

А.І. КУРСІН, зав. лаб. НТУ "ХПІ", Харків

МЕТОД НОРМАЛІЗАЦІЇ СЛОВОФОРМ ЗІ СЛОВНИКОМ ПОЧАТКОВИХ ФОРМ СЛІВ БЕЗ ГРАМАТИЧНОЇ ІНФОРМАЦІЇ

Запропоновано оригінальний метод нормалізації словоформ флексивних мов з використанням словника початкових форм слів, в якому слова не мають кодів типу словозміни. Такий алгоритм хоча і використовує словник, але не потребує граматичного кодування слів у ньому, хоча і вираховує можливі початкові форми за таблицею псевдо-закінчень, але досягає більшої точності, перевіряючи результати за словником. Табл.: 2. Бібліогр.: 8 назв.

Ключові слова: нормалізація словоформ, словник початкових форм.

Постановка проблеми і аналіз літератури. Неодмінною складовою більшості систем автоматичної обробки природномовної текстової інформації є етап морфологічного аналізу словоформ. На цьому етапі встановлюється початкова форма слова, його граматичний клас та граматична інформація щодо форми слова. Для великого класу задач, наприклад, інформаційного пошуку, буває достатньо лише встановлення початкової форми слова (ця операція ще називається нормалізацією словоформ або лемматизацією).

Традиційно алгоритми морфологічного аналізу розділяються на словникові і безсловникові. Словниковий алгоритм [1 – 3] оперує словником слів або основ з приписаною їм інформацією щодо граматичних категорій слова та типу словозміни [4]. Одним з суттєвих недоліків такого підходу є великі трудозатрати на створення такого словника, бо граматичне кодування слів попри всі спроби автоматизації виконується в значній мірі вручну. Безсловниковий алгоритм [5, 6] замість словника використовує списки флексій, квазі-флексій, квазі-суфіксів, притаманних словам даної мови, і робить аналіз, перевіряючи кінець словоформи за цими списками. Безумовною перевагою таких алгоритмів є те, що множина слів, які можуть бути ними оброблені, не обмежується рамками словників, що використовуються. Але безсловникові алгоритми зазвичай відрізняє менша точність аналізу, більша кількість хибних варіантів, хоча й дещо більша швидкість роботи.

Мета статті – описати оригінальний алгоритм нормалізації словоформ, який поєднує властивості обох типів алгоритмів морфологічного аналізу та частково позбавлений недоліків їх обох. Такий алгоритм хоча і використовує словник, але не потребує граматичного

кодування слів у ньому, хоча і вираховує можливі початкові форми за таблицею псевдо-закінчень, але досягає більшої точності, перевіряючи результати за словником.

Опис задачі. Скажемо декілька слів про задачу, яка дала поштовх до розробки алгоритму. Створювалася програма підтримки електронних словників з функцією пошуку в них довільних форм слів. Виявилось, що словники, якими мала бути наповнена система, містять багато слів відсутніх у словникові наявної процедури морфологічного аналізу, а введення їх до цього словника разом з граматичним кодуванням було неприйнятним через стислі строки відведені для виконання робіт. Тоді з'явилася ідея, використовуючи морфологічні таблиці словникового алгоритму, генерувати всі можливі початкові форми для даної довільної форми слова, а потім "просіювати" їх по списку початкових форм слів словника.

Опис алгоритму. Класичний словниковий алгоритм морфологічного аналізу [1, 2] разом зі словником використовує морфологічні таблиці, зокрема таблицю парадигм. Кожний з рядків цієї таблиці відповідає певній парадигмі або типу словозміни, а кожна позиція в рядку – певній формі слова і містить відповідну флексію. На базі цієї таблиці ми створили власну, кожний рядок якої містить одну з флексій непрямих форм слів та всі флексії початкових форм, що зустрічаються спільно з нею в парадигмах. Фрагмент такої таблиці показаний в табл. 1.

