Padronização da Descrição de Produtos Comerciais utilizando NER

Laércio Lucchesi, Tatiana Escovedo, Marcos Kalinowski

Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro, RJ, Brasil

laercio.lucchesi@gmail.com, tatiana@inf.puc-rio.br, kalinowski@inf.puc-rio.br

Abstract. *Product descriptions of retail establishments are information used in market analysis, but normally, these descriptions are poorly structured, non-standardized, and vary a lot for the same product. This article proposes the use of natural language processing techniques, more specifically, a branch known as NER (Named Entity Recognition), to solve the automatic generation of standardized descriptions from retail product descriptions. As a result, the trained model proved to be adequate to extract the characteristic information of new products launched on the market and the consequent construction of their standardized descriptions.*

Resumo. *Descrições de produtos dos estabelecimentos varejistas são informações utilizadas em análises de mercado, mas normalmente, essas descrições são mal estruturadas, não-padronizadas e variam muito para um mesmo produto. Este artigo propõe o uso de técnicas de processamento de linguagem natural, mais especificamente, um ramo conhecido como NER (Named Entity Recognition), para resolver a geração automática de descrições padronizadas a partir de descrições de produtos de varejo. Como resultado, o modelo treinado mostrou-se adequado para extrair as informações características de novos produtos lançados no mercado e a consequente construção de suas descrições padronizadas.*

1. Introdução

Tipicamente, a descrição de produtos feita por varejistas é pouco estruturada e apresenta grandes variações para o mesmo produto. Por outro lado, produtos são identificados de maneira única pelo código EAN[1]. Uma boa descrição de produto deve ser capaz de identificar suas principais características, tais como marca, embalagem, qualidade, tipo, dimensões, peso e volume. A combinação das características do produto com outras informações como quantidade, preço, local da venda e identificação do consumidor possibilita gerar valiosas análises de mercado sobre fornecedores e consumidores. Assim, para se ter uma base de dados de produtos que seja útil para estas análises de mercado, é ideal que a descrição do produto seja padronizada. A geração automática da descrição de novos produtos pode reduzir mão de obra e tornar o processo mais eficiente e menos vulnerável a erros. O presente trabalho baseou-se em dados reais de uma empresa que faz análises de mercado de produtos de varejo.

O principal objetivo deste artigo é propor uma solução para o problema de gerar automaticamente descrições padronizadas para novos produtos a partir de uma base de

[1] *European Article Number*

dados de produtos existentes composta por três atributos: EAN, descrição do produto no estabelecimento e descrição padronizada do produto. Para tanto, propõe-se utilizar um ramo do processamento de linguagem natural (NLP[2], em inglês) conhecido como NER[3] para construir um modelo que permita localizar e classificar em categorias predefinidas as características do produto presentes em texto não-estruturado. Este artigo possui mais 6 seções. A Seção 2 apresenta as bases do uso do NER e o índice de similaridade entre sequências de texto. A Seção 3 discorre sobre trabalhos correlatos. As Seções 4 e 5 descrevem o problema e a solução proposta. A Seção 6 mostra os resultados obtidos e a Seção 7 tece considerações finais e indica trabalhos futuros.

2. Fundamentação Teórica

O objetivo desta seção é apresentar os conceitos básicos referentes à NER e à similaridade de palavras.

2.1. NER

O Processamento de Linguagem Natural é um ramo da inteligência artificial que visa possibilitar que os computadores entendam e processem a linguagem como as pessoas o fazem. A extração de informação é uma etapa importante de qualquer sistema de NLP, pois é capaz de extrair automaticamente informações estruturadas de documentos [Singh 2018]. NER é uma técnica de processamento de linguagem natural que identifica entidades de um texto, classificando-as em categorias predefinidas [Li, Sun, Han and Li 2018]. Neste artigo, um modelo será treinado para extrair informações específicas presentes nas descrições do varejo.

2.2. Similaridade entre Pares de Sequência de Texto

Uma parte da solução do problema descrito na Seção 5 deste artigo faz uso de uma medida de similaridade entre pares de sequência de texto. O objetivo desta medida é obter um índice real que varie entre 0 e 1, onde 1 significa que as duas sequências são idênticas e 0 que as duas sequências são completamente diferentes. O índice de similaridade é calculado de acordo com a equação *índice de similaridade = 2*M/T*, onde M é a soma dos tamanhos das subsequências coincidentes e T é o comprimento total de elementos em ambas as sequências [Ratcliff and Metzener 1988].

