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, ФОРМУЛИРУЕМЫХ В ВИДЕ ИМЕННЫХ ГРУПП

AUTOCOMPLETION OF RUSSIAN LANGUAGE USER QUERIES FORMULATED AS NOUN PHRASES

В работе исследуется задача автоматического дополнения поисковых запросов для информационных систем с интерактивным пользовательским естественно-языковым интерфейсом и подробно описывается алгоритм ее эффективного решения для одного из наиболее частотных типов запросов, формулируемых в виде именных групп.

К л ю ч е в ы е с л о в а: *автодополнение*; *поисковый запрос*; *предиктивный ввод*; *лингвистический процессор*; *именная группа*.

The paper investigates the problem of automatic search query completion for information systems with an interactive natural-language user interface and provides detailed description of the algorithm for its effective solution for one of the most frequent types of user search queries, formulated in the form of noun phrases.

K e y w o r d s: *query type-ahead*; *search query*; *predictive query input*; *linguistic processor*; *noun phrase*.

Предиктивный ввод пользовательского запроса является необходимой функциональностью развитых информационных систем с интерактивным пользовательским естественно-языковым (ЕЯ) интерфейсом. Наиболее распространенные решения задачи автодополнения запроса опираются на автоматически фиксируемую в процессе эксплуатации этих систем историю проведенного поиска в виде списков наиболее частотных пользовательских запросов. В данном исследовании речь идет об одном актуальном частном

случае рассматриваемой задачи, когда существует возможность предварительного, т. е. до эксплуатации системы, формирования «истории», но не уже осуществленного, а предполагаемого поиска, в виде множества P так называемых подсказок, заранее автоматически распознаваемых в самой полнотекстовой базе данных (ПБД), поскольку именно ей будут адресованы пользовательские запросы. Постановка такой задачи и принципиальная схема ее решения представлены в [1]. Там же дана классификация синтаксических структур, распознавание которых в текстах из ПБД обеспечивает построение множества P. Показано, что доступный нам многоязычный базовый лингвистический процессор (БЛП) известной информационной системы IHS Goldfire [2] в наибольшей степени соответствует требованиям, предъявляемым к автоматическому лингвистическому анализу текстов из ПБД с целью решения указанной задачи распознавания.

Проведенный анализ показал, что наиболее подходящей структурой, используемой в качестве основы для формирования будущих подсказок, можно считать результат этапа семантико-синтаксического анализа текста, а именно так называемое САО (Субъект – Акция – Объект) – отношение, каждый элемент которого может иметь свои атрибуты [3; 4; 5]. Что касается текстов на русском языке, то по результатам анализа морфологических и синтаксических особенностей этого языка было решено каждому их предложению формировать расширенный формат этого отношения. Рассмотрим его на примере следующего предложения:

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

В результате обработки этого предложения БЛП автоматически распознает в нем следующие три САО-отношения:

	Поля	САО 1	САО2	САО3
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	вызываемое	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.

Конкретное наполнение различных полей САО-структуры является в нашем случае тем исходным материалом, из которого, на основе его анализа и учета приведенной в [1] классификации синтаксических структур подсказок, последние могут быть получены автоматически, и это, очевидно, влечет за собой необходимость дополнительной функциональности лингвистического процессора. Прежде всего, учитывая, что одним из основных критериев значимости подсказки для автоматического завершения запроса является ее информативность как отдельного запроса или его части было исследовано наполнение пользовательских запросов, представленных в свободном доступе в сети Интернет [6; 7]. Оно позволило в итоге разработать набор правил и основанных на них независимых друг от друга базовых процедур (БП) фильтрации САО-отношений с точки зрения их информативности, а также преобразования отдельных компонентов САО-отношений. При этом особое внимание уделяется «именной группе (ИГ)» как одному из самых распространенных типов пользовательских запросов [1].

