ПРОБЛЕМЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

УДК 81:004+004.912

Голяк Юлия Дмитриевна
соискатель кафедры
прикладной лингвистики,
Белорусский государственный университет;
специалист по компьютерной лингвистике
ООО «АйЭйчЭс Глобал»
г. Минск, Беларусь

Julia Haliak
Degree Seeking Applicant of the Department
of Applied Linguistics
Belarusian State University
Specialist in applied linguistics
at IHS Markit ltd.
juliaholiak@gmail.com

Совпель Игорь Васильевич
доктор технических наук, профессор;
заместитель генерального директора
по информационным технологиям
ООО «АйЭйчЭс Глобал»
г. Минск, Беларусь

Igor Sovpel
Habilitated Doctor of Technical Sciences,
Professor
Deputy General Director
for Information Technologies
at IHS Markit ltd.
igor.sovpel@gmail.com

АВТОМАТИЧЕСКОЕ ДОПОЛНЕНИЕ РУССКОЯЗЫЧНЫХ ПОЛЬЗОВАТЕЛЬСКИХ ЗАПРОСОВ НА ОСНОВЕ ПОДСКАЗОК ТИПА «ГЛАГОЛЬНАЯ ГРУППА», «ГРАММАТИЧЕСКАЯ ОСНОВА ПРЕДЛОЖЕНИЯ» И «ЛЕКСИКОН»

AUTOCOMPLETION OF RUSSIAN LANGUAGE USER QUERIES BASED ON SUGGESTIONS OF THE "VERB PHRASE", "SENTENCE SUBJECT AND PREDICATE" AND "LEXICON" TYPES

В работе рассматривается задача автоматического дополнения поисковых запросов пользователя на этапе их ввода, а также постадийно описан алгоритм построения подсказок типа «глагольная группа», «грамматическая основа предложения» и «лексикон», составляющих основу ее решения.

К л ю ч е в ы е с л о в а: *автодополнение; поисковый запрос; лингвистический процессор; глагольная группа; грамматическая основа предложения.*

The paper investigates the problem of automatic user search query completion on the input stage and provides stepwise description of the algorithm for the synthesis of autocompletion suggestions of the "verb phrase", "sentence subject and predicate" and "lexicon" types.

K e y w o r d s: *query type-ahead; search query; linguistic processor; verb phrase; setence subject and predicate.*

Данная статья в определенной степени является продолжением нашей предыдущей работы [1], где задача автодополнения рассматривается для наиболее частотных пользовательских запросов, формулируемых в виде именных групп. Там же дана постановка задачи автодополнения, решение которой основано на множестве автоматически распознаваемых в полнотекстовой базе данных так называемых подсказок. Такое распознавание оказалось возможным благодаря использованию базового лингвистического про-

цессора (БЛП) известной информационной системы IHS Goldfire [2], точнее – получаемого им результата семантико-синтаксического анализа текста в виде множества САО (Субъект – Акция – Объект) – отношений. Рассмотрим в качестве примера следующее предложение:

Антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой, предотвращает повреждения металла, вызываемые коррозией.

Выполнив лингвистический анализ данного предложения, БЛП распознает в нем следующие САО-отношения:

	Поля	САО 1	САО 2	САО 3
1.	ОС	–	повреждения металла	антикоррозийное покрытие аэрозольного нанесения
2.	С	–	–	–
3.	П	антикоррозийное покрытие аэрозольного нанесения	коррозия	антикоррозийное покрытие аэрозольного нанесения
4.	Ск	предотвращать	вызывать	создавать
5.	ИчС	–	–	–
6.	ПД	повреждения металла	повреждения металла	защитный слой
7.	ДДП	–	–	–
8.	ДТП	–	–	–
9.	Пр	–	–	–
10.	КДП	–	–	–
11.	О	–	–	–
12.	ВК	–	–	–
13.	ОПС	предотвращает	вызываемые	создающее

Здесь ОС – Опорное слово (компонент, через который одно САО-отношение предложения связывается с другим, если таковое имеет место), С – Союз, П – Подлежащее, Ск – Сказуемое (в форме инфинитива), ИчС – Именная часть Сказуемого, ПД – Прямое дополнение, ДДП – Дополнение в дательном падеже, ДТП – Дополнение в творительном падеже, Пр – Предлог, КДП – Косвенное дополнение с Предлогом, О – Обстоятельство, ВК – Вводная конструкция, ОПС – Оригинальное представление сказуемого в тексте.