3. Trabalhos Relacionados

O presente artigo possui semelhanças com alguns trabalhos ligados à extração de informações para produtos na Internet. [Putthividhya and Hu 2011] apresentam uma aplicação NER para extrair atributos de listas de produtos do eBay. A extração destes atributos enfrentou três desafios: (a) perda de estrutura gramatical em curtas descrições com muitos substantivos, (b) presença de erros tipográficos, abreviações e siglas e (c) falta de informações contextuais. Apesar destas limitações, o modelo foi capaz de identificar 300 novas marcas com uma precisão de 90,3%. [Sidorov 2018] traz um bom exemplo de como o NER pode melhorar significativamente a qualidade dos resultados de pesquisa no varejo. Este trabalho apresenta uma implementação de NER para

[2] *Natural Language Processing*
[3] *Named Entity Recognition*

detecção de atributos na descrição de produtos da BestBuy. A descrição dos produtos na forma de texto simples, sem qualquer estruturação, também foi um desafio. O modelo treinado atingiu um F-score de 79,8%. [Bhange et al. 2020] apresentam um trabalho de aplicação de NER no site da Home Depot para identificar as entidades de marca e tipo de produto. O melhor modelo treinado atingiu um F-score de 93,3%. Um dos problemas relacionados a aplicação de NER na descrição de produtos nos sites de comércio eletrônico é a falta de conjuntos de dados anotados. Nesta situação, reconhecer novos tipos de entidades é um desafio para métodos NER supervisionados. Para resolver esse problema, [Zhang et al 2020] apresentam um algoritmo de aprendizado não rotulado. O modelo treinado atingiu um F-score de 72,0%.

A contribuição científica deste trabalho é a aplicação de NER na padronização automática das descrições de produtos do varejo brasileiro. Foi necessário atingir simultaneamente três características: (1) identificação de várias entidades por descrição, (2) elevados índices de métricas de desempenho e (3) ter um modelo treinado específico para o varejo brasileiro. Os trabalhos das referências [Sidorov 2018], [Putthividhya and Hu 2011], [Zhang et al 2020] e [Bhange et al 2020] falham em ter estas três características simultaneamente atendidas.

4. Descrição do Problema

A geração da descrição de novos produtos é realizada pela empresa de maneira não automática. A quantidade de novos produtos é da ordem de 1000 por semana. Para cada novo produto, é necessário agrupar as descrições dos estabelecimentos pelo seu EAN e montar manualmente a descrição padronizada. A automação desta atividade reduz a necessidade de mão de obra e torna o processo mais eficiente e menos vulnerável a erros. A base de dados a ser utilizada neste artigo contém os seguintes atributos: (1) EAN do produto, (2) descrição do produto no estabelecimento e (3) descrição padronizada do produto. Quando um novo produto é lançado no mercado, os estabelecimentos comerciais criam descrições não padronizadas.

5. Descrição da Solução

A solução do problema consiste em obter um modelo NER que consiga extrair as principais características do novo produto a partir das descrições feitas pelos estabelecimentos comerciais de varejo. As principais características do produto irão compor a descrição padronizada do produto. Este modelo será treinado a partir das descrições do estabelecimento da base de dados. A Tabela 1 mostra a identificação para a descrição de estabelecimento “ARR ARBOR T JOAO PREM TP1 PCT 5KG”.

Tabela 1. Identificação das entidades

Entidade	*String*	*Posição Início*	*Posição Fim*
Produto	ARR	0	3
Produto_X	ARBOR	4	9
Marca	T JOAO	10	16
Marca_X	PREM	17	21
Tipo	TP1	22	25
Embalagem	PCT	26	29
Peso	5KG	30	33

Este enriquecimento da base de dados deve ser feito para cada uma das instâncias da base de dados. Fazer este enriquecimento de maneira manual é muito custoso, pois a quantidade de instâncias por tipo de produto pode chegar a centenas de milhares. A descrição padronizada do produto existente servirá de fonte primária para o enriquecimento. A quantidade de instâncias das descrições padronizadas é da ordem de poucos milhares; então, é mais viável fazer o enriquecimento manual das descrições padronizadas e, usando a similaridade de texto, identificar as entidades nas descrições do estabelecimento de maneira automática. A solução de enriquecimento usou Microsoft Excel simplificando muito todo o processo.

