НАПИСАНИЕ ОБЩИХ ПРАВИЛ ДЛЯ ИЗВЛЕЧЕНИЯ ПРЕДЛОЖНЫХ ИМЕННЫХ ГРУПП В ТОМИТА-ПАРСЕРЕ

Работа имеет **целью** написание правил для извлечения предложных именных групп в Томита-парсере.

Инструмент работы

Извлечение предложных именных конструкций выполнялось с помощью Томита-парсера.

Томита-парсер – разработка компании «Яндекс». Парсер служит для извлечения структурированных данных из текста на естественном языке. Томита-парсер извлекает сущности при помощи контекстно-свободной грамматики и словаря ключевых слов.

Томита применяется для извлечения таких фактов, как даты, родственные связи, адреса, должности и т.п. Парсер используется в сервисах «Яндекс.Новости» и «Яндекс.Работа».

Парсер включает в себя три стандартных лингвистических процессора: токенизатор (для разбиения на слова), сегментатор (для разбиения на предложения) и морфологический анализатор (*mystem*, разработка компании «Яндекс»).

Для извлечения фактов из текста парсеру необходим файл конфигурации с расширением *.proto*, где указаны: 1) используемый словарь ключевых слов (газеттир) с расширением *.gzt* (обязателен всегда); 2) грамматика с множеством правил, описывающих синтаксическую структуру выделяемых цепочек, с расширением *.cxx*; 3) файл с описанием фактов с расширением *.proto*.

Правила по извлечению цепочек пишутся на языке контекстно-свободных грамматик. Для формирования цепочки-правила необходимо в левой части указать нетерминал (обозначения конечных конструкций), а в правой части - его терминалы (то, из чего строится конструкция). Таким образом, простейшее правило формируется следующим образом:

Нетерминал -> Терминал;

Нетерминалы могут выступать в качестве терминалов. Нетерминал, ни разу не встречающийся в правой части, является вершиной грамматики.

В файле фактов описываются их названия и подполя (названия, номер).

Для извлечения фактов в грамматике прописываются правила интерпретации имеющихся цепочек как фактов соответствующего типа.

Лингвистическое описание целевых конструкций

Объектом извлечения стали предложные именные группы.

В данной работе именной группой считается словосочетание, в котором вершиной является имя существительное. Именная группа может состоять как из одного существительного, так и из существительного и зависимых от него слов и конструкций: прилагательных, модифицирующих их наречий, причастий и причастных оборотов, клауз и предложных групп.

Предложной именной группой считается сочетание предлога и именной группы. Так как в состав именной группы также может входить сочетание предлога и существительного, такие конструкции могут рассматриваться как две отдельные предложные именные группы или как одна предложная именная группа, в которой выделяются главная и зависимая части. Так, например, в описании синтаксических конструкций рабочей группы АОТ предложная именная группа трактуется как цепочка «предлог+ИГ», а описание предложных именных групп проекта OpenCorpora содержит указание на вложенные в них предложные группы.

С целью оптимизации вывода в данной работе была принята упрощённая трактовка предложных именных групп: предложная именная группа есть сочетание предлога и именной группы, не содержащей в себе предложную группу.

Параметры пользовательского корпуса

В качестве пользовательского корпуса использовалась выборка контекстов с целевой конструкцией из Хельсинского Аннотированного Корпуса (ХАНКО). Для поиска интересующих контекстов был задан параметр *([Prep])*, что соответствует поиску по контекстам, содержащим в себе предлог. В ХАНКО было обнаружено 4484 предложения с соответствующим параметром. В выборку попали первые 1000 предложений, многие из которых содержат более чем одну предложную именную группу. Выборка представлена в файле *input_hanco.txt*.

С целью исследовать работу парсера по извлечению предложных именных групп с отдельными особенностями строения был также сформирован

небольшой рабочий корпус, использовавшийся на этапе первичного тестирования. Туда вошли контексты, описанные на сайтах группы АОТ, проекта OpenCorpora, отдельные контексты из основного пользовательского корпуса, содержимое файла *input.txt* из папки *examples/sample*, предоставленной разработчиками Томита-парсера, а также искусственно созданные контексты. Тестовый корпус представлен в файле *input_test.txt*.

