Akademia Górniczo-Hutnicza
Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej

Autoreferat Rozprawy doktorskiej

Semi-automatyczna kontekstowa analiza i korekta tekstów z wykorzystaniem specjalistycznych grafów lingwistycznych

mgr inż. Marcin A. Gadamer

Promotor:

dr hab. Adrian Horzyk, prof. AGH

Kraków 2019

W dzisiejszym świecie można łatwo zauważyć bardzo szybki postęp technologiczny, jak również związaną z nim ewolucję w wielu dziedzinach. Takiemu przekształceniu uległ m. in. sposób komunikacji międzyludzkiej oraz związana z nim przemiana języka. Dzisiaj tradycyjne środki komunikacji odeszły na dalszy plan, oddając miejsce nowoczesnej technologii.

Obecnie bardzo ważnym elementem jest informacja oraz sposób jej przetwarzania. Dzięki informacji możliwy jest ciągły rozwój ludzkości. Niewątpliwie największym źródłem informacji jest Internet. W przeciągu kilku ostatnich lat powstała ogromna liczba nowych stron internetowych, forów, blogów, na których ludzie mogą wymieniać się informacjami. Do tego rozwoju przyczyniło się także stworzenie, na początku XXI wieku, wielojęzycznego projektu internetowej encyklopedii działającej w oparciu o zasadę otwartej treści – Wikipedii – gdzie każdy człowiek może zamieszczać informacje na dowolny temat. Wyzwaniem, z jakim należy się zmierzyć nie jest już dostęp do informacji, ale sposób jej powiązania i przetwarzania. Ważną kwestią jest jakość treści oraz możliwość powiązania tej samej informacji występującej na kilku lub nawet kilkudziesięciu różnych stronach internetowych. Nie ma jednego sposobu, aby dane te ze sobą połączyć i skojarzyć. Zazwyczaj do tego celu wykorzystywane jest pasywne indeksowanie oraz ręczna praca człowieka. Osoba, która przegląda kolejne strony internetowe, może z łatwością połączyć podobne dane z wielu stron, dzięki swojej inteligencji i wiedzy. Wiele podobnych informacji w Internecie jest wyszukiwanych na zasadzie konkurencyjnej treści i podobieństwa, ale rzadko kiedy uzupełniają się one wzajemnie.

W dzisiejszych czasach powstaje bardzo duża liczba elektronicznych dokumentów, w których często występują błędy językowe. Istnieje więc potrzeba opracowania bardziej inteligentnej i kontekstowej korekty dla tekstu, który został wprowadzony z różnego rodzaju błędami. Na pojawianie się tych błędów ma wpływ wiele różnorodnych czynników. Najczęściej błędy powstają z winy użytkownika, a czasami generuje je sam program komputerowy. Do najważniejszych błędów użytkownika można zaliczyć m. in.:

- brak znajomości zasad konstrukcji poprawnych zdań,
- brak pełnego przekazu treści (skróty myślowe, zdrobnienia),
- brak staranności przy wpisywaniu tekstu,
- pośpiech użytkownika podczas wprowadzania tekstu.

Do błędów, braków i niedoskonałości aplikacji komputerowych zalicza się natomiast :

- brak specjalistycznych algorytmów do obsługi języka polskiego,
- brak badania kontekstu wprowadzanego tekstu (kontekst wyrazów i zdań),
- brak badania wprowadzenia przypadkowo poprawnych/błędnych wyrazów,
- automatyczna korekta, która nie zawsze działa poprawnie w różnych procesorach tekstów.

Tematyka sposobu korekty tekstu jest szeroko znana, lecz nadal brakuje metod potrafiących dokonać automatycznej jego poprawy, biorąc pod uwagę semantykę. Obecnie stosowane są do tego celu różnorakie słowniki, wyznaczane są dla słów odległości edycyjne oraz różne algorytmy, których skuteczność jest jednak nadal bardzo ograniczona. Modele językowe oparte na n-gramach słów (także wzbogacone o tagi POS – ang. *Part Of Speech tagging*) są dobrze znane od ponad 20 lat i używane w celu rozpoznawania mowy. Pośród metod służących budowaniu modeli do analizy tekstu można wymienić modele wykorzystujące gramatyki formalne i transakcyjne, systemy regułowe oraz systemy oparte na analizie korpusów tekstu. Niemniej jednak, w znacznej części tych modeli brakuje analizy kontekstu wprowadzanych słów. Może się zdarzyć, że wprowadzone słowa są poprawne gramatycznie, lecz w zadanym kontekście okazują się nieprawidłowe.

