STUDIA INFORMATICA 2004
Volume 25 Number 2 (58)

Piotr R. KASPRZYK
Politechnika Śląska, Instytut Matematyki

INTELIGENTNY SYSTEM AUTOMATYCZNEJ POPRAWY BŁĘDÓW PISOWNI

Streszczenie. Artykuł przedstawia koncepcję inteligentnego systemu poprawy błędów pisowni języka polskiego. System ma brać pod uwagę kontekst błędów, typ dokumentu oraz gramatykę i semantykę języka polskiego. Pewne elementy systemu bazują na rozwiązaniach wykorzystywanych w systemach już istniejących na Politechnice Śląskiej.

Słowa kluczowe: sprawdzanie pisowni, przetwarzanie języka naturalnego.

INTELLIGENT SYSTEM FOR AUTOMATICAL SPELLING CORRECTION

Abstract. The article presents a conception of an intelligent system for spelling correction in Polish language. The system should correct errors using the information of the document type and the context in terms of various aspects of grammar and semantics of Polish language. Certain elements of the system are to be based on solutions actually used in tools developed in Silesian University of Technology.

Keywords: spelling correction, natural language processing

1. Wprowadzenie

Stykając się ze słowem pisanym tworzonym w silnych ograniczeniach czasowych, spotykamy się z błędami pisowni. Na błędy te natrafiamy częściej niż kilka czy kilkanaście lat temu. Autor spotkał się nawet ze stwierdzeniem, że istnieje „prawo publikacji" mówiące, że w każdej publikacji jest co najmniej jedna literówka. Słowo pisane w erze Internetu – to nie tylko publikacje; na co dzień mamy do czynienia z mailami (czyżby renesans epistolografii?) i rozmowami przez komunikatory internetowe – tutaj swoboda języka jest

większa. Powszechność stosowania programu, niestety, wiąże się z częstszymi błędami pisowni.

Wraz z upowszechnianiem się składu komputerowego coraz popularniejsze się stają programy do sprawdzania pisowni, większość – jeżeli nie wszystkie – z edytorów tekstu ma lepszy czy gorszy sposób kontroli pisowni użytkownika. Wyposażenie komunikatorów internetowych w system sprawdzania pisowni jest prawdopodobnie kwestią czasu. Istniejące narzędzia do sprawdzania pisowni mają jednak sporo wad. Oto kilka z nich:

- konieczność ścisłej interakcji użytkownika – systemy sprawdzania pisowni na ogół wskazują tylko wyraz, który zdaniem systemu, napisany jest błędnie i w jakiś sposób wyróżniają go graficznie (np. podkreślają falistą czerwoną linią), a użytkownik ma w jakiś sposób zareagować,
- uboga lista propozycji – generowane przez systemy propozycje wyrazów, na który można by zamienić wyraz napisany błędnie bazują na niewielu kryteriach podobieństwa, na ogół na podobieństwie w sensie maszynopisania,
- brak dalszej analizy listy propozycji – lista propozycji nie jest w żaden sposób analizowana w kontekście wyrazów sąsiednich, nie bierze się pod uwagę ani formy ani treści, tj. ani gramatyki, ani semantyki języka, w którym pisany jest tekst.

W związku z tym zasadny wydaje się pomysł stworzenia systemu, który z jednej strony byłby w miarę możliwości automatyczny – to znaczy ograniczałby do minimum konieczną interakcję użytkownika, a z drugiej byłby inteligentny, działałby w oparciu o kontekst i wykorzystywałby specyfikę języka polskiego.

2. Składniki systemu

System, o którym myślimy, działać powinien wieloetapowo. Etapy te są w dużym stopniu niezależne od siebie, a zatem warto system podzielić na kilka podsystemów realizujących poszczególne zadania.

Poszczególne etapy działania to:

- wykrycie błędu,
- wygenerowanie listy propozycji poprawy,
- analiza morfologiczna propozycji (może być już wykonana w poprzednim etapie),
- analiza składniowa zdania, w którym znaleźliśmy błąd, przy uwzględnieniu każdej z propozycji,
- analiza syntaktyczno–generatywna przy uwzględnieniu każdej z propozycji,
- analiza kontekstu znaczeniowego,
- wybór najwłaściwszej propozycji.

