Э. В. Нагарнович

АВТОМАТИЧЕСКАЯ КОРРЕКЦИЯ ОДНОБУКВЕННЫХ ОПЕЧАТОК ВО ФРАНКОЯЗЫЧНЫХ ЗАПРОСАХ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ

Ошибки до сих пор остаются одним из самых малоисследованных лингвистических объектов, хотя почти все ученые, занимающиеся вопросами речевой деятельности, сталкиваются с ними и фиксируют их. В письменном тексте могут быть допущены орфографические ошибки, в результате которых образуются комбинации символов, не являющиеся словами данного языка; семантические ошибки, когда правильные слова переходят в другие

правильные слова, но совсем иные по смыслу; грамматические ошибки, когда слова переходят из одной грамматической формы в другую; пропуски отдельных слов и групп слов; ошибки в знаках препинания и полиграфическом оформлении текста; формульно-цифровые ошибки. Поэтому кроме средств орфографической проверки каждая система автоматической обработки текста должна быть оснащена средствами автоматизированного или автоматического исправления выявленных ошибок.

В статье представлена процедура моделирования автоматической коррекции орфографических ошибок (опечаток) в поисковых запросах франкоязычных пользователей без учета контекста. Материалом исследования послужили однословные и многословные запросы, взятые из лога запросов франкоязычных пользователей системы *Goldfire*, предоставленного компанией *IHS Global*. Данная система ориентирована на семантический поиск технической информации, сведений и данных в области химии, физики, биологии, медицины, строительства и транспорта.

В процессе анализа лога запросов были выделены наиболее типичные ошибки, приведенные в таблице 1. Достаточно распространенным типом ошибок, имеющим вид «замена буквы», является пропуск диакритических знаков в однословных и многословных запросах: **é** – в 115 словах из 800 слов проанализированных запросов; **è** – в 9 словах, в том числе в слове *syst**é**me* вм. *syst**è**me*, **ô** – в 13 словах, в том числе в слове *c**ö**lon* вм. *c**ô**lon*, **â** – в 3 словах. Например, *d**e**clencheur magnétique* вм. *d**é**clencheur, augmentation de la conductivité electrique de poly-m**e**re* вм. *polym**è**re, contr**o**le dimensionnel de courbes* вм. *contr**ô**le, distance mini-male entre les c**a**bles secondaire et les barres de puissance* вм. *c**â**bles.*

Для создания эталонного словаря из лога системы *Goldfire* были выбраны франкоязычные запросы, не содержащие ошибок. Количество единиц в эталонном словаре составило 180 запросов со средней длиной слова, равной 12 буквам. В использованный для распознавания символов алфавит были включены буквы французского алфавита (*A-Z, Éé, Ââ, Êê, Îî, Ôô, Ûû, Àà, Èè, Ùù, Ëë, Ïï, Üü, Ÿÿ, Çç, œ*), а также цифры (*0 – 9*). Размер алфавита составил 51 символ. Эталонный словарь вошел в лингвистическую базу данных системы автоматической коррекции орфографических ошибок в запросах пользователей.

Таблица 1

Типичные ошибки, выделенные из лога запросов франкоязычных пользователей системы *Goldfire*

№ п/п	Тип ошибки	Количество слов в поисковом запросе	Пример
1.	Вставка буквы	одно	*operation => opération*
		два	*constraste image* *constraste => contraste*
		три и более	*élément chauffant pplat sur sol gel* *pplat => plat*

2.	Замена буквы	одно	*polyphosphotiques => polyphosphoriques*
		два	*poluution bactérienne* *poluution => pollution*
		три и более	*diagraphie non nucleaire* *nucleaire => nucléaire*
3.	Перестановка двух соседних букв	одно	*porject => project*
		два	*marjoliane grange* *marjoliane => marjolaine*
		три и более	*contrôle d'accès à un bâtiment par potrail* *potrail => portail*
4.	Пропуск буквы	одно	*maitenance => maintenance* *sol-el => sol-gel*
		два	*lamele colle* *lamele => lamelle*
		три и более	*tranformation d'un liquide en solide* *tranformation => transformation*

