E-commerce products
Project Proposal for NLP Course, Winter 2022

Paweł Golik, Mateusz Jastrzebiowski and Aleksandra Muszkowska
Warsaw University of Technology
pawel.golik.stud@pw.edu.pl,
mateusz.jastrzebiowski.stud@pw.edu.pl,
aleksandra.muszkowska.stud@pw.edu.pl

supervisor: Anna Wróblewska
Warsaw University of Technology
anna.wroblewska1@pw.edu.pl

Abstract

In this project, we will explore modern methods used in the product matching problem, which is a generalization of the entity matching problem. These tools are mainly based on deep neural networks that encode individual offerings into vectors called embeddings representing specific knowledge. We will analyze the embedded space and conduct an attempt to develop probing tasks aimed at investigating the properties of embeddings in this domain. In addition, we will check whether encoded vectors of offers for the same products will show high similarity and, on the contrary, offers of unrelated products will not be similar. We believe our work will help understand black-box knowledge representations (embeddings) and shed light on the similarity properties of embedded vectors.

1 Introduction

The e-commerce sector has seen much growth in recent years, compounded by the global coronavirus pandemic. Customers were previously limited to the offerings of local sellers. However, the growth of the Internet and delivery services can open up new sources of goods provided by e-commerce sites such as Allegro.pl and Amazon. The vast number of offers published daily by vendors leads to new challenges in efficiently finding offers of potential interest to customers. Unfortunately, many offers are presented in different formats and with different variations of product names which makes it challenging to build automated tools for matching offers of the same product. Having a tool capable of comparing two different offers with each other would allow for solving many problems, such as offers matching, but also suggesting similar offers in the absence of offers related to the selected product or detecting offers misclassified by vendors as a given product based on too little similarity to other, valid offers. Such a tool would not just answer whether two offers are for the same product but would be able to determine the similarity of two offers on some continuous scale. This approach provides a more general scope of application, such as casting the product matching problem as a zero-shot learning problem and avoiding re-training the model on introducing new products or product categories. In addition, it would be helpful to explain the representation of knowledge and the similarity between the two offers.

2 Related Work

Modern state-of-the-art product matching methods rely on deep learning techniques using Transformer models, which allow for creating embeddings from input representing specific knowledge about the encoded entities (Możdżonek et al., 2022), (Tracz et al., 2020). The embeddings map words to real-valued vectors, which reveal semantic aspects, for example, if words are related in meaning or belong to the same topic. Creating such an embedding means enriching as well as filtering out information. As far as we know, most of the research in product matching focuses on building classifiers on top of the extracted embedding representation (Możdżonek et al., 2022). In the training phase, the encoder learns to transform the input into the embedding space, which serves well for the classification task. A slightly different approach is presented in (Tracz et al., 2020), where the embeddings are directly compared and passed to the Loss function assessing their similarity. The Loss function is then minimized for the embedded inputs, which causes the weights of the encoder to change accordingly. This approach may probably ensure more suitable embed-

ding space for future examininig of the similarity of the embeddings as it already imposes a similarity constraint during the training phase. As far as we know, we would be the first to describe probing tasks for embeddings in the product-matching domain. Probing has been extensively described in (Şahin et al. , 2020), but it focuses mainly on probing word embeddings. In our case, we need to probe the embeddings of the offers created from texts consisting of many words. Lindstrom et al. (2020) propose novel probing tasks for the visual-semantic case (pairing images and text), defining three classification tasks relating to the images and text from which the embeddings were created. We can take inspiration from this approach and create similar probing tasks corresponding to the product-matching domain.

3 Our research proposition

While the power of embeddings comes from the distillation and enrichment of information through machine learning, their inner workings are poorly understood, and there is a shortage of analysis tools. To address this problem, we aim to examine the embedding space with probing tasks tailored to the product-matching case. This will allow us to shed light on the embedding space and answer questions about the information embedded in the vectors. For example, suppose embeddings can provide information about the number of words in an input text (representing a particular offer). In that case, it may indicate that the method is vulnerable to misclassifying two offers of the same product but of a different description length.

Additionally, we want to examine the similarity between a given pair of embeddings using similarity measures such as the generalized cosine similarity function (Duffner et al., 2021). We anticipate that the offers of the same product will resemble very high similarity, and those of similar products will have a slightly lower but still high similarity value. On the other hand, we believe that offers of non-related products will not be similar.

