PESC
Programa de Engenharia de Sistemas e Computação

Reconhecimento de Inferência Textual

Gabriel Garcia de Almeida

Tópicos

- **Visão geral:**
  - Definição;
  - Motivação;
  - Abordagens;
  - *Datasets*;
  - *Recursive Auto-Encoder*;
  - Competição *RepEval*;
- **Experimentos:**
  - Modelos baselines;
  - Matriz de confusão;
  - Varredura de hiperparâmetros;
  - Curva de Aprendizado;
  - Exploração;
  - Análise dos erros;
- **Ideias de passos futuros.**

PESC
Programa de Engenharia
de Sistemas e Computação

Visão Geral

Definição de Inferência Textual

- **Métodos de Inferência Textual**: *reconhecer*, gerar ou extrair pares de expressões tal que um humano que lê e confia no primeiro elemento, também possivelmente assumirá que o segundo elemento é verdade;
- **Métodos de Paráfrase**: reconhecer, gerar ou extrair frases que transmitem quase a mesma informação;

1. *Wonderworks Ltd*. **construiu** a nova ponte.
2. A nova ponte **foi construída** pela *Wonderworks Ltda*.
3. *Wonderworks Ltd*. **é** a construtora da nova ponte.

Androutsopoulos, Ion, e Prodromos Malakasiotis. “A Survey of Paraphrasing and Textual Entailment Methods”. *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Definição usando lógica

P : Leonardo da Vinci pintou a Mona Lisa.
Φ_P : **éPintorDe**(DaVinci, MonaLisa)

H : Mona Lisa é um trabalho de Leonardo da Vinci.
Φ_H : **éTrabalhoDe**(MonaLisa, DaVinci)

ψ : $\forall x \; \forall y$ **éPintorDe**(x, y) $\Rightarrow$ **éTrabalhoDe**(x, y)

$$(\Phi_P \wedge \psi) \models \Phi_H$$
$$(\Phi_H \wedge \psi) \not\models \Phi_P$$

Androutsopoulos, Ion, e Prodromos Malakasiotis. “A Survey of Paraphrasing and Textual Entailment Methods”. *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Motivações

- ***Question Answering:***
  - Motor de busca + reconhecimento de inferência textual

(17) Who sculpted the Doryphoros?

(18) The Doryphoros is one of the best known Greek sculptures of the classical era in Western Art. The Greek sculptor *Polykleitos* designed this work as an example of the "canon" or "rule", showing the perfectly harmonious and balanced proportions of the human body in the sculpted form. The sculpture was known through the Roman marble replica found in Herculaneum and conserved in the Naples National Archaeological Museum, but, according to *Francis Haskell* and *Nicholas Penny*, early connoisseurs passed it by in the royal Bourbon collection at Naples without notable comment.

(19) Polykleitos/Francis Haskell/Nicholas Penny sculpted the Doryphoros.

Androutsopoulos, Ion, e Prodromos Malakasiotis. "A Survey of Paraphrasing and Textual Entailment Methods". *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Motivações (2)

- **Sumarização:**
  - Remoção de sentenças redundantes;
  - Escolha da melhor sumarização;

(20) Mother Catherine, 82, the mother superior, will attend the hearing on Friday, he said.
(21) Mother Catherine, 82, the mother superior, will attend.

- **Geração de texto:**
  - Avaliação de diferentes possibilidades de frases;
    - Evitar palavras iguais, melhorar coerência do texto, aderência a estilos de escrita…
    - Simplificar termos técnicos;
- **Correção automática de provas;**
- **...**

Androutsopoulos, Ion, e Prodromos Malakasiotis. “A Survey of Paraphrasing and Textual Entailment Methods”. *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Abordagens *(reconhecimento)*

- **Baseadas em Lógica:**
  - Necessários: *parsers* (representações lógicas), *motor de inferência* e bases de "*senso comum*":
    - WordNet, VerbNet, FrameNet...

