Generating E-commerce Product Titles in Portuguese

Livy Real[1], Karina M. Johansson[2], Júlio C. S. Mendes[3], Bianca M. Lopes[2], Marcio T. I. Oshiro[1]

[1]B2W Digital - São Paulo, São Paulo, Brazil

[2]Federal University of São Carlos - São Carlos, São Paulo, Brazil

[3]University of São Paulo - São Paulo, São Paulo, Brazil

livy.coelho,marcio.oshiro{@b2wdigital.com}

karina.mayumi,bianca.lopes{@estudante.ufscar.br}

julioc.silvamendes@hotmail.com

Abstract. *This paper explores how Natural Language Processing techniques can be integrated to solve real-world problems in the e-commerce scenario. We address the issue of having high quality information products offered to customers in a marketplace platform, composed by thousands of sellers producing original content in multiple languages, following different SEO and cultural assumptions. We propose an NLP pipeline to generate high quality titles products in Portuguese.*

Resumo. *Este trabalho explora como integrar técnicas de Processamento de Linguagem Natural para solucionar um problema real. Exploramos o domínio do e-commerce, em especial de marketplaces, que oferecem a seus consumidores produtos de milhares de vendedores. A diversidade de vendedores e de seus backgrounds culturais faz com que a qualidade da informação disponível em sites de e-commerce seja normalmente pobre e não uniforme. Neste trabalho, propomos um fluxo de geração de títulos de produtos à venda em marketplaces que garante qualidade ao catálogo de um dos maiores marketplaces brasileiros.*

1. Introduction

Nowadays, e-commerce marketplaces play an important role in the consumer journey, both to customers and sellers. A marketplace is a web platform that sells products from different sellers. In the pandemic situation, we saw a growth of 73,88% in the Brazilian e-commerce[1]. Also, in Brazil, it became more usual to buy offshore products, since often offshore sellers offer a larger product assortment and a better price than national sellers for specific types of products, as wearable and kitchenware. In this paper, we explore the case of *Americanas Mundo*[2], the international marketplace of *Americanas*[3], one of the largest e-commerce platforms of Latin-America.

[1]https://www.ecommercebrasil.com.br/noticias/e-commerce-brasileiro-cresce-dezembro/

[2]https://www.americanas.com.br/hotsite/americanas-mundo

[3]https://www.americanas.com.br

Americanas Mundo offers products from all over the world, specially from China, to Brazilian customers. Nowadays, it offers around 20MI products, sold by thousands of sellers, who also have different technology setups. Since many of these sellers are not able to upload product content in different languages, mainly considering how fast new products appear in the market, offshore sellers who want to use *Americanas Mundo* marketplace need to provide information about their products already in Portuguese. In this context, the quality of the product information that the platform displays is extremely poor, because product titles and descriptions are often translated by sellers using a free general domain automatic translator. Considering the variety of source languages that the products are described and how different the background technology of each seller is, the task of improving the products information quality is not feasible through a machine translation approach.

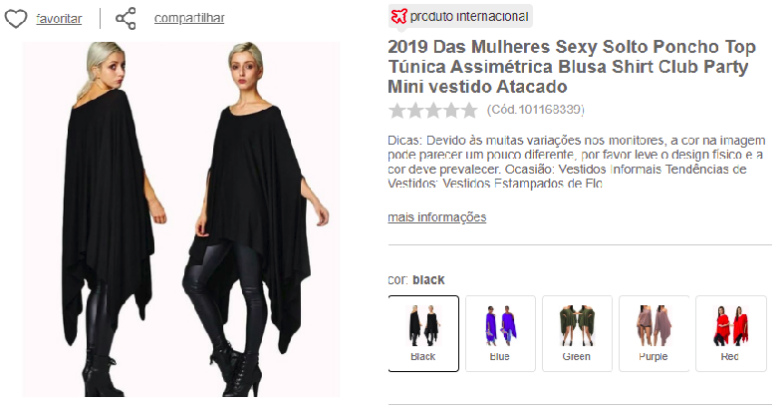

Figure 1. Example of fashion product in Americanas Mundo

In this paper, we show how we tackle this issue, breaking it in two Natural Language Processing (NLP) tasks: named entity recognition (NER) and template-based generation (TBG). Our main goal is to have a quickly implemented pipeline that assures quality to the **fashion** products titles offered by offshore sellers in *Americanas Mundo* marketplace.

2. Related Work

Using NLP techniques to improve the quality of information on e-commerce webpages became a hot topic in the applied NLP community in the last years. Recent advances include title generation [Mathur et al. 2018], description generation [Zhangming Chan 2019] and product summarization [Peng Yuan and Zhou 2020].

However, the use of NER in the e-commerce context is not new [Mahesh Joshi 2015]. Although using NER for improving products catalog is a feasible and successful task, how to scale and keep updated models running are still the targets of many recent works [Huimin Xu 2020, Zhang et al. 2020]. Recently some works also tackle the issue through a multi-modal approach [Najmi 2019].

