NORMALIZAÇÃO TEXTUAL DE CONTEÚDO GERADO POR USUÁRIOS

Thales Bertaglia
Orientadora: Maria das Graças Volpe Nunes
Universidade de São Paulo/
Instituto de Ciências Matemáticas e de Computação

Roteiro

- **Definição**
- **Motivação**
- **Introdução**
- **Normalização Textual Clássica**
- **Normalização Textual de CGU**
- **Projeto**
- **Considerações Finais**
- **Referências**

Normalização Textual

- **Transformação** de palavras **fora do padrão** em um texto para palavras **normais** (Sproat *et al*, 2001).

Normalização Textual

- **Transformação** de palavras **fora do padrão** em um texto para palavras **normais**
  - Expansão de abreviações
  - Expansão de números
  - Correção de erros ortográficos
  - ...

Motivação

Perguntas respondidas Mostre-me outra »

Como poim contato no websaite?

olá gente me ajuda eu tem um websaite da impreza meshfriends ai eu fiss uma pagina de contato como eu poim contato na pagina pq pq eu trabalha emeventos eu tem saite so qui o saite eu fis del serto agora eu fis um websaite como poim contado na pagina homo do websaite cuando o vizitante chegar no websaite ele quer marcar o evento ele va lá no contado escreve o nome telefone imail etc como eu poim essas funcios no websaite meshfriends me ajuda gente meu imail

Motivação

Perguntas respondidas

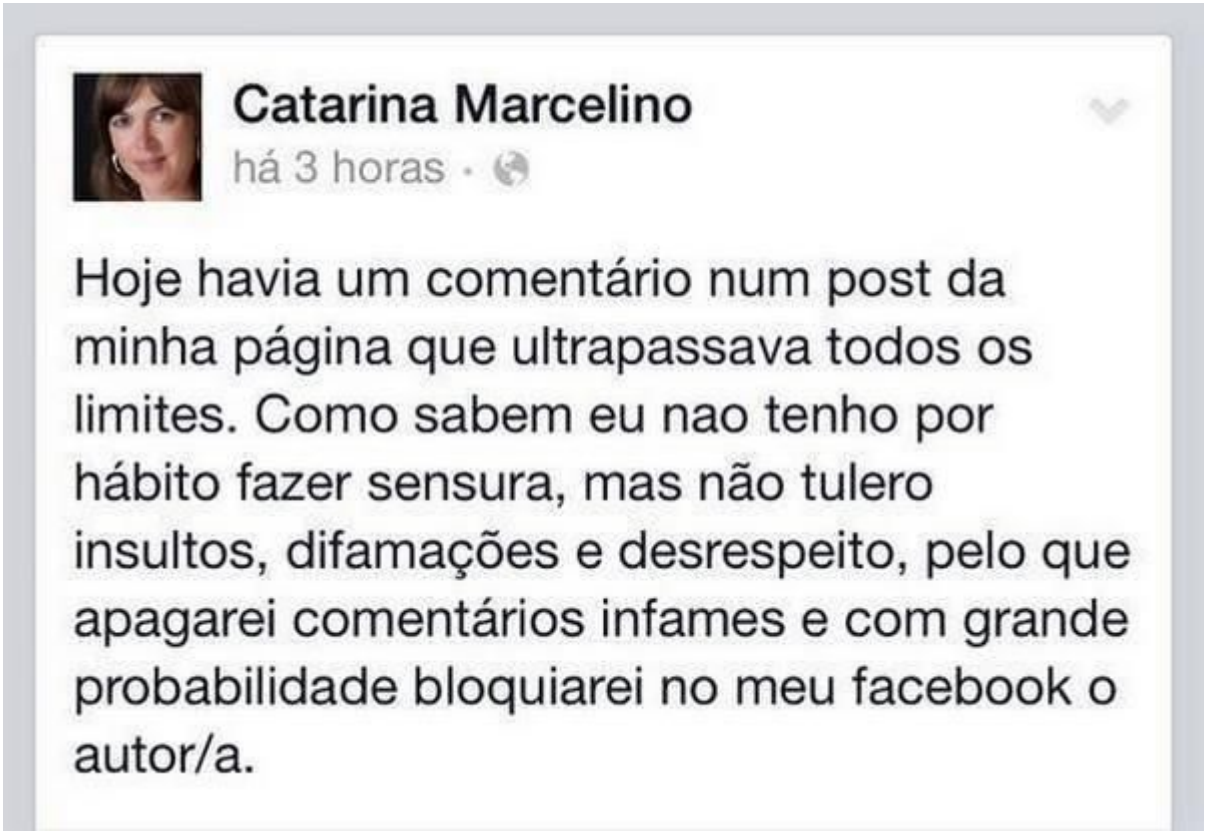
Mostre-me outra »

Como poim contato no websaite?

olá gente me ajuda eu tem um websaite da impreza meshfriends ai eu fiss uma pagina de contato como eu poim contato na pagina pq pq eu trabalha emeventos eu tem saite so qui o saite eu fis del serto agora eu fis um websaite como poim contado na pagina homo do websaite cuando o vizitante chegar no websaite ele quer marcar o evento ele va lá no contado escreve o nome telefone imail etc como eu poim essas funcios no websaite meshfriends me ajuda gente meu imail

Erros ortográficos, aglutinação, estrangeirismos

Motivação


Catarina Marcelino
há 3 horas
Hoje havia um comentário num post da minha página que ultrapassava todos os limites. Como sabem eu nao tenho por hábito fazer sensura, mas não tulero insultos, difamações e desrespeito, pelo que apagarei comentários infames e com grande probabilidade bloquiarei no meu facebook o autor/a.