Именно конкретное наполнение полей САО-структуры является исходным материалом, на основе анализа которого и учета приведенной в [3] классификации синтаксических структур подсказок осуществляется их автоматическое построение для наиболее частотных пользовательских запросов. Далее дается постадийное описание разработанных нами алгоритмов указанного построения для следующих наиболее актуальных после «именных групп» типов подсказок: «глагольная группа», «грамматическая основа предложения», «лексикон».

I. Построение списка подсказок для пользовательского запроса типа «глагольная группа»

Как и в случае с подсказками типа «расширенная именная группа с предложно-падежной зависимой конструкцией» [1], в данном случае речь также идет о двусоставных конструкциях. Первая их часть представлена собственно сказуемым, а вторая, зависимая часть, – дополнениями различных видов. Источником формирования первой части являются наполнения полей «Сказуемое» и «Оригинальное представление сказуемого в тексте» САО-отношений. Благодаря тому, что в поле «Сказуемое» хранится изначально канонизированный глагол, то есть его инфинитивная форма, а поле «Оригинальное представление сказуемого в тексте» содержит исходную форму сказуемого, использование этих полей позволяет формировать два типа глагольных групп:

1) глагольные группы с инфинитивом глагола;
2) глагольные группы с глаголом в третьем лице как единственного, так и множественного числа настоящего времени.

Формирование обоих типов подсказок обусловлено тем, что, как показывает анализ, это две наиболее частотные формы глагола, фигурирующие в запросах пользователей информационно-поисковых систем благодаря их сочетанию с такими, например, вопросительными словами: *как, где, зачем, когда, кто, почему* (*как заменить насос, где приобрести телефон, зачем делать раннюю прививку щенку, кто осуществляет взыскание налоговых выплат, кто выращивает кофе, почему перегревается двигатель*).

Что касается второй части подсказки, то источником ее формирования является наполнение таких полей САО-отношений, как «Опорное слово», «Прямое дополнение», «Дополнение в дательном падеже», «Дополнение в творительном падеже», «Косвенное дополнение с предлогом», т.е. именные группы.

Стадия 1.1. Фильтрация наполнения глагольных групп в зависимости от сказуемого

На основе проведенного анализа корпусов запросов нами был разработан набор следующих э к с п е р т н ы х п р а в и л, позволяющих заранее исключить формирование кандидатов в подсказки рассматриваемого типа в зависимости от сказуемого, представленного в САО-отношениях.

1. Если наполнение поля «Оригинальное представление сказуемого в тексте» является причастием, то исключается формирование кандидата подсказки с данным полем.

2. Если сказуемое содержит отрицательную частицу *не*, то соответствующие кандидаты в подсказки формируются с использованием его утвердительной формы, т.е. без частицы *не*. Во-первых, утвердительная форма, очевидно, вполне имеет право на существование в реальных запросах, а, во-вторых, в случае намерения пользователя все же использовать отрицательную форму, от него потребуются минимальные усилия – только набрать частицу *не* в предлагаемой ему подсказке.

3. Исключается формирование кандидатов в подсказки с теми сказуемыми, основной смысловой глагол которых не удовлетворяет критерию информативности, в нашем случае – является низкочастотным согласно анализу корпуса пользовательских запросов (например, глаголы *быть, являться, оказываться, стать* и др.).

В силу приведенных выше правил в используемом нами сквозном примере будет исключено (в соответствии с правилом 1) формирование подсказок с использованием наполнения поля «Оригинальное представление сказуемого в тексте» из САО 2 и САО 3.

Стадия 1.2. Преобразование составного глагольного сказуемого

Если в САО-отношении имеет место составное глагольное сказуемое, то на данной стадии фиксируется дополнительно еще одно сказуемое в виде его основного смыслового глагола и кандидаты в подсказки формируются как с использованием исходного сказуемого, так и производного. Например, в случае наличия «Сказуемого» *помогать предотвратить* оно будет дополнительно усечено до *предотвратить*.

Стадия 1.3. Формирование глагольных групп с прямым и косвенными дополнениями, а также с их комбинациями

На данной стадии наполнение полей «Сказуемое» и «Оригинальное представление сказуемого в тексте» объединяется последовательно с наполнением таких полей, как:

1) «Прямое дополнение»
2) «Дополнение в дательном падеже»
3) «Дополнение в творительном падеже»
4) «Косвенное дополнение с предлогом»

Далее дополнительно осуществляется объединение «Сказуемого» с наполнением каждого из указанных полей, но без согласуемых атрибутов, далее – с их главным словом, объединение «Сказуемого» одновременно с наполнением нескольких зависимых из множества указанных полей, в том числе и с их представлением без согласуемых атрибутов и в виде главных слов (подробнее эти процедуры усечения наполнения перечисленных полей представлены в [1]). Ниже в соответствующем порядке приводятся примеры кандидатов в подсказки, получаемые на данной стадии из рассматриваемого примера:

1) *предотвращать повреждения металла;*
2) *вызывать повреждения металла;*
3) *создавать защитный слой;*
4) *предотвращает повреждения металла;*
5) *предотвращать повреждения;*
6) *предотвращает повреждения;*
7) *вызывать повреждения;*
8) *создавать слой.*

Стадия 1.4. Формирование глагольных групп с прямым дополнением и причастным оборотом

На данной стадии кандидаты в подсказки формируются путем объединения наполнения полей отдельно «Сказуемое» и «Оригинальное представление сказуемого в тексте» с наполнением поля «Прямое дополнение», представленным именной группой, и зависимым от него причастным оборотом. Формируются также варианты подсказок с именными группами без атрибутов. Подробное описание процедуры объединения именной группы с причастным оборотом приведено в [1]. Для предложения, используемого нами в качестве сквозного примера, здесь будут получены следующие кандидаты в подсказки:

1) *предотвращать повреждения металла, вызываемые коррозией;*
2) *предотвращает повреждения металла, вызываемые коррозией;*
3) *предотвращать повреждения, вызываемые коррозией;*
4) *предотвращает повреждения, вызываемые коррозией.*

Стадия 1.5. Контрольная проверка полученного списка кандидатов и добавление их в базу подсказок

В силу изложенного, на выходе данного этапа для нашего сквозного примера будет получен следующий список подсказок:

1) *предотвращать повреждения металла;*
2) *вызывать повреждения металла;*
3) *создавать защитный слой;*
4) *предотвращает повреждения металла;*
5) *предотвращать повреждения;*
6) *предотвращает повреждения;*
7) *вызывать повреждения;*
8) *создавать слой;*
9) *предотвращать повреждения металла, вызываемые коррозией;*
10) *предотвращает повреждения металла, вызываемые коррозией;*
11) *предотвращать повреждения, вызываемые коррозией;*
12) *предотвращает повреждения, вызываемые коррозией.*

II. Построение списка подсказок для пользовательского запроса типа «грамматическая основа предложения»

Еще одним актуальным типом пользовательских запросов являются запросы типа «предложение», классифицированные на подгруппы по синтаксическому признаку (простые, сложносочиненные, сложноподчиненные), а также по цели высказывания (вопросительные, утвердительные и их комбинация в виде двух простых предложений, нередко без знака препинания между ними) [1].

Данный тип запроса является достаточно сложным с точки зрения решения целевой задачи, поэтому в процессе разработки соответствующего алгоритма было принято решение не формировать распространенные и сложные предложения в виде непрерывной подсказки, а ограничиться построением подсказки типа «*грамматическая основа предложения*», под которой пони-

мается объединение наполнений полей «Подлежащее» и «Оригинальное представление сказуемого в тексте» САО-отношений. Это «позволяет» алгоритму подбора подсказки в процессе набора пользователем своего запроса предложить ему наиболее подходящее в каждый конкретный момент времени продолжение запроса в виде зависимого члена предложения, выбранного из числа подсказок другого типа, или же второй грамматической основы.

Рассмотрим постадийно процедуру получения подсказок рассматриваемого типа.

Стадия 2.1. Фильтрация САО-отношений с неинформативными и несогласованными подлежащим и сказуемым

На данной стадии анализу подвергаются наполнения полей «Подлежащее» и «Сказуемое». При этом из дальнейшей обработки исключаются те САО-отношения, в которых:

- «Сказуемое» содержит отрицание, модальные слова (*мочь, должен, можно, нужно, нельзя*) или представляет собой глаголы речи (*говорить, сказать, рассказывать*), глаголы *быть, являться, иметь, состоять* и некоторые другие, как наиболее редкие в запросах рассматриваемого типа;
- «Оригинальное представление сказуемого в тексте» выражено причастием или деепричастием;
- главное слово «Подлежащего» не находится в именительном падеже, если оно склоняемо (как в случае оригинальности такой формы в тексте, так и как результат фильтраций);
- по какой-либо причине главное слово «Подлежащего» и «Оригинальное представление сказуемого в тексте» не совпадают по числу.

Отметим, что из всех САО-отношений, указанных ранее для нашего сквозного примера, данную стадию успешно преодолеют только САО 1.

