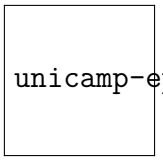

UNIVERSIDADE ESTADUAL DE CAMPINAS
Faculdade de Engenharia Elétrica e de Computação

Johannes Von Lochter

Resolução automática neuroinspirada para o fenômeno das palavras fora de vocabulário

Campinas
2018

Johannes Von Lochter

Resolução automática neuroinspirada para o fenômeno das palavras fora de vocabulário

Tese de Doutorado apresentada à Faculdade de Engenharia Elétrica e de Computação da Universidade Estadual de Campinas como parte dos requisitos exigidos para a obtenção do título de Doutor em Engenharia Elétrica, na Área de Automação.

Orientador: Prof. Dr. Akebo Yamakami
Coorientador: Prof. Dr. Tiago Agostinho de Almeida

ESTE EXEMPLAR CORRESPONDE À VERSÃO FINAL DA TESE DE DOUTORADO DEFENDIDA PELO ALUNO JOHANNES VON LOCHTER E ORIENTADA PELO PROF. DR. AKEBO YAMAKAMI.

Campinas
2018

Agência(s) de fomento e nº(s) de processo(s): CNPq, 141089/2013-0
ORCID: http://orcid.org/0000-0001-6687-8981

Ficha catalográfica
Universidade Estadual de Campinas
Biblioteca da Área de Engenharia e Arquitetura
Luciana Pietrosanto Milla - CRB 8/8129

Si38n

Silva, Renato Moraes, 1988-
Da Navalha de Occam a um método de categorização de textos simples, eficiente e robusto / Renato Moraes Silva. – Campinas, SP : [s.n.], 2017.

Orientador: Akebo Yamakami.
Coorientador: Tiago Agostinho de Almeida.
Tese (doutorado) – Universidade Estadual de Campinas, Faculdade de Engenharia Elétrica e de Computação.

1. Aprendizado de máquina. 2. Reconhecimento de padrões. 3. Comprimento mínimo de descrição (Teoria da Informação). I. Yamakami, Akebo,1947-. II. Almeida, Tiago Agostinho de. III. Universidade Estadual de Campinas. Faculdade de Engenharia Elétrica e de Computação. IV. Título.

Informações para Biblioteca Digital

Título em outro idioma: From the Occam's Razor to a simple, efficient and robust text categorization approach
Palavras-chave em inglês:
Machine learning
Pattern recognition
Minimum description length (Information Theory)
Área de concentração: Automação
Titulação: Doutor em Engenharia Elétrica
Banca examinadora:
Akebo Yamakami [Orientador]
Sahudy Montenegro González
João Paulo Papa
Levy Boccato
Romis Ribeiro de Faissol Attux
Data de defesa: 14-03-2017
Programa de Pós-Graduação: Engenharia Elétrica

COMISSÃO JULGADORA — TESE DE DOUTORADO

Candidato: Johannes Von Lochter **RA:** 181828
Data da Defesa: 20 de fevereiro de 2019

Título da Tese: "Resolução automática neuroinspirada para o fenômeno das palavras fora de vocabulário"

Prof. Dr. Akebo Yamakami (presidente, FEEC/UNICAMP)
Prof. Dr. Sahudy Montenegro González (DC/UFSCAR – Campus de Sorocaba)
Prof. Dr. João Paulo Papa (FC/UNESP/Bauru)
Prof. Dr. Levy Boccato (FEEC/UNICAMP)
Prof. Dr. Romis Ribeiro de Faissol Attux (FEEC/UNICAMP)

A ata de defesa, com as respectivas assinaturas dos membros da Comissão Julgadora, encontra-se no processo de vida acadêmica do aluno.

Dedico esta tese ao meus amados pais, David e Maria, e aos meus queridos irmãos, Leandro e Bruno.

Agradecimentos

Agradeço,

à Deus por abençoar minha vida;

ao meu orientador e co-orientador, Profs. Akebo Yamakami e Tiago A. Almeida, por compartilharem seu conhecimento, pela paciência, críticas construtivas, orientações e motivações que foram fundamentais para a realização deste trabalho e ajudaram a ampliar imensamente minha bagagem profissional;

aos meus pais David e Maria pela educação, por terem me ensinado o valor do trabalho e a correr atrás dos meus sonhos, por acreditarem mais em mim do que eu mesmo, por terem sempre me apoiado e incentivado e pelo amor e carinho que me deram em todas as etapas da minha vida;

aos meus irmãos Leandro e Bruno pelo apoio, pelo companheirismo e por serem sempre meus melhores amigos;

à minha namorada Simone pelo carinho, apoio, paciência e compreensão;

à prof.ª Tatiana A. Pazeto por ter me incentivado a entrar na pós-graduação e pelos seus ensinamentos e troca de experiências que foram fundamentais para minha trajetória no mestrado e doutorado;

ao Prof. Ricardo C. L. F. Oliveira pelo exemplo de humildade e pelas conversas valiosas durante os cafés diários que ocorreram por toda minha pós-graduação;

aos meus colegas do Departamento de Sistemas e Energia pela ótima convivência e troca de experiências;

aos professores Jacques Wainer, Fernando José Von Zuben, Romis Ribeiro de Faissol Attux, Ricardo R. Gudwin, Akebo Yamakami e Takaaki Ohishi pelas ótimas disciplinas oferecidas durante meu mestrado e doutorado;

à agência de fomento CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) pelo apoio financeiro concedido durante todo o período de doutorado;

à Faculdade de Engenharia Elétrica e de Computação e a Universidade Estadual de Campinas pela ótima estrutura que oferece aos estudantes e pesquisadores.

