Rozpoznawanie jednostek nazwanych i ekstrakcja informacji z dokumentów medycznych

Amelia Bieda

Promotor: dr inż. Daniel Kucharczyk

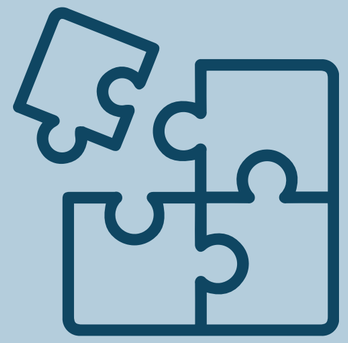
Plan prezentacji

1) Motywacja i cel pracy

2) Problem tokenizacji NER

3) Architektura Transformer

4) Modele typu BERT



5) Wykorzystane dane

6) Fine-tuning

7) Ocena wytrenowanych modeli

8) Wnioski

9) Dalsza praca i możliwości jej rozbudowania

Cel i motywacja pracy

Problem: ogromne ilości nieustrukturyzowanego tekstu, którego ręczna analiza jest niewykonalna.

Rozwiązanie: wykorzystanie modeli uczenia maszynowego, które realizują rozpoznawanie jednostek nazwanych i ekstrakcję informacji.

NER jako problem klasyfikacji tokenów [1]

Rozpoznawanie jednostek nazwanych

(ang. named entity recognition, NER) jest zadaniem, którego celem jest automatyczna identyfikacja i klasyfikacja predefiniowanych kategorii jednostek w nieustrukturyzowanym tekście.

Schemat etykietowania BIO

(ang. begining, inside, outside) - przykład zdania z kategoriami “choroba” i “lek”.

Leczenie	**cukrzycy**	**typu**	**pierwszego**	za	pomocą	**insuliny**	.
O	**B-CHOROBA**	**I-CHOROBA**	**I-CHOROBA**	O	O	**B-LEK**	O

[1] - Wang et al. “Cross-type biomedical named entity recognition with deep multi-task learning”. Bioinformatics, 1745–1752, 2018.

Architektura Transformer [2]

Rezygnacja z sieci rekurencyjnych dla **mechanizmu samo-uwagi**, który opiera się na trzech wektorach:

- zapytania (ang. query) - reprezentujący bieżący token,
- klucza (ang. key) - to jego kontekstowe cechy,
- wartości (ang. value) - zawiera faktyczną informację o tym tokenie.

Na tej podstawie model tworzy tzw. **rozkład wag uwagi**:

$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

gdzie Q, K, V to macierze zawierające odpowiednio wektory zapytań, kluczy i wartości dla wszystkich tokenów w sekwencji.

[2] Vaswani et al. “Attention is all you need”, 2017.

Modele BERT [3]

Model BERT (Bidirectional Encoder Representations from Transformers):

- analizuje tekst **dwukierunkowo**,
- jest trenowany na dwóch zadaniach:
  - Masked Language Modeling,
  - Next Sentence Prediction.

Istnieją jego wyspecjalizowane wersje, np. ClinicalBERT, BioBERT, PubMedBERT.

[3] - Devlin et al. “BERT: Pretraining of deep bidirectional transformers for language understanding”, 2019.

Dane BC5CDR [4]

- BioCreative V Chemical Disease Relation
- zbiór zawierający ludzkie adnotacje w kategoriach: substancje chemiczne i choroby
- stworzony z 1500 artykułów PubMed

[4] - Li et al. “BioCreative V CDR task corpus: a resource for chemical disease relation extraction”. Database, 2016.

Miary ewaluacji [5]

Tabela: Macierz pomyłek

	Prognozowana pozytywna	Prognozowana negatywna
Rzeczywista pozytywna	**Prawdziwie pozytywna (TP)**	**Fałszywie negatywna (FN)**
Rzeczywista negatywna	**Fałszywie pozytywna (FP)**	**Prawdziwie negatywna (TN)**

$$\text{Dokladnosc} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{FN}+\text{FP}+\text{TN}}$$

$$\text{Precyzja} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Czulosc} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

