

УДК 004.912

Ярмолюк Р.С.

Хмельницький національний університет, Україна

МЕТОДИ ПОШУКУ ТА КОРЕКЦІЇ ПОМИЛОК В ЗАПИСАХ ЕЛЕКТРОННОГО КАТАЛОГУ

Представлено огляд та аналіз основних методів пошуку та корекції помилок в електронному каталозі. Проаналізовано статистику виникнення символічних спотворень в записях електронного каталогу.

The review and analysis of the main methods of finding and correcting errors in the electronic catalog. Analyzed the statistics of symbolic distortions in the electronic catalog records.

Постановка проблеми. Проблема корекції спотворень, як технічна проблема, зв'язана з ЕОМ, має давню історію. Перші роботи були присвячені виправленню спотворень, які отримувались в результаті передачі символів і зчитування їх оптичними пристроями [1]. Задача полягає у відшуканні та корекції символічних спотворень в записах електронного каталогу.

Аналіз останніх досліджень і публікацій. В даний час найбільший інтерес представляє можливість корекції машиночитаних текстів на природній мові. Хоча автоматично виконується лише виявлення помилок і тільки орфографічного типу, а відповідно корекція ведеться звичайно за участі людини, навіть при такому обмеженому автоматизмі точність і продуктивність вивірки текстів значно зростає [1,2]. Теоретичні і практичні основи проблеми автоматичного пошуку та корекції помилок в записах електронного каталогу розробляли Вершинин М. И., Белоногов Г. Г., Бабко-Малая О. Б., Крауш А. С., Randall B. N., Ballard T, та інші.

Формулювання цілей статті та актуальність досліджень.

На даний час в Україні та за кордоном розроблено чимало автоматизованих бібліотечних систем (АБІС) різного рівня складності та масштабу. Серед таких систем можна виділити УФД/Бібліотека, ІРБІС, МАРК-SQL, КАБІС, UNILIB, LIBER, ALEPH, Руслан. Чимало АБІС є open-source продуктами, зокрема, Koha, ISIS, CDS Invenio, OpenBiblio, Evergreen. Однак аналіз описових можливостей переважної більшості перелічених АБІС показав, що в них відсутні ефективні засоби верифікації інформації в електронних каталогах.

Тому проблема пошуку та виправлення помилок у бібліографічних записах електронного каталогу є досить актуальною [3].

Виклад основних матеріалів дослідження. Аналіз різних джерел [1,4,5] дозволяє встановити, що проблема корекції помилок включає в себе наступні аспекти:

- корекція помилок слів, випадків перетворення правильного слова в інше правильне слово;
- корекція помилок структури слів – пропуск одного чи більше слів, стрічок, абзаців, перестановка слів;
- корекція помилок зв'язку слів, знаків пунктуації;
- корекція символічних помилок, помилок оператора (вставка, заміна, видалення, перестановка одного чи більше символів).

На даний час задачі автоматизації корекції спотворень розглядаються тільки для випадків символічних помилок і помилок пунктуації. Далі під словом спотворення ми будемо розуміти лише символічні спотворення.

При розробці засобів корекції текстових помилок потрібно виділити три основні задачі [1]:

- вивчення характеристик спотворень;
- визначення способу представлення знань;
- розробка алгоритму корекції, що перетворює спотворений текст в покращений.

До характеристик спотворень відносять статистику різних помилок в тексті, що групується за трьома параметрами:

- кількість спотворених символів;
- позиція спотвореного символу в слові;
- тип помилки.

Аналіз спотворень, в тому числі в електронних каталогах бібліотек (ЕКБ) проведений різними дослідниками, дозволяє запропонувати наступну типологію помилок [1]:

- заміна однієї літери на іншу;
- пропуск літер (найчастіше голосних);
- подвоєння літер (найчастіше приголосних);
- заміна літери на подібну по звучанню;
- перестановка літер;
- вставка зайвих літер;
- вставка лишніх пробілів перед чи після слів;
- нетипові помилки;
- комбінація попередніх помилок.

Вивчення статистики помилок показує:

- в середньому в записах ЕКБ частота помилок складає 0,1% в тому числі одно літерні помилки складають 85-95%;

- найбільш ймовірне спотворення початку слова; для слів довжини 3-8 символів найбільш ймовірні помилки в трьох-чотирьох позиціях;

- приблизний розподіл помилок: пропуск літер – 30-40%, вставка – 25-35%, заміна – 15-20%, перестановка – 10-15%;

- помилки в голосних (вставка і пропуск) зустрічаються частіше ніж в приголосних;

- найбільш ймовірні помилки в початкових лексичних одиницях полів бібліографічного запису.

Крім того характеристики спотворень повинні орієнтуватись на природну мову, предметну область ЕКБ і на конкретного оператора, тобто на певне представлення знань.

Представлення знань звичайно є способом опису всієї мови, її синтаксису, морфології, лексики чи певної частини мови, як, наприклад, представлення знання з конкретного документу, призначеного для перевірки. Існує багато різних форм представлення знань, основними з яких є морфологічні аналізатори, словники основ, словоформ і словосполучень різного типу, масиви n-грам (вибірki n-символів з множини всіх символів слів) і статистичні моделі мови, звичайно представлені у вигляді ймовірностей переходу, тобто, ймовірність того, що з'явиться дана літера, якщо відома попередня послідовність літер (ланцюги Маркова) [1].

Алгоритм корекції завжди базується на певному способі представлення знань, але не завжди використовує статистичні характеристики спотворень.