"It's common for people to believe that they're late to their field, that the tallest trees have already been felled. Personally, I think humanity's best works are still ahead, in every single field, and this is likely to remain true for thousands of years"

(François Chollet - @fchollet)

Resumo

Categorização de textos é um problema que tem recebido muita atenção nos últimos anos devido ao aumento expressivo no volume de informações textuais. O processo manual de categorizar documentos de texto é cansativo, tedioso, demorado e muitas vezes impraticável quando o volume de dados é muito grande. Portanto, existe uma grande demanda para que esse processo seja realizado de maneira automática através de métodos computacionais. Embora vários métodos já tenham sido propostos, muitos sofrem com o problema da maldição da dimensionalidade ou apresentam alto custo computacional, inviabilizando seu uso em cenários reais. Diante disso, esta tese apresenta um método de categorização de texto baseado no princípio da descrição mais simples, nomeado `MDLText`, que é eficiente, rápido, escalável e multiclasse. Ele possui aprendizado rápido, incremental e é suficientemente robusto para evitar o problema de sobreajustamento, o que é altamente desejável em problemas reais, dinâmicos, *online* e de grande porte. Experimentos realizados com bases de dados reais, grandes e públicas, seguidos por análises estatísticas dos resultados, indicam que o `MDLText` oferece um excelente balanceamento entre poder preditivo e custo computacional. Diante desses bons resultados, foi proposta uma generalização inicial do método para lidar também com problemas não-textuais, o que resultou em um método de classificação, nomeado `MDLClass`, que é simples, rápido e pode ser aplicado em problemas binários e multiclasse. A análise estatística dos resultados indicou que ele é equivalente à maioria dos métodos considerados o estado-da-arte em classificação.

Palavras-chaves: aprendizado de máquina; reconhecimento de padrões; categorização de texto; classificação; princípio da descrição mais simples.

Abstract

Text categorization has received attention in recent years because of the ever-increasing volume of text information. For a large number of documents, a manual classification is tiresome, tedious, time-consuming, and impractical, making computational methods attractive to deal with this task. The available methods that address this problem suffer from their computational burden and the curse of dimensionality, undermining their applicability in real scenarios. To overcome this limitation, we propose a simpler, faster, scalable and more efficient classification method based on the minimum description length principle, named `MDLText`. Its incremental and faster learning process makes it suitable to cope with data overfitting, which is desirable for real and large-scale problems. Experiments performed on real, public, and large-scale datasets followed by statistical analyses indicate that the `MDLText` provides an excellent trade-off between predictive capability and computational cost. Motivated by these results, we propose a generalized method, named `MDLClass`, to encompass non-textual problems. Similar to `MDLText`, this extension is simple and fast, and can also be applied to binary and multiclass classification problems. Statistical analyses show that `MDLClass` is equivalent to most of the state-of-the-art classification methods.

Keywords: machine learning; pattern recognition; text categorization; classification; minimum description length principle.

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

ABC	atributos baseados no conteúdo
ABL	atributos baseados nos *links*
ABLT	atributos baseados nos *links* transformados
χ^2	*chi-square* (chi-quadrado)
BIC	*Bayesian information criterion* (critério de informação Bayesiana)
B.NB	*naïve* Bayes Bernoulli (Bayes ingênuo Bernoulli)
CART	*classification and regression trees* (árvores de classificação e regressão)
CF	*confidence factors* (fatores de confidência)
CSV	*comma separated values* (valores separados por vírgulas)
DFS	*distinguishing feature selector* (seletor de atributos distintivos)
DT	*decision trees* (árvores de decisão)
GSS	Galavotti-Sebastiani-Simi
IG	*information gain* (ganho de informação)
IM	*instant messages* (mensagens instantâneas)
IQR	*interquartile range* (amplitude interquartil)
KNN	*k-nearest neighbours* (k-vizinhos mais próximo)
MAP	*maximum a posteriori* (máximo a *posteriori*)
MCC	Matthews *correlation coefficient* (coeficiente de correlação de Matthews)
M.NB	*naïve* Bayes multinomial (Bayes ingênuo multinomial)
MDL	*minimum description length* (descrição mais simples)
MML	*minimum message length* (mínima mensagem)
NGL	Ng-Goh-Low
NB	*naïve* Bayes (Bayes ingênuo)
OGD	*online gradient descent* (gradiente descendente *online*)
OHE	*one-hot encoding* (codificação de um *bit* por estado)
OR	*odds ratio* (razão de chances)
PDF	*portable document format* (formato de documento portátil)
RF	*random forest* (floresta aleatória)
ROMMA	*relaxed online maximum margin algorithm* (algoritmo de margem máxima *online* relaxada)
SMS	*short message service* (serviço de mensagens curtas)
SGD	*stochastic gradient descent* (gradiente descendente estocástico)
SPIM	*spam over instant messaging* (spam de mensagens instantâneas)
SVM	*support vector machines* (máquinas de vetores de suporte)
TF	*term frequency* (frequência do termo)

TF-IDF	*term frequency-inverse document frequency* (frequência do termo-frequência inversa dos documentos)
XML	*extensible markup language* (linguagem de marcação extensível)
YSC	YouTube Spam *Collection*

Lista de símbolos

$\mathcal{D}^\star$	conjunto de treinamento
$\lvert\mathcal{D}^\star\rvert$	quantidade de instâncias de treinamento
d_i	i-ésima instância. No contexto de categorização de texto, é o i-ésimo documento de texto
$c(d)$	classe do documento d
$\lvert\mathcal{C}\rvert$	quantidade de classes
c_j	j-ésima classe
$\mathfrak{mc}_{i,j}$	valor da i-ésima linha e da j-ésima coluna da matriz de confusão e que corresponde ao número de exemplos classificados como pertencentes à i-ésima classe, mas que na verdade pertencem à j-ésima classe.
vp_j	verdadeiros positivos relativos à j-ésima classe
fp_j	falsos positivos relativos à j-ésima classe
vn_j	verdadeiros negativos relativos à j-ésima classe
fn_j	falsos negativos relativos à j-ésima classe
$p_\star$	probabilidade total de concordância entre a classe verdadeira e a classe predita pelo classificador
$p_\circ$	probabilidade de concordância ao acaso entre a classe verdadeira e a classe predita pelo classificador
$\mathfrak{mc}_{i,\cdot}$	soma dos elementos da i-ésima linha da matriz de confusão
$\mathfrak{mc}_{\cdot,i}$	soma dos elementos da i-ésima coluna da matriz de confusão
$\mathcal{D}$	conjunto de dados
$\lvert\mathcal{D}\rvert$	quantidade de instâncias
$\mathcal{T}$	conjunto de atributos obtidos após aplicar OHE. No contexto de categorização de texto, é o conjunto de termos
t_i	i-ésimo atributo após aplicar OHE. No contexto de categorização de texto, é o i-ésimo termo
$\lvert\mathcal{T}\rvert$	quantidade de atributos obtidos após aplicar OHE. No contexto de categorização de texto, é a quantidade de termos
$w(t_i, d)$	peso do i-ésimo termo do documento d
$\hat{w}(t_i, d)$	peso do i-ésimo termo do documento d após a normalização L2
$\overline{TF}(t_i, d)$	$\log(1 + TF(t_i, d))$, isto é, a frequência do i-ésimo termo normalizada
$TF(t_i, d)$	frequência do i-ésimo termo
DF_{t_i}	número de documentos de treinamento que contém o i-ésimo termo
X	variável aleatória discreta
$\mathcal{X}$	alfabeto