Предложенная в статье формальная модель системы коррекции ошибок (опечаток) в запросах франкоязычных пользователей ориентирована, в основном, на однобуквенные искажения слов. Кроме однобуквенных опечаток известен ряд часто встречающихся не однобуквенных ошибок, например перестановка согласных через гласную (*é**mil**ination*). В рамках разработанной формальной модели обрабатываются некоторые не однобуквенные опечатки, однако их конкретные типы не рассматриваются, поскольку при практической реализации системы эти типы могут выбираться произвольно. Общая стратегия выбора того или иного способа поиска и исправления ошибок зависит от количества слов в запросе пользователя. Если запрос состоит из одного слова, то оно сопоставляется с каждым словом как однословных, так и многословных запросов из базы данных. Например, если введен однословный запрос *log**o**ciel* (вм. *log**i**ciel*), то сначала компьютер проверит вхождение этого слова в список однословных запросов. Если совпадения нет, компьютер сопоставит введенный запрос со словами списка многословных запросов. В эталонном корпусе он найдет запрос *logiciel de travail collaboratif* и исправит *log**o**ciel* на *log**i**ciel*. В случае ввода пользователем двухсловного запроса он сопоставляется со всеми двухсловными запросами базы данных. Если совпадения не найдены, компьютер ищет наличие каждого слова запроса в словаре двусловий. Такая процедура помогает достичь более высокой точности коррекции опечаток. Для запросов, состоящих из трех и более слов, была выбрана стратегия исправления ошибок, используемая при восстановлении однословных запросов. При этом каждое слово запроса сопоставляется с каждым словом лингвистической базы данных. Например, в запросе *c**oom**ent éviter la cristallisation d'urée* ошибку в слове компьютер исправит, опираясь на эталонный запрос *c**om**ment réduire le cholestérol*.

Отмеченные выше правила были положены в основу создания алгоритма автоматического обнаружения и коррекции орфографических ошибок в запросах франкоязычных пользователей. Рассмотрим несколько конкретных примеров, иллюстрирующих работу разработанной формальной модели. Например, в процессе обработки запроса *contrôle d'accès à un bâtiment par*

potrail, последовательно сравнивая слова запроса с единицами эталонного словаря, компьютер классифицирует слово *po**t**rail* как ошибочное, поскольку такое слово в словаре отсутствует. Он определяет длину данного слова (семь символов). Следующее действие системы связано с выделением первого и последнего символов слова. Ими будут, соответственно, *p* и *l*. Осуществляя поиск по словарю, компьютер выделяет из него слово длиной в семь символов, начинающееся и заканчивающееся указанными выше буквами. Это слово *po**r**tail*. Следующее действие системы заключается в сравнении анализируемого слова с найденным словом эталонного словаря слева направо и справа налево. При этом компьютер определяет количество несовпадающих букв. В данном случае оно равно двум. Если в словаре есть слова с таким же количеством символов и аналогичными начальной и конечной буквами, например, *parasol*, система выбирает слово, в котором количество не совпавших букв меньше. Сравнив слова *parasol* и *potrail,* где количество не совпавших символов равно пяти, компьютер определяет слово *portail* как единственно правильный вариант и заменяет им слово с ошибкой.

Рассмотрим процедуру автоматической обработки однословного запроса *lame**l**e*, в котором пропущена буква. Такое слово в эталонном словаре отсутствует, поэтому считается ошибочным. Длина слова равна шести символам, а первая и последняя буквы, соответственно, *l* и *e*. В словаре могут быть слова с такими начальными и конечными параметрами, но количество не совпавших букв будет больше исходного. Поэтому, уменьшив количество букв в анализируемом слове на единицу (пять символов) компьютер находит в эталонном словаре слово с аналогичными параметрами. Это слово *laque*. Оно заносится в список 1. Следующим действием системы будет увеличение количества букв в анализируемом слове на единицу (семь символов). С такими параметрами в эталонном словаре есть слово *lamelle*. Оно заносится в список 2. Система сравнивает анализируемое слово *lamele* со словами из списка 1 и списка 2. В результате сравнения остается вариант *lamelle*, и компьютер заменяет слово с ошибкой на найденный в эталонном словаре вариант.

В следующем запросе *diagraphie non nucleaire* система выявляет ошибочное слово *nucleaire*. Определив длину этого слова (девять символов), а также первую и последнюю буквы (*n* и *e)*, компьютер с опорой на данные признаки выделяет из эталонного словаря слова *nucléaire, normative, naissance, nombreuse*. Далее система сравнивает анализируемое слово с каждым из перечисленных выше четырех слов слева направо и справа налево, определяя при этом количество несовпавших букв. В случае со словом *nucléaire* количество несовпавших букв равно единице, со словом *normative* – шести, со словом *naissance* – шести, со словом *nombreuse* – семи буквам. Компьютер определяет слово *nucléaire* как единственно верное и заменяет им ошибочный вариант.