Таблиця 1. Флексії непрямой форми та початкових форм

Флексія непрямой форми	Флексії початкових форм
-ет	-ть, -ять, -ти, -уть, -вать, -ать, -ить, -отъ, -ь
-ете	-ть, -ять, -ти, -уть, -вать, -ать, -ить, -отъ, -ь
-етесь	-ться, -яться, -утся, -атся, -отся, -тись, -ваются, -иься, -ься
-ется	-ться, -яться, -утся, -атся, -отся, -тись, -ваются, -иься, -ься
-ех	-и
-ехот	-иста
-ехстах	-иста
-ец	-цы, -ьца, -ца, -ьцо
-ешъ	-ть, -ять, -ти, -уть, -вать, -ать, -ить, -отъ, -ь
-ешься	-ться, -яться, -утся, -атся, -отся, -тись, -ваются, -иься, -ься

Для певної словоформи алгоритм знаходить в таблиці всі можливі для неї флексії непрямих форм. Та будує потенційні варіанти початкових форм слова, використовуючи флексії з відповідних рядків таблиці.

Приклад такого списку для російської словоформи "девочкой" наведений в табл. 2.

Таблиця 2. Список для словоформи "девочкой"

девочкий	девочкоить	девочкойя
девочкий	девочкить	девочка
девочкая	девочкетъ	девочко
девочкое	девочкоять	

Потім проводиться пошук початкових форм зі списку в індексі початкових форм слів словника. Звісно, згенерований список початкових форм містить багато хибних варіантів. Але зазвичай лише один з них збігається з якою-небудь початковою формою зі словника. Ми були самі здивовані результатами попередніх тестів, які показали, що алгоритм "не сходиться", тобто дає більше одного можливого результату, або не дає жодного збігу, лише для 6% російських словоформ і 4% українських. До цих відсотків входили також і дійсні омоформи, тобто такі випадки, де і людина не може зробити однозначну нормалізацію, не вдаючись до контексту.

Згодом морфологічна таблиця даного алгоритму була вдосконалена за матеріалами граматичного словника Залізняка [4]. Всі непрямі закінчення довжини менш ніж 4 були розширенні до цієї довжини додаванням останніх літер основ, з якими вони вживаються. Таке ж розширення було проведене і з деякими окремими закінченнями, які давали суттєву кількість хибних результатів. Короткі слова (переважно службові слова, займенники, тощо) були винесені в окремий список стоп-слів (300 – 400 одиниць). Також, зважаючи на задачу, в якій застосовувався даний алгоритм, виявилось доцільним ввести певний пріоритет закінчень у таблиці. Наприклад, непрямі форми дієприкметників (напр., "пораненими") спершу приводяться до чоловічого роду однини ("поранений"), і лише за відсутності такої форми в словнику, проводиться пошук неозначеної форми дієслова ("поранити").

В результаті зазначених вдосконалень кількість випадків "несходимості" алгоритму зменшилася для російської мови – до 3,3%, з яких 2,7% складають дійсні омоформи.

Зауваження щодо реалізації. Будь-який алгоритм, що стосується пошуку інформації у словнику і претендує на практичну реалізацію, мусить мати достатньо швидку пошукову частину. В нашому випадку кількість пошукових операцій у словнику зростає завдяки великій кількості хибних початкових форм, які приймають участь у пошуку (в середньому в 15,1 рази для російської мови). Тому треба було вжити

заходи для забезпечення достатньої швидкості пошуку. Ми використовували індексацію слів словника за допомогою В-дерева [7] з ключами мінливої довжини зі стисненням методом Купера [8]. Така структура поєднує можливість швидкого пошуку строкових ключів з компактним об'ємом індексу. Збільшення кількості пошукових ключів для однієї словоформи нівелюється пакетним пошуком в індексі відразу всього списку початкових форм. Швидкість нормалізації в тестах складала приблизно 25 мс на тисячу словоформ (AMD Athlon 2.2 ГГц).

