

Similaridade de Sentenças com Embeddings e Redes Siamesas

CPE 727 - Aprendizado de Profundo

Felipe Fink Grael

11 de dezembro de 2025

Introdução

“O cachorro está correndo na praça”

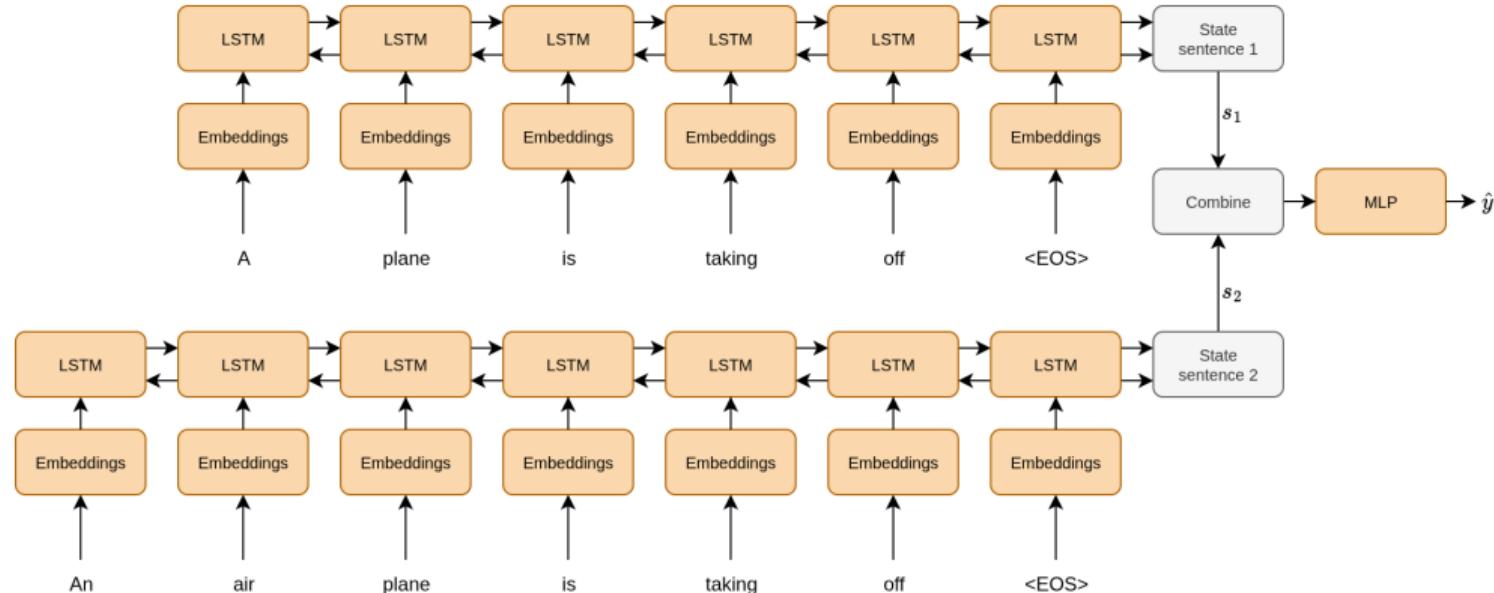
“O cão corre no parque”

- Como medir a similaridade **semântica** entre essas sentenças?
- Métodos esparsos, como TF-IDF são insuficientes, especialmente em textos curtos.
- **Objetivo:** usar métodos de Deep Learning para aprender representações semânticas e medir similaridade entre sentenças.

