NLP - dokumentacja wstępna

Skład zespołu

- Kamil Dąbrowski
- Michał Iskra
- Damian Kolaska

Definicja problemu

Celem projektu jest stworzenie modelu, dla zadania określania podobieństwa semantycznego zdań w języku angielskim, opartego o model Sentence BERT.

Problem określania podobieństwa semantycznego (iSTS)

Na wejście modelu podajemy parę zdań w języku angielskim. Wyjście modelu powinno pozwolić określić stopień podobieństwa zdań.

Miary podobieństwa:

- ocena podobieństwa (ang. similarity score);
- typ dopasowania (ang. alignment type);

Ocena podobieństwa Ocena podobieństwa jest miarą liczbową ze skali od 0 do 5 określającą stopień podobieństwa bądź relacji między fragmentami zdania.

- ocena 5 - znaczenie obu fragmentów jest takie samo;
- oceny 4-3 - znaczenie jest bardzo podobne;
- oceny 2-1 - znaczenie jest podobne w niewielkim stopniu;
- ocena 0 - brak podobieństwa pomiędzy fragmentami;

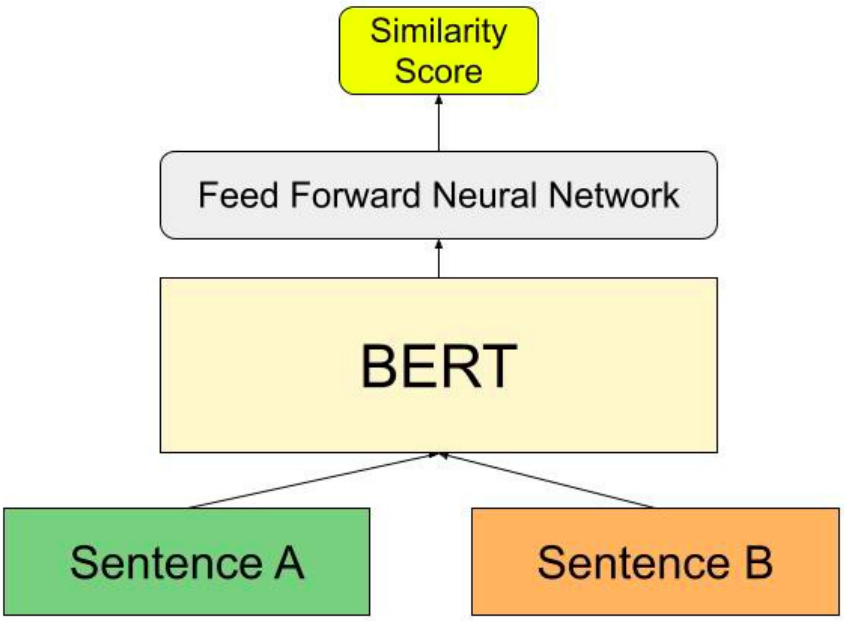
Typ dopasowania Typ dopasowania jest rozszerzeniem skali numerycznej. Typ dopasowania wyjaśnia, dlaczego fragmenty zostały ocenione jako powiązane bądź nie. Wyróżnia się osiem typów dopasowania:

- EQUI - oba fragmenty mają to samo znaczenie pod względem semantycznym;
- OPPO - fragment zdania pierwszego jest przeciwieństwem fragmentu ze zdania drugiego;
- SPE1 - znaczenie obu fragmentów jest podobne, przy czym pierwszy z nich jest bardziej szczegółowy;
- SPE2 - podobnie do SPE1, przy czy czym to fragment drugi zawiera więcej szczegółów;
- SIMI - porównywane fragmenty mają podobne znaczenie i dzielą podobne atrybuty, przy czym nie można ich przydzielić do żadnej z powyższych relacji (EQUI, OPPO, SPE1, czy SPE2);
- REL - podobnie do SIMI, przy czym fragmenty nie dzielą wspólnych atrybutów. Istnieje jednak pewna bliska relacja między nimi;