Разработанная формальная модель системы автоматической коррекции ошибок в поисковых запросах франкоязычных пользователей была запрограммирована на языке *Ruby*. В ходе компьютерного эксперимента в систему было введено 200 франкоязычных поисковых запросов, состоящих из разного количества слов и содержащих разные типы ошибок в разных позициях слова. Фрагменты компьютерного эксперимента представлены на рис. 1–5.

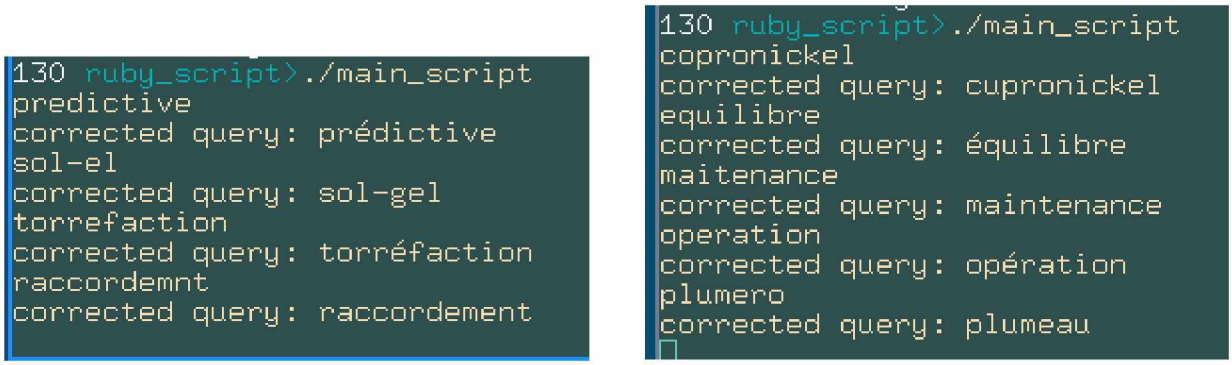

Рис. 1. Исправление ошибок в однословных запросах

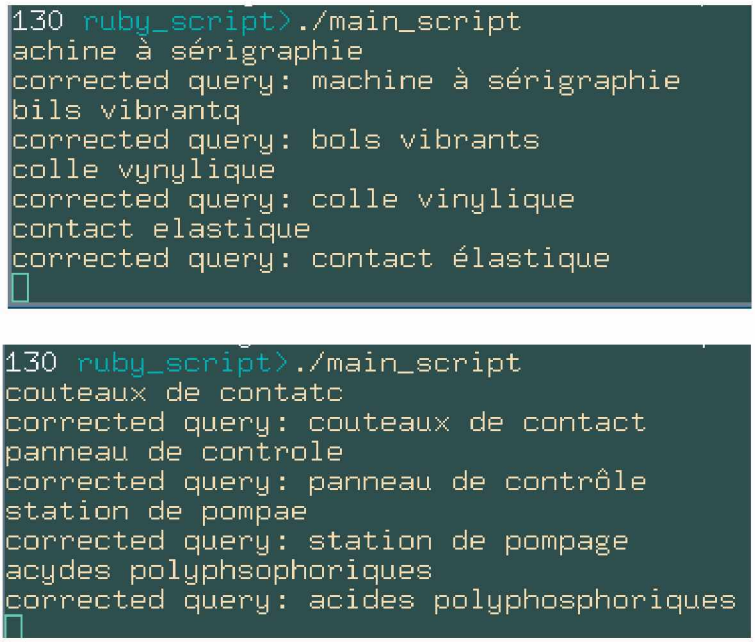

Рис. 2. Исправление ошибок в двухсловных запросах

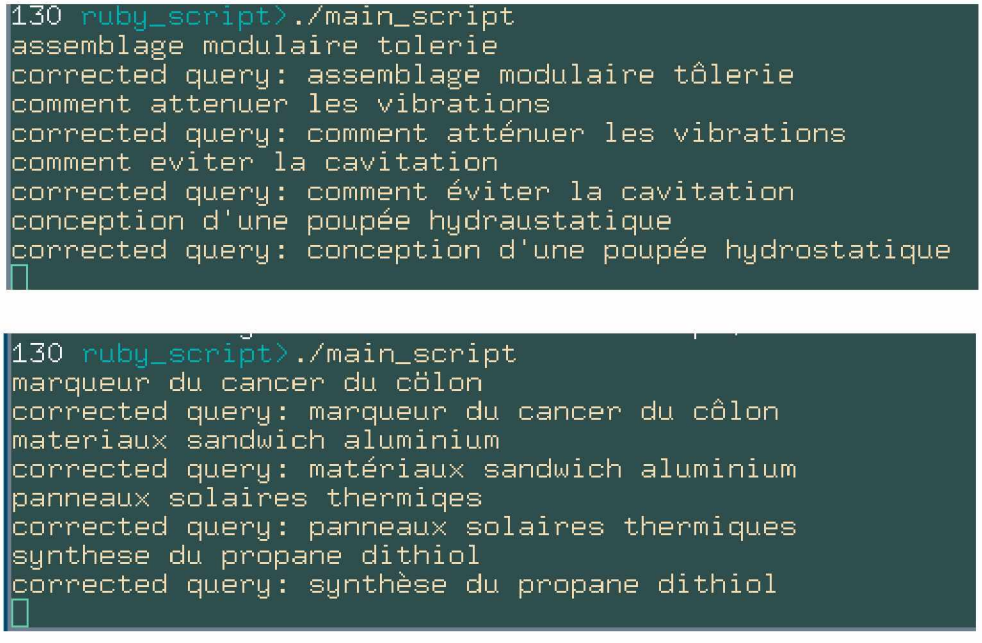

Рис. 3. Исправление ошибок в многословных запросах

```
130 ruby_script>./main_script
ooxydasion
corrected query: oxydation
cholasterol
corrected query: cholestérol
consummation
corrected query: consommation
serigrapie
corrected query: sérigraphie
```

```
130 ruby_script>./main_script
anjectio
corrected query: injection
energetique
corrected query: énergétique
vibrassions
corrected query: vibrations
declenchémant
corrected query: declenchémant
```

Рис. 4. Исправление нескольких ошибок в одном слове

```
130 ruby_script>./main_script
instrumntation nucleaire
corrected query: instrumentation nucléaire
etagement de la combustio
corrected query: étagement de la combustion
pannneau de controle
corrected query: pannneau de contrôle
poluution bacterienne
corrected query: pollution bactérienne
```

```
130 ruby_script>./main_script_2
coment reduire le cholesterol
comment réduire le cholestérol
compositio des pudres luminescentes
composition des poudres luminescentes
comment tmesurer la toxicite d'un fumigene
comment mesurer la toxicité d'un fumigène
amortisseur de presion
amortisseur de pression
coomment éviter la cristalysation d'uree