Celem pracy było uzyskanie wiedzy na temat możliwości przeprowadzenia efektywnej, semi-automatycznej, kontekstowej korekty różnych tekstów. Cel ten osiągnięto poprzez opracowanie innowacyjnego mechanizmu służącego do asocjacyjnego gromadzenia, kompresowania, a następnie przetwarzania tekstu (zdań) oraz zbudowanie algorytmów do korekty teksów wprowadzonych z różnego rodzaju błędami.

Badania miały na celu wykorzystać dostępne z różnych źródeł teksty do opracowania, a następnie zaimplementowania algorytmów, które będą w stanie znaleźć, przetworzyć i powiązać relacje pomiędzy kolejnymi słowami w taki sposób, aby „zrozumieć” czytane zdanie w danym kontekście słownym. Następnie wykonane algorytmy wraz z zaproponowaną strukturą, służącą zapisowi dla tak „zrozumianego” kontekstowo tekstu, miały służyć automatycznej korekcie wprowadzanego tekstu.

Algorytm kontekstowej korekty tekstu wraz z zaproponowanym modelem zapisu zdań stały się warstwą logiki dla aplikacji internetowej, w której użytkownik może wprowadzać tekst z błędami, a program wykorzystując opracowane algorytmy, koryguje go. Poprawa ta odbywa się w sposób automatyczny lub semi-automatyczny. Taki sposób korekty tekstu oznacza, że ostateczną decyzję co do wyboru najkorzystniejszego rozwiązania, spośród zaproponowanych przez algorytm podejmuje użytkownik. Wynika to z faktu, że algorytm korygujący tekst nie zna intencji piszącego i jedynie autor jest w stanie zdecydować, która z zaproponowanych korekt koresponduje z jego zamiarami. W pełni automatyczna korekta tekstu nie zawsze jest więc możliwa.

W rozprawie sformułowano następującą tezę:

Możliwe jest zbudowanie specjalistycznych grafów lingwistycznych na podstawie korpusów tekstów oraz algorytmów ich efektywnej analizy, pozwalających na przeprowadzenie poprawnej semi-automatycznej kontekstowej korekty różnych tekstów opartej na wiedzy zebranej w tych grafach, wykorzystując w trakcie ich tworzenia liczbę wystąpień poszczególnych słów w kontekście korygowanego tekstu.

W rozprawie zaprezentowano następujące zagadnienia:

- Dokładnie opisano badania własne oparte na wykorzystaniu grafu jako struktury danych, która dzięki swoim właściwościom, może służyć do zapisu zgromadzonych tekstów. Wprowadzone zostało pojęcie Grafu Przyzwyczajeń Lingwistycznych, jako modelu zapisu kontekstu słownego.
- Szczegółowo przedstawiony został, opracowany przez autora, innowacyjny sposób przechowywania słów. Scharakteryzowano wykorzystane rodzaje połączeń, które między tymi słowami występują tak, aby możliwe było odtworzenie kontekstu słownego wprowadzonego zdania.
- Przedstawiono i szczegółowo opisano proces korekty tekstu z wykorzystaniem opracowanych na bazie Grafu Przyzwyczajeń Lingwistycznych semi-automatycznych metod analizy i korekty tekstu.
- Następnie przeprowadzono porównanie otrzymanych wyników z innymi aplikacjami, które służą korekcie tekstu. Przedstawiono przykłady zarówno dla języka polskiego, jak i

angielskiego, wykazując, iż zastosowana struktura grafowa wraz z wykonanymi metodami może służyć do korekty tekstu w różnych językach.