- NOALI - gdy dla fragmentu z jednego ze zdań nie istnieje odpowiadający mu fragment w drugim zdaniu.
- ALIC - typ podobny do NOALI, jednak w tym przypadku brak dopasowania spowodowany jest ograniczeniem w przypisywaniu fragmentów 1:1. Bez tego ograniczenia istnieje możliwość przyporządkowania danego fragmentu;

Model BERT

BERT (ang. Bidirectional Encoder Representations from Transformers) jest modelem mającym zastosowanie w zadaniach przetwarzania języka naturalnego. [4] Został pierwszy raz zaprezentowany w 2018 przez badaczy z firmy Google. Innowacyjną cechą modelu jest fakt, że w przeciwieństwie do poprzedników analizuje tekst w obu kierunkach, co pozwala na uzyskanie głębszego zrozumienia tekstu. Model BERT nie nadaje się jednak do zadania poszukiwania pary zdań o największym podobieństwie w zbiorze (ang. semantic similarity search).
Cytując pozycję [1] *Finding in a collection of n = 10 000 sentences the pair with the highest similarity requires with BERT n · (n−1)/2 = 49 995 000 inference computations. On a modern V100 GPU, this requires about 65 hours.*

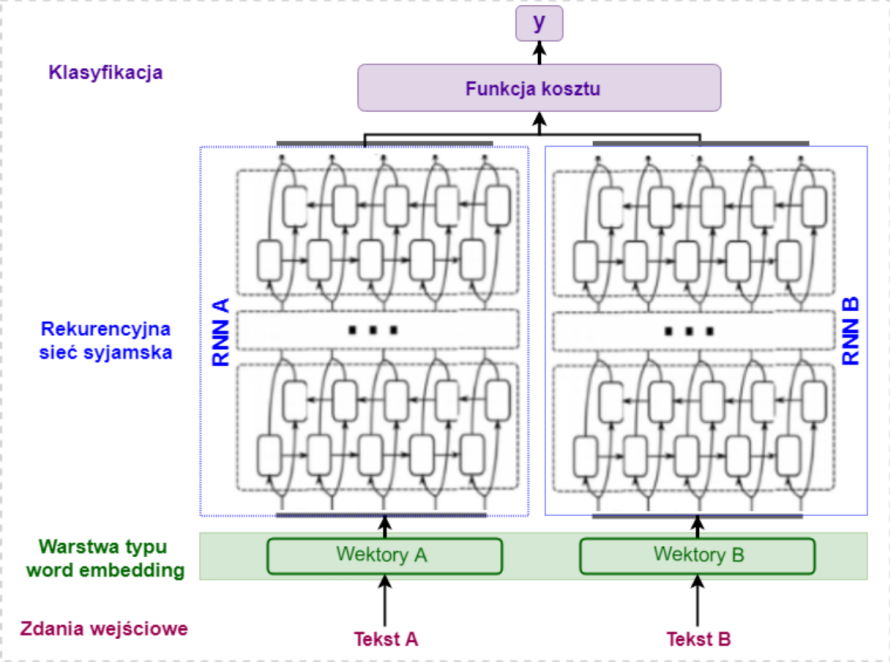


Rysunek 1. Zastosowanie modelu BERT dla zadania oceny podobieństwa. [5]

Porównując model BERT do XLNet, RoBERTa i DistilBERT, gdzie XLNet i RoBERTa poprawiają wydajność, podczas gdy DistilBERT poprawia szybkość wnioskowania. Dokładne porównanie znajduje się poniżej w tabeli.

Syjamska sieć neuronowa

Syjamska sieć neuronowa jest zbudowana z dwóch lub wiecej identycznych podsieci. W takiej sytuacji każda z podsieci ma taką samą konfigurację oraz wagi. Zbiór trenujący jest złożony z trójek (x1, x2, y), gdzie x1 i x2 to porównywane frazy, a y jest miarą dopasowania (oceną podobieństwa lub typ dopasowania). Pierwsza poróznywana fraza jest wejściem do pierwszej sieci, a fraza druga do sieci drugiej. Sieć ta opiera się na wielowarstwowych perceptronach MLP (ang. MultiLayer Perceptron), ponieważ jest to algorytm wstecznej propagacji błędu, który pozwala na stosunkowo proste trenowanie tej sieci. Trenowanie ma za zadanie maksymalizacje odległości w przestrzeni wektorów reprezentacji pomiędzy parami podobnymi przy braku podobieństwa oraz minimalizacja odległości w tej odległości przy podobieństwie. Każda podsieć łączy się jedną wspólną warstwą z pozostałymi podsieciami. W tej wartwie obliczana jest funkcja kosztu np.miara dystansu. Sieć ta została wykożystana do rozpoznawania podobieństwa semantycznego. Schemat architektury tego systemu znajduje się poniżej. [3]