БП 1. Исключение определенных САО-отношений и их полей из списка кандидатов для формирования подсказок

Из списка кандидатов для формирования подсказок исключаются:

- САО-отношения, содержащие в совокупности количество слов больше некоторого устанавливаемого экспертным путем порогового значения μ_0, например, в нашем случае оно равно 50;
- поля, содержащие служебные символы, специфические математические знаки, элементы алфавита ЕЯ, отличного от русского языка;
- поля, содержащие фрагменты искаженного текста;
- поля, содержащие определенного вида сокращения, например, *дис., рис., табл., гр., гл., им., проф., ср.*

Отметим, что данную процедуру успешно преодолеют все САО-отношения приведенного в качестве примера предложения и все поля каждого из этих отношений.

БП 2. Фильтрация именных групп САО-отношений

Из именных групп, составляющих наполнение таких полей САО-отношений, как «Опорное слово», «Подлежащее», «Именная часть сказуемого», «Прямое дополнение», «Дополнение в дательном падеже», «Дополнение в творительном падеже», «Косвенное дополнение с предлогом», удаляются части (в приводимых ниже примерах они выделены курсивом и жирным шрифтом), представляющие собой:

- вводные конструкции, например, *вне* ***всякого сомнения***, *в* ***большинстве случаев***;
- числительные, а также числовые обозначения, например, ***20-процентный, семнадцатый*** *чертеж*;
- описание отрезка времени или число повторений, например, на ***прошлой неделе, пять раз***;
- различного рода сокращения, например, ***и т.д., и пр., т.н;***
- буквенные обозначения объектов, например, *рисунок* ***A***, *график* ***G***;
- общие уточнения образа действия, например, ***таким способом***, ***этим образом***;
- местоимения, например, ***его*** *шляпа,* ***несколько*** *двигателей* и т.д.*;*
- неинформативные вне контекста атрибуты в виде прилагательных и причастий, например, ***указанный*** *перечень,* ***вышеперечисленные*** *пункты* и т.д. и ряд других.

Из указанных ранее САО-отношений для приведенного в качестве примера предложения БЛП распознает следующие приведенные к канонической форме попарно различные ИГ, все из которых также успешно преодолеют описанную процедуру БП2: *антикоррозийное покрытие аэрозольного нанесения, защитный слой, повреждения металла, коррозия.*

БП 3. Обработка именных групп с зависимой именной частью в косвенных падежах

Данной процедуре подвергаются те ИГ, для которых сформулированное условие имеет место. При этом в ИГ вычленяются зависимые от главного слова несогласованные части в косвенных падежах и добавляются искусственные разделители (метка *DELIMITER*) на стыке главной и зависимой частей.

В нашем списке условиям процедуры БП3 удовлетворяют ИГ: *антикоррозийное покрытие DELIMITER аэрозольного нанесения, повреждения DELIMITER металла.* Далее размеченные именные группы разделяются по метке на части, зависимые из них приводятся к канонической форме. Таким образом, в нашем случае на выходе мы получим следующий список ИГ: *антикоррозийное покрытие, аэрозольное нанесение, повреждения, металл.*

БП 4. Обработка именных групп с согласованными атрибутами

Именные группы с согласованными атрибутами сами по себе являются качественными кандидатами в формируемую базу подсказок, но анализ показал, что вероятность подсказать пользователю нужный суффикс запроса существенно повышается, если эта база будет дополнительно содержать эти же ИГ, но без их атрибутов. Поэтому данная реализованная нами процедура также включена в число БП решения целевой задачи: *антикоррозийное покрытие → покрытие, аэрозольное нанесение → нанесение.*

БП 5. Определение главных слов именных групп

Аналогично, взятые отдельно главные слова ИГ, обладая высокой информативностью, также являются качественными кандидатами в базу подсказок. Соответствующая процедура уже входит в состав функциональ-

ности БЛП, и, таким образом, она просто включается в состав БП. Каждое получаемое ею главное слово при этом приводится к канонической форме: *регулярная обработка антикоррозийным покрытием → обработка*, *двухфазная очистка воды → очистка.*

Как уже отмечалось раннее, в нашем случае речь идет о формировании списка подсказок для пользовательских запросов одного из самых частотных типов – «именные группы». В соответствии с представленной в [1] классификацией синтаксических структур таких запросов имеют место «простые именные группы» и «расширенные именные группы». Во втором случае именные группы осложнены несогласованными зависимыми частями, такими, как предложно-падежные конструкции, а также причастными оборотами.