Motivação

Em 2015, o PIB brasileiro já havia registrado uma retração de 3,8%. Com a previsão de um novo "encolhimento" do PIB neste ano, essa também será a primeira vez que o país registra dois anos seguidos de queda no nível de atividade da economia – a série histórica oficial, do IBGE, tem início em 1948.

Motivação

Próximas arquiteturas CUDA [editar | editar código-fonte]

A próxima geração de arquiteturas CUDA (codename: "Fermi") que vira por padrão na Geforce serie 400 (GTX 480 estará disponível a partir de 2010) a GPU é desenvolvida para suportar nativamente mais linguagens de programação como C++. É esperado que tenha um desempenho 8 vezes maior na performance de pontos flutuantes se comparada com a geração atual Nvidia Tesla. E terá a introdução de novas características como:

- Mais de 512 núcleos CUDA e 3 bilhoes de transistores
- NVIDIA Parallel DataCache technology
- NVIDIA GigaThread engine
- Suporte a Memoria ECC
- Suporte nativo ao Visual Studio

Motivação

- **Voltando à definição:** *Transformação* de palavras *fora do padrão* em um texto para palavras *normais*

- O que é padrão? E normal? Por que normalizar?

Motivação

- Exemplo de uma *review* extraída do Buscapé:

  comprei uma porcaria, como e que a lg bota uma *** dessa a venda.***!
  O que gostei: é lindo .Porem
  O que não gostei: a bateria e uma ***,nao vale nada,poderia durar pelo menos 3 dias sem carregar ,fazendo apenas o basico.Odiei!

- É bom o bastante para ferramentas de PLN?

Motivação

- Teste 1: tradução no Google Tradutor.

  bought crap like that and LG boot one *** this sale. ***!
  What I liked: it's beautiful .Porem
  What I did not like: the battery and a ***, not worth anything, could last at least 3 days without charge, making just basico.Odiei!

- Difícil compreender a tradução

Motivação

- Teste 2: *tagging* no LX Suite.

  a (DA) bateria (CN) e (CJ) uma (UM) ** (ADJ) *(CN) , (PNT) nao (CN) vale (V) nada (IND) , (PNT) poderia (V) durar (V) pelo (PREP+DA) menos (LADV3) 3 (DGT) dias (CN) sem (PREP) carregar (V) ,(?) fazendo (V) apenas (ADV) o (DA) basico.Odiei (CN) ! (PNT)

- Muitos erros devidos ao ruído

Motivação

- Vamos propor uma normalização:

  Comprei uma porcaria, como é que a LG bota uma dessa à venda.
  O que gostei: é lindo . Porém
  O que não gostei: a bateria é uma , não vale nada, poderia durar pelo menos três dias sem carregar , fazendo apenas o básico. Odiei!

- Correção ortográfica, remoção de caracteres, espaçamento etc.

Motivação

- Teste 3: tradução normalizada no Google Tradutor.

  Bought crap, how the LG boot this one for sale.
  What I liked: it's beautiful. however
  What I did not like: the battery is not worth anything, it could last at least three days without charge, doing just the basics. I Hated It!

- Ainda possui erros, mas lidou bem com a remoção de *

Motivação

- Teste 4: *tagging* normalizado no LX Suite.

  a (DA) bateria (CN) é (SER) uma (UM) , (PNT) não (ADV) vale (V) nada (IND) , (PNT) poderia (V) durar (V) pelo (PREP+DA) menos (LADV3) três (CARD) dias (CN) sem (PREP) carregar (V) , (PNT) fazendo (V) apenas (ADV) o (DA) basico (ADJ) . (PNT) Odiei (V) ! (PNT)

- Corrigiu vários erros e deixou de cometer alguns

Introdução

- Definir **o que** e **como** normalizar depende do domínio e da aplicação

- A normalização surgiu como uma solução local para tarefas de PLN

Introdução

- A dependência entre o normalizador e sua aplicação dificulta a consolidação de uma área de pesquisa

- Mas um normalizador não precisa efetivamente **transformar** palavras

Introdução

- Podemos redefinir normalização textual como **identificação** de palavras fora do padrão em um texto e **sugestão** de palavras normais para substituição

- Esse conceito permite um estudo mais amplo sobre normalização

Normalização Textual Clássica

Computer Speech and Language (2001) **15**, 287–333
doi:10.1006/csla.2001.0169
Available online at http://www.idealibrary.com on IDEAL®

Normalization of non-standard words

Richard Sproat,†∗ Alan W. Black,‡ Stanley Chen,§ Shankar Kumar,¶ Mari Ostendorf‖ and Christopher Richards∗∗

†AT&T Labs–Research, Florham Park, NJ, U.S.A., ‡Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, U.S.A., §IBM T. J. Watson Research Center, Yorktown Heights, NY, U.S.A., ¶Electrical and Computer Engineering Dept., Johns Hopkins University, Baltimore, MD, U.S.A., ‖Electrical Engineering Dept., University of Washington, Seattle, WA, U.S.A., ∗∗Department of Computer Science, Princeton University, Princeton, NJ, U.S.A.