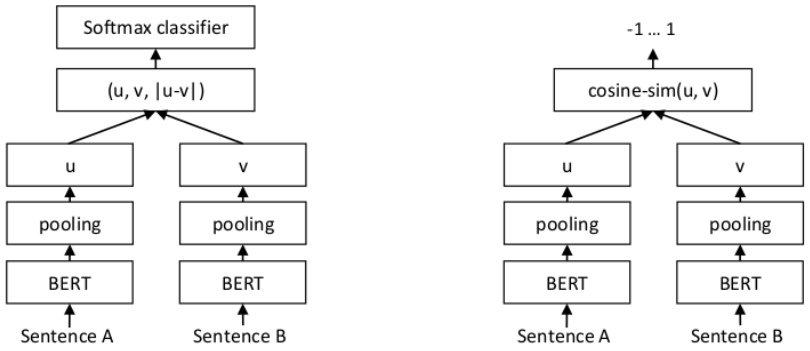


Rysunek 2. Architektura systemu z wykorzystaniem syjamskiej dwukierunkowej rekurencyjnej sieci neuronowej.[3]

Model Sentence BERT (SBERT)

SBERT jest rozszerzeniem modelu BERT umożliwiającym zastosowanie modelu w zadaniach, w których zastosowanie modelu BERT byłoby niepraktyczne, np.

- poszukiwanie pary zdań o najlepszym dopasowaniu (ang. semantic similarity search)
- grupowanie (ang. clustering)



Rysunek 3. Architektura modelu SBERT w trakcie uczenia (po lewej) oraz przy ocenie podobieństwa (po prawej). [1]

Architekturę modelu SBERT przedstawia Rysunek 1.2. Widzimy, że używa on dwóch kopii modelu BERT. Taką konfigurację nazywamy siecią syjamską. W uproszczeniu oznacza to, że modele dzielą wagi, które są aktualizowane przez wspólny gradient. Do obliczania straty używana jest funkcja *triplet loss*. Funkcja, aby obliczyć stratę, najpierw losowo ze zbioru wybiera wartość referencyjną, nazywaną kotwicą. Następnie szukana jest jedna wartość należąca do tej samej klasy co kotwica, tzw. pozytyw oraz jedna wartość należąca do innej klasy, tzw. negatyw. Odległość pozytywu od kotwicy jest minimalizowana, a odległość kotwicy od negatywu maksymalizowana. Aby zastosować funkcję SBERT używa specjalnej metryki do określania pozytywów oraz negatywów. Taki sposób obliczania straty stanowi główną siłę modelu SBERT, ponieważ sprawia, że model można nauczyć tylko na podstawie jednego przejścia po zbiorze uczącym bez konieczności porównania wszystkich kombinacji zdań. [5] [6]

Studia literaturowe

Do zadania iSTS dostępnych jest wiele modeli. Najpopularniejsze z nich do BERT, RoBERT oraz XLNet oraz SBERT. Model SBERT od innych modeli wyróżnia przede wszystkim wydajność. Jak widzimy a rysunku 5. model RoBERT jest bardzo złożonym modelem, przewyższa pod tym względem nawet model BERT, od którego SBERT obiecuje być o rzędy wielkości szybszy.

Model SBERT osiąga wydajność podobną do metod bezkontekstowych tworzenia zanurzeń takich jak GloVe (Rysunek 6.).