4 Dataset and EDA

We focus on Web Data Commons - Training Dataset and Gold Standard for Large-Scale Product Matching dataset (WDC for short) prepared by the staff of the University of Mannheim (Primpeli et al., 2019). The dataset contains offers in four categories - Cameras, Computers, Watches, and Shoes. Additionally, each offer is linked to a specific product (cluster_id) and contains textual attributes such as title, description etc. Each observation is a pair of such offers and a label indicating whether these two offers are for the same product (a positive pair) or not (a negative pair). Even in the case of a negative pair, both offers belong to the same category (but different clusters/products).

The training datasets are available in different sizes, varying from small to extra large. In every dataset, the ratio between positive and negative pairs is 1:3. Table 1 presents the exact sizes for each dataset. The proportions between the different collections within one category are as follows: 1 – small, 3 – medium, 15 – large, and 50 – extra-large (xlarge). We do not consider the extra-large datasets as our computational and time resources are limited.

The Gold Standard is verified manually and should be used for testing purposes. Each product contains highly similar negative pairs (complex cases) and less similar negative pairs (easy cases). Table 2 depicts the statistics for the Golden Standard dataset per each category.

The WDC dataset has already been used for product-matching tasks (Możdżonek et al., 2022) and the results of this research is depicted in Table 3.

5 Approach and research methodology

5.1 Textual representation

The representation of each offer is a concatenated title, attribute values, descriptions, attribute names, and units, which are then lower-cased. We may decide to get rid of some of these attributes if they deteriorate the model performance.

5.2 Encoder architectures

We will use state-of-the-art architecture to test product matching methods. Citing the work of (Możdżonek et al., 2022), we will test the pretrained mBERT and XLM-RoBERT models on a WDC dataset with two output classes to distinguish corresponding and non-corresponding pairs of the offer. The input data will be prepared in the following form:

1. token [CLS] - needed for the classification task,

Table 1: WDC datasets sizes. Source: (Peeters et al., 2022)

Category	Size	Positive	Negative	Total
Cameras	Small	486	1,400	1,886
	Medium	1,108	4,147	5,255
	Large	3,843	16,193	20,036
	xLarge	7,178	35,099	42,277
Computers	Small	722	2,112	2,834
	Medium	1,762	6,332	8,094
	Large	6,146	27,213	33,359
	xLarge	9,690	58,771	68,461
Watches	Small	580	1,675	2,255
	Medium	1,418	4,995	6,413
	Large	5,163	21,864	27,027
	xLarge	9,264	52,305	61,569
Shoes	Small	530	1,533	2,063
	Medium	1,214	4,591	5,805
	Large	3,482	19,507	22,989
	xLarge	4,141	38,288	42,429

2. data of the first offer,
3. token [SEP] - separator,
4. data of the second offer,
5. token [SEP] - marking the end.

In contrast to the work of (Możdżonek et al., 2022), we will use not only the titles of the offerings but also attribute values, descriptions, attribute names, and units.

5.3 Probing

The main task of this project is product matching. However, the models proposed are so-called black boxes from which it is difficult to deduce why such decisions were made. So the second crucial task will be to explain the model and find the relationship between the embeddings and the input data. To this end, we propose to use the probing method. Its task is to examine what information is contained in the word embeddings created by the NLP model.

The general outline of probing (well described in (Belinkov , 2021)) is to take a model trained on some task, product matching in our case. Then generate representations using the model and train another classifier that takes the representations and predicts some properties. If the classifier performs well, we say the model has learned information relevant to the property.

In our case, the probing classifier's input will be the offers' embeddings. We will examine whether they have learned a relationship with certain properties and check whether they contain information unrelated to the task. Examples of probing tasks could be to explore the relationship of the length of the concatenated input (offer, i.e., title + description +, etc.) to the embedding of the offer. Another example of a probing task is examining the impact of keyword content in the offer description.

5.4 Similarity measures

After learning the models, the goal is to extract the embeddings of each offer to test the similarity of embeddings from similar ones. To do this, we will use various metrics, including the cosine metric (Duffner et al., 2021). In particular, we consider two offers to match when, given the representation of the two offers in the embedded space, they will be close to each other at a cosine distance.

Moreover, if the hypothesis that similar offerings are close to each other in the embedding space is correct, it will be easy to classify new offers by calculating their embeddings. What is more, this way of classifying new offers can also be used for new product offers that have not appeared in the training dataset. The testing similarity of embeddings can also be considered a probing task.

5.5 Dataset split

We preserve the split of data presented by the authors of the WDC dataset (Peeters et al., 2022).

6 Work plan

Our project consists of 3 main parts: product matching, probing, and similarity measures. To begin with, we will study the data in depth to learn about the relationships in the data. The next task is to train our dataset's mBERT and XLM-RoBERT models. An important task will be to extract the embeddings of each offering using our models properly.

