Częstochowa, dn. 20 listopada 2019 r.

dr hab. inż. Rafał Scherer, prof. uczelni
Katedra Inteligentnych Systemów Informatycznych
Wydział Inżynierii Mechanicznej i Informatyki
Politechnika Częstochowska
al. Armii Krajowej 36
42-200 Częstochowa

Recenzja

rozprawy doktorskiej mgra inż. Marcina Gadamera, pt.: Semi-automatyczna kontekstowa analiza i korekta tekstów z wykorzystaniem specjalistycznych grafów lingwistycznych.

Promotor: dr hab. Adrian Horzyk, prof. AGH

Niniejszą recenzję opracowano na zlecenie Dziekana Wydziału Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej AGH, dr hab. inż. Ryszarda Sroki, prof. n., z dnia 04.10.2019 r.

1. Charakterystyka tematu, celu i tezy badawczej rozprawy

Analiza, reprezentacja oraz autokorekta i autouzupełnianie tekstu jest jedną z najważniejszych dziedzin informatyki. Nie jest prosto stworzyć metody dla wielu grup języków, dla różnych kontekstów czy różnych styli pisania. Obecnie zaczynają dominować rozwiązania oparte o głębokie uczenie maszynowe. Autor oparł się temu trendowi i postawił sobie za cel stworzenie modeli i narzędzi opartych o grafy. Rozwiązanie takie będzie miało na pewno zaletę w postaci lepszej interpretowalności, niż metody oparte o sieci neuronowe, które są zazwyczaj tzw. czarną skrzynką. Doktorant stawia 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.*

2. Zawartość rozprawy

Recenzowana praca mgra inż. Marcina Gadamera składa się z sześciu rozdziałów oraz bibliografii. Dokument liczy 145 stron.

Rozdział pierwszy zawiera wprowadzenie do problematyki i jej genezy. Wymienia typy błędów językowych z jakimi możemy się zetknąć w danych tekstowych. Doktorant identyfikuje problemy występujące przy automatycznej korekcie tekstów, szczególnie pisanych w języku polskim. Przedstawia cel pracy polegający na stworzeniu automatycznej korekty tekstu za pomocą grafów lingwistycznych oraz tezę pracy. Zawiera również omówienie poszczególnych rozdziałów pracy.

Rozdział 2 omawia problematykę przechowywania danych. Rozpoczyna od przedstawienia przyrostu ilości informacji z jaką mamy do czynienia we współczesnym świecie, co ma bardzo duży związek z rozwojem Internetu i urządzeń cyfrowych. Dalej Doktorant omawia sposoby przechowywania informacji – pliki płaskie, relacyjne bazy danych, różne typy nierelacyjnych baz danych, oraz inne nowatorskie sposoby organizacji danych w systemach komputerowych i ich połączenia.

Rozdział 3 dotyczy przetwarzania języka naturalnego. Doktorant wprowadza czytelnika do możliwych zastosowań takiej analizy oraz problemów napotykanych w takich badaniach. Między innymi, są to: niejednoznaczność i wieloznaczność wyrazów, używanie specyficznych słów, slangu, a nawet piktogramów, występowanie idiomów, neologizmów, archaizmów oraz skomplikowanych nazw. Dalej Autor omawia metody przetwarzania języka naturalnego: oparte na regułach, prawdopodobieństwie, uczeniu maszynowym i głębokim uczeniu. Omawia również i wyjaśnia prawo Zipfa. Następnie przedstawia etapy analizy tekstu: podział na zdania, podział na tokeny, normalizacja tekstu, rozpoznanie bytów nazwanych, ujednolicenie form wyrazów, analiza semantyczna. Dalej następuje zestawienie narzędzi służących do analizy tekstu, czyli wyrażeń regularnych, segmentacji zdań, drzew decyzyjnych, minimalna odległość edycyjna. Następnie, Autor omawia metody modelowania tekstu: probabilistyczne, wektorowe, łańcuchy Markowa, klasyczne, splotowe i rekurencyjne sieci neuronowe, osadzanie słów, word2vec, modele generatywne GAN.

Rozdział 4 omawia grafy. Autor wprowadza czytelnika do tematyki, a następnie opisuje początkowe badania oparte na przechowywaniu trójek wyrazów w relacyjnej bazie danych. Metoda ta była sprawdzeniem założeń i na jej bazie powstała autorska metoda nazwana Grafem przyzwyczajeń lingwistycznych, który jest inspirowany sposobem pobudzania neuronów w ludzkim mózgu. Jest zbudowany z kilku rodzajów elementów, np. wierzchołki odpowiadające literom, wierzchołki dla słów oraz dla początki i końce zdania. Ponadto, wierzchołki mogą posiadać parametry oznaczające występowanie znaków diakrytycznych oraz dużych liter. Jest to struktura używana w dalszej części pracy do różnych metod autokorekty. Autor wprowadził cztery sposoby połączeń pomiędzy wyrazami wraz z wagami odpowiadającymi za siłę połączeń. Graf będzie służył do korekty różnego rodzaju błędów, takich jak błędy fleksyjne, składniowe, słownikowe, frazeologiczne, słowotwórcze, gramatyczne czy ortograficzne. W dalszej części rozdziału Autor omawia stworzone metody korekty tekstu na bazie struktury grafu: semi-automatyczne metody analizy i korekty tekstów, metody statyczne, zmodyfikowaną metodę odległości edycyjnej, zmodyfikowaną metodę n-gramów oraz metodę wykorzystująca pobudzenia asocjacyjne.