x_i	i-ésimo elemento do alfabeto $\mathcal{X}$
$H(X)$	entropia de X
$p(x_i)$	probabilidade de x_i
$I(x_i)$	valor da informação de x_i
Z	sequência de símbolos
$\mathcal{C}(x_i)$	código correspondente à x_i
$\mathcal{A}$	alfabeto binário
$\mathcal{A}^*$	todas as sequências que podem ser formadas usando o alfabeto $\mathcal{A}$
$L_{\mathcal{C}}(x_i)$	tamanho de $\mathcal{C}(x_i)$
M_i	i-ésimo modelo
$\mathcal{C}$	conjunto de todas as possíveis classes
$L(d\|c_j)$	tamanho da descrição de d em relação à j-ésima classe
$\hat{S}(d, c_j)$	penalidade dada para a classe c_j
$\|d\|$	quantidade de termos no documento d
$L(t_i\|c_j)$	tamanho da descrição do i-ésimo termo em relação à j-ésima classe
$K(t_i)$	penalidade aplicada ao i-ésimo termo
n	variável que armazena a soma dos pesos de cada termo em $\mathcal{T}$ para cada classe em $\mathcal{C}$
$\hat{n}$	variável que armazena a soma dos pesos em n para cada classe em $\mathcal{C}$
$\|\Omega\|$	porção do tamanho de descrição para termos que nunca apareceram em $\mathcal{D}^\star$
$S(d, \overline{c}_j)$	similaridade de cosseno
$\overline{c}_j$	vetor protótipo da j-ésima classe
$\|\hat{\mathcal{D}}_{c_j}\|$	quantidade de documentos de treinamento pertencentes à j-ésima classe
$\hat{w}(:, d)$	todos os pesos normalizados dos termos do documento d
$\|\hat{\mathcal{D}}\|$	variável que armazena o número de documentos para cada classe em $\mathcal{C}$
μ	constante usada para evitar que o denominador seja 0
$F(t_i)$	pontuação de contribuição do i-ésimo termo
ϕ	frequência de cada termo em $\mathcal{T}$ para cada classe em $\mathcal{C}$
$\|\bar{d}\|$	quantidade média de termos por documento
τ	índice da classe mais frequente
$p(c_j\|t_i)$	probabilidade condicional da j-ésima classe, dada a presença do i-ésimo termo
$p(\bar{t}_i\|c_j)$	probabilidade condicional da ausência do i-ésimo termo, dada a j-ésima classe
$p(t_i\|\bar{c}_j)$	probabilidade condicional do i-ésimo termo
$p(\bar{t}_i\|\bar{c}_j)$	probabilidade condicional da ausência do i-ésimo termo
ψ	mediana de termos por documento
k	quantidade de modelos ou grupos avaliados na análise estatística
q	quantidade de observações consideradas na análise estatística

R_j	média dos *rankings* do j-ésimo modelo ou grupo avaliado na análise estatística
α	intervalo de confiança
$\mathcal{B}$	conjunto de atributos categóricos
$\|\mathcal{B}\|$	quantidade de atributos categóricos
b_i	i-ésimo atributo categórico
$\|b_i\|$	quantidade de possíveis valores do i-ésimo atributo categórico
$v_i(b_j)$	i-ésimo valor do j-ésimo atributo categórico
e	parâmetro que ajusta a velocidade de crescimento da função de penalidade K no método `MDLClass`

Sumário

1 Introdução

Nos primeiros 2 anos de vida, as crianças normalmente fazem avanços substanciais com o aprendizado da linguagem falada, incorporando as palavras em seu léxico, permitindo que se tornem capazes de fazer inferências sobre variações das palavras incorporadas e combinações das mesmas. Para as crianças, o léxico é o vocabulário aprendido durante sua vida, enquanto que o léxico de uma linguagem é todo o vocabulário da linguagem em si (BONVILLIAN *et al.*, 1983).

Por mais simples que uma palavras possa parecer, todas elas são profundamente importantes dentro de sua linguagem. Uma palavra pronunciada é uma complexa sequência de articulações e sons, a qual além da forma escrita, está conectada a um conjunto rico de atributos semânticos e propriedades sintáticas como os componentes morfológicos. Desde 1980, aproximadamente, pesquisas têm sido conduzidas para responder à pergunta, "como esse conjunto complexo de informações é aprendido?" (SAMUELSON; MCMURRAY, 2017).

Questionamentos também têm sido levantados a respeito do domínio da linguagem escrita na sociedade moderna e como essa dificuldade propaga da infância à vida adulta. De maneira geral, tem sido observado que falta competência de leitura para a execução dos direitos plenos de cidadania. Essas dificuldades são derivadas dos métodos de alfabetização e, principalmente, por não ser claro nem mesmo na Neurociência como que o cérebro aprender a ler e, por consequência, como produzir um processo melhor (MORAIS *et al.*, 2013).

O impacto de compreender melhor como o cérebro lida com a tarefa de aprendizagem com respeito às palavras em uma linguagem motivou diferentes frentes de pesquisa, mas ainda não há consenso de como isso é feito. O processamento de linguagem natural é uma subárea de Inteligência Artificial que desenvolve esforços para entender as relações de linguagem natural, como textos escritos e falados, bem como armazená-los, processá-los, e automatizar tarefas que apenas humanos seriam capazes de fazer sem o desenvolvimento dessa área de conhecimento. São exemplos dessas tarefas: análise de sentimento (LOCHTER *et al.*, 2016), reconhecimento de entidades (JUNG, 2012), verificação de notícias falsas (CARDOSO *et al.*, 2018) e verificação de mensagens de *spam* (ALMEIDA; YAMAKAMI, 2012).