Then, we will analyze the resulting embeddings for similarity to see if similar offerings occur close to each other in different metrics.

The most important task will be the novel idea of explaining the embeddings of the offers using probing methods. This task is divided into two parts. First, a review of the data and creative and

Table 2: Gold Standard Statistics per category. Source: (Primpeli et al., 2019)

Category	#positive	#negative	#combined	title	s:description	spec Table
Computers	150	400	550	100%	88%	21%
Cameras	150	400	550	100%	79%	5%
Watches	150	400	550	100%	77%	5%
Shoes	150	400	550	100%	88%	3%

Table 3: Current state-of-the-art F1 scores in product matching task for models trained on English WDC datasets. Mean value and standardized error (confidence level 95%) for each dataset were calculated from 4 samples. Source: (Możdżonek et al., 2022)

Category	Size	mBERT [x]	XLM-RoBERTa [y]	Ditto [z2]	WDC-Deepmatcher [z]
Cameras	Small	82.13(±4.70)	81.96(±7.75)	80.89	68.59
	Medium	87.86(±2.04)	88.11(±4.22)	88.09	76.53
	Large	90.88(±2.28)	92.36(±0.76)	91.23	87.19
	xLarge	-	-	93.78	89.21
Computers	Small	86.43(±3.69)	81.10(±13.40)	80.76	70.55
	Medium	90.13(±1.89)	88.69(±2.19)	88.62	77.82
	Large	92.48(±2.33)	93.71(±0.77)	91.70	89.55
	xLarge	-	-	95.45	90.80
Watches	Small	79.20(±7.89)	74.98(±13.36)	75.89	73.86
	Medium	84.11(±3.40)	81.30(±8.21)	82.66	79.48
	Large	90.28(±2.36)	91.26(±2.09)	88.07	90.39
	xLarge	-	-	90.10	92.61
Shoes	Small	87.31(±1.64)	83.78(±4.38)	85.12	66.32
	Medium	91.17(±4.21)	89.50(±3.69)	91.12	79.31
	Large	93.52(±2.63)	93.62(±0.67)	95.69	91.28
	xLarge	-	-	96.53	93.45

researcher work to obtain relevant probing tasks. Second, conducting a numbers of experiments to answer whether offers embeddings explain the information contained.

The project will be created in Python using ML libraries such as Hugging Face, Scikit-learn, and more.

References

[Belinkov 2021] Yonatan Belinkov. 2021. *Probing Classifiers: Promises, Shortcomings, and Advances*. Computational Linguistics, 48(1):207–219.

[Lindström et al. 2020] Lindström, Adam & Björklund, Johanna & Bensch, Suna & Drewes, Frank. 2021. *Probing Multimodal Embeddings for Linguistic Properties: the Visual-Semantic Case*. *In Proceedings of the 28th International Conference on Computational Linguistics*, pages 730–744, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[Şahin et al. 2020] Şahin, Gözde Gül and Vania, Clara and Kuznetsov, Ilia and Gurevych, Iryna 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. *Computational Linguistics*. **46**, 335-385 (2020,6), https://doi.org/10.1162/coli

[Możdżonek et al.2022] Możdżonek, Michał & Wróblewska, Anna & Tkachuk, Sergiy & Łukasik, Szymon 2022. *Multilingual Transformers for Product Matching – Experiments and a New*

Benchmark in Polish.. 1-8. 10.1109/FUZZ-IEEE55066.2022.9882843.

[Duffner et al.2021] Duffner, Stefan & Garcia, Christophe & Idrissi, Khalid & Baskurt Atilla 2021. *Similarity Metric Learning. Multi-faceted Deep Learning - Models and Data*

[Tracz et al.2020] Tracz, Janusz & Wójcik, Piotr Iwo & Jasinska-Kobus, Kalina & Belluzzo, Riccardo & Mroczkowski, Robert & and Gawlik, Ireneusz 2020. *BERT-based similarity learning for product matching. In Proceedings of Workshop on Natural Language Processing in E-Commerce, pages 66–75, Barcelona, Spain. Association for Computational Linguistics.*

[Primpeli et al.2019] Primpeli, A., Peeters, R., & Bizer, C. 2019. *The WDC Training Dataset and Gold Standard for Large-Scale Product Matching. Companion Proceedings of The 2019 World Wide Web Conference.*

[Peeters et al.2022] Peeters, Ralph & Bizer, Christian 2022. *Cross-language learning for product matching. WWW Companion, 2022a*