Правила для целевых групп

Поскольку задание подразумевало как извлечение всех предложных именных групп из текста, так и сортировку их по входящим в их состав предлогам, было принято решение создать две грамматики: грамматика *pp_unsorted.cxx* используется для извлечения целевых конструкций без сортировки, а в грамматике *pp_sorted.cxx* прописаны правила формирования цепочек с отдельными предлогами и их интерпретации как фактов. Для использования той или иной грамматики задействуются файлы *config_unsorted.proto / config_sorted.proto* и *mydic_unsorted.gzt / mydic_sorted.gzt* соответственно. Файл с описанием фактов *facttypes.proto* задействован только во второй грамматике.

Для извлечения предложных именных групп из текста были сформированы следующие правила выделения входящих в их состав именных групп:

- NP -> Word<gram='NUM'>* Adv* (Adj<gnc-agr[3]>) (Comma) (Adj<gnc-agr[2]>) ('и') (Comma) Adj<gnc-agr[1]>* Noun<gnc-agr[1], gnc-agr[2], gnc-agr[3]>;

Соответствует конструкции со следующей структурой:

1) токены с тегом «числительное», встречающиеся ноль или более раз;

2) токены с тегом «наречие», встречающиеся ноль или более раз;

3) согласованный с существительным токен с тегом «прилагательное», встречающийся ноль или один раз;

4) запятая, встречающаяся ноль или один раз;

5) согласованный с существительным токен с тегом «прилагательное», встречающийся ноль или один раз;

6) союз «и» или запятая, встречающиеся ноль или один раз;

7) согласованный с существительным токен с тегом «прилагательное», встречающийся ноль или один раз;

8) токен с тегом «существительное», с которым согласованы предыдущие прилагательные.

Правило работает для конструкций типа *«со ста двадцатью пятью очень ласковыми, пушистыми и мягкими рыжими котами»* (максимальное протестированное наполнение) и их составляющих.

- NP -> Adv* Adj<gnc-agr[1]>* Noun<gnc-agr[1]>* Word<h-reg1, gnc-agr[1]>+;

Правило имеет следующую структуру:

1) токены с тегом «наречие», встречающиеся ноль или более раз;

2) токены с тегом «прилагательное», согласованные с существительным, в количестве ноль или более раз;

3) токены с тегом «существительное», с которыми согласованы предыдущие прилагательные, в количестве ноль или более раз;

4) токены, начинающиеся со строчной буквы, согласованные с предыдущими существительными, в количестве один и более раз.

Правило сформировано для конструкций типа *«с генеральным консулом Ивановым Иваном Ивановичем»*.

- NP -> Adv* Adj<gc-agr[1]> ('и') Adv* Adj<gc-agr[2]> Noun<gc-agr[1], gc-agr[2]>;

Структура правила:

1) токены с тегом «наречие», встречающиеся ноль или более раз;

2) токен с тегом «прилагательное», согласованный с существительным;

3) союз «и» в количестве ноль или один раз;

4) токены с тегом «наречие», встречающиеся ноль или более раз;

5) токен с тегом «прилагательное», согласованный с существительным;

6) токены с тегом «существительное», с которыми согласованы предыдущие прилагательные, в количестве один или более раз.

Правило было задействовано для извлечения конструкций типа *«в совсем забытых и ещё популярных источниках»*, где согласование происходит по роду и падежу, но не по числу.

- NP -> Adv* Adj<c-agr[1], c-agr[2]> Noun<c-agr[1]> 'и' Noun<c-agr[2]>;

Правило имеет следующую структуру:

1) токены с тегом «наречие», встречающиеся ноль или более раз;

2) согласованный с последующими существительными токен с тегом «прилагательное»;

3) токен с тегом «существительное», с которым согласуется прилагательное;

4) союз «и»;

5) токен с тегом «существительное», с которым согласуется прилагательное.

Правило используется для извлечения конструкций типа *«с очень усталыми мамой и папой»*, где согласование происходит по падежу, но не по роду и числу.

