Primeiros experimentos com o SICK-BR

Bruno Ferrari Guide

Orientador: Marcos Lopes
Universidade de São Paulo

bruno.guide@usp.br

3 de dezembro de 2018

Tópicos

Introdução

Introdução

- Trabalho desenvolvido majoritariamente com o grupo responsável pela criação do SICK-BR
- Apresentação 2 em 1:
  - Primeiras rodadas de testes com modelos simples diversos usando esse corpus
  - Replicar os testes conduzidos por Fonseca, 2018 em um corpus diferente.

O Corpus

SICK-BR

pair_ID	sentence_A	sentence_B	entailment_label	relatedness_score	entailment_AB	entailment_BA	sentence_A_original	sentence_B_c	sentence_	sentence_	SemEval_set
1	Um grupo de crianças está brincanc	Um grupo de meninos em um qu	NEUTRAL	4.5	A_neutral_B	B_neutral_A	A group of kids is play	A group of bo	FLICKR	FLICKR	TRAIN
2	Um grupo de crianças está brincanc	Um grupo de crianças está brinc	NEUTRAL	3.2	A_contradicts_B	B_neutral_A	A group of children is	A group of kid	FLICKR	FLICKR	TRAIN
3	Os meninos jovens estão brincando	As crianças estão brincando ao a	ENTAILMENT	4.7	A_entails_B	B_entails_A	The young boys are pl	The kids are p	FLICKR	FLICKR	TRAIN
4	Os meninos jovens estão brincando	Não tem nenhum menino brinca	CONTRADICTION	3.6	A_contradicts_B	B_contradicts_A	The young boys are pl	There is no bo	FLICKR	FLICKR	TRIAL
5	As crianças estão brincando ao ar li	Um grupo de crianças está brinc	NEUTRAL	3.4	A_neutral_B	B_neutral_A	The kids are playing o	A group of kid	FLICKR	FLICKR	TRAIN
6	Não tem nenhum menino brincand	Um grupo de crianças está brinc	NEUTRAL	3.3	A_neutral_B	B_neutral_A	There is no boy playin	A group of kid	FLICKR	FLICKR	TEST
7	Um grupo de meninos em um quint	Os meninos jovens estão brincan	NEUTRAL	3.7	A_neutral_B	B_neutral_A	A group of boys in a y	The young bo	FLICKR	FLICKR	TEST
8	Um grupo de crianças está brincanc	Os meninos jovens estão brincan	NEUTRAL	3	A_neutral_B	B_contradicts_A	A group of children is	The young bo	FLICKR	FLICKR	TEST
9	Os meninos jovens estão brincando	Um grupo de crianças está brinc	NEUTRAL	3.7	A_neutral_B	B_neutral_A	The young boys are pl	A group of kid	FLICKR	FLICKR	TRAIN
10	Um cachorro castanho está atacan	Um cachorro castanho está atac	ENTAILMENT	4.9	A_entails_B	B_neutral_A	A brown dog is attack	A brown dog i	FLICKR	FLICKR	TEST
11	Um cachorro castanho está atacan	Um cachorro castanho está ajud	NEUTRAL	3.665	A_neutral_B	B_neutral_A	A brown dog is attack	A brown dog i	FLICKR	FLICKR	TEST
12	Dois cachorros estão lutando	Dois cachorros estão lutando e s	NEUTRAL	4	A_neutral_B	B_neutral_A	Two dogs are fighting	Two dogs are	FLICKR	FLICKR	TRAIN
13	Dois cachorros estão lutando e se a	Não tem nenhum cachorro lutan	CONTRADICTION	3.3	A_contradicts_B	B_contradicts_A	Two dogs are wrestlir	There is no dc	FLICKR	FLICKR	TEST
14	Um cachorro castanho está atacan	Dois cachorros estão lutando	NEUTRAL	3.5	A_neutral_B	B_neutral_A	A brown dog is attack	Two dogs are	FLICKR	FLICKR	TRAIN
15	Um cachorro castanho está atacan	Não tem nenhum cachorro lutan	NEUTRAL	2.7	A_neutral_B	B_neutral_A	A brown dog is attack	There is no dc	FLICKR	FLICKR	TEST
16	Dois cachorros estão lutando e se a	Um cachorro castanho está atac	NEUTRAL	2.9	A_neutral_B	B_neutral_A	Two dogs are wrestlir	A brown dog i	FLICKR	FLICKR	TEST
17	Dois cachorros estão lutando e se a	Um cachorro castanho está ajud	NEUTRAL	2.3	A_neutral_B	B_neutral_A	Two dogs are wrestlir	A brown dog i	FLICKR	FLICKR	TEST
18	Um cachorro castanho está atacan	Dois cachorros estão lutando e s	NEUTRAL	3.2	A_neutral_B	B_neutral_A	A brown dog is attack	Two dogs are	FLICKR	FLICKR	TRAIN
19	Uma pessoa de blusa preta está faz	Um homem com uma jaqueta pr	ENTAILMENT	4.9	A_entails_B	B_entails_A	A person in a black jac	A man in a bla	FLICKR	FLICKR	TEST
20	Não há nenhum homem de jaqueta	Uma pessoa de blusa preta está	CONTRADICTION	3.6	A_contradicts_B	B_contradicts_A	There is no man in a b	A person in a l	FLICKR	FLICKR	TEST
21	Uma pessoa de blusa preta está faz	Uma pessoa em uma motocicleta	NEUTRAL	3	A_neutral_B	B_neutral_A	A person in a black jac	A person on a	FLICKR	FLICKR	TEST
22	Uma pessoa habilidosa está andan	Uma pessoa está andando de bic	ENTAILMENT	4.3	A_entails_B	B_neutral_A	A skilled person is ridi	A person is rid	FLICKR	FLICKR	TEST
23	Ninguém está dirigindo uma bicicle	Uma pessoa está andando de bic	CONTRADICTION	4.1	A_contradicts_B	B_contradicts_A	Nobody is riding the b	A person is rid	FLICKR	FLICKR	TEST
24	Uma pessoa de blusa preta está faz	Uma pessoa habilidosa está andi	NEUTRAL	3.4	A_neutral_B	B_neutral_A	A person in a black jac	A skilled persc	FLICKR	FLICKR	TRIAL
25	Ninguém está dirigindo uma bicicle	Uma pessoa de blusa preta está	NEUTRAL	2.8	A_contradicts_B	B_neutral_A	Nobody is riding the b	A person in a l	FLICKR	FLICKR	TRAIN
26	Uma pessoa está andando de bicicl	Um homem com uma jaqueta pr	NEUTRAL	3.7	A_neutral_B	B_neutral_A	A person is riding the	A man in a bla	FLICKR	FLICKR	TRAIN
27	Uma pessoa está andando de bicicl	Não há nenhum homem de jaqu	NEUTRAL	3.2	A_neutral_B	B_neutral_A	A person is riding the	There is no m	FLICKR	FLICKR	TEST
28	Uma pessoa em uma motocicleta p	Uma pessoa está andando de bic	NEUTRAL	3.4	A_neutral_B	B_neutral_A	A person on a black m	A person is rid	FLICKR	FLICKR	TRAIN