Построение списка подсказок для пользовательского запроса типа «простая именная группа»

В силу изложенного ранее, полагая, что на вход поступают предварительно обработанные с помощью БЛП тексты из ПБД, имеет место следующая постадийная схема алгоритма решения данной задачи:

→ БП 1 → БП 2 → БП 3 → БП 4 → БП 5 → список подсказок

Построение списка подсказок для пользовательского запроса типа «расширенная именная группа с предложно-падежной зависимой конструкцией»

Примерами данного типа запросов могут служить: *добыча руды в шахтах, осложнения после лазерной коррекции зрения, просмотр сериалов без регистрации* и другие. Источником для их формирования являются две группы полей САО-отношений, обеспечивающих наполнение двух частей таких подсказок. Первая часть – главная именная группа – формируется из наполнения полей типа «Подлежащее», «Прямое дополнение», «Дополнение в дательном падеже», «Дополнение в творительном падеже». Фактически же будет достаточно использовать поле «Опорное слово», поскольку зависимость косвенного дополнения от другого именного члена предложения будет отражена через формирование отдельного САО-отношения с главным словом в данном поле. Вторая же часть подсказки – предложно-падежная группа – формируется из наполнения полей «Предлог» и «Косвенное дополнение». При этом речь идет как об отдельном САО-отношении, полученном БЛП, так и САО-отношении, синтезированном, при необходимости, из нескольких САО-отношений, распознанных БЛП в предложении. Построение списка подсказок данного типа осуществляется, как и в первом случае, постадийно, при этом все именные группы используемых полей САО-отношений предварительно подвергаются фильтрации (БП 2). Наиболее простым здесь является случай, когда кандидат в подсказки синтезируется путем непосредственного объединения наполнения полей «Опорное слово», «Предлог» и «Косвенное дополнение с предлогом» в пределах одного САО-отношения, полученного БЛП: *ржавчина (ОС) на (Пр) внешних поверхностях механизмов (КДП)*.

Более сложным является случай многословных составных предлогов, таких как *в соответствии с, в надежде на, в связи с* и др., поскольку их

объединение в единую структуру не заложено в функциональности БЛП на этапе построения САО-отношений. Дело в том, что определить, является ли такое сочетание составным предлогом или же самостоятельным знаменательным членом предложения, позволяет, семантический анализ предложения (ср. «*Ожидается перенос мероприятия* ***в связи с*** *пандемией»* и «*Наблюдаются перебои* ***в связи с*** *отдаленными населенными пунктами»*). В качестве иллюстрации предложенного нами решения проблемы рассмотрим первое из приведенных предложений. При его автоматическом лингвистическом анализе БЛП распознает следующие три САО-отношения:

	Поля	САО 1	САО2	САО3
1.	ОС	–	перенос мероприятия	связи
3.	П	перенос мероприятия	–	–
4.	Ск	ожидаться	–	–
5.	ИчС	–	–	–
6.	ПД	–	перенос мероприятия	связи
9.	Пр	–	в	с
10.	КДП	–	связи	пандемией
13.	ОПС	ожидается	–	–

Все остальные поля САО-отношений выше и далее не приводятся в случаях, если они пусты.

На основе проведенного анализа используемого в наших исследованиях текстового материала были составлены словарь составных предлогов и специальные словари лексической сочетаемости с целью разрешения семантической многозначности для рассматриваемой проблемы. В данном случае в соответствии с ними будет осуществлена линковка (по определенному правилу), двух САО-отношений (САО 2а и САО 3а) в одно специальное (САО 4а), в результате которой в его поле «Предлог» будет помещен составной предлог *в связи с:*

САО 4а	
1. ОС	*перенос мероприятия*
6. ПД	*перенос мероприятия*
9. Пр	*в связи с*
10. КДП	*пандемией*