2.1. Detekcja błędu

Ponieważ system ma służyć poprawie błędów pisowni, pierwszym etapem jego działania musi być ustalenie, co jest błędem, który będziemy poprawiać. Powszechnie stosowaną metodą jest wykorzystanie słownika. Każde słowo w sprawdzanym tekście jest wyszukiwane w leksykonie słów poprawnych. Jeśli nie zostanie znalezione – uznawane jest za błędne. W języku polskim – z uwagi na wielość form gramatycznych – wyszukiwanie musi zostać poprzedzone ustaleniem formy podstawowej słowa testowanego (w przeciwnym wypadku mielibyśmy do czynienia z leksykonami ogromnych rozmiarów) i błąd wykrywany jest już na tym etapie. Oprócz słownikowych, stosowane są też inne metody: badanie budowy wyrazów, wykrywanie nietypowych kombinacji głosek czy liter. Wszystko to jest robione przy założeniu, że wyraz, który zaklasyfikujemy jako istniejący w języku, jest słowem napisanym poprawnie. Pomija to pewien rodzaj błędów polegający na zniekształceniu słowa poprawnego w inne słowo poprawne leksykalnie. W języku polskim błędem tego rodzaju jest niepoprawna łączna lub rozdzielna pisownia. (Według statystyk F. Nowaka [6], zła łączna/rozdzielna pisownia jest najpopularniejszym błędem ortograficznym wśród uczniów w Polsce).

Istnieją też metody [2], które wykorzystują statystyki występowania poszczególnych słów w otoczeniu innych; przy stosowaniu tych metod można ustalić próg wrażliwości, powyżej którego traktujemy słowo bardzo rzadkie jako błędne, a poniżej – poprawne. Próg ten zależeć powinien od rodzaju tekstu, a dokładniej od „zaufania", jakim darzymy jego autora.

2.2. Lista kandydatów

Mając dane słowo, które uważamy z pewnych przyczyn za błędne, naszym zadaniem jest stwierdzić, jakie słowo było intencją piszącego. Analizujemy tutaj wyłącznie błędy pisowni, nie zajmując się stylem wypowiedzi autora, w związku z tym zakładamy, że mamy do czynienia ze słowem *w jakiś sposób* podobnym do zamierzonego. I tutaj natrafiamy na pierwszy problem – rodzaj błędu w dużym stopniu zależy od sposobu, w jaki powstał tekst pisany. W tekstach pisanych na klawiaturze komputera najczęściej (wg [1] i [7] 80%–95%) występują błędy polegające na zamianie, wstawieniu lub usunięciu litery, wynikające z naciśnięcia klawisza sąsiedniego, niedociśnięcia jakiegoś klawisza lub też zamianie kolejności klawiszy. W tekstach pochodzących z aplikacji OCR częściej mamy do czynienia z zamianą znaku na znak podobny optycznie, zamianą kilku znaków na jeden lub odwrotnie. W aplikacjach, w których mamy do czynienia z tekstowym zapisem rozmów w czasie rzeczywistych, mamy często (oprócz błędów klawiatury) do czynienia z błędami ortograficznymi, wynikającymi z niewiedzy piszącego, w jaki sposób dane słowo powinno być napisane: najczęściej słowo jest zamieniane na podobnie brzmiące.

Wiedząc, jakie jest źródło analizowanego tekstu, możemy ograniczyć liczbę podobnych wyrazów do wyrazów podobnych w konkretny sposób. Gdy tego nie wiemy, musimy wziąć pod uwagę wszystkie możliwości powstania błędu. Następnie – dysponując odpowiednio poindeksowanym słownikiem – generujemy listę słów podobnych do danego, a poprawnych leksykalnie. Specyfika języka polskiego wymaga, żeby brać pod uwage wszystkie możliwe formy wystąpienia danego słowa, a zatem słownik nie może ograniczać się do form podstawowych lub musi istnieć metoda generowania innych form z formy podstawowej.

2.3. Analiza morfologiczna propozycji

W tym momencie większość programów sprawdzających pisownie kończy pracę – „mamy listę propozycji, a Ty, użytkowniku, wybierz sobie tę poprawną". My chcemy posunąć się dalej, sami określić, która z form jest poprawna, badając poprawność formalną i semantyczną. Więc jeśli nie mamy jeszcze tej informacji (pamiętajmy, że listę propozycji mamy wygenerowaną z pewnego słownika; w „porządnym" słowniku każde słowo jest opisane, chociażby w kryteriach części mowy i formy), musimy zbadać morfologię każdego z proponowanych wyrazów. Określamy zatem, z jaką częścią mowy mamy do czynienia oraz z którą jej formą. Ponieważ w języku polskim występują homonimy, analiza morfologiczna pojedynczego wyrazu może dać kilka możliwych interpretacji. Wszystkie je bierzemy pod uwagę.