Normalização Textual Clássica

- Grande influência para a consolidação da normalização textual como uma área de pesquisa

- Contribuições: uma taxonomia para palavras fora do padrão, córpus anotados, implementação de ferramentas. Tudo disponibilizado publicamente

Normalização Textual Clássica

- Baseado na análise de quatro córpus: *The North American News Text Corpus* (NANTC), *Classifieds*, pc110 e RFR.

TABLE II. Size of different corpora and number of detected non-standard word tokens

Corpus	NANTC	classifieds	pc110	RFR
total # tokens	4·3 m	415 k	264 k	209 k
# NSWs	377 k	180 k	72 k	46 k
% NSW	8·8	43·4	27·3	22·0

Normalização Textual Clássica

models

	EXPN	abbreviation	*adv, N.Y, mph, gov't*
alpha	LSEQ	letter sequence	*CIA, D.C, CDs*
	ASWD	read as word	*CAT*, proper names
	MSPL	misspelling	*geogaphy*
	NUM	number (cardinal)	*12, 45, 1/2, 0·6*
	NORD	number (ordinal)	*May 7, 3rd, Bill Gates III*
	NTEL	telephone (or part of)	*212 555-4523*
	NDIG	number as digits	*Room 101*
N	NIDE	identifier	*747, 386, I5, pc110, 3A*
U	NADDR	number as street address	*5000 Pennsylvania, 4523 Forbes*
M	NZIP	zip code or PO Box	*91020*
B	NTIME	a (compound) time	*3·20, 11:45*
E	NDATE	a (compound) date	*2/2/99, 14/03/87* (or US) *03/14/87*
R	NYER	year(s)	*1998, 80s, 1900s, 2003*
S	MONEY	money (US or other)	*$3·45, HK$300, Y20,000, $200K*
	BMONEY	money tr/m/billions	*$3·45 billion*
	PRCT	percentage	*75%, 3·4%*
	SPLT	mixed or "split"	*WS99, x220, 2-car* (see also SLNT and PUNC examples)
	SLNT	not spoken, word boundary	word boundary or emphasis character: *M.bath, KENT*RLTY, _really_*
M I S C	PUNC	not spoken, phrase boundary	non-standard punctuation: "***" in *$99,9K***Whites*, "..." in *DECIDE... Year*
	FNSP	funny spelling	*sllooooooww, sh*t*
	URL	url, pathname or email	*http://apj.co.uk, /usr/local, phj@tpt.com*
	NONE	should be ignored	ascii art, formatting junk

Normalização Textual Clássica

- Os quatro córpus foram anotados manualmente. Cada *token* identificado como fora do padrão é associado a uma *tag* de acordo com a taxonomia

- Um *token* é considerado fora do padrão quando não consta no léxico ou na lista de exceções

Normalização Textual Clássica

```
<P>
AAA INVESTMENTS SO SHORE,<W NSW="SPLT"><WS NSW="NUM"> 40</WS>
<WS NSW="EXPN" PRON="plus">+</WS></W> modern
<W NSW="EXPN" PRON="brick"> brk</W><W NSW="EXPN" PRON="apartments">apts</W>
on<W NSW="SPLT"><WS NSW="NUM"> 4</WS><WS NSW="EXPN" PRON="plus">+</WS></W> acres,
<W N SW="EXPN" PRON="individual"> indiv</W><W NSW="EXPN" PRON="heating"> ht.</W>
Income<W NSW= "MONEY"> $400K.</W> Ask<W NSW="MONEY"> $2,975,000</W>
<W NSW="SPLT"><WS NSW="EXPN" PRON="with"> w/</WS><WS NSW="MONEY">$750K</WS></W>
down. ROBERT<W NSW="LSEQ">L.</W> TENNEY REALTY<W NSW="PUNC"> (</W>
<W NSW="NTEL">617</W><W NSW="PUNC">)</W><W NSW="NTEL"> 472-0629472-0630</W>
</P>
```

Normalização Textual Clássica

- Além da taxonomia e da anotação de córpus, um sistema de normalização é proposto

- Os autores modelam o problema de expansão de palavras fora do padrão como a predição da sequência mais provável de palavras **w** dada uma sequência *de tokens* **o** que contém ruídos

Normalização Textual Clássica

- Ou seja, a palavra expandida **ŵ** é a que maximiza $P(w \mid o)$

- Na prática, a modelagem corresponde a *um noisy channel*

Normalização Textual Clássica

- Uma sequência de palavras **w** é gerada com $P(w)$ por uma fonte.