- NP -> (AnyWord<wff=/[0-9]+/>) Noun+ AnyWord<wff=/[0-9]+/, cut>;

Структура правила:

1) число, состоящее из цифр от 0 до 9, в количестве ноль или один раз;

2) токены с тегом «существительное», встречающиеся один или более раз;

3) число, состоящее из цифр от 0 до 9, не включающееся в выделяемую конструкцию.

Правило было сформировано с целью очистить вывод от конструкций типа «к командиру 18», «в итоге 20», выделяемые Томитой по причине отсутствия каких-либо тегов у чисел.

- NP -> AnyWord<wff=/[0-9]+/> (Adj<gnc-agr[1]>) Noun<gnc-agr[1]>;

Структура правила:

1) число, состоящее из цифр от 0 до 9;

2) токен с тегом «прилагательное», согласованный с существительным, в количестве ноль или один раз;

3) токен с тегом «существительное», с которым согласуется прилагательное.

Правило нацелено на выделение конструкций типа «в 151 стрелковой дивизии».

Для извлечения всех предложных именных групп, соотносимых с вышеописанными структурами, было написано общее правило:

- PP -> Prep NP;

PP является вершиной грамматики; *Prep* соответствует токену с тегом «предлог»; *NP* соответствует предшествующим правилам выделения именных групп. Правило задействовано в грамматике *pp_unsorted.cxx*.

Для сортировки предложных именных групп в грамматике *pp_sorted.cxx* был создан блок однообразных правил:

- PP_*PREDLOG* -> ‘*предлог*’ NP;

Правила были написаны для 116 предлогов, вошедших в частотный список лемм служебной лексики из «Нового частотного словаря русской лексики» под редакцией О.Н. Ляшевской и С.А. Шарова.

Далее были сформированы правила интерпретации предложных цепочек как фактов:

- PP -> PP_*PREDLOG* interp (PrepPhrase.*PREDLOG*);

PP в данной грамматике является вершиной. Терминал *PP_PREDLOG* интерпретируется в подполе *PREDLOG* факта *PrepPhrase.*

Факт *PrepPhrase* имеет 116 подполей, соответствующих рассматриваемым предлогам, расположенных в алфавитном порядке:

- import "base.proto";
- import "facttypes_base.proto";
- message PrepPhrase: NFactType.TFact

{

optional string ALA = 1;

optional string BEZ = 2;

optional string BEZO = 3;

и т.д. }

Ход экспериментов

Написанные правила были протестированы сначала на рабочем корпусе, а затем на пользовательском корпусе контекстов из ХАНКО. Тестовый корпус пополнялся по ходу работы в случае выявления ошибок извлечения или недостаточного извлечения целевых конструкций.

Ввиду встроенных ограничений на количество интерпретируемых фактов (25 согласно выводу командной строки) далеко не все контексты объёмного пользовательского корпуса были обработаны Томита-парсером при извлечении предложных именных групп с разбивкой по предлогам. Так, парсер игнорирует 1971 найденный контекст и распределяет всего 195 контекстов по полям факта. В сумме это даёт 2166 контекстов, соответствующих правилам. При общем выделении предложных именных групп (без выделения фактов) обнаруживается 2169 контекстов, соответствующих правилам.

Анализ полученных результатов

С полученными результатами можно ознакомиться с помощью файлов *PrettyOutputUnsorted.html* и *PrettyOutputSorted.html,* а также менее удобного для чтения файла *facts_sorted.txt.*

Разница в 3 контекста между выводом общей грамматики и выводом сортирующей грамматики остаётся необъяснённой. Существует вероятность, что данные контексты содержат предлоги, не описанные в сортирующей грамматике.