A descrição padronizada de produto relacionada à Tabela 1 é *"ARROZ ARBORIO TIO JOAO PREMIUM TIPO 1 PACOTE 5KG"*. Usando o índice de similaridade conforme descrito na Subseção 2.2, identificam-se as entidades e suas posições inicial e final. A Tabela 2 relaciona as entidades da descrição do produto no estabelecimento e da descrição padronizada do produto aos seus respectivos índices de similaridade. Foram considerados similares os pares de texto cujos índices de similaridade fossem maiores do que 0,5.

Tabela 2. Exemplo de índices de similaridade

Entidade	***Estabelecimento***	***Padronizada***	***Similaridade***
Produto	ARR	ARROZ	0,75
Produto_X	ARBOR	ARBORIO	0,83
Marca	T JOAO	TIO JOAO	0,86
Marca_X	PREM	PREMIUM	0,73
Tipo	TP1	TIPO 1	0,67
Embalagem	PCT	PACOTE	0,67
Peso	5KG	5KG	1,00

O treinamento do modelo utilizou a versão 3.0 do spaCy[4], uma biblioteca para NLP em Python. No treinamento as predições do modelo são comparadas com as anotações de referência para estimar o gradiente da perda que é usado na retropropagação. O modelo de aprendizado para processamento de linguagem natural utilizado pelo spaCy possui quatro etapas: Incorporação, Codificação, Atenção e Predição. A **incorporação** transforma tokens em vetores que são muito mais adequados para o processamento e carregam informações semânticas [Mikolov et al 2013]. Uma vez que os vetores da etapa de incorporação dos tokens fora do contexto são obtidos, o spaCy inclui as palavras vizinhas. A **codificação** transforma uma sequência de vetores de tokens em uma matriz [Schuster and Paliwal 1997]. O passo de **atenção** reduz a matriz da etapa anterior para um vetor [Vaswani et al 2017] concentrando a informação mais relevante. A etapa final executa a **predição** do rótulo com base no vetor produzido na etapa de atenção usando uma rede neural totalmente conectada [Goldberg 2016].

O modelo é avaliado comparando-se suas saídas com os dados de teste. As quantidades de falsos positivos (FP), falsos negativos (FN) e verdadeiros positivos (VP) são usados para calcular precisão, *recall* e *F-score.* A **precisão** é a porcentagem dos resultados que são reconhecidos corretamente e é calculada de acordo com a equação ***precisão = VP/(VP+FP)***. ***Recall*** é a porcentagem do total de entidades corretamente reconhecidas e é calculado de acordo com a equação ***recall = VP/(VP+FN)***. ***F-score*** é

[4] Disponível em https://spacy.io.

uma outra medida de desempenho que combina a precisão e o *recall* e é calculado de acordo com a equação $F_{score} = 2*(precisão*recall)/(precisão+recall)$. O *F-score* é uma boa medida quando se busca um equilíbrio entre precisão e recall e foi utilizado como a principal métrica.

As bases de dados, modelos, códigos, configurações e ferramentas associadas estão disponíveis em *https://github.com/laerciolucchesi/ner-applied-to-retail*.

6. Resultados e Avaliação da Solução

Como já descrito na Seção 5, a base de dados original do produto arroz foi enriquecida para identificar as entidades que representam as características do produto. A base de dados utilizada possui aproximadamente 125.000 instâncias e, para fins de treinamento, foi dividida aleatoriamente (80% treino e 20% teste). A Tabela 3 mostra o resultado da avaliação do desempenho do modelo treinado para cada entidade e também para o desempenho geral.

Tabela 3. Métricas de desempenho

Entidade	*Precisão*	*Recall*	*F-score*
Produto	99,41%	99,72%	99,57%
Produto_X	91,03%	94,28%	92,63%
Marca	97,35%	97,67%	97,51%
Marca_X	87,16%	82,64%	84,84%
Tipo	95,83%	97,22%	96,52%
Embalagem	86,88%	85,82%	86,34%
Peso	98,60%	98,85%	98,72%
Geral	**97,18%**	**97,63%**	**97,40%**

Observa-se na Tabela 3 que os índices F-score variam numa faixa de 84% a 99%. Os atributos com os menores índices são a Embalagem e a Marca_X pois a quantidade de informação é menor se comparada aos demais atributos. Estes valores mais baixos são compensados pelo fato de que para se compor uma descrição padronizada, utilizam-se várias descrições de estabelecimento. Desta forma, podemos considerar que os índices de desempenho alcançados são satisfatórios.