- **w** é transmitida por uma canal ruidoso, que a transforma na sequência *de* tokens **o** com probabilidades $P(t \mid w)$ e $P(o \mid t, w)$

Normalização Textual Clássica

- O sistema proposto inclui as seguintes tarefas:
  - Tokenização e identificação de palavras fora do padrão
  - Detecção e segmentação de *tokens* compostos (SPLT)
  - Predição da melhor sequência de *tags*
  - Expansão dos *tokens* para achar palavras candidatas
  - Busca pela melhor palavra para substituir

Normalização Textual Clássica

- Para avaliação do sistema, dois terços do córpus foi usado para treino e um terço para teste. Os resultados para a tarefa de segmentação constam abaixo:

	NANTC	classifieds	pc110	RFR
Recall	98·89	94·96	87·66	98·88
Precision	74·41	87·32	81·68	89·51
Split correct	92·54	85·99	74·11	89·54
Total correct	98·45	95·19	92·97	98·40

Normalização Textual Clássica

- Em geral, os resultados do sistema superaram o estado da arte na época
- Ferramentas disponibilizadas publicamente
- A normalização foi resolvida?

Normalização Textual Clássica

- Problemas: córpus bem comportados. Ruído previsível

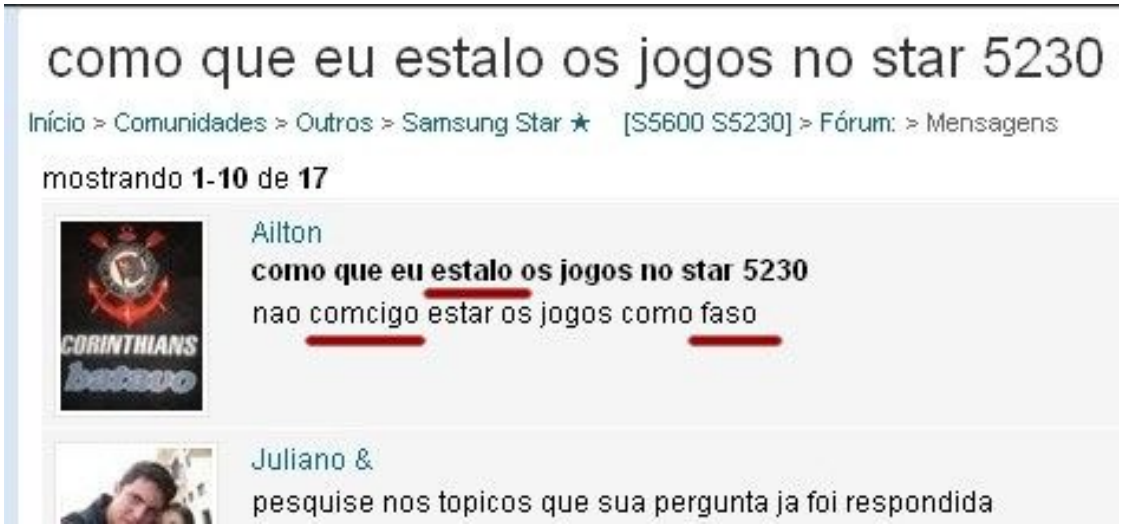
- Fim da década de 90, antes do auge da Internet para todos

Normalização Textual de CGU

- Explosão da Internet: qualquer um produz conteúdo: Web 2.0
- Nova língua, novos erros

Normalização Textual de CGU

- Discussão descentralizada, compartilhamento de opiniões livre



Normalização Textual de CGU

- Surgimento de linguagem nova e efêmera

Normalização Textual de CGU

- Mudança de conceitos. Que linguagem é considerada normal?

Normalização Textual de CGU

- Textos produzidos por escritores de todo nível cultural e social, alguns deles descompromissados com as regras da língua culta, e que refletem fortemente a oralidade, contendo, assim, inúmeros ruídos que devem ser tratados anteriormente ao uso de outras ferramentas (Avanço, 2015).

Normalização Textual de CGU

- Retorno à dependência de aplicação
- Muita variação entre domínios, portanto é difícil unificar o processo de normalização
- Para exemplificar, apresentaremos um normalizador de UGC focado em *reviews* de produtos

Normalização Textual de CGU

A Normalizer for UGC in Brazilian Portuguese

Magali Sanches Duran
NILC - Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
magali.duran@uol.com.br

Lucas Avanço
NILC - Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
avanco89@gmail.com

M. Graças Volpe Nunes
NILC- Center for Computational Linguistics
São Paulo University (USP)
São Carlos-SP, Brazil
gracan@icmc.usp.br

DURAN, M. S.; AVANÇO, L.; NUNES, M. G. V. A Normalizer for UGC in Brazilian Portuguese. Workshop on Noisy User/ generated Text. 2015.

Normalização Textual de CGU

- Focada na normalização de *reviews* do córpus do Buscapé (Hartmann *et al*, 2014) e do Mercado Livre

- O córpus é caracterizado por textos não muito longos, cada um deles variando bastante quanto ao nível de qualidade referente às normas da língua

Normalização Textual de CGU

- muita tequinologia é demais depois de adiqirir vc não vai querer outro

  O que gostei: exelente

  O que não gostei: nada declarar

- Quando decidi que era este o produto, eu já estava satisfeita com as suas funcionalidades e passei a comparar preço nas lojas. Quando recebi em casa, o produto me encantou ainda mais. É mais compacto do que eu imaginava, seus botões e imagens são realmente intuitivos e o manual é tão completo que é preciso conter a vontade de partir logo para a utilização.