- **Baseado em Similaridade de String:**
  - **Levenshtein** (Distância de Edição), **Bleu** (usada em tradução de máquina) ...;
  - Problemas com negações;
  - Problemas com discrepância entre o tamanho de ambos textos;

Androutsopoulos, Ion, e Prodromos Malakasiotis. "A Survey of Paraphrasing and Textual Entailment Methods". *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Intelligent Machines

An AI with 30 Years' Worth of Knowledge Finally Goes to Work

An effort to encode the world's knowledge in a huge database has sometimes seemed impractical, but those behind the technology say it is finally ready.

by Will Knight March 14, 2016

Having spent the past 31 years memorizing an astonishing collection of general knowledge, the artificial-intelligence engine created by Doug Lenat is finally ready to go to work.

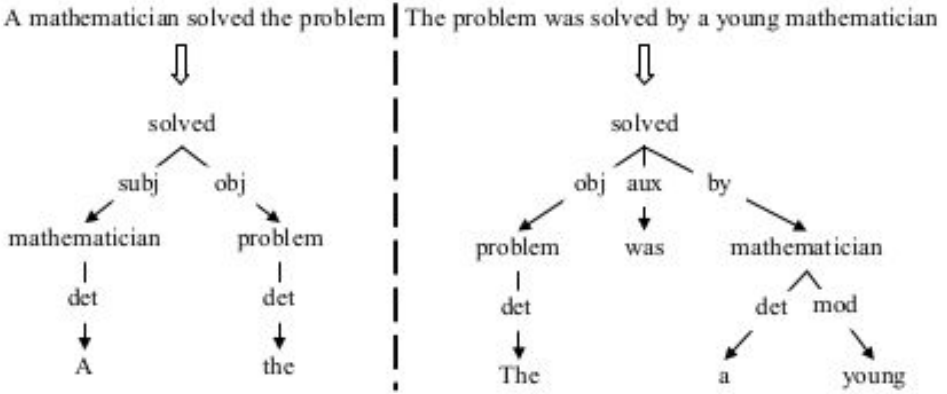
Lenat's creation is Cyc, a knowledge base of semantic information designed to give computers some understanding of how things work in the real world.

https://www.technologyreview.com/s/600984/an-ai-with-30-years-worth-of-knowledge-finally-goes-to-work/

Abordagens (2)

- **Baseado em Similaridade Sintática:**
  - Parsers de dependência + medidas de similaridade de árvores;

- **Baseado em *Aprendizado de Máquina*:**
  - Problema de classificação binária ou multiclasse (confirmação/neutro/contradição);
  - Pode ser usado como forma de agregar os demais métodos;
  - Dependência de dataset;



Androutsopoulos, Ion, e Prodromos Malakasiotis. “A Survey of Paraphrasing and Textual Entailment Methods”. *arXiv:0912.3747 [cs]*, 18 de dezembro de 2009. doi:10.1613/jair.2985.

Dataset 1: SNLI

- **S**tanford **N**atural **L**anguage **I**nference (2015);
- Existência apenas de datasets pequenos ou gerados (semi-)automaticamente;
- **Premissas** são legendas do *Flickr30k corpus*:
  - *Labels*: *contradiction*, *neutral* e *entailment*;
  - Duas tarefas para multidão:
    - Coleta de 3 possíveis **hipóteses**;
    - **Validação** de ~10% hipóteses;
- Hipóteses falam sempre do mesmo evento, visto de diversos ângulos:
  - Manter consistência;

- A person on a horse jumps over a broken down airplane.
  - A person is training his horse for a competition.
  - A person is at a diner, ordering an omelette.
  - A person is outdoors, on a horse.

- Children smiling and waving at camera
  - They are smiling at their parents
  - There are children present
  - The kids are frowning

Data set sizes:	
Training pairs	550,152
Development pairs	10,000
Test pairs	10,000
Sentence length:	
Premise mean token count	14.1
Hypothesis mean token count	8.3

General:	
Validated pairs	56,951
Pairs w/ unanimous gold label	58.3%
Individual annotator label agreement:	
Individual label = gold label	89.0%
Individual label = author's label	85.8%
Gold label/author's label agreement:	
Gold label = author's label	91.2%
Gold label ≠ author's label	6.8%
No gold label (no 3 labels match)	2.0%

Bowman, Samuel R., Gabor Angeli, Christopher Potts, e Christopher D. Manning. "A large annotated corpus for learning natural language inference". *arXiv:1508.05326 [cs]*, 21 de agosto de 2015. http://arxiv.org/abs/1508.05326.

