> SUBUNIT
> DOMAIN
> PROTEOLYTIC AUTO-CLEAVAGE
> ACTIVE SITES
> SITES
> DISRUPTION PHENOTYPE
  > MUTAGENESIS OF VAL-118; ARG-168; SER-309
```

(*б*)

Рис. 1 Примеры описаний в поле «RP»: (*а*) полное описание, (*б*) атомарные факты

базы данных UniProt и полям Full, Short, Name, Synonyms в них. Суммарный объем словника составил 308 370 словосочетаний.

Некоторые названия белков совпадают с общезначимыми словами английского языка. Для того чтобы исключить ошибки второго рода в таких случаях, из словника имен белков были удалены все слова, являющиеся словами английского языка. Для этой фильтрации был использован словник общеупотребительной лексики американского английского языка [15] объемом 99 171 словоформа, содержащий все падежные формы слов.

3.3 Методы

Для сегментации текста на слова был использован токенизатор, сохраняющий все знаки пунктуации, включая дефисы, как отдельные токены. Токенизатор был разработан на основе пакета re языка Python [16].

Для построения СОКС-грамматик был использован парсер Эрли с проходом снизу вверх из пакета nltk [17] для языка Python.

Для построения частотных списков использовались средства shell script и сопутствующие программы текстовой обработки из базового комплекта операционной системы GNU: cat, sort, uniq, grep, sed, head, tail, less.

4 Алгоритм разработки онтологии с помощью контекстно-свободных грамматик

Задача алгоритма состоит в том, чтобы за наименьшее время преобразовать наибольшую часть заранее заданного корпуса фактов, представленного в виде полуструктурированных текстовых данных, в онтологическое представление.

Основная идея алгоритма состоит в итерационном применении и пополнении КС-грамматики. После каждого применения грамматики предложения корпуса преобразуются в гетерогенную последовательность из токенов и нетерминальных символов грамматики. Полученный корпус гетерогенных последовательностей используется для того, чтобы определить, какое правило нужно добавить в корпус для получения наибольшего прироста количества предложений, разбор которых доведен до нетерминала-вершины.

При построении КС-грамматики терминальными символами грамматики являются токены из корпуса, множество нетерминальных символов является объединением из множества типов сущностей в онтологии и множества вспомогательных нетерминальных символов.

Входными данными для построения онтологии являются:

- корпус разбираемых текстов;
- базы данных и словники, позволяющие выделять в тексте релевантные именованные сущности.

Алгоритм состоит из пяти шагов:

1. Подготовить начальную грамматику.
2. Применить к корпусу текстов правила грамматики, заменив покрытые правилами фрагменты текста соответствующими нетерминалами.
3. Оценить покрытие корпуса текстов нетерминалами и выбрать метод пополнения грамматики (см. ниже).
4. Пополнить грамматику новым правилом (см. ниже).
5. Перейти на шаг 2.

Начальная грамматика содержит заранее определенный нетерминал-вершину; множество нетерминальных символов, состоящее только из нетерминала-вершины; множество терминальных символов, совпадающее с множеством токенов корпуса; множество правил, являющееся пустым.

Оценка покрытия может производиться одним из двух способов.

1. Выбрать из корпуса случайным образом 100 предложений, среди них найти наиболее частую синтаксическую конструкцию или тип именованной сущности, который еще не покрыт правилами грамматики.
2. Построить частотный список предложений, выбрать из них наиболее частое, для которого может быть написано правило СОКС-грамматики, не имеющее ложных срабатываний.

В результате оценки должно быть порождено правило одного из трех видов:

(1) синтаксическое упрощение;
(2) создание или пополнение газетира;
(3) семантическое правило.

Синтаксическими упрощениями называются правила грамматики, которые не отображаются в результирующей онтологии, но обобщают однородные конструкции и упрощают последующее расширение грамматики.

К этому типу правил относятся, например,

```
and -> 'AND' | ',' | ',' 'AND' | ';' | ';' 'AND'
и
det -> 'A' | 'AN' | 'THE'
```

Необходимость создания или пополнения газетира возникает в тех случаях, когда наиболее частым

не покрытым нетерминалами явлением в корпусе оказываются названия именованных сущностей, принадлежащие к одному классу.

Например, в предложении