Полученное таким образом новое САО-отношение уже подчиняется приведенной ранее схеме построения кандидата в подсказки для простого случая: *перенос мероприятия (ОС) в связи с (Пр) пандемией (КДП)*. Заметим, что во втором из приведенных предложений, с использованием словарей такая линковка будет запрещена, и в результате сформируются два кандидата в подсказки *перебои в связи* и *связь с отдаленными населенными пунктами.*

Процедуре синтеза составных предлогов в интересах решаемой нами целевой задачи будут подвергнуты и те составные предлоги, которые не содержат второго краткого непроизводного предлога (*в, на, с* и др.), но определенные части которых при формировании САО-отношений БЛП по умолчанию отнесены к полю «Косвенное дополнение с предлогом»: *выплата (ОС) в (Пр) пользу основного истца (КДП) → выплата (ОС), в пользу (Пр) основноо истца (КДП).* Таким образом, составной предлог *в пользу* целиком оказывается в поле «Предлог», а в поле «Косвенное дополнение» – семантически корректная именная группа *основного истца.*

В дополнение, к главной и зависимой именным частям только что сформированных кандидатов в подсказки применяется процедура определения главного слова (БП 5), в результате чего формируются их «укороченные» варианты. Таким образом, например, для кандидата в подсказки *повреждения металла на внешних поверхностях механизмов* будет дополнительно синтезирован вариант *повреждения на поверхностях.* Что касается приведенного выше кандидата в подсказки *выплата в пользу основного истца,* то здесь на основании сгенерированного САО-отношения будет дополнительно синтезирован вариант *выплата в пользу истца,* что более приемлемо, чем вариант, возможный в случае использования исходного отношения – *выплата в пользу.*

Проведенные исследования показали, что наряду с подсказками в виде ИГ с зависимыми предложно-падежными конструкциями релевантными автодополнению пользовательского запроса могут оказаться подсказки в виде отдельно взятых предложно-падежных конструкций, то есть комбинации только полей «Предлог» и «Косвенное дополнение с предлогом», а также – «Предлог» и главное слово «Косвенного дополнения с предлогом». Эти комбинации наполнений указанных полей фактически являются зависимыми частями полученных выше кандидатов в подсказки, поэтому далее осуществляется вычленение этих зависимых частей и включение их в формируемый список кандидатов в подсказки: *в пользу основного истца, в пользу истца, в связи с пандемией, на внешних поверхностях механизмов, на поверхностях.*

Построение списка подсказок для пользовательского запроса типа «расширенная именная группа с причастным оборотом»

Примерами данного типа запросов могут служить запросы: *покрытие, создающее защитный слой; повреждения, вызванные коррозией; поломки, устранимые подручными средствами; товары, приобретенные ребенку.* Как и в предыдущем случае, подсказки рассматриваемого типа также состоят из двух частей: главной ИГ и собственно причастного оборота. Процедура получения первой части подробно представлена выше, поэтому далее будет описан только процесс формирования из полей САО-структуры причастного оборота. Заметим, что БЛП при автоматическом лингвистическом анализе текста само причастие в своей оригинальной форме помещает в поле ОПС.

Именно это поле, наряду с полями ПД, ДДП, ДТП, КДП, является источником для формирования второй части подсказок рассматриваемого типа. Подчеркнем также, что все ИГ используемых при этом полей САО-отношений предварительно подвергаются описанной ранее процедуре фильтрации БП2.

Причастные обороты для кандидатов в подсказки рассматриваемого типа автоматически формируются путем объединения наполнений таких непустых полей САО-отношений, как ОПС и ПД (*создающие защитный слой*), ОПС и ДДП (*препятствующие появлению ржавчины*) и ОПС и ДТП (*оборудованный современной техникой*). В последнем случае учитывается и ситуация, когда в поле *ОПС* находится страдательное причастие, а наполнение поля *ОС* совпадает с наполнением поля *ПД*. Примером такого САО-отношения служит САО 2, сформированное БЛП в результате трансформации пассивного залога из оригинального предложения в активный. Такая конфигурация позволяет провести обратную конвертацию – привести наполнение поля *П* к творительному падежу, а затем объединить с ним причастие для получения причастного оборота (*вызываемые коррозией*). Кроме того, причастные обороты формируются путем объединения наполнений полей *Пр* и *КДП* (*встроенный в обновленный механизм, шлифованный с помощью алмазных дисков*). При этом используется уже описанная ранее процедура распознавания и синтеза многословных составных предлогов.

