*Prof. dr hab. Stanisław Matwin*

*Instytut Podstaw Informatyki PAN*
*oraz*
*Faculty of Computer Science, Dalhousie University, Kanada*

Halifax, 8 wrzesnia 2018

# Recenzja

Przedmiotem recenzji jest rozprawa doktorska Pana **mgr inż. Karola Grzegorczyka pt. „ Vector representations of text data in deep learning”.** Promotorem rozprawy jest Pan prof. zw. Witold Dzwinel. Recenzję opracowałem na wniosek Rady Wydziału Informatyki, Elektroniki i Telekomunikacji Akademii Górniczo-Hutniczej.

[Uwaga: terminologia polska dotycząca sieci neuronowych, użyta w niniejszej recenzji, pochodzi częściowo z tłumaczenia monografii Y. Bengio , A. Courville , I. Goodfellow, „Deep Learning - Systemy uczące się”, Wydawnictwo Naukowe PWN, Wwa 2018. W kilku miejscach, tam gdzie termin polski nie wydaje się być powszechnie używany, podano też termin angielski]

## 1. Dziedziny nauki związane z tematyką rozprawy

Tematyka rozprawy doktorskiej, która jest przedmiotem tej recenzji, dotyczy dyscypliny Informatyka. Zgodnie z taksonomią dyscyplin naukowych wykorzystywaną przez Narodowe Centrum Nauki (NCN) – odpowiada ona panelowi dziedzinowemu ST6 (Informatyka i Technologie Informacyjne), w szczególności pod-tematom ST6_7 (Sztuczna inteligencja, systemy inteligentne i wieloagentowe) oraz ST6_11 (Uczenie maszynowe, statystyczne przetwarzanie danych i zastosowanie w przetwarzaniu sygnałów) i ST6-9 („Interakcja człowiek – komputer, rozpoznawanie i synteza mowy, przetwarzanie języka naturalnego”).

## 2. Istotność podejmowanej tematyki, jakość postawionych tez, nowatorstwo i inne walory pracy

### 2.1. Istotność

Praca poświęcona jest budowie reprezentacji danych tekstowych w oparciu o techniki tzw. „głębokich sieci neuronowych”, którym słusznie przypisuje się w ostatnich latach zasługi w postępach Sztucznej Inteligencji i Maszynowego Uczenia Się. Szczególna zaleta tych metod polega na ich zdolności uczenia się z „surowych” danych reprezentacji właściwej dla danego

zadania. Ten właśnie aspekt metod uczenia się reprezentacji, szczególnie w odniesieniu do danych tekstowych, stanowi tematykę pracy. Wkład pracy w dziedzinie Maszynowego Uczenia Się polega na

1) autorskiej metodzie pozwalającej na połączenie metody Semantycznego Indeksowania Mieszanego („semantic hashing") z Binarną Wektoryzacją Paragrafów (Binary Paragraph Vectorization).

   oraz

**2)** opracowaniu Ujednoznaczniającej Metody Zagnieżdżania Słów (Disambiguating Skip-gram Embedding), potrzebnej we wszechobecnym w analizie tekstów i lingwistyce komputerowej zadaniu WSD.

W mojej ocenie oba osiągnięcia pracy stanowią istotny wkład do dziedziny badań, w której umiejscowiona jest praca.

### 2.2. Teza Pracy

Główną tezę pracy możnaby przedstawić jako stwierdzenie, że

> *możliwa jest o wiele bardziej efektywna realizacja semantycznego ineksowania mieszającego, używająca współczesnych metod reprezentacji danych, niż ta zaproponowana w przełomowej pracy Hintona i Salakhutdinova z 2009 r.*

Dodatkowo, praca przedstawia tezę, że

> *można zbudować kontekstualizującą reprezentację słów w korpusie z wbudowanym ujednoznacznianiem słów.*

### 2.3. Nowatorstwo

Praca ma nowatorski charakter. Jej wkład, opisany powyżej, jest inowacyjny i oryginalny. Zostało to niejako potwierdzone akceptacją artykułu przedstawiającego wyniki zawarte w pracy przez wysoce selektywną konferencję Empirical Methods in Natural Language Processing 2018.

### 2.4. Inne walory pracy

Praca jest jasno i klarownie napisana. Lektura pracy przedstawia Autora jako wysokiej klasy eksperta w dziedzinie sieci neuronowych. Praca jest poprawna metodologicznie i ciekawa w warstwie przedstawiającej intuicje i opinie Autora. Jest to szczególnie istotne w pracy, której wyniki są głownie empiryczne, oparte na bardzo licznych i z natury rzeczy selektywnych doświadczeniach. Właściwe zaprojektowanie tych eksperymentów i ich logiczny ciąg są więc

ważnym kryterium oceny pracy. W pracy podjęto wiele decyzji dotyczących np. szczegółów ewaluacji proponowanej metody na dużych danych (RCV1-v2 oraz Wikipedia) – wszystkie one są szczegółowo opisane i sensownie uzasadnione. W mojej opinii ta warstwa poracy jest jej zdecydowanym atutem: eksperymenty i dane są przemyślane, wyniki są jasno i uczciwie przedstawione i przekonywująco przedyskutowane.