Para fins ilustrativos, supõe-se o lançamento de uma nova marca chamada *NOVAMARCA*. Os varejistas o descreveriam tipicamente conforme a seguir: *"ARR PARB T 1 NOVAM PCT 1K", "ARRO NMARCA PARBO TIPO 1 PREM 1 KG", "AR NOVAMARC TIPO 1 1KG PROMOCAO", "AROZ NOVAMARCA T 1 PREMI PCT 1KL" e "ARROS PARBOLIZADO NOVMAR PREM T1 PACOTE 1 KL"*. Tal qual a aplicação descrita em [Putthividhya and Hu 2011], essas descrições apresentam perda de estrutura gramatical em curtas descrições com muitos substantivos, erros tipográficos, abreviações e siglas e falta de informações contextuais. Usando o modelo treinado, obtém-se a Tabela 4.

Tabela 4. Sumário de identificação de entidades

Entidade	*Identificações da entidade*
Produto	ARR, ARRO, AR, ARROZ, ARROS
Produto_X	PARB, PARBOLIZADO
Marca	NOVAM, NMARCA, NOVAMARCA, NOVAMARCA, NOVMAR
Marca_X	PREM, PREMI, PREM
Tipo	T 1, TIPO 1, TIPO 1, T1
Embalagem	PCT, PCT, PACOTE
Peso	1K, 1 KG, 1KG, 1KL, 1 KL

Usando-se a medição de similaridade da Subseção 2.2; constrói-se a descrição padronizada. O resultado encontrado para a descrição do exemplo é *"ARROZ PARBOILIZADO TIPO 1 NOVAMARCA PREMIUM PACOTE 1KG"*. Os resultados foram replicados com sucesso para outros produtos além do arroz.

7. Conclusão

Este artigo aplicou a técnica de NLP conhecida como NER para o problema de geração automática de descrição padronizada de produtos. Um modelo foi construído para classificar em categorias as características de produto presentes em texto não-estruturado, que compuseram a descrição padronizada. O modelo treinado atingiu um F-score de 97,4% e provou ser adequado para o objetivo de gerar descrições padronizadas. O processo descrito neste artigo, aplicado para o produto arroz, pode ser aplicado também para outros produtos, com a diferença de que cada tipo de produto terá seu próprio conjunto de entidades específicas. O uso de NER mostrou-se viável para esta aplicação. Possibilidades de trabalhos futuros: (a) treinamento de modelos com bases de dados com baixo percentual de anotação; (b) avaliação do uso de outras bibliotecas de NLP, tais como, NLTK, PyTorch, Flair e Gensim; e (c) utilização de web scraping para complementar informações de produtos que tenham descrições insuficientes.

Referências

Ratcliff, J. and Metzener, D. (1988) "Pattern Matching: The Gestalt Approach", Dr. Dobb's Journal, Issue 46, July.

Mikolov, T., Chen, K., Corrado, G. and Dean, J. (2013) "Efficient Estimation of Word Representations in Vector Space", Proceedings of Workshop at ICLR.

Li, J., Sun, A., Han, J. and Li, C. (2018) "A survey on deep learning for named entity recognition", arXiv preprint arXiv:1812.09449.

Schuster, M. and Paliwal, K. (1997) "Bidirectional recurrent neural networks", Signal Processing, IEEE Transactions on. 45. 2673 - 2681. 10.1109/78.650093.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. and Polosukhin, I. (2017) "Attention is all you need", preprint arXiv:1706.03762.

Goldberg, Y. (2016) "A Primer on Neural Network Models for Natural Language Processing", Journal of Artificial Intelligence Research 57 (2016) 345-420

Sidorov, M. (2018) "Attribute extraction from eCommerce product descriptions", Final CS229 project report. Stanford University

Putthividhya, D. and Hu, J. (2011) "Bootstrapped Named Entity Recognition for Product Attribute Extraction", Proceedings of 2011 Conference on Empirical Methods in Natural Language Processing, p.1557–1567, Edinburgh, July 27–31.

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", Proceedings of 3rd Workshop on e-Commerce and NLP (ECNLP 3), p.1–6, July 10.

Bhange, B. R., Chengy, X., Bowden, M., Goyaly, P., Packery, T. and Javedy, F. (2020) "Named Entity Recognition for E-Commerce Search Queries", March 8.

Singh, S. (2018) "Natural Language Processing for Information Extraction", arXiv e-prints, arXiv:1807.02383, July 1st.