Do przeprowadzenia badań wykorzystano teksty pochodzące z różnych źródeł, na ich podstawie uzyskując konteksty wypowiedzi na różne tematy. Podczas analizy dostępu do korpusów słownych skorzystano z m. in. serwisu Projektu Gutenberg (https://www.gutenberg.org/). Oferuje on ponad 59 tysięcy darmowych e-booków. Teksty pochodzące z takiego źródła posiadają kilka zalet – są one powszechnie znane, są dobrze napisane (nie powinny w nich wystąpić błędy), co jest niezmierne ważne podczas „uczenia" grafu poprzez zasilanie go nowymi tekstami, dodatkowo istnieje zazwyczaj wiele tekstów jednego autora. Dzięki tak dużej liczbie książek możliwe było uzyskanie bardzo bogatego i różnorodnego kontekstu wypowiedzi.

W kolejnym kroku przeanalizowano, a następnie przetworzono pozyskane teksty i zapisano relacje pomiędzy kolejnymi słowami w taki sposób, aby „zrozumieć" czytane zdanie w danym kontekście słownym. Dodatkowo opracowano budowę specjalistycznych grafów lingwistycznych, które zostały nazwane Grafami Przyzwyczajeń Lingwistycznych (grafami LHG). W celu zapisu tekstów do grafu, opracowano metody, które umożliwiły automatyczne pozyskiwanie tekstów z plików różnego formatu. Podstawową zaletą skorzystania z tych specjalistycznych grafów jest posiadanie informacji o kontekście wypowiedzi, poprzez zastosowanie specjalistycznych połączeń asocjacyjnych oraz o ilości wystąpień poszczególnych słów w danym kontekście.

Zastosowanie Grafu Przyzwyczajeń Lingwistycznych umożliwia badanie fleksyjno-częstotliwościowego kontekstu słownego. Jak zauważono wykorzystanie teorii grafów dynamicznie wzrasta w dziedzinach takich, jak chemia, lingwistyka, geografia, czy architektura. Ponadto grafy są potężnym narzędziem, jeśli chodzi o możliwość wizualizacji danych w nich zebranych.

Jedną z kluczowych inspiracji podczas tworzenia modelu języka opartego na strukturze grafowej był model ludzkiego mózgu, który dzięki swojej konstrukcji jest niezwykle wydajny. Podobnie jak dla ludzkiego mózgu proces uczenia, tak dodawanie nowych kontekstów do grafu jest operacją bardzo czasochłonną. Podczas czytania tekstów, dla każdego zdania tworzona jest najpierw tablica ze słowami oraz ich cechami. Każdy element tablicy jest dodawany do grafu. Słowo traktowane jest jako nowy wierzchołek (neuron słowny), a znaki interpunkcyjne stanowią o jego właściwościach. Dla każdego wierzchołka tworzona jest krawędź (asocjacja sekwencyjna) do następnego wierzchołka (neuronu słownego). Jeśli dla jakiegokolwiek wierzchołka będącego w

grafie występuje niejednoznaczność (z wierzchołka wychodzą lub do wierzchołka wchodzą dwie lub więcej krawędzi danego typu asocjacji), to tworzona jest krawędź asocjacji kontekstowej wyższego rzędu. Podczas analizowania kolejnych zdań zdarza się, że dodane krawędzie powodują powstanie niejednoznaczności dla poprzednio dodanych zdań. W związku z tym algorytm po analizie ostatnio dodanego zdania może kilka razy przeanalizować raz jeszcze niektóre z poprzednich zdań w poszukiwaniu kolejnych niejednoznaczności. Tak więc, gdy w grafie zapisywanych jest coraz więcej zdań, coraz częściej zdarza się wystąpienie niejednoznaczności, która w późniejszym etapie powinna zostać poprawiona. Gdy graf zostanie zasilony odpowiednią ilością tekstów, można w bardzo szybki sposób z niego skorzystać. Wystarczy wyszukać odpowiednie słowo (które jest dostępne w stałym czasie), a następnie poruszać się po poszczególnych połączeniach w grafie, których liczba (wchodząca lub wychodząca z określonego neuronu) jest skończona. Podobną właściwość można zauważyć w ludzkim mózgu, którego połączenia pomiędzy neuronami są wielokrotnie aktywowane w ciągu sekundy. Kolejnym podobieństwem do mózgu jest fakt, że Graf Przyzwyczajeń Lingwistycznych nie jest strukturą zamkniętą, która raz utworzona nie może być rozwijana w przyszłości. W każdym momencie graf LHG można rozszerzyć o kolejne zdania, a jego rozbudowa automatycznie poprawia wyniki korekty tekstów. W tym celu wystarczy uruchomić procedurę odpowiedzialną za przetworzenie kolejnych korpusów tekstu. Z metody tej skorzystano podczas korekty tekstu. Gdy użytkownik wprowadzi zdanie, które zostanie wstępnie poprawione poprzez zasugerowanie kilku opcji do wyboru, a użytkownik wybierze jedną z propozycji (tym samym uzna, że zaproponowana korekta jest prawidłowa), zdanie takie zostaje dodane do grafu jako kolejny prawidłowy kontekst. Daje to niesamowitą zdolność dodawania nowych zdań, które nie zostały do tej pory przeanalizowane, a są poprawne i mogą zasilić opracowany model.