Стадия 2.2. Объединение наполнения полей «Подлежащее» и «Оригинальное представление сказуемого в тексте»

На этой стадии наполнения полей «Подлежащее» и «Оригинальное представление сказуемого в тексте», успешно преодолевших фильтрацию САО-отношений, объединяются в единую структуру, представляющую собой так называемую грамматическую основу предложения. Причем, помня о том, что запросы типа «предложение» по цели высказывания могут быть как утвердительными, так и вопросительными, целесообразно сформировать в этот момент кандидатов в подсказки в двух вариантах с точки зрения очередности полей: в прямом и в обратном порядке. Как показывает анализ, обратный порядок подлежащего и сказуемого характерен в русском языке для вопросительных предложений (ср. *Двигатель перегревается* и *Почему перегревается двигатель*?).

Таким образом, на данной стадии для нашего сквозного примера будут получены следующие кандидаты в подсказки:

1) *антикоррозийное покрытие аэрозольного нанесения предотвращает;*
2) *предотвращает антикоррозийное покрытие аэрозольного нанесения.*

Стадия 2.3. Формирование подсказок с усеченным наполнением поля «Подлежащее»

Здесь из полученных на предыдущей стадии кандидатов в подсказки дополнительно формируются таковые за счет усечения именных групп, составляющих наполнение поля «Подлежащее». Это усечение осуществляется путем применения процедур обработки именных групп с согласованными атрибутами и распознавания их главных слов, подробно представленных в [1].

Таким образом, будут дополнительно получены следующие кандидаты в подсказки:

1) *покрытие аэрозольного нанесения предотвращает;*
2) *покрытие предотвращает;*
3) *предотвращает покрытие аэрозольного нанесения;*
4) *предотвращает покрытие.*

Стадия 2.4. Контрольная проверка списка кандидатов в подсказки и добавление их в базу подсказок

В нашем примере на выходе этой стадии получим следующий список подсказок:

3) *антикоррозийное покрытие аэрозольного нанесения предотвращает;*
4) *предотвращает антикоррозийное покрытие аэрозольного нанесения;*
5) *покрытие аэрозольного нанесения предотвращает;*
6) *покрытие предотвращает;*
7) *предотвращает покрытие аэрозольного нанесения;*
8) *предотвращает покрытие.*

III. Построение списка подсказок типа «лексикон»

Еще одним, дополнительным, но, как оказалось, важным типом подсказок, формируемых для решения целевой задачи, являются подсказки условно названного типа «лексикон». К ним относятся однословные подсказки разных частей речи. Для их получения все поля САО-отношений, прошедших все упомянутые ранее виды фильтрации, проходят через ряд однотипных стадий, целью которых является распознавание в этих полях по лексико-грамматическим категориям слов, задаваемых в виде тегов, их частей речи, приведение слов к начальной форме (с сохранением оригинального числа слов для тех частей речи, у которых имеется такая грамматическая категория) и добавление в базу подсказок. Полный перечень таких стадий выглядит следующим образом.

Стадия 3.1. Атрибутивная

Здесь в полях САО-отношений распознаются прилагательные, причастия, порядковые числительные и некоторые определительные местоимения, которые затем приводятся к именительному падежу исходного рода и числа. В нашем примере на этой стадии сформируются следующие кандидаты в подсказки:

1) *антикоррозийное покрытие аэрозольного нанесения -> антикоррозийное, аэрозольное;*
2) *вызываемые -> вызываемые;*

3) *защитный слой -> защитный;*

4) *создающее -> создающее.*

Стадия 3.2. Адвербиальная

На данной стадии распознаются наречия в начальной форме или сравнительной степени (форма сравнительной степени сохраняется). В нашем примере наречия не присутствуют, и выход этой стадии будет пустым, но в качестве иллюстрации можно привести следующие именные группы и результат их обработки: *мягко очищать поверхность -> мягко; приобрести товары дешевле -> дешевле.*

Стадия 3.3. Глагольная

Здесь распознаются и приводятся к инфинитиву глагольные формы (без сохранения исходной формы):

1) *предотвращать, предотвращает -> предотвращать;*

2) *вызывать -> вызывать;*

3) *создавать -> создавать.*

Стадия 3.4. Стадия числительных

В данном случае распознаются и канонизируются количественные и собирательные числительные (при этом сохраняется также их исходная форма). Оригинальная форма числительных, в отличие от большинства других частей речи, сохраняется в виде отдельного кандидата, поскольку в большей части именных групп числительные отфильтровываются при построении списков подсказок рассмотренных ранее типов. Поэтому только такой подход позволяет сохранить в базе подсказок максимальное разнообразие форм числительных. Ввиду отсутствия числительных в рассматриваемом предложении, приведем такие примеры: *провести анализ пяти экспериментов -> пять, пяти; шестеро взрослых и трое детей -> шестеро, трое.*