A representação de texto adotada pelos sistemas computacionais por muitos anos teve aspecto distributivo, evidenciado principalmente na tradicional representação *bag of words*. A representação distributiva é caracterizada por um vetor para cada amostra de texto, onde cada posição do vetor está diretamente relacionada a uma entrada de um dicionário previamente coletado. Enquanto essa representação tem deficiências bem

conhecidas, tais como perda de localidade das unidades de texto e esparsidade na representação do vetor, ainda assim traz resultados satisfatórios para tarefas que não sofrem tanto com essas deficiências, como a verificação de mensagens de spam. No entanto, algumas tarefas têm o desempenho severamente prejudicado por conta dessas deficiências, como reconhecimento de entidade nomeada.

Para contornar as dificuldades associadas à representação distributiva foi proposta a representação distribuída, onde cada termo é representado por um vetor de tamanho fixo que captura características semânticas do texto (TSHITOYAN *et al.*, 2019), e as amostras são representadas por um vetor resultante de uma função de composição que assume como entrada os vetores referentes a cada termo presente na amostra. Embora a representação distribuída tenha endereçado as maiores dificuldades da representação distributiva, mesmo as representações mais evoluídas desta categoria ainda não endereçam de forma adequada o fenômeno das palavras desconhecidas (GARNEAU *et al.*, 2018).

As palavras desconhecidas para um sistema computacional, assim como para os humanos, são as palavras sem referência prévia, as quais não fazem parte do vocabulário, sendo denominadas palavras fora do vocabulário (do inglês, "*out of vocabulary* – oov"). Esse fenômeno afeta qualquer tarefa de processamento de linguagem natural e é mais comumente investigado para linguagem falada, no entanto esse trabalho investiga esse fenômeno na linguagem escrita.

Conforme observado através de evidências neste trabalho, o fenômeno das palavras fora do vocabulário traz impacto negativo para o processamento de linguagem natural em sistemas computacionais, pois uma palavra fora do vocabulário é ignorada durante o processamento da amostra, ignorando sua existência. De maneira similar, também é possível perceber os efeitos desse fenômeno até mesmo para os humanos. Admita como exemplo, o seguinte trecho retirado do clássico "Alice no País das Maravilhas" de Lewis Caroll, no qual foram removidos termos arbitrariamente e substituídos pela palavra "removido".

> Concordo inteiramente com [removido] - disse a Duquesa. - E a moral disso é: 'Seja o que você [removido] ser'. Ou se você [removido] isso dito de uma maneira mais simples: '[removido] se imagine como não sendo outra coisa do que aquilo que poderia parecer aos outros que aquilo que você foi ou poderia ter sido não fosse outra coisa do que o que você poderia ter sido parecia a eles ser outra coisa'.
>
> - Acho que eu poderia entender isso melhor - disse [removido] de maneira muito educada - se estivesse tudo escrito. Mas, desse jeito, eu [removido] consigo entender o que você quer dizer.

Para efeito de comparação, a seguir é dado o mesmo trecho sem as palavras removidas, a qual o leitor pode perceber como a compreensão é comprometida com a

ausência de alguns termos.

> Concordo inteiramente com **você** - disse a Duquesa. - E a moral disso é: 'Seja o que você **pareceria** ser'. Ou se você **preferir** isso dito de uma maneira mais simples: '**Nunca** se imagine como não sendo outra coisa do que aquilo que poderia parecer aos outros que aquilo que você foi ou poderia ter sido não fosse outra coisa do que o que você poderia ter sido parecia a eles ser outra coisa'.
>
> - Acho que eu poderia entender isso melhor - disse **Alice** de maneira muito educada - se estivesse tudo escrito. Mas, desse jeito, eu **não** consigo entender o que você quer dizer.

Nos trechos mostrados anteriormente procurou-se remover aleatoriamente um termo por sentença, o qual já é possível perceber o efeito negativo que esse fenômeno produz, prejudicando a semântica da amostra ao não ter-se todas as palavras disponíveis para leitura. Também foi investigada a remoção de mais termos e são apresentadas evidências no decorrer deste texto de que o efeito negativo se acentua à medida que mais termos são desconhecidos durante o processamento da amostra.

Para contornar esse fenômeno e procurar representações adequadas para os termos que não são conhecidos, esse trabalho investigou os efeitos das palavras de vocabulário utilizando duas abordagens baseadas nas hipóteses da (1) representação distribuída, onde termos similares estão similarmente próximos no espaço dimensional, e (2) na aquisição e inferência humana, a qual é cunhada como neuro-inspirada no restante deste trabalho. Enquanto a primeira não trouxe resultados satisfatórios, a abordagem neuro-inspirada é apresentada com evidências de que o processo traz melhorias ao tratar as palavras fora de vocabulário em tarefas de análise de sentimento e reconhecimento de entidade nomeada.

A organização do restante desta tese é descrita a seguir.

- Capítulo 2 - Representação de texto traz a fundamentação teórica sobre representação de texto de forma distributiva e distribuída, e responde a pergunta de pesquisa: "A representação de texto utilizando o sistema vetorial aprendido a partir de grandes corpos de texto se beneficia quando o corpo de texto tem características semelhantes ao domínio da aplicação?";
- Capítulo 3 - Expansão semântica aborda a técnica de expansão semântica para enriquecer amostras curtas de texto e responde a pergunta de pesquisa: "No cenário de textos curtos e ruidosos, a expansão semântica traz melhores resultados para a tarefa de classificação?"

- Capítulo 4 - Resolução automática por agrupamento de palavras desconhecidas trata de técnicas de agrupamento para resolver o fenômeno das palavras desconhecidas automaticamente, além de responder a pergunta de pesquisa: "O agrupamento de diferentes sentenças pode ser utilizado para encontrar um candidato substituto a uma frase com palavra(s) desconhecida(s)?";

- Capítulo 5 - Resolução automática neuro-inspirada de palavras desconhecidas aborda a combinação de algumas técnicas de inteligência artificial e propõe um fluxo para tratar de palavras desconhecidas, além de trazer resultados para responder a pergunta de pesquisa: "O processo neuro-inspirado de resolução de palavras desconhecidas apresenta resultados favoráveis em meio computacional?"; e

- Capítulo 6 traz as conclusões e trabalhos futuros referentes ao tema dessa tese e posiciona o leitor frente aos desafios e questionamentos apontados durante o texto.