```
PALMITOYLATION AT CYS-11, AND MUTAGENESIS
 OF SER-2; ARG-6 AND CYS-11.
```

четыре раза встречаются названия конкретных аминокислотных остатков в белке, представленные как название аминокислоты и номер ее позиции, записанные через дефис. В тот момент, когда в корпусе такие случаи становятся самыми частотными из неразобранных, необходимо пополнить газетир списком названий аминокислот.

Третий вариант действий состоит в том, чтобы пополнить СОКС-грамматику *семантическим правилом*. Для этого необходимо выявить самую частотную конструкцию, такую что в ней нет токенов, которые могли бы войти в именованную сущность; в ней нет лексики, играющей исключительно синтаксическую роль; она не сведена к нетерминалу, являющемуся вершиной онтологии.

Такая конструкция может являться предложением целиком, в этом случае из нее будет образовано новое правило для СОКС-грамматики, в левой части которого будет находиться вершина онтологии:

```
feature -> 'STRUCTURE' 'BY' method
feature -> modification 'AT' range
```

Пример предложений, использующих приведенный фрагмент грамматики:

```
STRUCTURE BY ELECTRON MICROSCOPY
 (9.4 ANGSTROMS).
PHOSPHOPANTETHEINYLATION AT SER-37.
```

Такая конструкция может одновременно быть предложением и сводиться к нетерминалу, который при этом не является вершиной онтологии, например:

```
feature -> method
feature -> interaction
```

Пример предложений, использующих приведенный фрагмент грамматики:

```
IDENTIFICATION BY MASS SPECTROMETRY.
CALMODULIN-BINDING.
```

Такая конструкция может являться частью предложения, в этом случае нетерминал в левой части правила не будет являться вершиной онтологии, например:

```
interaction -> interaction 'WITH' protein
```

Пример предложений, использующих приведенный фрагмент грамматики:

```
INTERACTION WITH MPK6
```

4.1 Преобразование деревьев синтаксического разбора в онтологическое представление данных

В результате работы СОКС-парсера предложения исходного текста преобразуются в деревья синтаксического разбора. Например, предложение

```
FUNCTION, AND INTERACTION WITH RBM8A;
NXF1 AND THE EXON JUNCTION COMPLEX.
```

после разбора преобразуется в следующее дерево:

```
(description
 (feature
  (feature FUNCTION)
  (and , AND)
  (feature
    (interaction
     (interaction INTERACTION)
     WITH
     (protein
      (protein (protein RBM8A) (and ;)
       (protein NXF1))
      (and AND)
      (protein (det THE)
       (protein (words EXON JUNCTION)
        COMPLEX))))))
 .)
```

Такое дерево содержит набор связей, которые в точности соответствуют онтологическим. Помимо таких связей в дереве имеются связи и узлы, имеющие синтаксическую роль (сочетание и детерминанты). Кроме того, связи, отвечающие за сочетание, представлены здесь не как однородные связи внутри одного объекта, а как вложенная рекурсивная цепочка связей.

Для преобразования деревьев такого вида в онтологические факты необходимо:

- заменить текстовое описание именованных сущностей на идентификатор базы данных (например, заменить `RBM8A` на `Q9Y5S9; RBM8A` является названием белка, общего для многих видов, база данных UniProt содержит 64 белка с идентичным названием, текст данного предложения получен из описания белка, извлеченного из h.sapiens; следовательно, нас интересуют и белки `RBM8A` только из h.sapiens, такой белок только один);
- нормализовать числовые значения (например, заменить на 4.2 поддерево

  ```
  (float (digits 4) . (digits 2)));
  ```

– раскрыть случаи сочетания необходимым для данного онтологического класса способом;
– удалить нетерминалы, играющие синтаксическую роль (например, поддерево: (det THE));
– в случаях, когда несколько аргументов обозначаются одним и тем же нетерминалом, дать аргументам различные имена;
– преобразовать правила грамматики в объявление онтологических классов, отношений класс–подкласс и объявлений свойств;
– преобразовать газетиры в объявление онтологических индивидов и отношений класс–индивид;
– преобразовать дерево разбора в объявление набора онтологических индивидов, объявление их отношения к соответствующим онтологическим классам и отношений часть–целое и атрибут для этих индивидов.

Для приведенного примера фрагмент грамматики (вместе с вставленными в него для наглядности фрагментами необходимых газетиров) выглядит следующим образом:

```
description -> feature '.'
feature -> feature and feature
feature -> interaction
feature -> 'FUNCTION'
interaction -> interaction 'WITH' protein
interaction -> 'INTERACTION'
protein -> protein and protein
protein -> words 'COMPLEX'
protein -> det protein
protein -> Q9Y5S9 | Q9UBU9
and -> 'AND' | ',' | ',' 'AND' | ';' |
 ';' 'AND'
```

Он однозначным образом преобразуется в набор определений (здесь авторы используют OWL2 functional notation [18]:

```
Declaration(Class(:Description))
Declaration(Class(:Feature))
Declaration(Class(:Function))
Declaration(Class(:Interaction))
Declaration(Class(:Protein))
Declaration(ObjectProperty
 (:InteractionWith))
ObjectPropertyDomain(:InteractionWith
 :Protein)

SubClassOf(:Feature :Description)
SubClassOf(:Interaction :Feature)
SubClassOf(:Function :Feature)

Declaration(NamedIndividual(:Q9Y5S9))
ClassAssertion(:Protein :Q9Y5S9)
Declaration(NamedIndividual(:Q9UBU9))
ClassAssertion(:Protein :Q9UBU9)
```

При этом приведенное описание трансформируется в набор онтологических объектов:

```
Declaration(NamedIndividual(:function1))
ClassAssertion(:Function :function1)

Declaration(NamedIndividual(:interaction1))
ClassAssertion(:Interaction :interaction1)

ObjectPropertyAssertion(:InteractionWith
:interaction1 :Q9Y5S9)
ObjectPropertyAssertion(:InteractionWith
:interaction1 :Q9UBU9)
```

```
> [X - RAY CRYSTALLOGRAPHY [1 . 80 ANGSTROMS]resolution OF [44 - 480]range
OF [WILD - TYPE AND MUTANTS [TYR - 118 ; ARG - 168 AND ALA - 309]range]variant
IN [ACTIVE AND RESTING STATES]form AND IN [COMPLEX WITH [PEPTIDE SUBSTRATE]chemical]chemenv]feature
> [FUNCTION]feature
> [CATALYTIC ACTIVITY]feature
> [ENZYME REGULATION]feature
> [SUBSTRATE SPECIFICITY]feature
> [SUBUNIT]feature
> [DOMAIN]feature
> [PROTEOLYTIC AUTO - CLEAVAGE]feature
> [ACTIVE SITES]feature
> [SITES]feature
> [DISRUPTION PHENOTYPE]feature
  > [MUTAGENESIS OF [VAL - 118 ; ARG - 168 ; SER - 309 AND GLN - 338]range]feature AND
```

Рис. 2 Разбор описания

