Ю. Д. Голяк
Минск, ООО «АйЭйчЭс Глобал»

ЛИНГВИСТИЧЕСКИЙ ПРОЦЕССОР ДЛЯ АВТОМАТИЧЕСКОГО ДОПОЛНЕНИЯ ПОЛЬЗОВАТЕЛЬСКИХ ЗАПРОСОВ НА РУССКОМ ЯЗЫКЕ

В работе предлагается постановка задачи автодополнения пользовательского поискового запроса на русском языке для информационных систем с интерактивным пользовательским естественно-языковым интерфейсом и принципиальная схема ее решения с учетом предполагаемой «истории» поиска на базе текстов предметной области. Приводится классификация типов пользовательских запросов и их синтаксических структур, описывается функциональность базового лингвистического процессора, необходимого для решения целевой задачи, а также требования к дополнительному модулю формирования подсказок для предполагаемых запросов.

Постановка задачи и принципиальная схема ее решения

Автодополнение (автоматическое завершение, предиктивный ввод) пользовательского запроса, в дальнейшем QTA (Query Type-Ahead), является необходимой функциональностью практически в любых современных информационных системах с интерактивным пользовательским естественно-языковым (ЕЯ) интерфейсом. Суть этой процедуры состоит в следующем. Пусть вводимый пользователем запрос в определенный момент времени представляет собой цепочку

$W_1\ W_2\ \ldots\ W_n$, (1),

где каждое W_i; $i = 1, n - 1$, $n > 1$ является словом естественного языка, а W_n, $n \geq 1$ – словом либо префиксом слова. Задача автодополнения запроса вида (1), в определенной степени уже отображающего информационную потребность пользователя, заключается в автоматическом формировании списка таких, в идеале, грамматически и семантически корректных цепочек слов естественного языка, которые включают в себя члены цепочки (1), т.е. речь идет не только о последовательном завершении строки вводимого запроса, рассматриваемого в качестве его префикса, но о более сложной процедуре, предполагающей, в общем случае, погружение уже набранной пользователем части запроса в контекст, а именно дополнение ее в начале, в конце и даже внутри. При этом предлагаемые цепочки могут содержать не все слова из (1). В любом случае они должны быть отранжированы по их релевантности относительно исходной цепочки. Кроме того, процедура автодополнения может применяться к вводимому запросу неоднократно (фрагментарно). Большинство современных поисковых систем, реализующих автодополнение запроса, опираются на автоматически неявно фиксируемую в процессе их эксплуатации историю проведенного поиска в виде списков

наиболее частотных пользовательских запросов и/или их наиболее информативных фрагментов. Как оказалось, актуальным для рассматриваемой задачи является случай, когда поисковое пространство, во-первых, заранее «известно» поисковой системе и, во-вторых, представляет собой относительно постоянную полнотекстовую базу данных (ПБД), например, БД текстов некоторой предметной области. Проведенные исследования показали, что эффективным решением задачи здесь представляется идея предварительного, т.е. до эксплуатации системы, формирования «истории», но не уже осуществленного, а предполагаемого поиска, в виде множества P так называемых подсказок, автоматически распознаваемых в самой ПБД, поскольку именно ей будут адресованы пользовательские запросы. Таким образом, может быть предложена следующая принципиальная схема решения задачи построения множества подсказок для автодополнения пользовательских запросов в рассматриваемом случае:

– экспертный анализ на основе общедоступных источников наиболее частотных поисковых запросов и классификация их типов;

– классификация синтаксических структур запросов этих типов;

– автоматический лингвистический анализ ПБД с целью распознавания в ее текстовых документах полученных на предыдущем этапе синтаксических структур и выбор из ПБД всевозможных соответствующих им лексических наполнений, что и составляет множество P.

Информационным источником для исследования в нашем случае послужили списки поисковых запросов к базе Topic Explorer, представленные в [1] (в количестве 8 тысяч запросов); списки запросов, предложенные в [2] для решения задачи вложенной сегментации запроса (в количестве 10 тыс. запросов), и базы запросов, предоставленные в качестве данных для решения задач из области ранжирования запроса на соревновательной платформе kaggle.com [3] (в количестве 10 тыс. запросов). Информация для анализа на собственно русскоязычном материале была почерпнута в процессе работы с сервисом wordstat.yandex.ru [4], позволяющим получить статистику запросов в Яндексе, включающих некоторое заранее заданное слово или словосочетание, и похожих на них запросов.