  O que gostei: Barato e fácil de usar.

  O que não gostei: Nada.

Normalização Textual de CGU

- As palavras fora do padrão contidas no córpus foram identificadas por meio de um léxico e categorizadas de acordo com o tipo de ruído

- As categorias definidas são: X (erros ortográficos), SI (siglas), NP (nomes próprios), AB (abreviações), IN (internetês), ES (estrangeirismo), UM (unidades de medida) e SC (sem categoria)

Normalização Textual de CGU

- O exemplo abaixo, extraído de (Hartmann *et al*, 2014), ilustra a categorização:

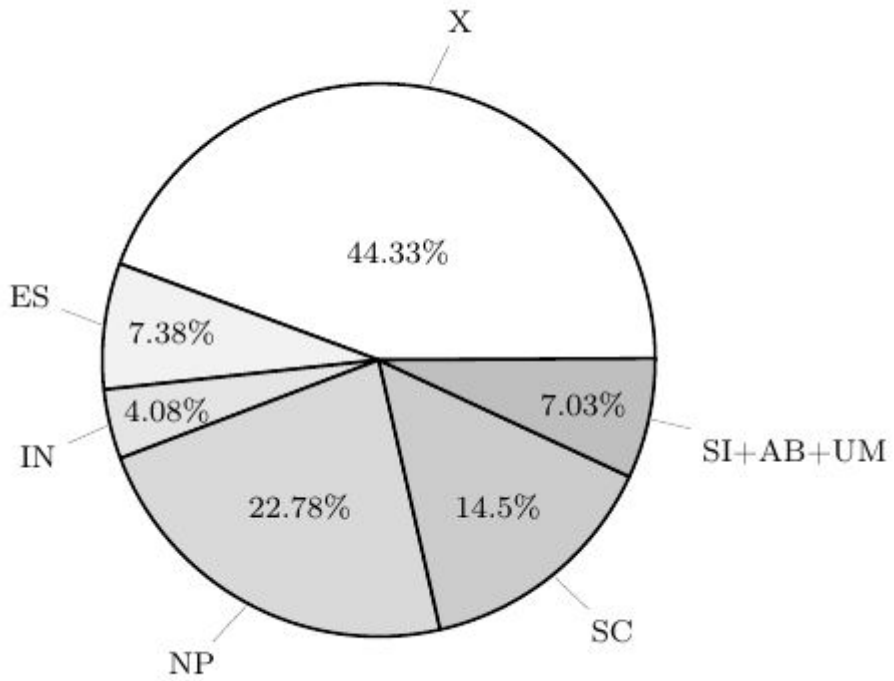
ela e **[X: é]** muito escura quando vc **[AB: você]** esta [X: está] deitado se vc **[AB: você]** tiver **[IN: estiver]** sentado ela e **[X: é]** boa mas deitada nao **[X: não]** e **[X: é]** muito nao **[X: não]**. A Samsung inova o mercado de tv's **[AB: televisões]** com uma grande obra de arte que se adequa **[X: adéqua]** à qualquer ambiente. Esta tv **[AB: televisão]** possui excelente imagem quando ligada a uma fonte de dvd **[SI: DVD]** com hdmi **[SI: HDMI]** e na tv **[AB: televisão]** a cabo (Digital). O som é perfeito quando é personalizado pelo usuário. Ou seja, MENU, SOUD **[ES: SOUND]**, EQUALIZAR E ENTER [ES].

Gostei dimais **[X: demais]** dessa câmera, comprei outra!Além de uma excelente e reconhecida marca, essa câmera tem um design **[ES]** super inovador e mtu **[IN: muito]** atraente...uma resolução mtu **[IN: muito]** boa e **[X: é]** td **[AB: tudo]** o que uma boa câmera SONY tem que ter!! Até hj **[AB: hoje]** nunca me deixou na mão... recomendo!!!

muto **[X: muito]** bom para manuziar **[X:manusear]** quando vc **[AB: você]** ta **[AB: está]** trabalhano **[X: trabalhando]** com este produto ,nao **[X: não]** tenho que recramar **[X: reclamar]** gostei mesmo parabéns. RECOMENDO O PRODUTO, FÁCIL DE USAR, ADOREI !

Normalização Textual de CGU

- A distribuição de erros por categorias pode ser vista abaixo (Avanço, 2015):