На базі описаного алгоритму ми розробили процедури нормалізації для російської та української мов, а також для англійської мови. В останньому випадку замість використання таблиць флексій працює підпрограма, яка генерує можливі початкові форми слів за правилами англійської мови. Ці процедури нормалізації використовуються в ряді інформаційно-пошукових та лінгвістичних систем, як в своєму первинному варіанті – для пошуку довільної форми слова у списку початкових форм слів, так і у вигляді окремого модуля нормалізації, який комплектується достатньо великим словником слів відповідної мови. Перевага практичного використання даного алгоритму в тому, що укладачі словника позбавляються кодування ключових слів щодо типу їх словозміни. Проблема обробки незнайомих слів ми не вирішуємо кардинально, але процедура поповнення словника новими словами також значно спрощується. Завдяки використанню достатньо швидкого пошукового індексу, наша процедура нормалізації може замінити собою пошук ключових слів в індексі інформаційно-пошукової системи.

Висновки. Описаний нами алгоритм доводить, що достатньо точну нормалізацію словоформ флексивних мов можна проводити і без кодування слів у словнику щодо типу словозміни. Це робить перспективним застосування алгоритму в інформаційно-пошукових системах та системах обробки текстової інформації.

Список літератури: 1. Белоногов Г.Г. Алгоритм морфологического анализа русских слов / Г.Г. Белоногов, Ю.Г. Зеленков // *Вопр. инф. теории и практики.* – М.: 1985. – № 53. – С. 62–93. 2. Антошкина Ж.Г. Морфологический процессор русского языка / Ж.Г. Антошкина // *Бюллетень машинного фонда русского языка.* – М.: 1996. – Вып. 3. – С. 53–57. 3. Сегалович И. Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов / И. Сегалович, М. Маслов // *Труды международного семинара Диалог'98 по компьютерной лингвистике и её приложениям.* – Том 2. – 1998. – С. 547–552. 4. Зализняк А.А. Грамматический словарь русского языка / А.А. Зализняк. – М.: Русские словари, 2003. – 795 с. 5. Sheremetyeva S. Rapid Deployment Morphology / S. Sheremetyeva, W. Jin, S. Nirenburg // *Machine Translation.* – 1998. – Vol. 13. – № 4. – С. 239–268. 6. Ножов И.М. Прикладной морфологический анализ без словаря / И.М. Ножов // *КИИ-2000. Труды конференции.* – М.: 2000. – Т.1. – С. 424–429. 7. Bayer R. Organization and Maintenance of Large Ordered Indexes / R. Bayer, E. McCreight // *Acta*

Informatica. – 1972. – № 1 (3). – P. 173–189. 8. Cooper W.S. The storage problem / W.S. Cooper // Mechanical Translation. – 1958. – Vol. 5. – № 2. – P. 74–83.

Стаття представлена д.т.н., проф. НТУ "ХПІ" Серковим О.А.

УДК 81.322.3:81.366

Метод нормализации словоформ со словарём начальных форм слов без грамматической информации / Курсин А.И. // Вестник НТУ "ХПИ". Тематический выпуск: Информатика и моделирование. — Харьков: НТУ "ХПИ". – 2011. – № 17. – С. 76 – 80.

Предложен оригинальный метод нормализации словоформ флективных языков с использованием словаря начальных форм слов, в котором у слов отсутствуют коды типа словоизменения. Такой алгоритм хотя и использует словарь, но не нуждается в грамматической кодировке слов в нем, хотя и высчитывает возможные начальные формы за таблицей псевдо-окончаний, но достигает большей точности, проверяя результаты по словарю. Табл.: 2. Библиогр.: 8 назв.

Ключевые слова: нормализация словоформ, словарь начальных форм.

UDC 81.322.3:81.366

Word form normalization method using a dictionary of citation word forms without grammatical codes / Kursin A.I. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – №. 17. – P. 76 – 80.

An original method of word form normalization is proposed for inflectional languages. The method eliminates the necessity of inflection type codes for words in a dictionary used for normalization. Such algorithm uses a dictionary though, but does not need grammatical code of words in him, though calculates possible initial forms after the table of pseudo-end, but arrives at greater exactness, checking up results after a dictionary. Tabl.: 2. Refs.: 8 titles

Keywords: word form normalization, dictionary of initial forms.

Надійшла до редакції 22.04.2010