Model	STS12	STS13	STS14	STS15	STS16	STSb	SICK-R	Avg.
Avg. GloVe embeddings	55.14	70.66	59.73	68.25	63.66	58.02	53.76	61.32
Avg. BERT embeddings	38.78	57.98	57.98	63.15	61.06	46.35	58.40	54.81
BERT CLS-vector	20.16	30.01	20.09	36.88	38.08	16.50	42.63	29.19
InferSent - Glove	52.86	66.75	62.15	72.77	66.87	68.03	65.65	65.01
Universal Sentence Encoder	64.49	67.80	64.61	76.83	73.18	74.92	**76.69**	71.22
SBERT-NLI-base	70.97	76.53	73.19	79.09	74.30	77.03	72.91	74.89
SBERT-NLI-large	72.27	**78.46**	**74.90**	80.99	76.25	**79.23**	73.75	76.55
SRoBERTa-NLI-base	71.54	72.49	70.80	78.74	73.69	77.77	74.46	74.21
SRoBERTa-NLI-large	**74.53**	77.00	73.18	**81.85**	**76.82**	79.10	74.29	**76.68**

Table 1: Spearman rank correlation ρ between the cosine similarity of sentence representations and the gold labels for various Textual Similarity (STS) tasks. Performance is reported by convention as $\rho \times 100$. STS12-STS16: SemEval 2012-2016, STSb: STSbenchmark, SICK-R: SICK relatedness dataset.

Rysunek 4. Porównanie modelu SBERT z innymi modelami dla zadania STS. [1]

	BERT	RoBERTa	DistilBERT	XLNet
Size (millions)	**Base**: 110 **Large**: 340	**Base**: 110 **Large**: 340	**Base:** 66	**Base**: ~110 **Large**: ~340
Training Time	**Base**: 8 x V100 x 12 days* **Large:** 64 TPU Chips x 4 days (or 280 x V100 x 1 days*)	**Large:** 1024 x V100 x 1 day; 4-5 times more than BERT.	**Base**: 8 x V100 x 3.5 days; 4 times less than BERT.	**Large:** 512 TPU Chips x 2.5 days; 5 times more than BERT.
Performance	Outperforms state-of-the-art in Oct 2018	2-20% improvement over BERT	3% degradation from BERT	2-15% improvement over BERT
Data	16 GB BERT data (Books Corpus + Wikipedia). 3.3 Billion words.	160 GB (16 GB BERT data + 144 GB additional)	16 GB BERT data. 3.3 Billion words.	**Base**: 16 GB BERT data **Large**: 113 GB (16 GB BERT data + 97 GB additional). 33 Billion words.
Method	BERT (Bidirectional Transformer with MLM and NSP)	BERT without NSP**	BERT Distillation	Bidirectional Transformer with Permutation based modeling

Rysunek 5. Porównanie modelu BERT z innymi modelami. [7]

Model	CPU	GPU
Avg. GloVe embeddings	6469	-
InferSent	137	1876
Universal Sentence Encoder	67	1318
SBERT-base	44	1378
SBERT-base - smart batching	83	2042