Dzięki zastosowaniu wyspecjalizowanej struktury do przechowywania zdań, możliwe było przeprowadzenie w dalszym etapie semi-automatycznej kontekstowej korekty różnych tekstów z zastosowaniem wiedzy zebranej i powiązanej wcześniej. Jeden z algorytmów wykorzystuje równoczesne „pobudzenia” wielu wierzchołków poprzez wspomniane połączenia asocjacyjne. Każde z połączeń pobudza połączony z nim wierzchołek na taki czas, jakiego rzędu ono jest, z siłą odwrotnie proporcjonalną do jego rzędu. Na bazie cech słowa zapisanego w grafie został opracowany kolejny algorytm służący poprawie interpunkcji. Głównym celem tego algorytmu jest sprawdzanie, czy dane słowo może zaczynać się od małej bądź dużej litery, czy po danym wyrazie powinien wystąpić przecinek lub inny znak interpunkcyjny. Dodatkowo algorytm ten sprawdza, czy pierwsze słowo z zdaniu zaczyna się od dużej litery, a kończy się kropką, wykrzyknikiem lub

pytajnikiem. Następnym algorytmem, jaki został zaprojektowany i zaimplementowany jest algorytm służący do weryfikacji poprawności i kolejności występowania słów. W celu oceniania poprawności słów posłużono się badaniem częstości występowania słowa w zaproponowanym grafie, jak również informacjami pochodzącymi ze słownika danego języka. Dzięki temu, podczas wprowadzania tekstu możliwe jest oznaczanie słów, które występują w grafie oraz zostały odnalezione w słowniku jako poprawne, a słowa które nie zostały znalezione oznaczane są jako potencjalnie błędne.

W przeprowadzonych porównaniach z innymi aplikacjami, służącymi korekcie tekstu, wykazano, że metody oparte na Grafie Przyzwyczajeń Lingwistycznych dają bardzo dobre rezultaty dla różnych typów błędów. Zauważono również, że korekta tekstu jest zadaniem bardzo trudnym, z uwagi na stały rozwój języka oraz mnogość form, w jakich mogą wystąpić poprawne zdania, dlatego też Graf Przyzwyczajeń Lingwistycznych bazuje na różnych metodach, których połączenie daje lepsze wyniki niż zastosowanie poszczególnych algorytmów osobno.

W pracy wykazano, że możliwym jest zbudowanie efektywnej, innowacyjnej struktury reprezentującej model języka. Taką strukturę osiągnięto dzięki zastosowaniu modelu asocjacyjnego. Dodatkowo na podstawie korpusów tekstów oraz algorytmów ich efektywnej analizy, jak również z zastosowaniem wiedzy zebranej w Grafie Przyzwyczajeń Lingwistycznych, zostało opracowanych kilka metod, służących przeprowadzeniu semi-automatycznej kontekstowej korekty różnych tekstów. Porównanie ich efektywności działania z komercyjnie dostępnymi aplikacjami przyniosło zadowalające efekty.