Функціонально всі методи боротьби з помилками можна розділити на ті, що виявляють помилки і ті, що коректують.

Алгоритм виявлення полягає в «читанні» тексту і відборі неправильних слів. Ціль корекції полягає в спробі виправити помилковий текст, що може бути зроблено наступним способом [1]:

- послідовне виконання алгоритмів виявлення і корекції. В такому випадку для кожного слова, що визначено, як «неправильне», відбувається пошук одного або декількох кандидатів на виправлення;

- текст апріорі вважається спотвореним і виявлення та корекція слів в тексті відбувається одночасно.

Порівняльний аналіз показав, що перший підхід є більш ефективним, так, як в другому випадку збільшується ймовірність помилкової корекції.

За принципом функціональності методи пошуку та корекції умовно можна розділити на наступні групи [1]:

- методи порівняння, коли інформація з ЕКБ порівнюється з інформацією в словнику системи;

- методи породження, коли на основі інформації з ЕКБ і певної моделі породження слів відбувається генерація правильної порції інформації;

- комбіновані методи;

- контекстні методи, які використовують інформацію контексту слів.

В свою чергу, методи порівняння можна розділити на методи точного і наближеного порівняння.

Методи точно порівняння завжди засновані на різних представленнях словників і полягають в співставленні рядка символів виділеного із тексту з одиницями словника на предмет їх повного збігу. Даний метод завжди включає в себе виявлення неправильних слів, а алгоритм корекції зводиться до створення різних модифікацій спотвореного слова і пошуку цих варіантів за допомогою алгоритму виявлення.

Методи наближеного порівняння можуть ґрунтуватися на словниках і таблицях n -грам і включати в себе два етапи [1]:

- вибір словникових одиниць, найбільше схожих на рядок символів. Звичайно ця задача розв'язується за допомогою різних методів кодування одиниць словника;

- вибір єдиного кандидата із кількох відібраних на першому етапі. Ця задача вирішується за допомогою визначення відстані близькості між двома символічними стрічками і обрахунку цієї відстані між спотвореним словом і всіма іншими кандидатами з ціллю виявлення найменшої відстані.

Найбільш вивчені є n -грамні методи, що засновані на порівнянні n -грам спотвореного слова зі n -грамами зі словника. Кількість різних літер в кожній мові обмежена і літери можуть бути просто перераховані. Важливим характеристиками тексту є повторюваність літер, пар літер(біграм), трійок літер(триграм) і, в загальному випадку, n -грам. Сполучуваність літер є стабільною характеристикою для кожної мови, що заснована на алфавіті. Таким чином, знайдені в тексті n -грами, що не входять в таблицю n -грам, являються ознакою можливої помилки [1].

Методи породження відповідають другому способу розв'язку задачі корекції, а саме одночасному виявленню і корекції помилок. Вони базуються на статистичних моделях мови і гіпотезі про те, що

природна мова – ланцюг Маркова n -го порядку. Статистична модель зазвичай представляє собою таблицю з ймовірностями переходу і процес корекції складається в послідовному пошуку найбільш ймовірного символу при умові, що відомий попередній ланцюг довжини n [1].

Методи породження не показали хороших результатів і для підвищення ефективності були розроблені комбіновані методи. Одним із прикладів є послідовне виконання алгоритму породження і алгоритму порівняння.

Контекстні методи являються більш перспективними, оскільки без урахування контексту слів і інформації про зв'язки між словами 100% корекція не можлива.

Розробку засобів корекції потрібно розбити на декілька етапів [1]:

- вивчення характеристик спотворень в залежності від конкретних умов, включаючи статистику помилок оператора;
- представлення знань, включаючи лексику мови;
- розробка алгоритму виявлення, що базується на словниках основ чи словоформ;
- використання ефективного алгоритму на основі словникових методів співставлення.

Специфічні помилки в записах ЕКБ не можуть бути виправлені формальними методами, так як потребують аналізу змістової компоненти поля (підполя) [1]. Тільки деякі підполя, які мають характерні признаки, наприклад, числові, можуть бути проаналізовані автоматично на предмет наявності помилок.

Висновки. Таким чином, традиційні методи пошуку та виправлення помилок в ЕКБ потребують наявності великих словників, таблиць n -грам, достатньої кількості даних по статистиці помилок та постійного втручання оператора, що не завжди є можливим. Тому актуальною залишається проблема розробки нових ефективних методів та засобів верифікації інформації в ЕКБ.

Література

1. Вершинин М. И. Электронный каталог проблемы и решения / М. И. Вершинин. – СПб. : ПРОФЕССИЯ, 2007. – 233с.
- 2.Randall B. N. Spelling errors in data bases: shadow or substance? / B. N. Randall // Libr. Resources a. techn. Services. – 1999. – Vol. 43, №3. – P. 161 – 169.
- 3.Ярмолук Р. С. Основні типи та джерела помилок у записах електронного каталогу / Р. С. Ярмолук // Вісник Національного Університету «Львівська політехніка». Інформаційні системи та мережі, - 2010. - № 689. – С. 348-357.

4. Гельбух А. Ф. Исправление орфографических ошибок с помощью перебора, управляемого морфологическим словарем. / А. Ф. Гельбух // НТИ. Сер. 2. – 1993. - №5 – С. 23-30.

5. Dameran F. J. A technique for computer detection a correction of spelling errors / F. J. Dameran // Communications of the ACM. – 1964. - №7 – P. 171-176.