2 Representação de texto

O processamento de linguagem natural (PLN) é a intersecção da linguística com a ciência da computação. Enquanto a linguística contempla vários subcampos, para este trabalho é especificamente importante conhecer a sintaxe, a morfologia e a semântica, dentre elas a semântica como mais relevante. A sintaxe está diretamente ligada à posição dos termos em uma amostra e na formação da estrutura das sentenças para que se faça sentido. Morfologia, por sua vez, diz respeito á estrutura da palavra, em vez da sentença, e estuda a composição das mesmas por radical, prefixos e sufixos, demonstrando a formação das palavras e como elas se relacionam a outras palavras de uma mesma linguagem. A semântica estuda o significado, especificamente a semântica léxica estuda o significado das palavras e a semântica composicional para a formação de amostras maiores que palavras, como sentenças e documentos (BENDER, 2013).

A linguística está diretamente associada com o processamento de linguagem natural e auxilia, entre muitas coisas, na representação computacional adequada para texto capaz de transmitir especificamente o significado de uma amostra, apoiada no entendimento sintático, morfológico e semântico. Um exemplo de tal necessidade de representação são os modelos de linguagem que são ferramentas capazes de condensar características de um texto a favor de predizer a próxima palavra dado uma sentença ou documento (CHOMSKY, 1957). Para seu correto funcionamento, um modelo de linguagem deve desenvolver a capacidade sintática de predizer palavras que façam sentido na estrutura da sentença, organizar uma representação compacta da estrutura morfológica das palavras e o resultado precisa preservar a semântica do contexto.

A representação de texto no meio computacional foi construída a partir dos conceitos emprestados da linguística, onde a linguagem por si é simbólica e discreta, os caracteres compõem palavras, as quais denotam objetos, conceitos, eventos, ações ou ideias. Nesse sentido, tanto os caracteres quanto as palavras da linguagem escrita são simbólicos e discretos, e existe uma relação morfológica explorada na estrutura da palavra que a associa a palavras semelhantes. No entanto, muitas vezes uma relação semântica não se apresenta na estrutura da palavra, mas em seu significado, como “bola” e “esporte”. Enquanto as características morfológicas são mais simples de serem capturadas e intuitivas, capturar a relação semântica é um desafio em aberto para a representação de texto (GOLDBERG; HIRST, 2017).

Em processamento de linguagem natural, as menores unidades como caracteres e palavras, são utilizadas para construir o vocabulário que permeará a representação de texto em uma determinada tarefa de máquina. Esse vocabulário pode ser utilizado de diferentes maneiras para representar unidades de texto maiores como expressões, sentenças

e documentos. Em algumas situações específicas, para resolver uma tarefa pode ser necessário incluir n-gramas no vocabulário, quando duas unidades de texto co-ocorrem com muita frequência, por exemplo "*New York*". De todo modo, não há um consenso de qual seria a melhor forma para utilizar um vocabulário com caracteres, palavras ou n-gramas para representar sentenças, documentos ou quaisquer unidades de texto maiores do que as já citadas.

Entre as técnicas de representação estudadas e desenvolvidas na literatura estão as representações distributivas, por grupo e distribuídas. O restante deste capítulo está organizado da seguinte forma.

- Na Seção 2.1 são discutidos os detalhes sobre a representação distributiva e como funciona o método de representação mais clássico conhecido como "*bag-of-words*";
- Na Seção 2.2 são discutidas as implementações de representações por grupo e é feito destaque à mais comum e conhecida como agrupamento de Brown;
- Na Seção 2.3 são tratados os métodos que utilizam as representações induzidas por redes neurais e que tratam o vocabulário como um conjunto de vetores por palavra, conhecidos como "*word embeddings*" ou "*word vectors*".

2.1 Representação distributiva

Na representação distributiva de texto em um sistema computacional cada entrada no vocabulário ocupa exatamente uma dimensão na representação de um documento. Uma amostra poderia ser representada como a presença ou ausência de cada termo do vocabulário, resultando em um vetor esparso do tamanho do próprio vocabulário. Essa definição também caracteriza o método *bag-of-words* (BOW).

A primeira definição para BOW está no trabalho de Zellig Harris, publicado em 1954, quando o autor questiona se a linguagem tem uma estrutura distributiva. Na elaboração da sua justificativa, a linguagem pode ser estruturada de forma distributiva devido às partes de uma linguagem ocorrerem em posições relativas a outros elementos, respeitando uma sintaxe e preservando a semântica através disso (HARRIS, 1954).

A quantidade de palavras, ou até unidades de texto menores, incorporadas a um vocabulário pode ser expressiva e levar a uma representação esparsa pela definição do método BOW. Isso acontece devido às amostras terem poucos eventos (palavras distintas) se comparadas à composição total do vocabulário, resultando em um vetor com mais 0 do que 1. Para melhor representar o funcionamento do método de representação e suas deficiências, o método BOW é explicado a seguir com exemplo de sua utilização.

No método BOW, a primeira etapa é construir o vocabulário que será usado como referência para representar as amostras de texto. As amostras de texto que são

utilizadas para construir o vocabulário também são conhecidas como corpo de texto ou *corpus*. A representação da amostra se dá por um vetor do tamanho ao vocabulário, onde cada índice do vetor está relacionado a um índice do vocabulário, e cada índice tem uma informação a respeito da entrada do vocabulário, como ocorrência (binário), frequência (contagem) ou alguma relação de proporção.

Para construir o vocabulário é necessário estabelecer qual a menor unidade de texto para a aplicação ou tarefa, o que modifica a interpretação do que é uma entrada no vocabulário. Se a menor unidade de texto for um caractere, as entradas no vocabulário serão as letras e outros símbolos (como números e pontuação) presente no *corpus*. Caso a menor unidade seja uma palavra, o vocabulário será o conjunto das palavras observadas no *corpus*. Quando a menor unidade são n-gramas, cada entrada do vocabulário terá n palavras, como por exemplo "*New York*" para $n = 2$, "*The Washington Post*" para $n = 3$ e "*United States of America*" para $n = 4$. Suponha que a construção do vocabulário se dará a partir de um *corpus* composto apenas pelas amostras a seguir.