2.4. Poprawność składniowa

Od tego momentu zaczynamy badać kontekst. Pierwszy, który bierzemy pod uwagę, jest kontekst składniowy. Badamy mianowicie zdanie, w którym znajduje się błędny wyraz i staramy się dopasować do niego propozycje. Chodzi nam tutaj o dopasowanie formalne, czyli zgodność ze sobą poszczególnych części mowy – liczby, osoby czy też rodzaju – w sensie zasad składni języka polskiego. Jeśli zdanie z konkretnymi propozycjami nie byłoby poprawne formalnie, skreślamy te propozycje z listy.

2.5. Poprawność syntaktyczno–semantyczna

Na liście znajdują się tylko propozycje tworzące wraz z otoczeniem zdanie poprawne składniowo. Język polski ma jednak więcej ograniczeń w tworzeniu wypowiedzi. Dla czasowników istnieje rekcja – pewne wiążą się z dopełniaczem, a inne z biernikiem; dla każdego z nich istnieją przypadki użycia, które determinują istnienie i formy użycia innych określeń, a nawet kategorie znaczeniowe wyrazów określanych. Konteksty te wiążą się

z poszczególnymi słowami, nie zależą wyłącznie od ich formy. Podobnie jak w czwartym etapie, propozycje, dla których dana wypowiedź nie jest poprawna, odrzucamy z listy.

2.6. Kontekst znaczeniowy

Jeśli ciągle na liście mamy więcej niż jeden wyraz, powinniśmy w jakiś sposób określić, który z nich jest najlepszy. Pewien przyczynek do dokonania decyzji mogą nam dać wspomniane już powyżej metody statystyczne, wykorzystywane do wykrywania błędu będącego słowem poprawnym leksykalnie. Jeśli dysponujemy danymi statystycznymi dotyczącymi użycia poszczególnych słów w pobliżu innych, możemy przyjąć, że korzystniejsza jest dla nas propozycja, która w pewnym kontekście występuje częściej. Rolę kontekstu gra tutaj bezpośrednie otoczenie słowa poprawianego, bez uwzględnienia gramatyki.

2.7. Wybór właściwej propozycji

Mamy listę słów, wszystkie pasują do kontekstu. W jaki sposób wybrać „to jedyne", które autor miał na myśli? Coś o tych słowach wiemy – wiemy, że są podobne do słowa błędnie napisanego. Logiczne będzie tutaj założenie, że bardziej podobne słowo jest bardziej prawdopodobnym rozwiązaniem problemu. Najbardziej oczywistą miarą podobieństwa jest ilość zmian litera-litera czy też głoska-głoska w słowie poprawnym w celu otrzymania słowa błędnego. Jeśli żadna z propozycji nie przoduje w podobieństwie, a dysponujemy danymi statystycznymi dotyczącymi:

1) częstości występowania zdań, o konkretnym rozbiorze gramatycznym,
2) częstości występowania kontekstu orzeczenia zdania,
3) częstości występowania słów o takiej treści w kontekście słów sąsiednich,

możemy te dane wykorzystać i jeśli jakaś częstość jest zdecydowanie większa, wybrać propozycję jej odpowiadającą. Jeśli natomiast tymi danymi nie dysponujemy, cóż – taki nasz los – musimy spytać użytkownika, co sądzi o naszych propozycjach. Jeśli doceni naszą pracę, a takie założenie przyjmujemy, na pewno coś wybierze.

3. Zalety proponowanego rozwiązania

Jak widać – nie zakładamy stuprocentowego poprawnego działania systemu. Życie jest ciekawsze i niesie dużo więcej niespodzianek, niż może przewidzieć jakikolwiek program diagnostyczny, więc nie spodziewamy się, że nasz system będzie działał zawsze dobrze i będzie sobie radził z niuansami bardziej skomplikowanych fragmentów literatury polskiej.

Oczekujemy od niego tylko, żeby był lepszy od znanych rozwiązań i zorientowany na język polski, który zdaniem autora, niesie na tyle dużo informacji i jest na tyle skomplikowany, że trudno jest przez przypadek napisać dłuższe zdanie poprawnie. Wiedzą o tym na pewno obcokrajowcy mieszkający w Polsce...