Figura: Amostra dos dados contidos no SICK-BR

Baselines

Baseline Majority

- Esse modelo classifica o par de sentenças de acordo com a classe mais presente no corpus de treinamento.

Baseline Overlap

- Esse modelo extrai dois features do corpus: a quantidade de palavras exclusivas da primeira sentença e a quantidade de palavras exclusivas da segunda.
- Os features então alimentam uma SVM para fazer a classificação.

O classificador Bayesiano Ingênuo (Ou naive bayes)

Bayes

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Figura: O teorema (ou regra) de Bayes

NB - vetor de características

- Para atribuir a probabilidade de um par de frases (p) ter como relação uma categoria (c), ou seja, para calcular P(c|p), o modelo trata cada sentença como um vetor de traços ($\bar{p}$).
- Esses traços são as variáveis observadas, caso as palavras daquela sentença e o valor de similaridade entre as duas sentenças.
- Cada valor possível de cada traço tem uma probabilidade relacionada a cada categoria possível.
- Essas probabilidades são extraídas do corpus.

NB e SICK-BR: características

- dois grupos de Features foram testados nessa rodada. Um é o Bag-of-Words (BOW) e o outro a similaridade.
- BOW pois as duas sentenças foram consideradas um saco de palavras, informações sintáticas foram descartadas.
- No modelo BOW, Cada item do vocabulário (palavras únicas) do corpus de treinamento foi considerado um possível *feature*.
- As palavras contidas num par de sentenças qualquer entram como features para a classificação daquele par.

Infernal[1]

[1] Slides aproveitados de uma apresentação feita por Igor Câmara para o GLiC-USP

INFERNAL

- INFERence in NAtural Language. Desenvolvido na tese de doutorado de Erick Fonseca, defendida em 2018.
- Sistema de **engenharia de atributos**.
- No trabalho do Erick, o corpus utilizado foi o Assin, agora replico no SICK-BR.

INFERNAL: pré-processamento

Alguns procedimentos de pré-processamento

- Anotação de árvores sintáticas.
- Lematização das palavras (um dicionário de lemas + POS Tags para desambiguação)
- Detecção de entidades nomeadas (SpaCy)
- Alinhamentos lexicais (WordNet e PPDB - *Lexically-Expanded Paraphrase Database*)

INFERNAL: atributos

Atributos

- **BLEU** - métrica usada para avaliação de tradução. Calcula a sobreposição de n-gramas.
- **Sobreposição de dependências** - para identificar fenômenos como voz ativa e passiva.
- **Nominalização** - identifica a presença de um verbo e um substantivo derivado dele (exemplo: *correr* e *corrida*).
- **Proporção e tamanho** - proporção entre quantidade de tokens de ambas as sentenças.
- **Argumentos verbais** - correspondência de verbo com sujeito e objeto direto nas duas sentenças.