(A) "*I bought a crappy gift.*"

(B) "*I purchased a good gift.*"

A construção do vocabulário também está condicionada a certas restrições como termos que aparecem uma quantidade mínima de vezes no *corpus* ou quantidade máxima de entradas que um vocabulário pode suportar. Para este exemplo, considere que nenhuma restrição foi imposta. O vocabulário observado com granularidade $n = 1$ para estas amostras é dado por {*I*, *bought*, *purchased*, *crappy*, *good*, *a*, *gift*}.

Após a construção do vocabulário, a representação de uma amostra é dada por um vetor ou o conjunto de amostras é representado no formato de matriz (Tabela 3), onde as linhas correspondem às amostras e as colunas correspondem às entradas do do vocabulário. Assumindo que a relação seja a presença de uma entrada do vocabulário a uma determinada amostra, o valor de cada atributo (coluna) representa a presença ou ausência.

Tabela 3 – Exemplo de representação com *bag-of-words*.

	I	*bought*	*purchased*	*crappy*	*good*	*a*	*gift*
(A)	1	1	0	1	0	1	1
(B)	1	0	1	0	1	1	1

Na segunda linha, a amostra B não contém a palavra 'bought", portanto a ausência dela é indicada com o valor 0 nesse atributo, o qual está relacionado ao índice (segundo elemento) do dicionário. Não é difícil extrapolar a partir desse exemplo que, a medida que o vocabulário cresce dada a variedade de palavras conhecidas, a quantidade de atributos indicando presença é muito menor do que atributos indicando ausência , fazendo com que a representação se torne esparsa.

Esse método de representação ainda é amplamente utilizado, apesar de ter severas dificuldades em gerar compreensão mais complexas da semântica das amostras de texto, inviabilizando tarefas como ... . Além de tornar a representação esparsa, esse método também sofre da perda de localidade dos termos, e dos fenômenos de polissemia e sinonimia.

2.2 Representação por grupo

2.3 Representação distribuída

2.3.1 Word2vec

2.3.2 FastText

2.3.3 LSTM

2.3.4 Transformer

https://arxiv.org/pdf/1904.06707.pdf

3 Palavras fora do vocabulário

Pode ser frustrante para as pessoas se depararem com uma palavra desconhecida em meio a um texto. A palavra desconhecida, ou termo, pode gerar confusão acerca do conteúdo transmitido, dificultando e até impossibilitando o entendimento do todo. Esse fenômeno é conhecido como palavra fora do vocabulário dentro da área de processamento de linguagem natural.

As palavras fora de vocabulário são aquelas que o modelo se depara na etapa de teste sem ter tido conhecimento delas na etapa de treino. Assim, o vocabulário construído na etapa de treino não terá observado essas "palavras fora de vocabulário", portanto não terá representação para elas também.

Do mesmo modo que uma palavra desconhecida na leitura de uma pessoa pode comprometer o entendimento do conteúdo do texto, as palavras fora de vocabulário podem comprometer a interpretação das máquinas nos modelos de processamento de linguagem natural (GOLDBERG; HIRST, 2017).

Historicamente, o fenômeno das palavras fora de vocabulário tem sido mais explorado pelos pesquisadores de processamento de linguagem natural nas áreas de reconhecimento de entidades nomeadas e etiquetamento de classe gramatical ("*part of speech tagger*" no inglês). Embora as palavras fora de vocabulário sejam ruins para a maioria das tarefas de processamento de linguagem natural, especificamente nessas duas os danos são maiores para o modelo aprendido. Isso se deve ao fato de que em ambas as tarefas o erro é propagado para os termos adjacentes aos termos desconhecidos pelo vocabulário.

Para o contexto das representações distribuídas, uma palavra fora de vocabulário é um termo que não tem representação vetorial conhecida no modelo treinado. Isso implica na incapacidade do modelo de utilizar a informação do termo na amostra para representação de sentenças ou documentos.

Algumas alternativas foram adotadas na literatura para tratar as palavras fora de vocabulário encontrando uma palavra substituta conhecida no vocabulário ou uma representação nova para o termo desconhecido. Nesse trabalho as abordagens propostas são divididas e discutidas em dois grupos: abordagem com substituição direta e abordagem com inferência neural.

3.1 Abordagem com substituição direta

Na abordagem com substituição direta, as palavras fora do vocabulário são substituídas por variantes similares considerando o grau morfológico de semelhança. Por exemplo, uma palavra grafada incorretamente pode não constar no vocabulário, mas sua

forma canônica sim. Também é comum encontrar na metodologia de alguns trabalhos e discussões em fóruns públicos na Internet a utilização de representações fixas e aleatórias para os termos fora de vocabulário.

Goldberg & Hirst (GOLDBERG; HIRST, 2017) sugerem associar todas palavras fora de vocabulário a uma representação aleatória. Para isso, seria necessário empregar uma etapa de pré-processamento na qual as palavras inexistentes no dicionário são substituídas por um termo único. Esse termo é normalmente identificado como "[UNK]", a abreviação de *unknown*, que significa "desconhecido" em tradução livre do inglês. Desse modo, uma representação aleatória seria gerada para o termo "[UNK]", e toda vez que uma palavra fora de vocabulário aparecesse no teste, essa representação seria utilizada em seu lugar.

Outra proposta descrita no mesmo trabalho (GOLDBERG; HIRST, 2017) é representar cada termo desconhecido por um vetor diferente de valores aleatórios. Assim, evitaria que todos termos desconhecidos fossem representados pelo mesmo vetor. Da mesma maneira, também seria possível representar os termos por grupos, explorando características morfológicas. Todo termo desconhecido que terminasse em "*ing*" poderia ser representado pelo termo "[UNK_ing]". Esse agrupamento de termos morfologicamente semelhantes passa a ser representada por um vetor único para o grupo, evitando os dois cenários anteriores: (1) cada termo desconhecido tem uma representação aleatória única e (2) uma única representação aleatória é utilizada para todos termos desconhecidos.