Podsumujmy więc zalety systemu, który miałby powstać, w porównaniu rozwiązaniami istniejącymi:

- ograniczenie do minimum koniecznej interakcji użytkownika,
- generowanie propozycji uzależnione od rodzaju nośnika – czyli od możliwych rodzajów błędu,
- wykorzystanie silnych ograniczeń generowania prawidłowych wypowiedzi w języku polskim.

4. Co już jest, a czego jeszcze nie ma

Politechnika Śląska, w ramach pracy Instytutu Informatyki, stworzyła już pewne narzędzia analizy języka, które można wykorzystać przy budowie systemu. Są to: POLMORF – system do analizy morfologicznej języka polskiego, PolSyn – system do analizy składniowej. Ponadto były podjęte próby analizy poprawności wybranych klas błędów semantycznych. Wykorzystywany przy nich był „Słownik syntaktyczno-generatywny czasowników języka polskiego" Kazimierza Polańskiego, określający reguły tworzenia wypowiedzi w języku polskim (piąty etap działania naszego systemu).

Do dyspozycji mamy także – co jest bardzo ważne – sporą wiedzę i doświadczenie autorów systemów automatycznej poprawy pisowni tworzonych i działających na całym świecie. Problemy automatycznego poprawiania pisowni nie są nowe; w interesującej nas formie analizowane są od czasów powstania komputerowego składu tekstu, czyli od końca lat 80. XX w. Dużą pomocą jest sieć Internet, a w szczególności witryna ACLWEB, na której umieszczanych jest na bieżąco bardzo dużo artykułów dotyczących przetwarzania języka.

Pozostałe części systemu, które pozostały do wykonania, to:

- podsystem detekcji błędów bazujący na metodach słownikowych lub nie,
- podsystem generacji listy możliwych propozycji,
- podsystem analizy kontekstowej opierający się na metodach statystycznych.

Należy też rozszerzyć funkcjonalność systemów PolSyn i analizy semantycznej na większą klasę zdań, na razie ich działanie jest dość ograniczone. W ramach pracy nad metodami statystycznymi należy opracować dużą bazę uczącą, na podstawie której statystyki powinny być opracowywane. Widać więc dużo możliwych kierunków badań, ale – o czym

autor jest przekonany – sukces, czyli powstanie inteligentnie działającego systemu automatycznej poprawy pisowni, jest realny.

LITERATURA

1. Damerau F. J.: A technique for computer detection and correction of spelling errors. Communications of the A.C.M., vol. 7, pp.171–176, March 1964.
2. Golding A.: A Bayesian hybrid method for context–sensitive spelling correction. Proceedings of the Third Workshop on Very Large Corpora, p. 39–53, Boston, MA.
3. Hajic J., Drozd J.: Spelling-checking for Highly Inflective Languages. ACL Web archives, http://www.aclweb.org
4. Kukich K.: Techniques for Automatically Correcting Words in Text. ACM Computing Surveys, Vol. 24, No. 4, December 1992.
5. Mitton R.: Spellchecking by computer. Journal of Simplified Spelling Society, Vol. 20, No 1, 1996 pp 4–11.
6. Nowak F.: Metoda statystyczna (kwantytatywna) w nauczaniu ortografii. Wydawnictwa uczelniane WSP w Bydgoszczy, poz. 351, Bydgoszcz 1989.
7. Pollock J., Zamora A.: Automatic spelling correction in scientific and scholarly text. Communications of the A.C.M., vol. 27, no. 4, pp.358–368, April 1984.

Recenzent: Dr inż. Nina Suszczańska

Wpłynęło do Redakcji 12 lutego 2004

Abstract

The article presents concept of an intelligent system for spelling correction. The system should correct spelling errors in various types of documents in Polish language. Its processing should depend on the type of document and predicted types of errors. Unlike existing systems that work in Polish language it should analyse context of the erroneous word and after generating the list of suggested corrections it should eliminate those which are not suitable in the context in the aspects of Polish syntax and semantics. As the Silesian University of Technology (namely – Institute of Computer Science) has developed numerous tools for language analysis, some of them can be used, as a part of the system while functionality of some others must be extended.

Adres

Piotr R. Kasprzyk: Politechnika Śląska, Instytut Matematyki, ul. Kaszubska 23,
44–100 Gliwice, Polska, peter@polsl.gliwice.pl .