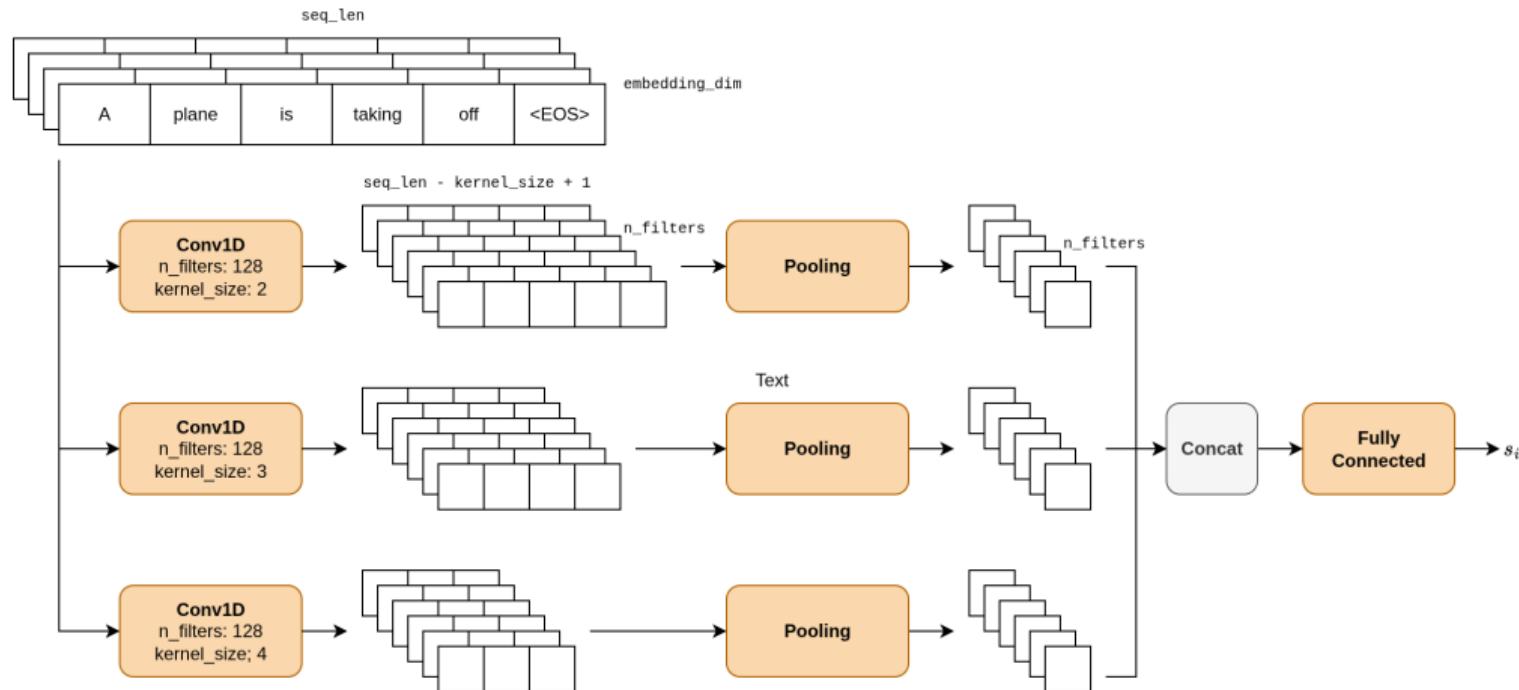
Trabalhos Relacionados

- Métodos de Word Embeddings, como Word2Vec [1] e FastText [2], são eficazes para capturar similaridade semântica entre palavras.
- Redes Recorrentes são amplamente utilizadas em diversas tarefas de NLP, de classificação a tradução automática [3]
- Arquitetura TextCNN aplica redes convolucionais para tarefas de classificação de texto [4]
- Arquitetura de redes siamesas para comparação de sentenças [5]

Redes Siamesas



Arquitetura TextCNN



Semantic Textual Similarity Benchmark

- Conjunto de dados STS-B [6] desenvolvido para o SemEval 2017.
- contém pares de sentenças anotados com similaridade semântica em uma escala de 0 a 5.
- Utilizado para treinar e avaliar modelos de similaridade de sentenças.
- Métrica de avaliação: Correlação de Pearson entre as similaridades previstas e anotadas.
- 7249 pares para treino e validação, 1379 para teste.
- Tipicamente avaliado com Correlação de Pearson.