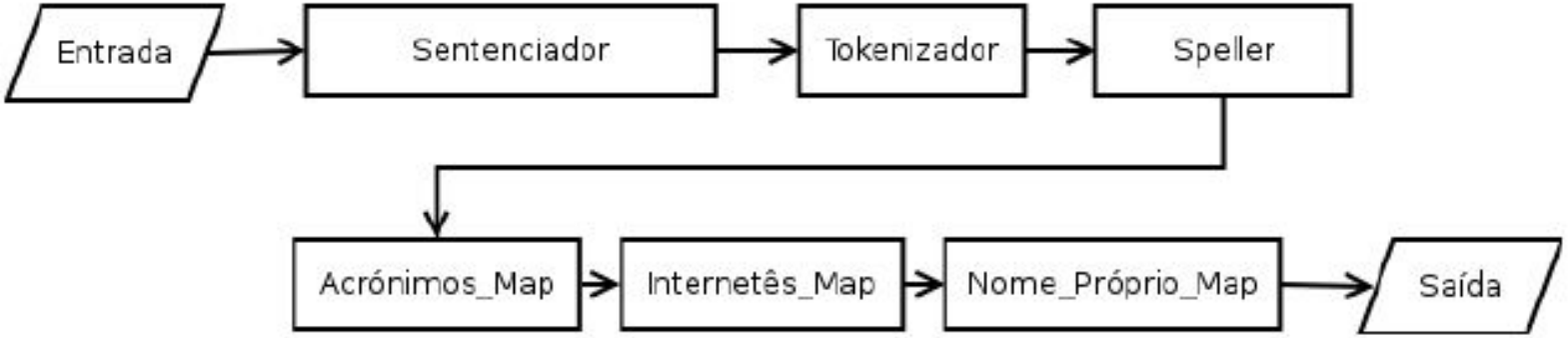
Normalização Textual de CGU

- Erros ortográficos são os mais comuns, seguido por erros de nomes próprios

- Os usuários não têm muita preocupação em utilizar a primeira letra maiúscula para se referir a entidades

Normalização Textual de CGU

- O sistema proposto, denominado UGCNormal, segue o seguinte fluxo (Avanço, 2015):



Normalização Textual de CGU

- O primeiro módulo consiste em aplicar o sentenciador proposto por (Condori e Pardo, 2015) para inserir pontos e, consequentemente, deixar iniciais maiúsculas

- O tokenizador é necessário para lidar com a linguagem própria da internet e é capaz de identificar emoticons, unidades de medida e URLs

Normalização Textual de CGU

- O módulo Speller é um *spellchecker* baseado em fonética para o português brasileiro

- No córpus analisado, se nota que a maioria dos erros gramaticais ocorre pela influência da língua falada

Normalização Textual de CGU

- Nesse contexto, o corretor proposto foi focado em corrigir erros motivados pela similaridade fonética

- O método obtém uma lista de candidatos à correção buscando palavras do léxico que estejam a uma distância de edição igual a 1 ou 2 da palavra errada

Normalização Textual de CGU

- Também são consideradas informações de similaridade fonética na geração de candidatos

- Vinte e uma regras fonéticas para português brasileiros são utilizadas no *spellchecker*

Normalização Textual de CGU

- Os demais módulos utilizam dicionários para realizar a substituição de palavras, sendo que os dicionários construídos possuem:
  - 21.699 nomes próprios
  - 432 formas de internetês
  - 248 palavras estrangeiras
  - 18 siglas para unidades de medida
  - 156 tipos de acrônimos genéricos

Normalização Textual de CGU

- Para exemplificar o funcionamento do normalizador, considere a sentença:

eleh eh mtt daora para vc mecher entao eu recomendu conprarem ese sangsung eleh otimo alem de ser barato eh dahora

Normalização Textual de CGU

- Saída após sentenciador:

Eleh eh mtt daora para vc mecher entao eu
recomendu conprarem ese sangsung
eleh otimo alem de ser barato eh dahora.

Normalização Textual de CGU

- Após tokenizador:

Eleh eh mtt daora para vc mecher entao eu
recomendu conprarem ese sangsung
eleh otimo alem de ser barato eh dahora .

Normalização Textual de CGU

- Após Speller:

Ele eh mtt daora para vc mexer então eu recomendo comprarem esse samsung ele ótimo além de ser barato eh dahora .

Normalização Textual de CGU

- Após Acrônimos_Map:

Ele eh mtt daora para vc mexer então eu recomendo comprarem esse samsung ele ótimo além de ser barato eh dahora .

Normalização Textual de CGU

- Após Internetês_Map:

Ele é muito da hora para você mexer então eu recomendo comprarem esse
samsung ele ótimo além de ser barato é da hora .

Normalização Textual de CGU

- Após Nome_Próprio_Map:

Ele é muito da hora para você mexer então eu recomendo comprarem esse Samsung ele ótimo além de ser barato é da hora .

Normalização Textual de CGU

- Avaliação do sistema:

Tipo de erro/ruído	Buscapé	Mercado Livre
Ortográfico Não-contextual	50 / 56 = 0,89	87 / 108 = 0,80
Ortográfico Contextual	15 / 39 = 0,38	24 / 76 = 0,31
Internetês	4 / 6 = 0,67	15 / 25 = 0,60
Caixa (Nomes próprios)	11 / 12 = 0,92	13 / 19 = 0,68
Caixa (início de sentença)	14 / 14 = 1,00	7 / 12 = 0,58
Palavras aglutinadas	0 / 2 = 0	2 / 6 = 0,33
Pontuação	44 / 47 = 0,94	58 / 79 = 0,73

Normalização Textual de CGU

- Funciona bem no domínio no qual foi delimitado
- Tem dificuldades para lidar com contexto (*real word errors*)
- O uso de dicionários limita demais a aplicação. Não escalável

Projeto

- Como projetar um normalizador mais genérico?
- Quais erros são mais frequentes e mais importantes?
- Corrigir é necessário?
- Quais técnicas para lidar com contexto?