```
Declaration(NamedIndividual(:protein1))
AnnotationAssertion( rdfs:comment
:protein1 "EXON JUNCTION COMPLEX" )
ClassAssertion(:Protein :protein1)
```

5 Результаты и обсуждение

В ходе работы была построена СОКС-грамматика, содержащая 179 правил (рис. 2).

5.1 Оценка покрытия

Для оценки была выбрана случайным образом тестовая выборка из 100 предложений, 96 из них уникальные. Тестовая выборка содержит 205 атомарных причин цитирования, 135 из них уникальные.

Задача построения газетиров находится за пределами настоящей работы, поэтому в тестовой выборке перед тестированием сущности, входящие в газетиры, были вручную заменены на соответствующие им нетерминалы. Дополнительно в грамматику были добавлены правила, позволяющие обрабатывать такие предобработанные входные данные.

В тех случаях, где в тестирующей выборке одна и та же сущность могла быть описана более длинной или более короткой цепочкой, использовалась более короткая цепочка. Таким образом вручную были размечены классы: белок, вещество, болезнь, лекарство, химическая модификация.

Полученные в результате тестирования оценки покрытия представлены в таблице.

Результаты тестирования покрытия

Тестирование	Доля
Все атомарные причины цитирования	73%
Уникальные атомарные причины цитирования	43%
Все предложения	54%
Уникальные предложения	52%

Следует обратить внимание на значительный (в 1,7 раза) прирост покрытия при отключении процедуры удаления дубликатов из корпуса атомарных причин цитирования. Это является косвенным следствием большого числа дубликатов, которые, в свою очередь, являются следствием ограниченности выбранного языка (он использует только именные группы) и его лексической ограниченности (кураторы следуют инструкции, регламентирующей используемую лексику). Такие ограничения приводят к значительному объему дублирования в корпусе. Это дает возможность при меньшем числе правил в грамматике добиваться более высокого покрытия корпуса, что и предложено в настоящей статье.

Очевидно, что более сложные конструкции обладают бóльшим разнообразием и, следовательно, меньшей степенью дублирования, что и продемонстрировано на оценке покрытия полных предложений. Таким образом, для более сложных или менее ограниченных языков кажется осмысленным в качестве предобработки выделять наиболее узко лишь такие конструкции, которые имеют сущности, значимые для составляемой онтологии. Для построения онтологий, описывающих объекты и их свойства, такой предобработкой может служить выделение именных групп.

6 Заключение

В работе поставлена актуальная задача разработки новых онтологий на основе корпусных данных и предложен подход к ее решению. Для составления онтологий в работе дано определение и представлен алгоритм составления семантически ориентированных КС-грамматик. Важным аспектом подхода является использование в качестве материала для построения онтологии корпуса предложений на ограниченном подмножестве естественного языка.

Алгоритм опробован для текстов именных групп ограниченного языка, используемого в базе UniProt для описания причин цитирования статьи, в результате чего составлена грамматика и разработан синтаксический анализатор таких причин цитирования.

Литература

1. *Doğan R. I., Leaman R., Lu Zh.* Ncbi disease corpus: A resource for disease name recognition and concept normalization // J. Biomed. Inform., 2014. Vol. 47. P. 1–10. doi: 10.1016/j.jbi.2013.12.006.
2. *Li Ch., Song R., Liakata M., Vlachos A., Seneff S., Zhang X.* Using word embedding for bio-event extraction // 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015) Proceedings. — Beijing, China: ACL, 2015. P. 121–126.
3. *Kim J., Nguyen N., Wang Yu., Tsujii J., Takagi T., Yonezawa A.* The genia event and protein coreference tasks of the BioNLP shared task 2011 // BMC Bioinformatics, 2012. Vol. 13. Suppl. 11:S1. doi:10.1186/1471-2105-13-S11-S1.
4. *Nédellec C., Bossy R., Kim Ji., Kim Ju., Ohta To., Pyysalo S., Zweigenbaum P.* Overview of BioNLP shared task 2013 // BioNLP Shared Task 2013 Workshop (BioNLP-ST 2013) Proceedings. — ACL, 2013. P. 1–7.
5. *Tanabe M., Kanehisa M.* Unit 1–12 using the KEGG database resource // Current protocols in bioinformatics. — John Wiley & Sons, Inc., 2012. P. 1.12.1–1.12.43. doi: 10.1002/0471250953.bi0112s38.

6. The UniProt Consortium. UniProt: A hub for protein information // Nucleic Acids Res., 2015. Vol. 43. P. D204–D212. doi: 10.1093/nar/gku989.
7. *Tonkon M. J., Miller R. R., DeMaria A. N., Vismara L. A., Amsterdam E. A., Mason D. T.* Multifactor evaluation of the determinants of ischemic electrocardiographic response to maximal treadmill testing in coronary disease // Am. J. Med., 1977. Vol. 62. Iss. 3. P. 339–346. doi: 10.1016/0002-9343(77)90830-0.
8. *Giaretta P., Guarino N.* Ontologies and knowledge bases towards a terminological clarification // Towards very large knowledge bases. — Amsterdam: IOS Press. P. 25–32.
9. *Jones D., Bench-Capon T., Visser P.* Methodologies for ontology development // IT&KNOWS Conference, XV IFIP World Computer Congress Proceedings. — Budapest, 1998.
10. *Reed S. L., Lenat D. B.* Mapping ontologies into Cyc // AAAI 2002 Conference Workshop on Ontologies For The Semantic Web, 2002. P. 1–6.
