

Recenzja rozprawy doktorskiej

Identification of hidden semantic relations in texts using relations' patterns

przedstawionej przez Michała Gałuszę

na Wydziale Cybernetyki, WAT

Rozprawa doktorska została przedstawiona jako samodzielny utwór oparty na publikacji

Michał Gałusza, Andrzej Walczak: Graph-based document-level relationship extraction for risk analysis: A transitive and dialog coherence approach. Expert Syst. Appl. 257: 124990 (2024)

(część zawartości rozprawy została przeniesiona wprost z publikacji, do czego zresztą autor rozprawy miał prawo na gruncie przepisów prawa autorskiego). Deklaracja zamieszczona w wyżej wymienionej publikacji wskazuje na kluczową rolę doktoranta w jej powstanie określając rolę współautora jako „supervision”. Tak więc przedstawione wyniki są w przeważającej części rezultatem pracy doktoranta.

Dokonując oceny, recenzent ma za zadanie stwierdzić czy:

„1. Rozprawa doktorska prezentuje ogólną wiedzę teoretyczną kandydata w dyscyplinie albo dyscyplinach oraz umiejętność samodzielnego prowadzenia pracy naukowej lub artystycznej.

2. Przedmiotem rozprawy doktorskiej jest oryginalne rozwiązanie problemu naukowego, oryginalne rozwiązanie w zakresie zastosowania wyników własnych badań naukowych w sferze gospodarczej lub społecznej albo oryginalne dokonanie artystyczne.”

Ad 2) Niewątpliwie rozprawa podejmuje istotny z praktycznego punktu widzenia problem naukowy i wskazuje na pewną oryginalną technikę NLP dla jego rozwiązania.

Ad 1) Rozprawa wykazuje ogólną wiedzę kandydata, zarówno teoretyczną jak i praktyczną, w omawianej dziedzinie.

Tak więc ogólne warunki do zaakceptowania rozprawy doktorskiej wydają się spełnione. Niemniej jednak obecny sposób prezentacji i techniczny poziom edycyjny nie pozwalają na ocenę zawartości merytorycznej, poprawności koncepcji i wyników. Z tego względu wnioskowałbym o doprowadzenie rozprawy do postaci możliwej do weryfikacji. Zastrzegam, że niemożność dokonania przeze mnie wiążącej oceny wynika poniekąd z faktu mojego ograniczonych kompetencji z zakresu Natural Language Processing. Część rozprawy związana z zagadnieniami algorytmów na grafach jest dla mnie jasna, ale część ta stanowi jedynie rutynowy pod względem technik fragment rozprawy

Wnioskowałbym o rozważenie dokonania następujących zmian:

1. Rozprawa nagminnie używa niezdefiniowanych pojęć, lub pojęć zdefiniowanych w sposób nie umożliwiający określenia ich jednoznacznego znaczenia. Każde pojęcie nie należące do zbioru pojęć podstawowych powinno być zdefiniowane w sposób umożliwiający rekonstrukcję na przykład przez programistę, pragnącego samodzielnie zrekonstruować procedury i algorytmy omawiane w rozprawie. Oczywiście, nie jest to zawsze możliwe w przypadku odwołań do technologii stosowanych w trybie „plug-and-play” ze względu na stopień skomplikowania materii. W tym przypadku sugerowałbym usunięcie namiastków definicji: w obecnej postaci nie wyjaśniają one nic osobie nie znającej pojęcia, zaś czytelnikowi zorientowanemu w materii utrudniają dotarcie do istotnych kwestii. W tym wypadku odwołanie do literatury i wskazanie wykorzystywanych własności powinno być wystarczające.
Sugerowałbym dosyć mechaniczne podejście polegające na inwentaryzacji pojęć i sprawdzanie czy w tekście poprzedzającym znajduje się odpowiednie jego wprowadzenie.
2. Na uwagę zasługują komentarze dotyczące złożoności obliczeniowej procedur. Nieco ostrożniej należy w tekście naukowym określać złożoność czasową. Na przykład konieczne jest uporządkowanie nieprawidłowo stosowanej notacji „O”.
Drobna uwaga algorytmiczna: w przypadku szukania ścieżek bardziej efektywne jest stosowanie algorytmów „all-pairs-shortest-path” niż rozwiązywanie zagadnienia oddzielnie dla każdej pary.
3. Warstwa językowa rozprawy nie może budzić zarzutów pod względem poprawności gramatycznej i czystości języka angielskiego. Niemniej jednak sugerowałbym modyfikacje polegające na uproszczeniu języka w kierunku jednoznaczności i jasności przekazu. W szczególności, wskazane byłoby używanie dokładnie jednego terminu na określone pojęcie/obiekt zamiast kilku synonimów. Tym samym, inaczej niż w języku naturalnym, rozprawa powinna przyjąć zasadę stosowaną w tekstach prawniczych, że jakakolwiek zmiana oznacza, że *token* (w sensie parsowania tekstu) posiada inne znaczenie.
Mam wrażenie, że tekst został poprawiony językowo przez osobę lub narzędzia nie rozumiejące znaczenia poprawianych zdań. Stąd praktyczny brak błędów językowych/edycyjnych, przy trudnym parsowaniu ich znaczenia. Sugerowane podejście, to stosowanie zasady Brzytwy Occama – wyjaśnienia kwestii technicznych jak najprostszym, choć może mało literackim językiem.