Projeto

- *Deep learning* pode ser utilizado
- Poucos trabalhos na área de normalização
- Necessidade de grande quantia de dados

Projeto

- Uma ideia inicial: explorar as propriedades semânticas/contextuais de *word embeddings*

Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de SMS SPAM

Raul Freire Aguiar*, Ronaldo Cristiano Prati*
Centro de Matemática, Computação e Cognição(CMCC)
Universidade Federal do ABC (UFABC)
Santo André, SP, Brasil
{f.raul,ronaldo.prati}@ufabc.edu.br

AGUIAR, R. F; PRATI, R. C. Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de SMS SPAM. II Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). 2015.

Projeto

- Utiliza *embeddings* como ***feature*** para identificar *spam* em SMSs
- Embeddings substituem léxicos
- Necessário verificar se é aplicável

Projeto

- Algumas arquiteturas de *deep learning*:

Neural Language Correction with Character-Based Attention

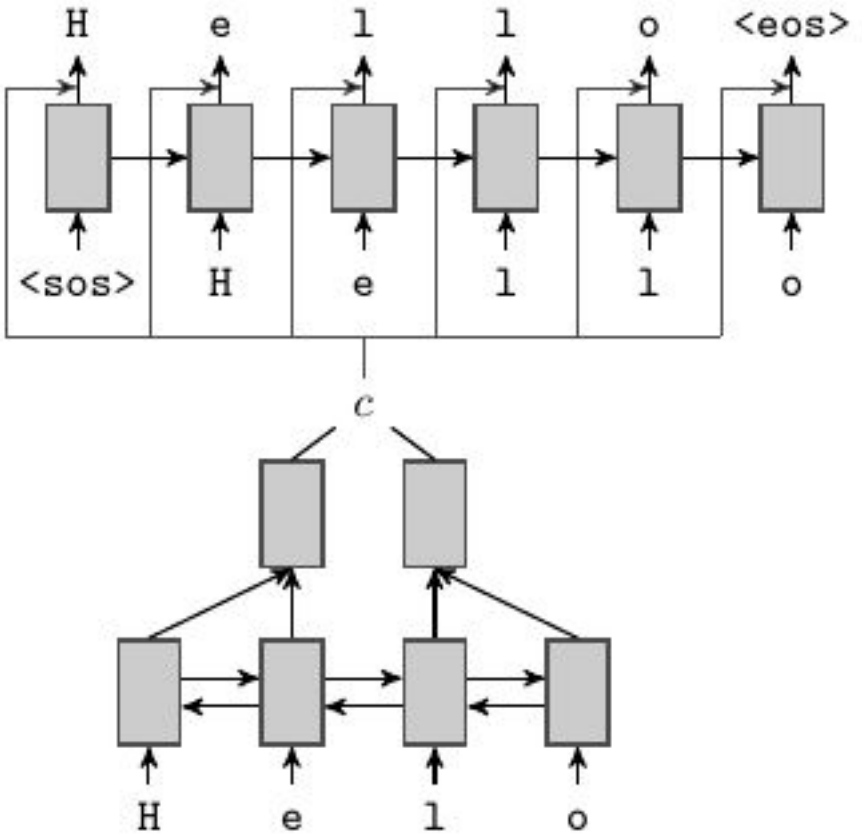
Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, Andrew Y. Ng
Computer Science Department, Stanford University
{zxie,avati,naveen67,ang}@cs.stanford.edu, jurafsky@stanford.edu

XIE, Z. et al. Neural Language Correction with Character/ Based Attention. 2016.

Projeto

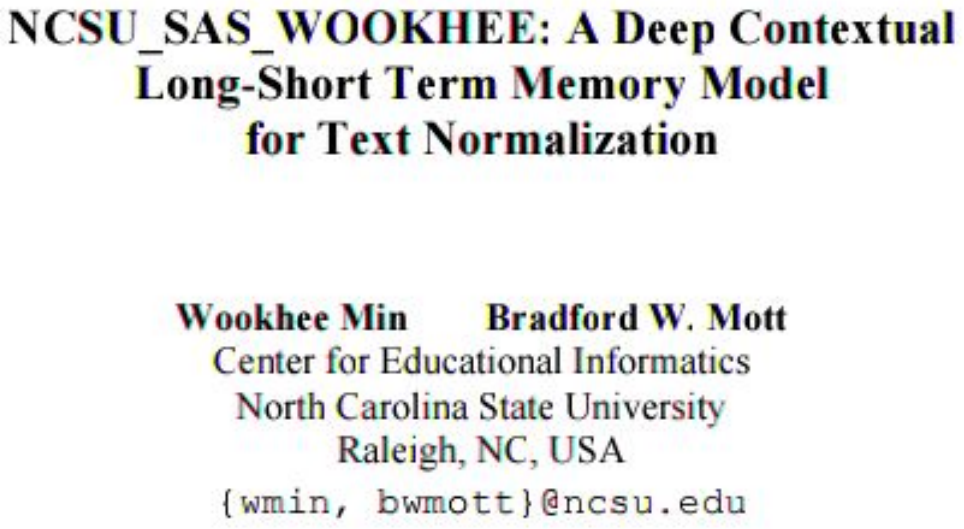
- Estado da arte em normalização
- Propõe uma arquitetura *encoder* e *decoder* que opera a nível de caractere
- Treinada em córpus totalizando mais de $550{,}000$ sentenças