SNLI - Leaderboard

Publication	Model	Parameters	Train (% acc)	Test (% acc)
Feature-based models				
Bowman et al. '15	Unlexicalized features		49.4	50.4
Bowman et al. '15	+ Unigram and bigram features		99.7	*78.2*
Other neural network models				
Rocktäschel et al. '15	100D LSTMs w/ word-by-word attention	250k	85.3	83.5
Pengfei Liu et al. '16a	100D DF-LSTM	320k	85.2	84.6
Yang Liu et al. '16	600D (300+300) BiLSTM encoders with intra-attention and symbolic preproc.	2.8m	85.9	85.0
Pengfei Liu et al. '16b	50D stacked TC-LSTMs	190k	86.7	85.1
Munkhdalai & Yu '16a	300D MMA-NSE encoders with attention	3.2m	86.9	85.4
Wang & Jiang '15	300D mLSTM word-by-word attention model	1.9m	92.0	86.1
Cheng et al. '16	300D LSTMN with deep attention fusion	1.7m	87.3	85.7
Cheng et al. '16	450D LSTMN with deep attention fusion	3.4m	88.5	86.3
Parikh et al. '16	200D decomposable attention model	380k	89.5	86.3
Parikh et al. '16	200D decomposable attention model with intra-sentence attention	580k	90.5	86.8
Munkhdalai & Yu '16b	300D Full tree matching NTI-SLSTM-LSTM w/ global attention	3.2m	88.5	87.3
Wang et al. '17	BiMPM	1.6m	90.9	87.5
Sha et al. '16	300D re-read LSTM	2.0m	90.7	87.5
Gong et al. '17	448D Densely Interactive Inference Network (DIIN)	4.4m	91.2	88.0
Chen et al. '16	600D ESIM + 300D Syntactic TreeLSTM (code)	7.7m	93.5	**88.6**
Wang et al. '17	BiMPM **Ensemble**	6.4m	93.2	**88.8**
Gong et al. '17	448D Densely Interactive Inference Network (DIIN) **Ensemble**	17.4m	92.3	**88.9**

http://nlp.stanford.edu/projects/snli/

Dataset 2: MultiNLI 0.9

- **Multi**-Genre **N**atural **L**anguage **I**nference;
- Objetivo: aumentar a **abrangência** do SNLI:
  - Somente **descrição** de cenas;
  - *Estado-da-arte* próximo da concordância humana;
- **Premissas** são de 10 gêneros textuais:
  - Ex.: Transcrições de conversas, relatos do 11/9, guias de viagem, textos de ficção e não-ficção…
  - Uso similar da multidão;
  - 392.702 pares de sentenças;
  - Cinco gêneros incluídos no conjunto de treino:
    - Avaliação da "adaptação entre-gêneros";

- One of our number will carry out your instructions minutely. (fiction)
  - A member of my team will execute your orders with immense precision.
- Fun for adults and children. (travel)
  - Fun for only children.
- Were they in there? (slate)
  - Were they supposed to be in there?

Statistic	SNLI	MultiNLI
Pairs w/ unanimous gold label	58.3%	58.2%
Individual label = gold label	89.0%	88.7%
Individual label = author's label	85.8%	85.2%
Gold label = author's label	91.2%	92.6%
Gold label ≠ author's label	6.8%	5.6%
No gold label (no 3 labels match)	2.0%	1.8%

Williams, Adina, Nikita Nangia, e Samuel R. Bowman. "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". *arXiv:1704.05426 [cs]*, 18 de abril de 2017. http://arxiv.org/abs/1704.05426.