Alternativas à adoção de representação aleatória são a utilização de dicionários e estratégias de busca para encontrar palavras candidatas cuja semelhança possa ser mapeada e utilizada como critério de busca. Em (HABASH, 2008), o autor propõe algumas técnicas para resolver palavras fora de vocabulário na tarefa de tradução de máquina, resolvendo palavras grafadas incorretamente, nomes de entidades desconhecidos no dicionário e substantivos compostos. Na data de publicação deste trabalho, a representação distribuída também não era explorada pelas limitações de hardware. Por conta disso, o autor propõe soluções que atuam diretamente nas palavras e não em suas representações: (1) preenchimento de tabelas com substantivos compostos mais comum e recombinação deles com palavras raras; (2) as palavras grafadas incorretamente são verificadas considerando apenas um erro possível, considerando trocas mais comuns, em vez de comparar a palavra incorreta com todo o dicionário; (3) um dicionário léxico é utilizado para enriquecer o vocabulário obtido na fase de treino e (4) uma lista de nomes próprios mais comuns entre dois idiomas.

3.2 Abordagem com inferência neural

Na abordagem com inferência neural são discutidas as propostas que utilizam redes neurais para inferir a representação ou uma palavra substituta para a palavra fora

do vocabulário. Por exemplo, a utilização de sub-palavras do método FastText de compor representações para palavras fora do vocabulário é um tipo de inferência neural (BOJANOWSKI *et al.*, 2016).

A quebra das palavras em sub-palavras também é amplamente empregada na tradução de máquina, onde é difícil ter uma expectativa do vocabulário por conta da saída estar em outro idioma. No trabalho de (SENNRICH *et al.*, 2016) os autores discutem a compressão do vocabulário utilizando a representação por sub-palavras, técnica especialmente empregada na área de tradução de máquina que deve ter um vocabulário aberto.

Além da sub-palavra, a estrutura por caractere também pode ser utilizada, conforme proposto em (PINTER *et al.*, 2017). Neste trabalho é proposta MIMICK, uma rede neural com arquitetura LSTM bidirecional é utilizada para relacionar os caracteres com o significado de cada termo. A partir de uma palavra desconhecida, seus caracteres são inseridos no modelo, o qual se torna responsável por inferir uma representação para a palavra desconhecida. Essa rede trabalha como uma regressão múltipla, em vez de fornecer uma saída discreta, como a escolha de uma palavra do dicionário. A utilização da representação pode ser feita diretamente ou uma busca pelo termo mais próximo à representação sugerida pode ser executada no dicionário.

No trabalho de (GARNEAU *et al.*, 2018), uma extensão COMICK é sugerida, capaz de combinar o contexto da palavra fora de vocabulário com os caracteres que a compõe para fornecer uma representação. Uma rede LSTM bidirecional fica responsável por fazer a composição dos caracteres, enquanto uma camada densa mais adiante recebe a soma dos vetores do contexto à esquerda, à direita e da própria saída da LSTM referente aos caracteres. A proposta do autor teve melhor desempenho que a arquitetura MIMICK e a substituição das palavras fora de vocabulário por vetores aleatórios.

A proposta de (Won; Lee, 2018) é muito semelhante à COMICK, mas os autores chamam a arquitetura proposta de Context-Char e também é uma extensão da arquitetura MIMICK. Nesse trabalho, os autores calculam um vetor médio das palavras e utilizam esse vetor médio para multiplicar as palavras do contexto e anexam essa informação apenas na saída de cada direção da LSTM bidirecional que processa os caracteres da palavra fora de vocabulário.

Enquanto arquiteturas recorrentes são mais conhecidas nas aplicações de texto, também há trabalhos na literatura que sugerem a utilização de redes convolucionais para representar adequadamente palavras fora do vocabulário, como a proposta em (HU *et al.*, 2019). Nesse trabalho, os autores utilizam uma estrutura hierárquica com mecanismo de atenção para relacionar a importância do contexto e as características morfológicas da palavra fora de vocabulário. Apesar da arquitetura ser diferente, a estratégia e a fonte de informação são semelhantes às arquiteturas COMICK e Context-Char já discutidas.

Apesar do contexto ser uma fonte rica de informações para associar um termo

desconhecido, do ponto de vista humano, normalmente é empregado algum dicionário para descobrir com exatidão o significado de alguma palavra. Infelizmente, termos novos podem não estar presentes no dicionário, principalmente no cenário de textos curtos e ruidosos. No entanto, quando existe uma definição para um determinado termo, essa definição pode ser explorada para gerar uma representação vetorial. Em (HILL *et al.*, 2015) é sugerida a utilização de redes neurais recorrentes para processar cada palavra e sua definição encontrada em um dicionário para um determinado termo desconhecido da amostra. Os resultados encontrados, tanto qualitativo como quantitativo, indicam que através da representação das palavras isoladas da definição é possível encontrar representações de qualidades para termos desconhecidos que tenham correspondente em algum dicionário.

No trabalho de (Liu; Xv, 2019) os autores propõem a utilização de uma arquitetura Transformer combinada a uma LSTM para fazer a sumarização do texto, e utilizam o próprio mecanismo para resolver o problemas das palavras fora de vocabulário, ao sumarizar o texto para palavras conhecidas no vocabulário.

Uma publicação[1] no Medium de Ed Rushton também tem uma abordagem interessante para contornar as representações aprendidas para palavras grafadas incorretamente, a qual o autor chama de vetor geral de tradução. Utilizando a relação de distância das *word embeddings*, ele relaciona um espaço de palavras grafadas incorretamente com sua versão canônica sem utilizar um dicionário. Para tal tarefa, o autor escolhe uma palavra do dicionário grafada incorretamente e busca pelas dez palavras mais próximas por distância de cosseno. Em seguida, calcula a diferença da palavra do dicionário para todas as dez palavras mais próximas. Esse vetor resultante é o vetor geral de tradução. Para encontrar a versão canônica de outra qualquer palavra, basta somar o vetor geral de tradução ao vetor da palavra grafada incorretamente e selecionar a palavra mais próxima.

3.3 Resultados

Baseado nos três níveis de percepção cognitiva sobre a inferência de palavras desconhecidas, foram conduzidos experimentos analisando o desemepnho de uma técnica para cada nível. As bases de dados foram manipuladas para a inserção artificial de palavras desconhecidas de maneira controlada, mas as bases originais também foram avaliadas aplicando as técnicas citadas. No primeiro cenário, o objetivo é ter uma visão otimista a respeito da aplicabilidade dos métodos, enquanto o segundo cenário procura validar a utilização das técnicas combinadas na solução de problemas de classificação.