Стадия 3.5. Стадия модальных слов

Здесь в качестве кандидатов в подсказки в полях САО-отношений распознаются модальные слова *можно, нужно, надо, необходимо, нельзя, невозможно* и т.д.: *необходимо провести подготовку -> необходимо.*

Стадия 3.6. Предложная

На данной стадии в качестве кандидатов в подсказки распознаются предлоги длиной более 4 символов, поскольку автодополнение таких длинных предлогов может сэкономить время ввода не менее, чем время ввода знаменательных слов: *поступать согласно правилам -> согласно; простудиться вследствие переохлаждения -> вследствие.*

Стадия 3.7. Контрольная проверка полученного списка кандидатов и добавление их в базу подсказок

Список полученных кандидатов в подсказки типа «лексикон», для рассматриваемого нами сквозного примера будет выглядеть так:

1) *антикоррозийное;*

2) *аэрозольное;*

3) *вызываемые;*

4) *защитный;*

5) *создающее;*
6) *предотвращать;*
7) *вызывать;*
8) *создавать.*

Обратим внимание, что в процессе формирования данного типа подсказок отсутствует стадия для существительных, поскольку источником для получения одиночных существительных служит процедура определения главных слов в процессе формирования подсказок типа «именная группа» [1].

В результате полного цикла целевой обработки рассматриваемого в качестве сквозного примера предложения с целью получения из него подсказок рассмотренных трех типов будет получен следующий их список:

1) *предотвращать повреждения металла;*
2) *вызывать повреждения металла;*
3) *создавать защитный слой;*
4) *предотвращает повреждения металла;*
5) *предотвращать повреждения;*
6) *предотвращает повреждения;*
7) *вызывать повреждения;*
8) *создавать слой;*
9) *предотвращать повреждения металла, вызываемые коррозией;*
10) *предотвращает повреждения металла, вызываемые коррозией;*
11) *предотвращать повреждения, вызываемые коррозией;*
12) *предотвращает повреждения, вызываемые коррозией;*
13) *антикоррозийное покрытие аэрозольного нанесения предотвращает;*
14) *предотвращает антикоррозийное покрытие аэрозольного нанесения;*
15) *покрытие аэрозольного нанесения предотвращает;*
16) *покрытие предотвращает;*
17) *предотвращает покрытие аэрозольного нанесения;*
18) *предотвращает покрытие;*
19) *антикоррозийное;*
20) *аэрозольное;*
21) *вызываемые;*
22) *защитный;*
23) *создающее;*
24) *предотвращать;*
25) *вызывать;*
26) *создавать.*

В заключение отметим, что постадийная процедура построения базы подсказок является эффективным инструментом управления наполнением этой базы в смысле включения в ее состав тех или иных типов подсказок в зависимости от предпочтений пользователя или особенностей предметной области. Что касается представленных во всех приведенных схемах стадий контрольной проверки полученного списка кандидатов в подсказки, то здесь,

кроме традиционных процедур экспертного контроля текстовых данных, могут, например, осуществляться процедуры автоматического сокращения списка кандидатов в подсказки путем установления для них некоторого порогового значения их частотности или информативности. Во втором случае указанная информативность может быть определена использованием достаточно эффективного метода, предложенного в [4] как раз для CAO-отношений и их компонентов.

Представленные результаты, так же, как и результаты, полученные в [1] были внедрены в состав упомянутой ранее информационной системы IHS Goldfire и показали свою актуальность и эффективность.

ЛИТЕРАТУРА

1. *Голяк, Ю. Д., Совпель, И. В.* Автоматическое дополнение русскоязычных пользовательских запросов, формулируемых в виде именных групп / Ю. Д. Голяк, И. В. Совпель // Вестник МГЛУ. Филология. Сер. 1. – 2021. – С. 101–110.

2. IHS Goldfire [Electronic resource]. – Mode of access : https://www.ihs.com/pdf/IHS-Goldfire-Platform- Whitepaper_140823110915517-432.pdf. – Date of access : 07. 05. 2021.

3. *Голяк, Ю. Д.* Автодополнение поискового запроса на основе автоматического извлечения подсказок из преиндексированных документов предметной области / Ю. Д. Голяк // Вестник БГПУ. Сер. 1. Педагогика. Психология. Филология. – 2018. – № 3. – С. 91–95.

4. *Воронков, Н. В.* Методы, алгоритмы и модели систем автоматического реферирования текстовых документов : дис... к-та техн. наук : 05.13.17 / Н. В. Воронков. – Минск, 2007. – 165 л.

Поступила в редакцию 18.06.2021