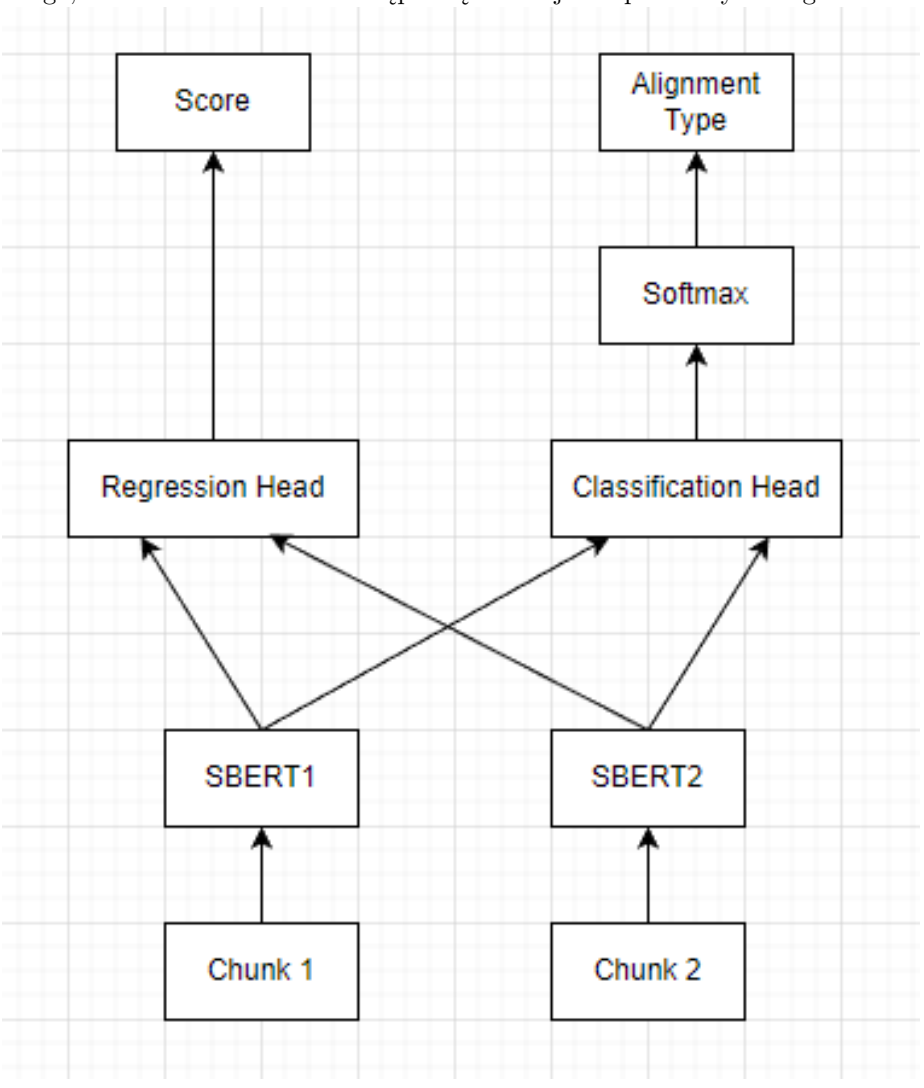
Rysunek 6. Porównanie wydajności (zdania na sekundę) różnych metod tworzenia zanurzeń [1]

Proponowane rozwiązanie

Model

Zanurzone wektory zdań będące rezultatem modelu SBERT podamy na wejście dwóch jednowarstowych perceptronów. Jednego służącego do regresji (ocena), a drugiego do klasyfikacji (dopasowanie).

Pomijamy kwestię dzielenia i dopasowywania fragmentów tekstu. Skorzystamy z tego, że dla zadań SemEval dostępne są dane z już dopasowanymi fragmentami.



Rysunek 5. Architektura proponowanego rozwiązania.

Funkcja kosztu

Regression Head - mean squared error (MSE)
Classification Head - binary cross-entropy (BCE)

Dwa podejścia wyliczania funkcji straty:

- ocena i typ uczone razem - suma wyliczonych strat, wspólny gradient;
- ocena i typ uczone oddzielnie - oddzielne funkcje straty, oddzielny gradient;

Dane

Model zostanie wytrenowany i przetestowany na danych z zadań SemEval z 2015 i 2016.

Literatura

[1] Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, Nils Reimers and Iryna Gurevych
[2] Interpretable semantic textual similarity of sentences using alignment of chunks with classification and regression
[3] Rozpoznawanie podobieństwa semantycznego z wykorzystaniem architektur uczenia głębokiego, Ewelina Grudzień
[4] Wykorzystanie rozwiązań opartych na modelu BERT w określaniu podobieństwa semantycznego, Aleksandra Budzyńska
[5] https://towardsdatascience.com/an-intuitive-explanation-of-sentence-bert-1984d144a868
[6] https://en.wikipedia.org/wiki/Triplet_loss
[7] https://towardsdatascience.com/bert-roberta-distilbert-xlnet-which-one-to-use-3d5ab82ba5f8
[8] https://www.quora.com/What-are-the-main-differences-between-the-word-embeddings-of-ELMo-BERT-Word2vec-and-GloVe