Baselines

TF-IDF

- 2000 tokens mais frequentes no conjunto de dados.
- Similaridade de Cosseno entre sentenças

$$r = 0.62$$

Autoencoder

- Parte da mesma tokenização do TF-IDF.
- Vetores com norma unitária
- Encoder (2000, 300, 100)
- Similaridade de Cosseno representações latentes

$$r = 0.49$$

Método

- Redes Siamesas LSTM e CNN
- Validação cruzada k-fold, k=5
- Combinação dos segmentos de rede é a concatenação de:
 - Cada uma das representações
 - Diferença absoluta entre as representações
 - Produto elemento a elemento entre as representações
- Rede final Fully Connected:
 - Entrada tem dimensão 4x a dimensão da representação de cada rede
 - Uma camada escondida com função de ativação ReLU
 - Camada de saída com ativação linear prevendo similaridade
- Embeddings FastText pré-treinados, mas ajustados durante o treinamento
- Exploração de hiperparâmetros com Optuna [7]

Arquiteturas Usadas

LSTM Siamesa

- 2 camadas de LSTM bidirecional, 32 unidades cada
- Dimensão 128 na camada escondida da fully connected
- Learning rate de 8.5e-3
- Dropout de 0.7
- Weight Decay de 1e-3

CNN Siamesa

- Kernels de tamanho 3, 4 e 5
- 512 filtros por kernel
- Pooling: max
- Dimensão 256 na camada escondida da fully connected
- Learning rate de 1.2e-4
- Dropout de 0.41
- Weight Decay de 4.7e-6

Resultados

Método	Correlação de Pearson (r)
TF-IDF (baseline)	0.62
Autoencoder (baseline)	0.49
LSTM Siamesa	0.46 ± 0.019
CNN Siamesa	0.41 ± 0.013

Médias e desvios para conjunto de validação (5-fold cross-validation).

Discussão

- Método é bastante sensível a features na concatenação final e hiperparâmetros
- Modelo TF-IDF, mesmo com limitações, captura boa parte da similaridade neste dataset
- Técnicas modernas superam o baseline, mas usam modelos de larga escala
- Trabalhos futuros:
 - Explorar pre-treino com mais dados
 - Explorar função custo contrastiva

Referências Bibliográficas

- [1] T. Mikolov, K. Chen, G. S. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *International Conference on Learning Representations*, 2013.
- [2] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *Transactions of the Association for Computational Linguistics*, vol. 5, 07 2016.
- [3] W. De Mulder, S. Bethard, and M.-F. Moens, “A survey on the application of recurrent neural networks to statistical language modeling,” *Computer Speech and Language*, vol. 30, no. 1, pp. 61–98, 2015.

Referências Bibliográficas

- [4] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (A. Moschitti, B. Pang, and W. Daelemans, eds.), (Doha, Qatar), pp. 1746–1751, Association for Computational Linguistics, Oct. 2014.
- [5] P. Neculoiu, M. Versteegh, and M. Rotaru, "Learning text similarity with Siamese recurrent networks," in *Proceedings of the 1st Workshop on Representation Learning for NLP* (P. Blunsom, K. Cho, S. Cohen, E. Grefenstette, K. M. Hermann, L. Rimell, J. Weston, and S. W.-t. Yih, eds.), (Berlin, Germany), pp. 148–157, Association for Computational Linguistics, Aug. 2016.

Referências Bibliográficas

- [6] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (S. Bethard, M. Carpuat, M. Apidianaki, S. M. Mohammad, D. Cer, and D. Jurgens, eds.), (Vancouver, Canada), pp. 1–14, Association for Computational Linguistics, Aug. 2017.
- [7] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, (New York, NY, USA), p. 2623–2631, Association for Computing Machinery, 2019.