Проведенный экспертный анализ показал, что наиболее частотные поисковые запросы можно разделить на следующие основные типы (подробно в [5]):

1) одно или несколько несогласованных ключевых слов: *купить машину Минск*;

2) согласованные словосочетания: *извержение вулкана в Исландии*;

3) вопросительные предложения: *как настроить будильник на телефоне?*;

4) утвердительные предложения: *ноутбук перегревается и шумит кулер*;

5) комбинация двух последних типов: *ноутбук перегревается и шумит кулер, что делать?*

Анализ запросов каждого типа (кроме 1) позволил дать следующую классификацию их синтаксических структур.

Словосочетания:

- именные группы: *лазерная хирургия*;
- именные группы с предложно-падежными зависимыми конструкциями: *добыча руды в шахтах*;
- именные группы с причастным оборотом: *картины, проданные за самые высокие цены;*
- глагольные группы с прямым объектом: *купить телефон*;
- глагольные группы с косвенным объектом: *подготовиться к ЦТ*;
- глагольные группы с прямым и косвенным объектами: *вывести пятно от вина*;
- глагольные группы с причастным оборотом: *встроить деталь, снабженную витой резьбой.*

Вопросительные предложения:

- простые: *чем предотвратить коррозию металла?*;
- сложносочиненные: *почему появляется ржавчина и темнеет металл?*;
- сложноподчиненные: *как открыть консервную банку, когда нет ножа?*

Утвердительные предложения:

- простые: *перегревается двигатель*;
- сложносочиненные: *шумит кулер и греется системный блок*;
- сложноподчиненные: *мигает индикатор, даже если батарея заряжена.*

Комбинация нескольких типов:

- сочетание двух простых предложений или предложения и словосочетания, нередко без знака препинания между ними: *протекает кран вызвать сантехника.*

Очевидно, что лексическое наполнение именно таких синтаксических структур в той или иной степени будет выражать информационную потребность пользователя с помощью его поискового запроса. И если такое наполнение «подтолкнет» его здесь к использованию лексических фрагментов из ПБД, это обеспечит на этапе поиска гарантированно релевантную реакцию системы.

Ориентируясь на полученную классификацию синтаксических структур запросов основных типов, можно дать их формальное определение в терминах металингвистических переменных с целью последующего построения паттернов для алгоритма автоматического распознавания потенциальных подсказок в предполагаемом поисковом пространстве и построения их списка, при условии, что существует или может быть построен лингвистический процессор, оперирующий при автоматическом анализе текстов из указанного пространства теми же металингвистическими переменными.

Проведенные исследования показали, что для решения целевой задачи в качестве такового может быть использован доступный нам известный многоязычный БЛП IHS Goldfire (далее просто БЛП), поддерживающий работу с текстами на шести языках: английском, немецком, французском, японском, китайском и русском [6].

Функциональность базового лингвистического процессора

Процедура автоматического лингвистического анализа с помощью указанного БЛП предполагает поэтапную обработку текста [7]:

1. Форматирование и нормализация текста;
2. Лексический анализ текста;
3. Лексико-грамматический анализ текста;
4. Синтаксический анализ текста;
5. Семантический анализ текста.

Рассмотрим для примера следующее предложение: *В большинстве случаев влагоотталкивающие покрытия помогают предотвратить повреждения металла, вызываемые коррозией.*

Результатом обработки этого предложения при помощи БЛП на выходе этапа 4 станут распознанные в каждом предложении синтаксические отношения, представленые, как правило, в виде функционального или синтаксического дерева. Здесь же фиксируются канонизированные именные группы (*большинство случаев, коррозия, повреждения металла, влагоотталкивающие покрытия*), а также глагольные группы с 14-компонентной структурой, в данном случае единственная такая группа. Для каждого словоупотребления указан назначенный ему автоматически на предыдущем этапе лексико-грамматический класс: *JPO* – прилагательное множественного числа в именительном падеже; *NCPO* – нарицательное существительное множественного числа в именительном падеже; *IN* – предлог и т.д.

1. Подлежащее 1 (с атрибутом)	влагоотталкивающие_JPO покрытия_NCPO
2. Сказуемое 1	помогают_V3P предотвратить_VB
3. Именная часть сказуемого 1	–
4. Прямое дополнение 1	повреждения_NCPA металла_NCMG
5. Предлог 1	в_IN
6. Косвенное дополнение 1	большинстве_NCNR случаев_NCPG
7. Обстоятельство 1	–
8. Подлежащее 2	–
9. Сказуемое 2	вызываемые_LPPA

10. Именная часть сказуемого 2	–
11. Прямое дополнение 2	–
12. Предлог 2	–
13. Косвенное дополнение 2	коррозией_NCFI
14. Обстоятельство 2	–

Последний пятый этап обработки обеспечит распознавание семантических отношений между концептами, выраженными именными группами, в рамках так называемой САО-структуры: Субъект – Акция (Предикат) – Объект. Причем каждый элемент в этой структуре может иметь атрибуты [8]. В общем случае структура такого отношения состоит из 7 полей (плюс служебное восьмое поле). Любое из полей может оставаться пустым, в зависимости от наполнения конкретного предложения. Структура двух семантических отношений САО из рассматриваемого примера выглядит следующим образом:

	САО 1	САО 2
Субъект	влагоотталкивающие покрытия	коррозия
Акция	помогать предотвратить	вызывать
Объект	повреждения металла	повреждения металла
Атрибут прилагательное	–	–
Предлог	в	–
Непрямой объект	большинстве случаев	–
Атрибут наречие	–	–
Оригинальное представление акции в тексте	помогают предотвратить	вызываемые

Эти структуры служат основой для формирования будущих подсказок, поскольку:

1) канонизированные **именные группы** в полях **субъект** и **объект** по сути являются уже готовыми подсказками: *коррозия, повреждения металла, влагоотталкивающие покрытия* (именная группа *большинство случаев* должна быть отфильтрована как нерелевантная);

2) из **прямого объекта 1**, **сказуемого 2** и **косвенного дополнения 2** можно «собрать» подсказку вида ***именная группа с причастным оборотом***: *повреждения металла, вызываемые коррозией;*

3) из **сказуемого** второй САО-структуры, где причастие *вызываемые* приведено к начальной глагольной форме *вызывать,* и **объекта** той же структуры конструируется подсказка вида ***глагольная группа с прямым объектом*** *вызывать повреждения металла*, которая станет подходящим дополнением к запросу, например, начинающемуся с префикса *'что может'.*

Вместе с тем, как показал анализ, формирование подсказок некоторых видов требует дополнительной функциональности БЛП:

1) **сказуемое 1** и **прямой объект 1** можно объединить для получения подсказки вида ***глагольная группа с прямым объектом*** *помогают предотвратить повреждения металла*, но личная форма глагола *помогают* кажется не уместной для запроса. Здесь требуется приведение глагола к форме единственного числа *'что' помогает предотвратить повреждения металла* или возможность оставить в подсказке только смысловой глагол из предиката *предотвратить повреждения металла;*

2) вторую САО структуру, за исключением служебного поля, целесообразно использовать в качестве подсказки вида ***простое утвердительное предложение*** *коррозия вызывать повреждения металла*, однако начальная форма глагола не позволяет сделать эту конструкцию согласованной и требует приведения глагола к нужной форме *вызывает*.

Кроме того, использование всех компонентов, безусловно, требует учета их информативности, т.е. в состав дополнительной функциональности БЛП необходимо включить возможность ранжировать элементы по информативности и определенным семантическим критериям. Все это в совокупности с БЛП и составляет лингвистический процессор решаемой задачи.

ЛИТЕРАТУРА

1. *Reidsma, M.* Summon Topic Explorer Results by Search Query [Data set]. Zenodo. http://doi.org/10.5281/zenodo.47723 [Electronic resource]. – Mode of

access : https://zenodo.org/record/47723#.XS4aS-gzaHs. – Date of access : 07.04.2020.
2. *Rishiraj*. Supplementary Material for Nested Segmentation of Web Search Queries [Data set]. Zenodo. http://doi.org/10.5281/zenodo.1137746 [Electronic resource]. – Mode of access : https://zenodo.org/record/1137746#. XS4iNegzaHs. – Date of access : 07. 04. 2020.
3. Kpapamih: search-queries [Electronic resource]. – Mode of access : https:// www.kaggle.com/kpapamih/search-queries#searchterms.txt. – Date of access : 07.04.2020.
4. Wordstat. Yandex [Electronic resource]. – Mode of access : https://wordstat. yandex.ru/. – Date of access : 07. 04. 2020.
5. «Вести БГПУ». Сер. 1. Педагогика. Психология. Философия. – 2018. – № 3. – С. 91–95.
6. IHS Goldfire [Electronic resource]. – Mode of access : https://www.ihs.com/ pdf/IHS-Goldfire-Platform- Whitepaper_140823110915517432.pdf. – Date of access : 07.04.2020.
7. *Апресян, Ю. Д.* Лингвистическое обеспечение системы ЭТАП-2 / Ю. Д. Апресян [и др.] // М. : Наука, 1989. – С. 296.
8. *Совпель, И. В.* Инженерно-лингвистические принципы, методы и алгоритмы автоматической переработки текста / И. В. Совпель. – Минск : Выш. шк., 1991. – 120 с.

The paper proposes a definition of the problem of automatic completion of a user search query in Russian language for information systems with an interactive user natural-language interface and a conceptual scheme for its solution considering the pre-assumed search "history" based on the texts of the domain area. The classification of user queries' types and their syntactic structures is provided, the functionality of the basic linguistic processor necessary for the target problem solving, as well as the requirements for an additional module for query suggestions generation are described.