POLITECHNIKA POZNAŃSKA

Łukasz Borchmann

Span Identification and Key Information Extraction Beyond Sequence Labeling Paradigm

Streszczenie rozprawy doktorskiej

Promotor:
prof. dr hab. Andrzej Marciniak

Poznań 2022

Identyfikacja fragmentów tekstu i ekstrakcja kluczowych informacji poza paradygmatem znakowania ciągu

(tytuł w języku polskim)

Metody oparte o znakowanie ciągu (ang. *sequence labeling*), w których sekwencji danych wejściowych $(y_1, \ldots, y_n)$ przypisuje się sekwencję etykiet $(x_1, \ldots, x_n)$, mają szerokie zastosowanie w dziedzinie przetwarzania języka naturalnego.

Chociaż co do zasady możliwe jest rozpatrywanie każdego elementu sekwencji niezależnie, problem ten, do czasu pojawienia się architektury BERT-a [1] i innych modeli powstałych na bazie Transformera [2], był na ogół adresowany z wykorzystaniem korelacji pomiędzy docelowymi etykietami [3].

Jak pokazano na przykładzie identyfikacji fragmentów tekstu realizujących techniki propagandowe, to porzucone kilka lat temu podejście wykorzystać można do poprawy skuteczności modelu (Rozdział 3). Zaproponowany model, dzięki połączeniu tej i innych tradycyjnych technik ze współczesną architekturą RoBERT-y [4] zdobył pierwszą i drugą lokatę w konkursie SemEval [5].

Najbardziej typowym problemem rozpozpatrywanym przez pryzmat przypisywania etykiet jest jednak rozpoznawanie bytów nazwanych (ang. *named entity recognition*), tj. lokalizowania podciągów tekstu, które reprezentują pewne wyróżnione, umowne obiekty, takie jak imiona, nazwiska, daty, nazwy organizacji czy nazwy geograficzne [6, 7].

Podejście do ich rozpoznawania bazujące na *sequence labelingu*, zakłada przypisywanie etykiety każdemu tokenowi (słowu graficznemu) w zdaniu, zakładając schemat, który umożliwia późniejszą identyfikację bytów wielowyrazowych*.

* Jego przykładem jest konwencja *BIO*, w której to etykietę `O` przypisuje się słowom nie wchodzącym w skład żadnego bytu, natomiast początek i kontynuacja interesującego fragmentu jest wyróżniona poprzez przypisanie etykiet `B` oraz `I`.

Ten, wydawałoby się, efektywny paradygmat traci jednak na adekwatności, kiedy dopuszczamy istnienie zagnieżdżonych bytów nazwanych — nazw takich jak *Rondo ONZ*, które posiadają inną nazwę jako swój podciąg (tu – w nazwie lokalizacji zawarta jest nazwa organizacji). Choć problem wciąż można zaadresować z sukcesem przy użyciu znakowania ciągu, o czym świadczy m.in. przedstawiony w rozdziale drugim model, to jednak często za cenę elegancji lub prostoty rozwiązania[†].

Ten stan odczytać można jednak jako zachęte do poszukiwania nowego paradygmatu — to jemu poświęcona jest zasadnicza część rozprawy, skupiona na propozycjach zastąpienia znakowania ciągu bardziej adekwatnymi metodami, w zastosowaniach gdzie byłoby ono użyte przez referencyjne rozwiązania.

Analizując przytoczone przykłady detekcji technik propagandowych i rozpoznawania bytów nazwanych, a także inne zastosowania znakowania ciągu w przetwarzaniu języka naturalnego, zauważyć można istnienie dwóch kategorii problemów, wyróżnionych ze względu na przeznaczenie ekstrahowanej informacji:

- kiedy wiedza o dokładnej lokalizacji w dokumencie jest konieczna ze względu na zastosowanie modelu (wyszukiwanie lub wspomaganie decyzji użytkownika, np. poprzez wskazanie gdzie występują potencjalnie „toksyczne" treści do wymoderowania lub w którym miejscu powinien zachować uwagę gdyż próbuje się wpłynąć na jego opinię);

- gdy wystarczające jest przypisanie metadanych do dokumentu, zaś dokładne wskazanie ich źródła w treści nie jest konieczne (przykładowo, ostatecznym celem ekstrakcji informacji z faktury może być odczytanie jaka kwota płatności została do uregulowania oraz na jakie konto dokonać należy opłaty).

Pierwsza kategoria, określana jako identyfikacja fragmentów tekstu (ang. *span identification*, por. [5]), rozpatrywana jest przede

[†] W tym miejscu przywołać można metody zakładające dynamiczne dodawanie warstw sieci lub przewidywanie hipergrafu etykiet, wykraczające swoją złożonością poza proste etykietowanie [8, 9]

wszystkim w związku z wyszukiwaniem zbliżonych semantycznie i funkcjonalnie klauzul tekstu prawnego w nieustrukturyzowanym tekście, na podstawie niewielkiej liczby przykładów. Ten problem, określany jako *contract discovery* wprowadzony został w Rozdziale 4.

Dla przytoczonego zastosowania zaproponowano zróżnicowane metody wykorzystujące neuronowe modele języka, m.in., wiążące ich reprezentacje z klasycznym algorytmem dyskretnej transformaty czasowej (ang. *dynamic time warping* [10]), który uogólniono do sytuacji z wieloma sekwencjami referencyjnymi (Rozdział 5). Nieadekwatność tradycyjnych modeli wykorzystywanych przy znakowania ciągu wynika tu przede wszystkim z niewielkiej liczby przykładów treningowych.