INFERNAL: atributos

- **Negação**: Se um verbo alinhado nas duas sentenças está negado em uma delas.
- **Quantidades**: quantidades (iguais ou diferentes) são usadas para se referir a uma palavra alinhada.
- **Cosseno das sentenças** - onde o vetor de cada sentença é a média dos vetores de word embedding das palavras que a constituem.
- **TED simples**: valor de TED simples, considerando o custo de toda operação como 1. TED normal; normalizado pelo tamanho de cada sentença.

INFERNAL: atributos

- **TED com distância de cosseno** - extensão da anterior que calcula custo de substituição como $1 - \cos(w_1, w_2)$.
- **Sobreposição de palavras** - Proporção de lemas em comum sobre o total de palavras de T e de H.
- **Sobreposição de sinônimos** - Igual ao anterior, mas considerando toda palavra alinhada, não só as com lema idêntico.
- **Sobreposição soft** - Idem ao anterior, mas considerando uma medida de similaridade, avaliado pelo cosseno das word embeddings.
- **Entidades nomeadas** - identifica entidades nomeadas alinhadas ou desalinhadas nas duas sentenças.

Resultados

Resultado - Baselines

Baselines		
	Accuracy	F1
Majority	0.6311	0.48
Overlap	0.5687	0.24

Figura: Resultados dos modelos de Baseline no SICK-BR

Resultado - Naive Bayes no SICK-BR

0.8-0.2	Precisão	Cobertura	Acurácia	F
Média Base	0,5191	0,0290	0,5790	0,0361
D_Padrão Base	0,0875	0,1277	0,0520	0,1545
Média BOW	0,5167	0,0013	0,5683	0,0026
D_Padrão BOW	0,0513	0,0004	0,0125	0,0007
Média BOW-Sim	0,8350	0,6274	0,7651	0,6881
D_Padrão BOW-Sim	0,1764	0,1160	0,0809	0,0491
Média Sim	0,9072	0,5794	0,7974	0,7070
D_Padrão Sim	0,0121	0,0160	0,0097	0,0126

Figura: Resultados das versões de Naive Bayes desenvolvidas como primeiro teste no corpus SICK-BR

Resultado - Infernal no Assin

Modelo	Validação		PT-BR		PT-PT		Geral	
	Acurácia	F_1	Acurácia	F_1	Acurácia	F_1	Acurácia	F_1
Naive Bayes	80,30%	0,68	79,05%	0,62	80,05%	0,68	79,55%	0,65
RL	85,50%	0,72	**87,30%**	**0,71**	85,75%	0,72	**86,52%**	**0,72**
RL, balanceado	85,20%	**0,74**	85,00%	0,69	84,60%	**0,74**	84,80%	**0,72**
Random Forest	85,20%	0,72	86,20%	0,67	**86,20%**	**0,74**	86,20%	0,71
Gradient Boost	**85,80%**	0,73	86,35%	0,67	86,10%	**0,74**	86,22%	0,71
SVM	85,60%	0,73	86,90%	0,70	85,75%	0,73	86,33%	**0,72**
SVM, balanceado	80,20%	0,69	79,20%	0,64	80,95%	0,71	80,08%	0,68
L2F/INESC-ID	—	—	85,85%	0,66	84,90%	0,71	—	—
Baseline	81,40%	0,69	82,80%	0,64	81,75%	0,7	82,27%	0,67

Figura: Resultados de modelos baseados nos features do Infernal no Assin

Infernal no SICK-BR

	Gradiant Boosting Classifier		Logistic Regression		Naive Bayes - Gaussian		Random Forest		SVM	
	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro	Accuracy	F1-macro
test	75.03%	0.708	72.77%	0.663	64.86%	0.621	74.46%	0.709	74.66%	0.696
trial	75.15%	0.708	72.53%	0.647	64.04%	0.6	76.36%	0.726	74.34%	0.689

Figura: Resultados de modelos baseados nos features do Infernal no SICK-BR

Conclusões e perspectivas

BOW e Similaridade

- Bag of Words é um modelo com desempenho fraco.
- A similaridade é uma boa variável para a tarefa, no entanto é uma medida *muito* problemática.

Assin e SICK

- Os resultados foram parecidos em ambos os corpora.
- Uma questão que aparece nessas análises iniciais é a de que o SICK-BR ganharia muito se fosse maior.
- De modo geral os resultados mostram que o SICK-BR é uma ferramenta tão boa quanto o Assin para a tarefa.

Próximos passos

- Analisar os erros e acertos do Infernal no Sick-BR
- Testar outros modelos.
- Desenvolver estratégias para aumentar o tamanho do Sick-BR de modo consistente, mantendo a qualidade do Corpus.

Obrigado!