Также имеет место объединение причастия одновременно с несколькими непустыми полями из их множества {*ПД, ДДП, ДТП, КДП*}. Таким образом могут быть получены причастные обороты: *работавший руководителем в технической команде, пересылающий сообщение адресату, сообщающий показания оператору посредством автоматического оповещения.* После того, как причастные обороты сформированы, происходит их объединение с соответствующими им главными ИГ: *антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой; повреждения металла, вызываемые коррозией.* Дополнительно для каждого из уже полученных кандидатов в подсказки формируется его «укороченный» вариант, содержащий только главную именную группу и причастие без зависимых от него полей (*антикоррозийное покрытие аэрозольного нанесения, создающее; повреждения металла, вызываемые*), а для случаев страдательного причастия и причастия, образованного от непереходного глагола, формируется каждый раз еще один кандидат в подсказки с обратным следованием указанных составных частей (*вызываемые повреждения металла*). Также в этом случае с использованием процедуры БП5 формируются кандидаты в подсказки путем усечения именной группы: *покрытие, создающее защитный слой; покрытие, создающее; повреждения, вызываемые.*

Таким образом, в соответствии с изложенным для приведенного в начале работы примера предложения в совокупности будут получены следующие

подсказки для автодополнения пользовательского запроса типа «именная группа»: *антикоррозийное покрытие аэрозольного нанесения; покрытие аэрозольного нанесения; аэрозольное нанесение; нанесение; повреждения металла; повреждения; металл; коррозия; защитный слой; слой; антикоррозийное покрытие аэрозольного нанесения, создающее защитный слой; повреждение металла, вызываемые коррозией; антикоррозийное покрытие аэрозольного нанесения, создающее; повреждения металла, вызываемые; покрытие, создающее защитный слой; вызываемые повреждения металла; покрытие, создающее; повреждения, вызываемые.*

В заключение отметим, что полученные результаты внедрены в состав упомянутой ранее информационной системы IHS Goldfire и доказали свою востребованность и эффективность при решении задачи информационного поиска.

ЛИТЕРАТУРА

1. *Голяк Ю. Д.* Автодополнение поискового запроса на основе автоматического извлечения подсказок из преиндексированных документов предметной области / «Вести БГПУ». Сер. 1. Педагогика. Психология. Филология. – Минск. – 2018. – №3. – С. 91–95.
2. IHS Goldfire [Electronic resource]. – Mode of access : https://www.ihs.com/pdf/IHS-Goldfire-Platform-Whitepaper_1408231109155174 32.pdf. – Date of access : 07. 05. 2021.
3. *Совпель, И. В.* Система автоматического извлечения знаний из текста и ее приложения / И. В. Совпель // Искусственный интеллект. – 2004. – № 3. – С. 668–679.
4. *Голяк Ю. Д.* Принципиальная схема решения задачи автодополнения пользовательских поисковых запросов на русском языке и ее анализ / Ученые записки ВГУ имени П. М. Машерова : сб. науч. трудов. – Витебск: ВГУ имени П. М. Машерова, 2020. – Т. 31. – С. 142–147.
5. *Чеусов, А. В.* Разработка алгоритмов и технологии построения многоязычного базового лингвистического процессора: дис. … канд. техн. наук: 05.13.17 / А. В. Чеусов. – Минск, 2013. – 116 с.
6. Kpapamih: search-queries. [Electronic resource]. – Mode of access : https://www.kaggle.com/kpapamih/search-queries#searchterms.txt. – Date of access : 07. 05. 2021.
7. Wordstat. Yandex. [Electronic resource]. – Mode of access : https://wordstat.yandex.ru/. – Date of access : 07. 05. 2021.

Поступила в редакцию 18.05.2021