As always, much work has been done for English and Chinese languages. As far as we know, this is the first work to address the generation of high-quality e-commerce information in Portuguese.

3. General Pipeline

Since what we have at hand is a huge amount of poorly translated product information and no trustful data in Portuguese, to solve the problem of how to offer a better-quality product information to our customers, we decided to follow a building-in blocks pipeline consisting in two different tasks: NER and TBG.

The NER module is responsible for extracting relevant attributes from the 'original' titles we had. For this, we used the MITIE library, detailed bellow. After the extracting phase, the TBG module generates a new title, considering which attributes to keep, their normalization, solving few post-edition translation issues and, finally, proposing an attribute order to the new titles. The different steps of this module will be discussed in section 3.2.

3.1. Named-entity Recognition

The task of named entity recognition (NER) involves the extraction and categorization of entities from an unstructured text. In general domains, entities such as localities, people and organizations are common. But here, the entities are attributes of fashion products, such as *model*, *size*, and *color*.

For the generation of the NER model, it was used the MIT Information Extraction Toolkit (MITIE) [4], a tool for the extraction of named entities and binary relations. MITIE is based on unsupervised learning to reduce corpora dimensionality and on supervised learning for the classification task. We used a corpus of 30,000 poorly translated titles of the fashion domain to train MITIE embeddings and produced a manually labeled corpus to classify the relevant attributes to our task. In section 3.1.1, we describe the annotation process we followed. Our larger dataset and the annotated training set were provided as input to MITIE. As output, the trained NER model was obtained and evaluated using the annotated test set. The results are shown in the Table 1.

Table 1. NER model evaluation

	precision	**recall**	**f1-score**
Color	0.67	0.40	0.50
Public	1.00	0.62	0.76
Material	0.80	0.71	0.75
Model	0.64	0.51	0.57
Occasion	0.76	0.61	0.67
Size	0.50	0.50	0.50
Product Type	0.88	0.80	0.84
avg	0.80	0.63	**0.71**

3.1.1. Annotated Corpora

To build a representative corpus for training and testing our classification module, we started from the 30,000 fashion titles we had. It was grouped applying the Agglomerative

[4] https://github.com/mit-nlp/MITIE

Clustering[5] algorithm. From the 24 clusters we got, the 400 most representative titles were selected, in other words, the most similar ones to theirs clusters. That selection was made by choosing the titles with the greater sums of cosine similarity with the other titles of the cluster. We discarded all titles with serious tokenatization issues and finally we had 387 titles to annotate. The generated dataset was split in 358 titles for training and 29 for testing.

The annotation work was performed by two linguists of the team, reaching an agreement of 0.7 of Cohen Kappa coefficient, which is considered a substantial agreement [Landis and Koch 1977]. After the first annotation round, a third linguist reviewed the annotation to solve any disagreement.

The annotation process was based on guidelines we internally created together with the *Americanas Mundo* Fashion Department. Our goal was to indicate how to construct an ideal structure for attributes and titles of fashion products. Hence, we annotated terms that would suit the following attributes: Product Type, Model, Public (gender and age), Brand, Occasion, Color, Material, and Size. In our dataset, there were not enough examples of Brand, maybe due to the nature of these products, therefore we left this attribute aside.

3.2. Title Generation

This phase consisted in producing a new title based on the extracted attributes. First, we needed to normalize the attributes we found. At this stage, we tried to use different lemmatizers available to Portuguese, as spaCy[Honnibal et al. 2020] and Stanza[Qi et al. 2020]. However, it was concluded that its use was not suitable for our domain. Many ambiguous nouns in Portuguese are not ambiguous in the e-commerce domain, but the general domain NLP tools are not able to prioritize the relevant word form to this work. An example of this problem is the noun "vestido" ("dress", it can also mean the participle "dressed"), which was lemmatized to "vestir" ("to dress").

Finally, our generation pipeline was composed by 1. a post-editing translation dictionary[6], 2. a singularizer, 3. the cleaning up of repeated information, 4. normalization script[7] and 5. an attribute ordering template.

For the template generation, we followed the structures proposed by the Search Engine Optimization (SEO) team, that considers both intelligibility and discovery of the products. The template which we followed is: `Product Type + Model + Public + Occasion + Color + Material + Size`. This quite simple pipeline allowed us to obtain more accurate titles. The Table 2 shows a few examples of generated titles.

To measure the impact of these new titles in the whole marketplace, we used as extrinsic evaluation the current machine learning based categorization tool developed internally. We evaluated the automatic categorization of 3,000 original and processed

[5]`https://scikit-learn.org/stable/modules/generated/sklearn.cluster.AgglomerativeClustering.html`

[6]It solved common translation mistakes, as the translation of "dress party" to "vestido **partido**" ("party" as political party) and not to "vestido festa" ("party" as an occasion).