Należy podkreślić biegłość autora w posługiwaniu się złożonymi narzędziami z bogatego repertuaru współczesnych, otwartych systemów typu TensorFlow, pytorch itp. I tak np. dyskusja na str. 40 motywująca wybór TensorFlow i mądre użycie opcji tego systemu jest bardzo ciekawa.

Wreszcie angielszczyzna pracy jest biegła i wartka. W pracy można spotkać pewne literówki i niewielkie poprawki redakcyjne będą potrzebne w ostatecznej wersji – sugestie w tek kwestii mogę przekazać Autorowi osobiście.

## 3. Elementy pracy, które mogłyby zostać ulepszone

Jedyną częścią pracy, pozostawiającą pewne poczucie niedosytu u czytelnika, jest dyskusja kosztów ceny proponowanych metod. O ile jest oczywiste, że kodowanie binarne jest 32 (albo 64!) razy mniej wymagające jeśli chodzi o pamięć, to koszt obliczeniowy jest o wiele mniej jasny. Pewne podstawowe informacje są podany na str. 67 jeśli chodzi o czas uczenia się reprezentacji, nie jest jasne jak podany tam liczby porównują się z podobnymi danymi dla którejś z metod z literatury, z którymi wyniki jakościowe są i tak porównywane, np. w rozdz. 4.2. Uwaga ta stosuje się zarówno do kosztu wytrenowania modelu, jak też i do kosztu użycia („inference"), jak o tym mowa na str. 29 i 78.

Na str. 23, gdy wspomina się metodę SVM, Autor stwierdza że „One of the factors that enabled SVMs to flourish is their relatively low computational and memory cost". O ile zgadzam się z tą opinią w kwestii pamięci, nie zgadzam się że SVMs mają niski koszt obliczeniowy, ponieważ metoda wymaga rozwiązania zadania optymalizacji kwadratowej. Ogólnie uważa się więc, że metoda SVM, przynajmniej w swej wersji klasycznej, nie jest skalowalna do danych typu Big Data.

## 4. Dyskusja

Dla czytelnika jest jasne, że Autor musiał był przeprowadzić bardzo znaczącą liczbę eksperymentów aby dojść do przedstawionych wyników. Jest to szczególnie istotne przy użyciu metod uczenia się w głębokich sieciach neuronowych, charakteryzujących się wielką liczbą parametrów i hiperparametrów, niejednokrotnie współzależnych. Obietnica odejścia od „inżynierii atrybutów" jest spełniona, ale za cenę „architekturyzacji"( „architecture engineering") na wielką skalę. Dlatego ciekawe byłoby usłyszeć od Autora choćby zgrubną

ocenę wysiłku niezbędnego do otrzymania przedstawionych w pracy wyników, przy czym nawet nie jest jasne jak mierzyć ten wysiłek.

Bardzo ciekawa jest dyskusja transferu poznawania („transfer learning") w rozdz. 3.2.7. Wyjaśnienie wyników wydaje się przekonywujące, ale byłoby jeszcze bardziej przekonywujące gdyby czytelnik dowiedział się, ile słów RCV1-V2 nie znajduje się w Wikipedii.

Wreszcie na koniec naiwne być może pytanie: czy Binarna Wektoryzacja Paragrafów mogła być użyteczna w zastosowaniach tekstowych innych niż wyszukiwanie informacji? Np. w zadaniach klasyfikacji lub grupowania tekstów?

## 5. Dorobek naukowy kandydata

Kandydat posiada dorobek naukowy zgodny ze stadium kariery naukowej, lub nawet przewyższający to stadium. W repozytorium DBLP znajduje się 7 prac w których Kandydat jest współautorem, plus praca przyjęta ostatnio na prestiżową konferencję EMNLP. Obie te liczby potwierdzają, że dorobek Kandydata jest znaczący, jak na tak wczesne stadium kariery naukowej.

## 6. Podsumowanie

Podsumowując przedstawianą opinię, stwierdzam ostatecznie, że praca **mgr inż. Karola Grzegorczyka** spełnia wymagania przewidziane dla rozpraw doktorskich w aktualnie obowiązującej ustawie *(Ustawa o stopniach naukowych i o tytule naukowym oraz o stopniach i tytule w zakresie sztuki z dnia 14 marca 2003 roku, Dziennik Ustaw Nr 65, poz. 595)*. Dlatego stawiam wniosek o **przyjęcie tej pracy jako rozprawy doktorskiej i o dopuszczenie Kandydata do jej publicznej obrony.**

*Prof. dr hab. Stanisław Matwin, prof. zw. IPI PAN*