Projeto



Projeto

- Arquiteturas propostas para o ‘desafio‘ de normalização lexical *de tweets* em inglês (https://noisy-text.github.io/norm-shared-task.html)



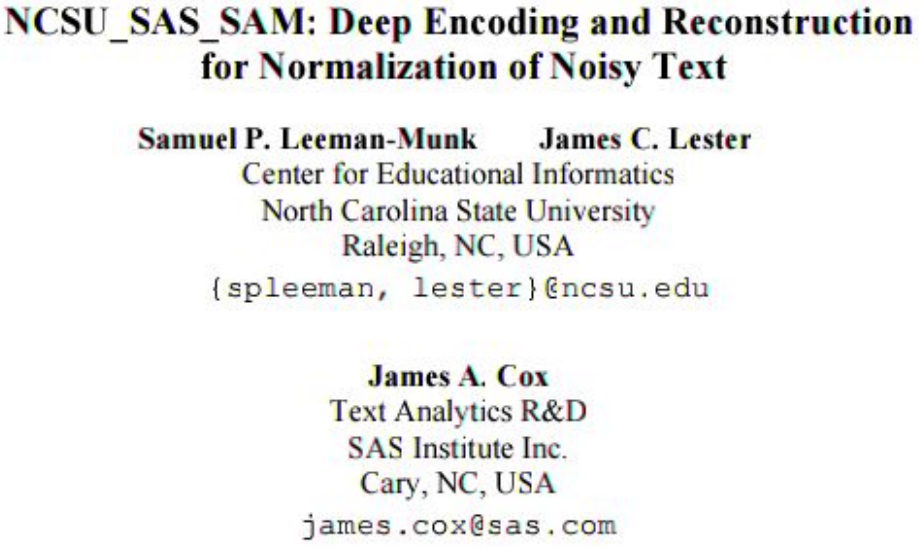
NCSU_SAS_WOOKHEE: A Deep Contextual Long-Short Term Memory Model for Text Normalization

Wookhee Min **Bradford W. Mott**
Center for Educational Informatics
North Carolina State University
Raleigh, NC, USA
{wmin, bwmott}@ncsu.edu

MIN, W.; MOTT, B.W. NCSU_SAS_WOOKHEE: A Deep Contextual Long/ Short Term Memory Model for Text Normalization. Workshop on Noisy User/ generated Text. 2015.

Projeto

- Arquiteturas propostas para o ‘desafio‘ de normalização lexical *de tweets* em inglês (https://noisy-text.github.io/norm-shared-task.html)



NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text

Samuel P. Leeman-Munk **James C. Lester**
Center for Educational Informatics
North Carolina State University
Raleigh, NC, USA
{spleeman, lester}@ncsu.edu

James A. Cox
Text Analytics R&D
SAS Institute Inc.
Cary, NC, USA
james.cox@sas.com

LEEMAN MUNK, S.P.; LESTER, J.C. NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text. Workshop on Noisy User/ generated Text. 2015.

Considerações Finais

- Muitos trabalhos desenvolvidos para tarefas bem específicas
- Foco em redes sociais (principalmente Twitter)
- Ausência de unificação e técnicas consolidadas na área

Referências

- SPROAT, R. et al. Normalization of nonstandard words. Journal of Computer Speech & Language. 2001.
- AVANÇO, L. Sobre normalização e classificação de polaridade de textos opinativos na web. Universidade de São Paulo. Dissertação de Mestrado. 2015.
- DURAN, M. S.; AVANÇO, L.; NUNES, M. G. V. A Normalizer for UGC in Brazilian Portuguese. Workshop on Noisy User/ generated Text. 2015.
- HARTMANN, N. et al. A Large Corpus of Product Reviews in Portuguese: Tackling Out Of Vocabulary Words. International Conference on Language Resources and Evaluation. 2014.
- CONDORI, R. E. L.; PARDO, T. A. S. Experiments on Sentence Boundary Detection in User Generated Web Content. 16th International Conference on Intelligent Text Processing and Computational Linguistics. 2015.

Referências

- AGUIAR, R. F; PRATI, R. C. Incorporação de representação vetorial distribuída de palavras e parágrafos na classificação de SMS SPAM. II Encontro Nacional de Inteligência Artificial e Computacional (ENIAC). 2015.
- XIE, Z. et al. Neural Language Correction with Character/ Based Attention. 2016.
- LEEMAN MUNK, S.P.; LESTER, J.C. NCSU_SAS_SAM: Deep Encoding and Reconstruction for Normalization of Noisy Text. Workshop on Noisy User/ generated Text. 2015.
- MIN, W.; MOTT, B.W. NCSU_SAS_WOOKHEE: A Deep Contextual Long/ Short Term Memory Model for Text Normalization. Workshop on Noisy User/ generated Text. 2015.