Парсер справился с выделением достаточно разнообразных по своему наполнению предложных именных групп. Так, в общем выводе встречаются конструкции следующих типов:

- предлог + существительное: *в итоге, по сведениям, о реабилитации*;
- предлог + (местоименное) прилагательное + существительное: *с внешним миром, для этих целей, в минувшем июне*;

- предлог + (местоименное) прилагательное/причастие + (местоименное) прилагательное + существительное: *за мировую шахматную корону, от этой неприятной темы;*
- предлог + прилагательное + «и» + прилагательное + существительное: *без биологических и прочих добавок, с немецкими и американскими спецслужбами;*
- предлог + прилагательное + «и» + прилагательное + существительное без согласования по числу: *в греческом и итальянском консульствах, на немецком и венгерском языках;*
- предлог + прилагательное + прилагательное + прилагательное + существительное: *за все эти долгие годы, с равной цивилизованной демократической державой;*
- предлог + прилагательное + прилагательное + «и» + прилагательное + прилагательное + существительное: *из самых главных и значимых отечественных рок-команд;*
- предлог + наречие + прилагательное + существительное: *в крепко сбитого молодца, на очень небольшом количестве;*
- предлог + числительное + существительное: *до двух встреч, накануне первого миллениума;*
- предлог + числительное + прилагательное + существительное: *с первых газетных полос, из пятидесяти двух минувших недель;*
- предлог + число + существительное: *с 1994 года, на 11 дней;*
- предлог + число + прилагательное + существительное: *в 151 стрелковой дивизии;*
- предлог + существительное + «и» + существительное: *между «правыми» и «левыми», с различными министерствами и ведомостями.*

Томита-парсер также справился с:

- дефисным написанием: *под дубль-елку, за финансово-экономическое обеспечение, от 700-тысячного еврейского населения;*
- нестандартной капитализацией: *в уРУС-МАРТАНЕ, в «мЕДИА-МОСТ»;*
- кавычками: *с «фантомом», в «рваном» ритме.*

Следует, однако, отметить, что кавычки не всегда были обработаны удачно. Так, если временами встречаются верно определённые, но неудачно «вырезанные» контексты вроде *на эстраде»,* то в выводе также можно найти

конструкции, выделенные кусочно, недостаточно, вроде *включая нашумевшее «дело*.

На выходе был также обнаружено некоторое количество неверно выделенных предложных именных групп.

В ряде случаев ошибки обусловлены неверными разборами встроенного в Томита-парсер морфоанализатора *mystem*. Так, неверно были определены как предложные именные группы следующие контексты: *в том, к другой, к тому, о том, у провожавшего их маршала*. Ошибка связана с тем, что *mystem* приписывает токенам несколько наборов граммем, и если один из них удовлетворяет условиям, описанным в правилах, то контекст принимается.

Одной из самых частотных ошибок является определение конструкции с числом как целой именной группы или включение числа в именную группу: *с 7 до, с 5, к президентским выборам 2000, к первым десятилетиям XI, в 1950*. Это предположительно связано с тем, что, как упоминалось выше, *mystem* не присваивает никакие теги числам, что затрудняет их исключение из выделяемых именных групп.

В нескольких случаях правило, допускающее построение расширенной именной группы с несколькими прилагательными, разделяемыми запятой, привело к выделению следующих конструкций как предложных именных групп: *на их, тюремщиков; в 1957-м, шведам; напротив, её недавний результат*. Отчасти это связано с ошибками разбора *mystem*. Также это может говорить о том, что правило слишком синтаксически сложно или неудачно построено, а потому подлежит доработке.

По неизвестной причине парсером были верно выделены следующие предложные именные группы, которым не соответствуют имеющиеся правила: *в первых числах января, вокруг романтической и идеальной фигуры шведа, до недавнего «исторического» решения генпрокуратуры, на свет компанией Goodoo*. Именные группы с вложенными генетивными группами сознательно не были прописаны в правилах, так как в ходе экспериментов оказалось, что их включение приводит к засорению вывода, несоразмерному приросту верно выделяемых конструкций. Похожим образом в вывод были включены генетивные конструкции, содержащие латиницу: *с редкой изобретательностью ARIALPHONE System, с падением NASDAQ.*

Результаты сортировки по входящим в предложные именные группы предлогам можно назвать успешными, поскольку все конструкции с конкретными предлогами были верно интерпретированы в соответствующие

подполя общего факта *PrepPhrase*. Проблемой остаётся только оптимизация оформления таблицы с извлечёнными фактами, так как чтение таблицы со 116 столбцами несколько затруднено.

Заключение

Правила подлежат дальнейшей оптимизации и расширению, однако цель работы была достигнута: по результатам работы имеются грамматики с набором правил для извлечения и группировки предложных именных групп.