Koncepcja merytoryczna proponowana w rozprawie

Zasadniczym problemem stawianym w rozprawie jest pytanie w jaki sposób przetwarzać automatycznie opisy incydentów napisane w języku naturalnym tak aby w efektywny sposób identyfikować konsekwencje podatności.

Autor rozprawy argumentuje, że oparcie się na uniwersalnych mechanizmach stosowanych w NLP nie przynosi oczekiwanych rezultatów. Przychylam się do tego stanowiska, jednak rozprawa mogłaby mocniej wspierać proponowane podejście, gdyby bardziej szczegółowo dokumentowała dlaczego podejścia standardowe dają niezadawalające rezultaty.

Autor rozprawy opiera się na prostym podejściu polegającym na wyszukiwaniu w tekstach informacji o podatnościach i ich bezpośrednich konsekwencjach. Oczywiście, w języku naturalnym rzadko kiedy konsekwencje takie wymienione są bezpośrednio w tym samym zdaniu. Autor rozprawy zauważa jednak, że związki takie pojawiają się w stosunkowo niewielkiej „odległości”, tak aby zachować ciągłość narracji autora. Założenie to wydaje się prawdziwe bez względu na strukturę języka naturalnego i jego wewnętrzną logikę. Autor dokonuje pewnych niezbędnych uproszczeń – co umożliwia automatyzację procesu kosztem pewnej liczby fałszywych rezultatów a także pominąć. Niemniej jednak zadaniem nie jest dokładna analiza treści określonego dokumentu ale przetworzenie dużej liczby dokumentów i odsianie z zawikłanych i nieustandaryzowanych tekstów takich potencjalnych związków.

Podstawowym mechanizmem mającym pomóc odnajdywanie rozproszonej informacji jest mapowanie tekstu na pewien graf a następnie analiza ścieżek w grafie mogących stanowić o potencjalnym związku logicznym. Potrzebne własności grafowe mogą być wykryte przy pomocy rutynowych algorytmów grafowych, a ze względu na wielkość analizowanych raportów (i odpowiadających im grafów) realistycznych do wykonania.

Pomysł opisany w rozprawie może znaleźć zastosowanie zapewne w wielu innych sytuacjach niż analiza raportów o incydentach bezpieczeństwa. Najistotniejszym elementem jest wykrywanie „hidden patterns”, które umykają płytkiej syntaktycznej analizie, zaś mogą być ewidentne i stanowią oś logiczną dla człowieka czytającego tekst.

Tak jak w wielu innych pracach z zakresu sztucznej inteligencji, prezentowany pomysł jest stosunkowo elementarny. Kluczowe jest jednak to w jakim stopniu pomysł ten faktycznie działa. Wymaga to sprawdzenia nie na „normalnych tekstach”, ale na specyficznych tekstach stanowiących motywację rozprawy, pochodzących z różnych obszarów kultury językowej, organizacyjnej i administracyjnej.

Konkluzja

Ogólna idea rozprawy i wkład w rozwój dziedziny skłaniają do konkluzji, że przedstawiona rozprawa jest wystarczającym osiągnięciem badawczym dla nadania stopnia doktora. Dla przedstawienia Komisji

pozytywnej

rekomendacji do nadania stopnia doktora na podstawie przedstawionej rozprawy niezbędne byłoby jednak przesłedenie i sprawdzenie detali proponowanych technik i eksperymentów na danych testowych. W obecnym stanie rozprawa nie osiągnęła jednak minimalnego wymaganego poziomu pod tym względem. Upřednio procedura pozwalała w takim przypadku na skierowanie rozprawy do uzupełnienia/poprawy, z korzyścią dla procedury i doktoranta. **Bez takich uzupełnień nie jestem w stanie przedstawić ostatecznej i miarodajnej rekomendacji dla Komisji.**

Mirosław Kutytowski

Mirosław Kutytowski