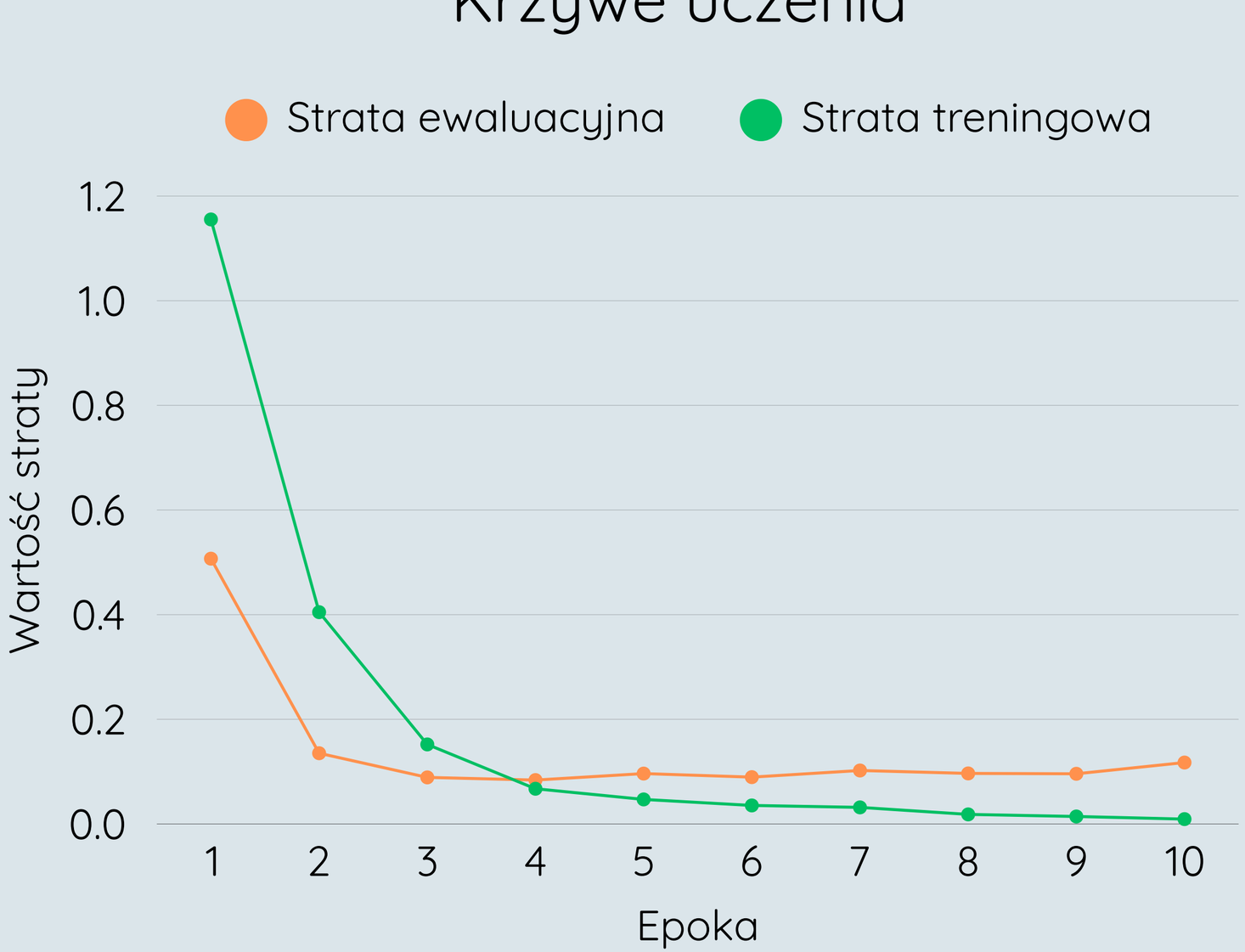
$$F_1 = \frac{2 \cdot \text{Czulosc} \cdot \text{Precyzja}}{\text{Czulosc} + \text{Precyzja}}$$

[5] - Sammut and Webb, editors. “Encyclopedia of Machine Learning and Data Mining.” 2 edycja 2017, str. 26, 497, 990, 991 .

Fine-tuning

Proces fine-tuningu:

- do ostatniej warstwy modelu dodajemy prostą warstwę klasyfikacyjną,
- trenujemy cały model na danych medycznych, oznaczonych etykietami BIO.



* dla fine-tuningu modelu PubMedBERT na danych BC5CDR

Efekty fine-tuningu

Modele typu BERT, po odpowiednim “dostrojeniu” mogą skutecznie analizować teksty medyczne, rozpoznając choroby i substancje chemiczne z wysoką dokładnością.

	Dokładność	Precyzja	Czułość	Wynik F1
BERT	96.08	79.93	85.82	82.77
ClinicalBERT	96.63	81.28	88.60	84.78
BioBERT	96.85	85.41	89.79	87.55
PubMedBERT	97.28	86.05	91.36	88.63

Kontynuacja i rozbudowa pracy

Ekstrakcja relacji (np. lek–leczy-chorobę)

Użycie nowszych i większych modeli

Analiza dokumentów klinicznych (nie tylko artykułów naukowych)

Rozpoznawanie jednostek wielojęzycznych

Stworzenie systemu do automatycznego streszczania danych medycznych

Literatura

1. Wang et al. “Cross-type biomedical named entity recognition with deep multi-task learning”. Bioinformatics, 1745–1752, 2018.
2. Vaswani et al. “Attention is all you need”, 2017.
3. Devlin et al. “BERT: Pretraining of deep bidirectional transformers for language understanding”. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, 2019.
4. Li et al. “BioCreative V CDR task corpus: a resource for chemical disease relation extraction”. Database, 2016.
5. Sammut and Webb, editors. “Encyclopedia of Machine Learning and Data Mining.” 2 edycja 2017, str. 26, 497, 990, 991 .

Dziękuję za uwagę