comment éviter la cristallisation d'urée
```

Рис. 5. Исправление ошибок в каждом слове многословного запроса

Основные результаты тестирования компьютерной системы сводятся к следующему.

1. 90 % запросов, содержащих орфографические ошибки, были обработаны правильно.

2. Примеры 10 % не обработанных запросов и связанные с этим действия компьютера отражены в табл. 2.

3. На поиск и исправление всех ошибок компьютер потратил почти 22 секунды, то есть приблизительно 0,1 секунды на коррекцию одного запроса.

Таблица 2

Примеры не обработанных системой запросов, содержащих ошибки

№ п/п	Действие компьютера	Пример	Кол-во запросов	Комментарий
1.	Исправление не произведено	*abbatement mécanique* *bils vibrants* *conductibilite thermique* *conseve vitamine* *constraste image* *declencheur magnétique* *echauffement excessif* *echauffement primaire* *electro précipitation* *emission acoustique* *etat de l'art contrôle non destructif* *marjoliane grange* *poluution bactérienne* *prevenir l'arthrose* *recuperer l'énergie* *rsique chimique* *systeme d'écriture* *acydes polyphosphoriques* *aviasion civile*	19	10 из 19 запросов на восстановление диакритиков
2.	Неверная замена	*comment detecter un arc serie dans une installation* => *comment détect-er un arc servie dans une installa-tion* (**надо:** *comment detecter un arc sé-rie dans une installation*)	1	В словаре есть пример, который повлиял на исправление ошибки: *comment pétrir une pâte et retir-er les outils ay-ant* ***servie*** au pé-trissage

Представленная формальная модель может быть использована в качестве дополнительного модуля интеллектуальной информационно-поисковой системы для обнаружения и коррекции орфографических ошибок и опечаток в словах поисковых запросов франкоязычных пользователей.