Rozdział 5 omawia aplikację Grafu Przyzwyczajeń Lingwistycznych w metodach autokorekty tekstu. Pierwszą metoda jest sprawdzanie poprawności wyrazów w słownikach. Dla celów

metod statycznych skorzystano ze słowników języka polskiego i angielskiego oraz morfologicznego słownika języka polskiego dla stworzenia grafu. Następnie omówiono metody bazujące na asocjacjach sekwencyjnych oraz kontekstowych dla wprowadzonego tekstu. Omówiono na przykładach metodę korekty tekstu, oraz sprawdzono działanie metody przez porównanie z aplikacjami dla języka polskiego i angielskiego oraz dla dzieł literackich. Autor stworzył aplikację internetową implementującą wyżej wymienione metody i dokonał porównania z wieloma istniejącymi rozwiązaniami.

Ostatnim rozdziałem jest Podsumowanie. Doktorant nie tylko podsumowuje w nim wykonane badanie, ale wyznacza wiele kierunków badawczych, takich jak stworzenie modelu osoby piszącej teksty, grupowanie słów podobnych, zastosowanie bibliotek do oznaczania części mowy, optymalizacja liczby połączeń grafu, dynamiczne zwiększanie systemu grafów czy wprowadzenie grafów dla częstych słów.

Pracę kończy rozbudowana bibliografia składająca się ze stu dwóch aktualnych pozycji.

Ogólnie zasadnicze i oryginalne rezultaty pracy można podsumować następująco:

- Opracowanie wprowadzenia do tematyki analizy tekstu i języka naturalnego.
- Stworzenie nowego modelu reprezentującego tekst jako grafy zawierające wyrazy, znaki interpunkcyjne i zdania.
- Stworzenie algorytmu uczenia grafów z wielu źródeł.
- Zastosowanie stworzonych algorytmów do korekty tekstu wraz z implementacją jako system internetowy.

Wymienione oryginalne metody przedstawione w pracy zostały opublikowane w kilku artykułach naukowych, dwóch w czasopismach i trzech w materiałach konferencyjnych. Zaprezentowany materiał pokazuje, że Doktorant zrealizował cel pracy.

3. Uwagi krytyczne i wskazówki dotyczące rozprawy

Praca napisana jest schludnie i przejrzyście. Praca obfituje w czytelne rysunki oraz schematy. Poniżej zamieszczam kilka uwag i pytań. Uwagi te nie umniejszają wartości naukowej rozważanej rozprawy doktorskiej.

Rozdziały nie mają tytułów, a podrozdziały nie są numerowane. Rozdziały nie mają również wprowadzenia, dlatego, na przykład, gdy rozdział 4 zaczyna się podrozdziałem „Graf jako struktura", początkowo nie wiadomo czego będzie dotyczył cały rozdział. Rozdział 4 zawiera podrozdział „Semi-automatyczne metody analizy i korekty tekstów". Rozdział 5 ma również podrozdziały dotyczące metod korekty tekstu i dopiero po dalszej jego lekturze można stwierdzić, że jest to już opis eksperymentów.

Autor używa stwierdzenia „analiza sentymentu", które jest tłumaczeniem często używanego określenia *sentiment analysis*. Według mnie, należałoby tłumaczyć je na analiza nastroju lub nastawienia.

Brakuje w pracy bezpośredniego odniesienia lub porównania do masowo stosowanych obecnie metod opartych na sztucznych sieciach neuronowych, a szczególnie na sieciach splotowych,

rekurencyjnych (np. LSTM) czy splotowych rekurencyjnych. Dlaczego istnieje obecnie przesunięcie uwagi środowiska na takie metody? Czy działają lepiej?

4. Wnioski końcowe recenzji

Podsumowując recenzję stwierdzam, że Pan mgr inż. Marcin A. Gadamer w rozprawie doktorskiej „Semi-automatyczna kontekstowa analiza i korekta tekstów z wykorzystaniem specjalistycznych grafów lingwistycznych":

- Zrealizował cel rozprawy,
- Uzyskał oryginalne rezultaty naukowe dotyczące reprezentacji dokumentów tekstowych za pomocą nowatorskich struktur grafowych,
- Dokonał ciekawego wprowadzenia do tematyki,
- Stworzył przegląd wybranych bieżących pozycji literaturowych dotyczących tematu,
- Stworzył nowe modele reprezentujące tekst na poziomie liter, słów oraz zdań na bazie grafów,
- Zastosował stworzone metody w narzędziu (aplikacji internetowej) do korekty tekstu, oraz porównał z wieloma istniejącymi rozwiązaniami,
- Wykazał się umiejętnością samodzielnej pracy badawczej, znajomością literatury światowej i wiedzą w dziedzinie uczenia maszynowego.

Należy również podkreślić działalność zawodową Doktoranta, dającą mu ogromne doświadczenie praktyczne, oraz organizację dużej konferencji GeeCON poświęconej językowi Java.

Recenzowana praca spełnia wymagania ustawy o tytule i stopniach naukowych w dyscyplinie naukowej automatyka i robotyka. Wnoszę o jej przyjęcie i dopuszczenie do publicznej obrony. Jednocześnie ze względu na wysoki poziom naukowy rozprawy, wnioskuję o wyróżnienie rozprawy.