Przeprowadzone badania i eksperymenty potwierdziły tezę, iż możliwe jest takie zbudowanie specjalistycznych grafów lingwistycznych na podstawie korpusów tekstów oraz algorytmów ich efektywnej analizy, żeby było możliwe przeprowadzenie semi-automatycznej kontekstowej korekty różnych tekstów z zastosowaniem wiedzy zebranej w tych grafach wykorzystując w trakcie ich tworzenia liczbę wystąpień poszczególnych słów w kontekście korygowanego tekstu.

Algorytmy wraz z zaproponowaną strukturą służącą zapisowi korpusów tekstu posłużyły do automatycznej korekty wprowadzanego tekstu, co obrano za cel podejmowanych w pracy badań.

Oryginalnymi osiągnięciami przeprowadzonych badań są:

- efektywny model grafowy służący do zapisu wyrazów oraz znaków interpunkcyjnych

występujących w badanym tekście;
- opracowane metody umożliwiające pozyskanie tekstów z kilku źródeł w celu zbudowania podanego grafu;
- opracowane i wdrożone niezależne metody służące do analizy i kontekstowej korekty tekstu;
- wykonana aplikacja webowa (serwis internetowy), która umożliwia skorzystanie z zaimplementowanych algorytmów.

Problem korekty tekstu jest obecnie szeroko znanym i ważnym problemem w świecie. Opracowane rozwiązanie pozwala wykorzystać go w wielu zagadnieniach np. podczas tworzenia tekstów, wspomagania automatycznego tłumaczenia języka, wspomaganie pracy redaktorów i korektorów, komunikacji z drugim człowiekiem, jak również komunikacji z robotami. Zaproponowana aplikacja stara się przeprowadzić korektę tekstu w sposób automatyczny lub semi-automatyczny. Wszystko po to, aby móc wyłączyć człowieka z manualnego sprawdzania dużej ilości tekstu, oferując wstępnie sprawdzony tekst i zaznaczając jedynie miejsca, w których należy wprowadzić pewne poprawki. Niemniej jednak, zagadnienie badania języka, próba jego zrozumienia i zamiana wiedzy uzyskanej w procesie uczenia na jednoznaczne reguły i procedury, które będą wykorzystywane przez algorytmy, są bardzo trudne, a czasami wręcz niemożliwe. Dlatego też choć prace badawcze nad poprawą korekty tekstu nadal trwają, to w chwili obecnej jedynie człowiek jest w stanie poprawić wszystkie błędy językowe.

Podsumowując, podjęta w pracy teza została potwierdzona: na podstawie korpusów tekstów oraz algorytmów ich analizy, możliwe jest zbudowanie specjalistycznych grafów lingwistycznych pozwalających na przeprowadzenie poprawnej semi-automatycznej kontekstowej korekty zdań, opartej na statystyce wystąpień tekstu w zadanym kontekście.

Najważniejsze osiągnięcia przedstawione w pracy:

- W wyniku przeprowadzonych prac uzyskano efektywny model grafowy służący do zapisu wyrazów oraz znaków interpunkcyjnych występujących w badanym tekście.
- Opracowano metody umożliwiające pozyskanie tekstów z kilku źródeł w celu zbudowania podanego grafu.
- Zaprojektowano i wdrożono niezależne metody służące do analizy i kontekstowej korekty

tekstu.

- Na podstawie powyższych osiągnięć wykonano aplikację webową (serwis internetowy), która umożliwia skorzystanie z zaimplementowanych algorytmów.
- Badania wykazały, że opracowany graf wraz z zaproponowanymi mechanizmami korekty tekstu jest dokładniejszy w porównaniu z istniejącymi rozwiązaniami. Graf Przyzwyczajeń Lingwistycznych potrafi w satysfakcjonującym stopniu odtworzyć kontekst zdań, które zostały wcześniej w nim zapisane.
- Osiągnięte rezultaty pozwalają na uznanie Grafu Przyzwyczajeń Lingwistycznych jako innowacyjnego rozwiązania, na bazie którego mogą zostać opracowane kolejne niezależne metody korekty tekstu w przyszłości.