Rozważania dotyczące problemu identyfikacji fragmentow tekstu są kontynuowane w Rozdziale 6. Tamże zaprezentowano metodę detekcji takich fragmentów dokumentu, których obecność jest kluczowa ze względu na funkcję kosztu modelu realizującego zadanie streszczania tekstu. Tym samym, zaproponowano model ekstrakcyjno-abstraktywny, w którym komponent ekstrakcji nie wymaga dodatkowych danych treningowych, jakich wymagałoby klasyczne znakowanie ciągu.

W ramach drugiej kategorii problemów, określanych jako ekstrakcja kluczowych informacji (ang. *key information extraction*, por. [11, 12]), skupiono się na zróżnicowanych zadaniach, zorientowanych na otrzymanie par klucz-wartość na podstawie dokumentu (lub odpowiedzi na pytania zadawane w języku naturalnym). Zaproponowane modele oparte o architekturę enkoder-dekoder, opisane w Rozdziale 7 i Rozdziale 8, są jak dotąd najbardziej skutecznymi spośród opisanych w literaturze. Skupiają się one kolejno na problemie ekstrakcji z dokumentów o bogatej strukturze graficznej oraz problemach, gdzie należy dokonać ekstrakcji wielu par klucz-wartość z jednego tekstu.

Motywacja dla porzucenia paradygmatu znakowania ciągu w tym miejscu wynika z czynników takich jak niedostępność anotacji na poziomie tokenu (dysponujemy jedynie wartościami przypisanymi do całego dokumentu) czy nieobecność wartości w treści (np. w skutek błędu OCR, niepoprawnie rozpoznanej kolejności tokenów lub zakładanej normalizacji).

Wspomnianą część wieńczy próba spojrzenia na ewaluację systemów dokonujących ekstrakcji kluczowych informacji oraz realizujących pokrewne zadania na rzeczywistych dokumentach o bogatej strukturze graficznej i formatowaniu. Przedstawione w Rozdziale 9 ujęcie i wybór zadań, skupia się na zapewnieniu takiej procedury ewaluacji, która w jak największym stopniu odpowiada rzeczywistym zastosowaniom z zakresu automatyzacji procesów biznesowych.

Zróżnicowanie typów zadań oraz nagromadzenie problemów wzmiankowanych przy okazji opisu ekstrakcji kluczowych informacji sprawia, że zaproponowanie w tym miejscu rozwiązania bazującego na znakowaniu ciągu wiązałoby się licznymi trudnościami. Wprowadzony w rozdziale model referencyjny oparty o architekturę enkodera-dekodera podobnym ograniczeniom nie podlega.

Mimo faktu, że paradygmat znakowania ciągu jest jak dotąd szeroko rozpowszechniony w zadaniach identyfikacji fragmentu tekstu i ekstrakcji kluczowych informacji, w świetle niniejszej rozprawy można przypuszczać, że nastąpi jego częściowe porzucenie. Oczekiwać go można przede wszystkim w drugim z wymienionych zastosowań, w związku z coraz mocniej zaznaczoną pozycją alternatywy modeli opartych o architekturę enkoder-dekoder.

Pod względem struktury i treści, na rozprawę składa się osiem wcześniej opublikowanych artykułów podzielonych na trzy rozdziały. Dwa z nich opublikowano w ramach krajowego (PolEval 2018) i międzynarodowego (SemEval 2000) warsztatu poświęconego ewaluacji modeli uczenia maszynowego. Jeden opublikowano w czasopiśmie (Expert Systems with Applications), natomiast pozostałe pięć w ramach międzynarodowych konferencji CoNLL 2020, EMNLP 2020, ICDAR 2021, NeurIPS 2021 oraz ACL 2022.

Bibliografia

[1] J. Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *NAACL-HLT*. 2019 (cytowane na stronie 1).

[2] Ashish Vaswani et al. "Attention is All you Need". In: *NeurIPS*. 2017 (cytowane na stronie 1).

[3] Nam Nguyen and Yunsong Guo. "Comparisons of Sequence Labeling Algorithms and Extensions". In: *ICML*. 2007 (cytowane na stronie 1).

[4] Yinhan Liu et al. "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *ArXiv preprint* (2019) (cytowane na stronie 1).

[5] Giovanni Da San Martino et al. "SemEval-2020 Task 11: Detection of Propaganda Techniques in News Articles". In: *SemEval*. 2020 (cytowane na stronach 1, 2).

[6] Vikas Yadav and Steven Bethard. "A Survey on Recent Advances in Named Entity Recognition from Deep Learning models". In: *COLING*. 2018 (cytowane na stronie 1).

[7] Archana Goyal, Vishal Gupta, and Manish Kumar. "Recent Named Entity Recognition and Classification techniques: A systematic review". In: *Comput. Sci. Rev.* 29 (2018) (cytowane na stronie 1).

[8] Meizhi Ju, Makoto Miwa, and Sophia Ananiadou. "A Neural Layered Model for Nested Named Entity Recognition". In: *NAACL*. 2018 (cytowane na stronie 2).

[9] Arzoo Katiyar and Claire Cardie. "Nested Named Entity Recognition Revisited". In: *NAACL*. 2018 (cytowane na stronie 2).

[10] Taras K. Vintsyuk. "Speech discrimination by dynamic programming". In: *Kibernetika* 4.1 (1968) (cytowane na stronie 3).

[11] Z. Huang et al. "ICDAR2019 Competition on Scanned Receipt OCR and Information Extraction". In: *ICDAR*. 2019 (cytowane na stronie 3).

[12] Tomasz Stanisławek et al. "Kleister: Key Information Extraction Datasets Involving Long Documents with Complex Layouts". In: *ICDAR*. 2021 (cytowane na stronie 3).