Existem domínios melhores que outros em Cross-Domain Sentiment Analysis?

ou, podemos prever o sentimento de avaliações de computadores a partir de avaliações de artigos para bebês?

Cross-domain sentiment analysis: o que é e por que importa

No mundo real a tarefa de análise de sentimento não é tão trivial pela dificuldade de acesso a datasets anotados. Os modelos têm que ser treinados em domínio diferente do que se quer predizer. Os estudos na área tentam reduzir a queda de desempenho nessas situações, que se dá pela diferentes semânticas e vocabulários em cada domínio.

Metodologia: dataset, pré-processamento e modelo

Foi utilizado o dataset Amazon Multidomain que apresenta avaliações de produtos de diversos diferentes domínios, dos quais foram selecionados 5. Pré processamento: remoção de stop words, lemmatização, parsing de datas e números e minimum document frequency = 5. O modelo utilizado foi uma **Convolutional Neural Network com MaxPooling.**

BASELINE

Melhor combinação fonte → alvo

81,80%
Grocery → Apparel

79,09%
Grocery → Baby

86,88%
Grocery → Comp&Games

83,57%
Comp&Games → Grocery

80,80%
Comp&Games→ Sports

Acurácia média por domínio fonte

Fonte	Acurácia média
Apparel	*77,22 +- 4,09*
Baby	***71,69 +- 5,74***
Comp & games	***81,61 +- 4,35***
Grocery	*77,95 +- 4,09*
Sports	*78,40 +- 4,40*

Acurácia média por domínio alvo

Alvo	Acurácia média	Diferença para o baseline (%)
Apparel	*76,71 +- 5,75*	*6,22*
Baby	***73,33 +- 3,26***	*7,28*
Comp & games	*78,48 +- 3,88*	***9,66***
Grocery	***81,60 +- 2,68***	***2,35***
Sports	*75,00 +- 4,26*	*7,17*

Comp&Games apresentou o melhor poder preditivo médio e **Baby**, o pior.

Baby é o domínio de mais difícil predição e **Grocery**, o mais fácil.

EXPERIMENTO 1

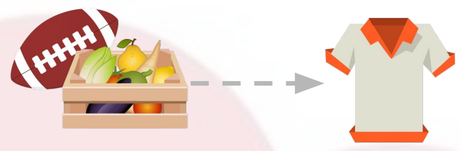
Fonte: combinação de todos os domínios, com exceção do domínio alvo

Alvo	Acurácia	Dif. p/ baseline (%)
Apparel	78,42	*-4,13*
Baby	67,80	*-14,27*
Comp &games	74,01	*-14,81*
Grocery	80,04	*-4,22*
Sports	77,82	*-3,69*

Em todos os casos, a combinação de domínios fonte **não melhorou** o baseline

EXPERIMENTO 2

Melhorar a pior dupla <fonte - alvo> por meio da adição de um domínio-fonte complementar (domínio baseline)



Fonte-Alvo	Acc.	Com fonte complementar	
		Acc.	Dif. para o baseline(%)
Sports - Apparel	72,66	78,69	*-2,61*
Apparel - Baby	66,52	74,36	*-9,09*
Baby - Comp &games	77,56	72,97	*-7,73*
Apparel - Grocery	74,69	74,96	*-8,36*
Baby - Sports	71,68	74,15	*-6,24*

O resultado da dupla foi **melhorado em três dos cinco casos**, entretanto, em nenhum caso a acurácia foi superior ao uso do melhor domínio-fonte.

Conclusão

- Cross-domain sentiment analysis é uma tarefa nada trivial; para resultados mais aceitáveis precisa de pelo menos uma amostra, anotada ou não, do domínio alvo já que a combinação de domínios-fonte não traz melhoras no baseline;
- A noção de "domínios parecidos" é enganosa e necessita de formalização para nortear decisões;
- É possível identificar domínios com mais poder preditivo, bem como domínio mais sensíveis à variação de domínio fonte.

Universidade Federal de Minas Gerais
DEpartamento de Ciência da Computação
Processamento de LInguagem Natural
Valéria Pereira de Souza 2016/118541
valeriaps@ufmg.br

UFMG
UNIVERSIDADE FEDERAL DE MINAS GERAIS