11. *Хомский Н.* Аспекты теории синтаксиса / Пер. В. А. Звегинцева. — М.: Изд-во Моск. ун-та, 1972. 258 с. (*Chomsky N.* Aspects of the theory of Syntax. — MIT Press, 1969. 261 p.)
12. Criteria used to assign the pe level of entries. http://www.uniprot.org/docs/pe_criteria.
13. Controlled vocabulary. http://www.uniprot.org/help/controlled_vocabulary.
14. *UniProt Consortium.* UniProt manual curation sop. http://www.uniprot.org/docs/sop_manual_curation.pdf.
15. *Beale A.* Spell checker oriented word lists. 1999–2015. http://wordlist.aspell.net/12dicts-readme.
16. *Van Rossum G.* Python programming language // USENIX Annual Technical Conference, 2007.
17. *Bird S., Klein E., Loper E.* Natural language processing with Python. — O'Reilly Media, 2009. 512 p.
18. *Horridge M., Patel-Schneider P. F.* OWL 2 Web Ontology Language Manchester Syntax. — 2nd ed. — W3C Working Group Note, 2009. http://www.w3.org/TR/owl2-manchester-syntax.

Поступила в редакцию 23.09.15

BioNLP ONTOLOGY EXTRACTION FROM A RESTRICTED LANGUAGE CORPUS WITH CONTEXT-FREE GRAMMARS

D. A. Alexeyevsky

National Research University Higher School of Economics; 20 Myasnitskaya Str., Moscow 101000, Russian Federation

Abstract: BioNLP is an emerging area of NLP that brings new challenging objects for language processing and new valuable resources for bioinformatics and medicine. One notable task in BioNLP is creating de-novo ontologies. This is generally a tedious process; however, in some cases, it is possible to automate it to some extent. One such case is when a corpus of texts in a restricted subset of natural language is available. This paper presents a simple approach to automate ontology creation in such cases. The approach is aimed to simplify mapping of entities in natural texts to predefined ontologies wherever possible. The paper discusses which properties of the corpus enable the approach presented.

Keywords: BioNLP; ontology creation; context-free grammar

DOI: 10.14357/19922264160111

Acknowledgments

The work was partly supported by the Russian Foundation for Basic Research (project 15-07-09306).

References

1. Doğan, R. I., R. Leaman, and Zh. Lu. 2014. Ncbi disease corpus: A resource for disease name and concept normalization. *J. Biomed. Inform.* 47:1–10. doi: 10.1016/j.jbi.2013.12.006.
2. Li, Ch., R. Song, M. Liakata, A. Vlachos, S. Seneff, and X. Zhang. 2015. Using word embedding for bio-event extraction. *2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015) Proceedings*. Beijing, China: ACL. 121–126.
3. Kim, J., N. Nguyen, Yu. Wang, J. Tsujii, T. Takagi, and A. Yonezawa. 2012. The genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics* 13(Suppl. 11:S1). doi: 10.1186/1471-2105-13-S11-S1.
4. Nédellec, C., R. Bossy, Ji. Kim, Ju. Kim, To. Ohta, S. Pyysalo, and P. Zweigenbaum. 2013. Overview of

BioNLP shared task 2013. *BioNLP Shared Task 2013 Workshop (BioNLP-ST 2013) Proceedings*. ACL. 1–7.

5. Tanabe, M., and M. Kanehisa. 2012. Unit 1–12 using the KEGG database resource. *Current protocols in bioinformatics*. John Wiley & Sons, Inc. 1.12.1–1.12.43. doi: 10.1002/0471250953.bi0112s38.
6. The UniProt Consortium. 2015. Uniprot: A hub for protein information. *Nucleic Acids Res*. 43:D204–D212. doi:10.1093/nar/gku989.
7. Tonkon, M. J., R. R. Miller, A. N. DeMaria, L. A. Vismara, E. A. Amsterdam, and D. T. Mason. 1977. Multifactor evaluation of the determinants of ischemic electrocardiographic response to maximal treadmill testing in coronary disease. *Am. J. Med*. 62(3):339–346. doi: 10.1016/0002-9343(77)90830-0.
8. Giaretta, P., and N. Guarino. 1995. Ontologies and knowledge bases towards a terminological clarification. *Towards very large knowledge bases*. Amsterdam: IOS Press. 25–32.
9. Jones, D., T. Bench-Capon, and P. Visser. 1998. Methodologies for ontology development. *IT&KNOWS Conference, XV IFIP World Computer Congress Proceedings*. Budapest. 62–75.
10. Reed, S. L., and D. B. Lenat. 2002. Mapping ontologies into Cyc. *AAAI 2002 Conference Workshop on Ontologies For The Semantic Web*. 1–6.
11. Chomsky, N. 1969. *Aspects of the theory of Syntax*. MIT Press. 261 p.
12. Criteria used to assign the pe level of entries. Available at: http://www.uniprot. org/docs/pe_criteria (accessed January 21, 2016).
13. Controlled vocabulary. Available at: http://www.uniprot.org/help/controlled_vocabulary (accessed January 21, 2016).
14. UniProt Consortium. Uniprot manual curation sop. Available at: http://www.uniprot.org/docs/sop_manual_curation.pdf (accessed January 21, 2016).
15. Beale, A. 1999–2015. *Spell checker oriented word lists*. Available at: http://wordlist.aspell.net/12dicts-readme (accessed January 21, 2016).
16. Van Rossum, G. 2007. Python programming language. *USENIX Annual Technical Conference*.
17. Bird, S., E. Klein, and E. Loper. 2009. *Natural language processing with Python*. O'Reilly Media. 512 p.
18. Horridge, M., and P. F. Patel-Schneider. 2009. OWL 2 Web Ontology Language Manchester Syntax. 2nd ed. W3C Working Group Note. Available at: http://www.w3.org/TR/owl2-manchester-syntax (accessed January 21, 2016).

Received September 23, 2015

Contributor

Alexeyevsky Daniil A. (b. 1983) — PhD student, Faculty of Humanities, National Research University Higher School of Economics; 20 Myasnitskaya Str., Moscow 101000, Russian Federation; dalexeyevsky@hse.ru