[7]It converts "para mulher" ("to woman") to "feminino" ("feminine").

Table 2. Examples of generated titles

'Original' Title	Camisa de mangas longas sexy lace na estação ferroviária Zhouzhou
Generated title	Camisa de manga longa[8]
'Original' Title	Gola ampla nostálgico duas mangas compridas confortável vestido meados de longa
Generated title	Vestido gola de manga comprida[9]
'Original' Title	2019 das mulheres sexy solto poncho top túnica assimétrica Blusa shirt club party mini vestido atacado
Generated title	Vestido poncho top túnica blusa camisa assimétrica solto feminino[10]

titles. The processed titles reach 99% of accurate categorization, while only 18% of the 'original' titles were classified in the right product category.

3.3. Error Analysis

In the present scenario, an error caused by the pipeline can be really serious, since it may produce a title that is not totally understandable by our customer and we assume that the processed title is always better than the original one. Therefore, having a proper error analysis of our pipeline is particularly relevant.

Analyzing the generated titles that still did not get the accurate categorization, we could note that the main flaw of our pipeline happens when our NER module detected more than one value for the `product_type` label. It is quite common that some sellers use synonyms or event slightly related words to describe their products aiming to get more relevance in search engines. However, if on one hand a title full of information increases the discoverability of the product, on the other hand, titles with too much information, sometimes redundant information, may cause confusion and misunderstanding about what the actual product is.

The last example from the Table 2 is an example of this issue in the fashion domain. This example contains six words ("poncho", "top", "túnica", "blusa", "shirt", "vestido") that are potential good `product_type` values. Even considering the image of the product shown in Figure 1 it is difficult to decide if the product is actually a poncho, a dress, or a shirt. For now, to avoid cases like that, we only generate new titles to products with less than three extracted values to the `product_type` NE label.

4. Conclusion and Future Work

This work showed how we address a common, but challenging, issue in marketplaces, the fact that product titles often have an awfully bad quality information. It impacts both on intelligibility and discovery products in such e-commerce platforms. We discussed the real scenario of *Americanas Mundo*, proposing a two steps solution for product titles generation. We first extract relevant attributes from product titles which were automatically

[8]In english, "Long sleeve shirt sexy lace in train station Zhouzhou" and "Long sleeve shirt".

[9]In english, "Wide collar nostalgic two long sleeves confortable dress mid long" and "Dress collar long sleeve".

[10]In english, "2019 women's sexy loose poncho top asymmetrical tunic blouse shirt club party mini dress wholesale" and "Dress poncho top tunic blouse shirt asymmetrical loose feminino".

translated and then we use a template-based strategy to generate more accurate and clean titles.

This approach may not be as scalable as we would like, but it was a possible and fast pipeline to test in the scenario we had. Our NER model achieved a F1 score of 0.71. For titles quality evaluation we use an extrinsic evaluation, based on automated product categorization. Our experiments showed that 99% processed titles fall under the right category, against 18% of accuracy in categorization of non-processed titles.

Our next step is to improve our NER model, both augmenting our training set and testing pre-trained models approaches, as [Souza et al. 2019]. As future work, we want to explore the attributes extraction approach proposed by [Huimin Xu 2020] and relating them directly to the users' queries, maybe following [Cheng et al. 2020].

References

Cheng, X., Bowden, M., Bhange, B. R., Goyal, P., Packer, T., and Javed, F. (2020). An end-to-end solution for named entity recognition in ecommerce search.

Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.

Huimin Xu, Wenting Wang, X. M. X. J. M. L. (2020). Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title. *Proceedings of the 57th Annual Meeting of the ACL.*

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Mahesh Joshi, Ethan Hart, M. V. J.-D. R. (2015). Distributed word representations improve ner for e-commerce. *Proceedings of NAACL-HLT 2015*.

Mathur, P., Ueffing, N., and Leusch, G. (2018). Multi-lingual neural title generation for e-commerce browse pages.

Najmi, A. (2019). Imputation of missing product information using deep learning: A use case on the amazon product catalogue. Master's thesis, TECHNISCHE UNIVERSITÄT MÜNCHEN.

Peng Yuan, Haoran Li, S. X. Y. W. X. H. and Zhou, B. (2020). On the faithfulness for e-commerce product summarization. *Proceedings of the 28th International Conference on Computational Linguistics*.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the ACL.*

Souza, F., Nogueira, R., and Lotufo, R. (2019). Portuguese named entity recognition using bert-crf. *arXiv preprint arXiv:1909.10649*.

Zhang, H., Hennig, L., Alt, C., Hu, C., Meng, Y., and Wang, C. (2020). Bootstrapping named entity recognition in e-commerce with positive unlabeled learning.

Zhangming Chan, Xiuying Chen, Y. W. J. L. Z. Z. K. G. D. Z. R. Y. (2019). Stick to facts: Towards fidelity-oriented product description generation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.