[1] https://blog.usejournal.com/a-simple-spell-checker-built-from-word-vectors

3.3.1 Nível 1 - Compreensão por grafia

3.3.2 Nível 2 - Compreensão por dicionário

3.3.3 Nível 3 - Compreensão por contexto

Referências

ALMEIDA, T. A.; YAMAKAMI, A. Facing the spammers: A very effective approach to avoid junk e-mails. *Expert Systems with Applications*, v. 39, n. 7, p. 6557 – 6561, 2012. ISSN 0957-4174. Disponível em: ⟨http://www.sciencedirect.com/science/article/pii/S0957417411017209⟩. Citado na página 18.

BENDER, E. M. *Linguistic Fundamentals for Natural Language Processing: 100 Essentials from Morphology and Syntax*. [S.l.]: Morgan Claypool Publishers, 2013. ISBN 1627050116. Citado na página 22.

BOJANOWSKI, P.; GRAVE, E.; JOULIN, A.; MIKOLOV, T. Enriching word vectors with subword information. *CoRR*, abs/1607.04606, 2016. Disponível em: ⟨http://arxiv.org/abs/1607.04606⟩. Citado na página 28.

BONVILLIAN, J. D.; ORLANSKY, M. D.; NOVACK, L. L. Developmental milestones: Sign language acquisition and motor development. *Child Development*, [Wiley, Society for Research in Child Development], v. 54, n. 6, p. 1435–1445, 1983. ISSN 00093920, 14678624. Citado na página 18.

CARDOSO, E. F.; SILVA, R. M.; ALMEIDA, T. A. Towards automatic filtering of fake reviews. *Neurocomputing*, v. 309, p. 106–116, 2018. Citado na página 18.

CHOMSKY, N. *Syntactic Structures*. The Hague: Mouton and Co., 1957. Citado na página 22.

GARNEAU, N.; LEBOEUF, J.-S.; LAMONTAGNE, L. Predicting and interpreting embeddings for out of vocabulary words in downstream tasks. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, 2018. p. 331–333. Disponível em: ⟨https://www.aclweb.org/anthology/W18-5439⟩. Citado 2 vezes nas páginas 19 e 28.

GOLDBERG, Y.; HIRST, G. *Neural Network Methods in Natural Language Processing*. [S.l.]: Morgan & Claypool Publishers, 2017. ISBN 1627052984, 9781627052986. Citado 3 vezes nas páginas 22, 26 e 27.

HABASH, N. Four techniques for online handling of out-of-vocabulary words in Arabic-English statistical machine translation. In: *Proceedings of ACL-08: HLT, Short Papers*. Columbus, Ohio: Association for Computational Linguistics, 2008. p. 57–60. Disponível em: ⟨https://www.aclweb.org/anthology/P08-2015⟩. Citado na página 27.

HARRIS, Z. Distributional structure. *Word*, v. 10, n. 23, p. 146–162, 1954. Citado na página 23.

HILL, F.; CHO, K.; KORHONEN, A.; BENGIO, Y. Learning to understand phrases by embedding the dictionary. *CoRR*, abs/1504.00548, 2015. Disponível em: ⟨http://arxiv.org/abs/1504.00548⟩. Citado na página 29.

HU, Z.; CHEN, T.; CHANG, K.; SUN, Y. Few-shot representation learning for out-of-vocabulary words. *CoRR*, abs/1907.00505, 2019. Disponível em: ⟨http://arxiv.org/abs/1907.00505⟩. Citado na página 28.

JUNG, J. J. Online named entity recognition method for microtexts in social networking services: A case study of twitter. *Expert Systems with Applications*, Pergamon Press, Inc., USA, v. 39, n. 9, p. 8066–8070, jul. 2012. ISSN 0957-4174. Disponível em: ⟨https://doi.org/10.1016/j.eswa.2012.01.136⟩. Citado na página 18.

Liu, X.; Xv, L. Abstract summarization based on the combination of transformer and lstm. In: *2019 International Conference on Intelligent Computing, Automation and Systems (ICICAS)*. [S.l.: s.n.], 2019. p. 923–927. Citado na página 29.

LOCHTER, J. V.; ZANETTI, R. F.; RELLER, D.; ALMEIDA, T. A. Short text opinion detection using ensemble of classifiers and semantic indexing. *Expert Systems with Applications*, Pergamon Press, Inc., USA, v. 62, n. C, p. 243–249, nov. 2016. ISSN 0957-4174. Citado na página 18.

MORAIS, J.; LEITE, I.; KOLINSKY, R. *Alfabetização no Século XXI: Como se aprende a ler e a escrever*. [S.l.]: Penso Editora, 2013. Citado na página 18.

PINTER, Y.; GUTHRIE, R.; EISENSTEIN, J. Mimicking word embeddings using subword RNNs. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, 2017. p. 102–112. Disponível em: ⟨https://www.aclweb.org/anthology/D17-1010⟩. Citado na página 28.

SAMUELSON, L. K.; MCMURRAY, B. What does it take to learn a word? *Wiley Interdisciplinary Reviews: Cognitive Science*, v. 8, n. 1-2, p. e1421, jan 2017. ISSN 1939-5078. Disponível em: ⟨https://onlinelibrary.wiley.com/doi/abs/10.1002/wcs.1421⟩. Citado na página 18.

SENNRICH, R.; HADDOW, B.; BIRCH, A. Neural machine translation of rare words with subword units. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, 2016. p. 1715–1725. Disponível em: ⟨https://www.aclweb.org/anthology/P16-1162⟩. Citado na página 28.

TSHITOYAN, V.; DAGDELEN, J.; WESTON, L.; DUNN, A. R.; RONG, Z.; KONONOVA, O.; PERSSON, K. A.; CEDER, G.; JAIN, A. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, v. 571, p. 95–98, 2019. Citado na página 19.

Won, M.; Lee, J. Embedding for out of vocabulary words considering contextual and morphosyntactic information. In: *2018 International Conference on Fuzzy Theory and Its Applications (iFUZZY)*. [S.l.: s.n.], 2018. p. 212–215. Citado na página 28.