MultiNLI - Baselines

Genre	#Examples Train	Dev.	Test	Agrmt.	Model Acc. ESIM	CBOW
SNLI	*550,152*	*10,000*	*10,000*	*89.0%*	*86.7%*	*80.6 %*
FICTION	77,348	2,000	2,000	89.4%	73.0%	67.5%
GOVERNMENT	77,350	2,000	2,000	87.4%	74.8%	67.5%
SLATE	77,306	2,000	2,000	87.1%	67.9%	60.6%
TELEPHONE	83,348	2,000	2,000	88.3%	72.2%	63.7%
TRAVEL	77,350	2,000	2,000	89.9%	73.7%	64.6%
9/11	0	2,000	2,000	90.1%	71.9%	63.2%
FACE-TO-FACE	0	2,000	2,000	89.5%	71.2%	66.3%
LETTERS	0	2,000	2,000	90.1%	74.7%	68.3%
OUP	0	2,000	2,000	88.1%	71.7%	62.8%
VERBATIM	0	2,000	2,000	87.3%	71.9%	62.7%
MultiNLI Overall	**392,702**	**20,000**	**20,000**	**88.7%**	**72.2%**	**64.7%**

CBOW = Continuous bag-of-words

ESIM = Enhanced Sequential Inference Model

Williams, Adina, Nikita Nangia, e Samuel R. Bowman. “A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference”. *arXiv:1704.05426 [cs]*, 18 de abril de 2017. http://arxiv.org/abs/1704.05426.

Competição *RepEval* 2017

- Usa o MultiNLI para avaliar modelos de **representação vetorial distribuída**;
  - Inferência textual seria uma forma de medir o *entendimento de linguagem natural*;

Team Name	Authors	Matched	Mismatched	Model Details
alpha (ensemble)	Chen et al.	**74.9%**	**74.9%**	STACK, CHAR, ATTN., POOL, PRODDIFF
YixinNie-UNC-NLP	Nie and Bansal	74.5%	73.5%	STACK, POOL, PRODDIFF, SNLI
alpha	Chen et al.	73.5%	73.6%	STACK, CHAR, ATTN, POOL, PRODDIFF
Rivercorners (ensemble)	Balazs et al.	72.2%	72.8%	ATTN, POOL, PRODDIFF, SNLI
Rivercorners	Balazs et al.	72.1%	72.1%	ATTN, POOL, PRODDIFF, SNLI
LCT-MALTA	Vu et al.	70.7%	70.8%	CHAR, ENHEMB, PRODDIFF, POOL
TALP-UPC	Yang et al.	67.9%	68.2%	CHAR, ATTN, SNLI
BiLSTM baseline	Williams et al.	67.0%	67.6%	POOL, PRODDIFF, SNLI

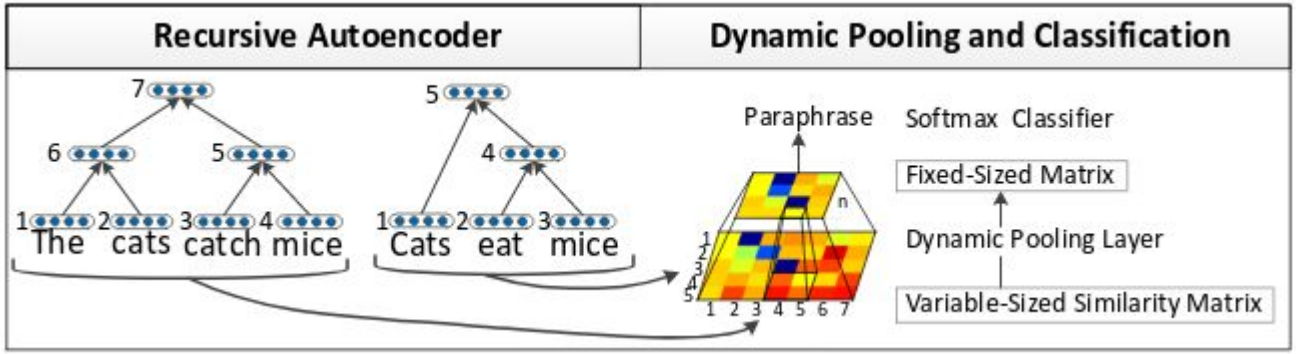
Table 3: RepEval 2017 shared task competition results. The Model Details column lists some of the key strategies used in each system, using keywords: STACK: use of multilayer bidirectional RNNs, CHAR: character-level embeddings, ENHEMB: embeddings enhanced with auxiliary features, POOL: max or mean pooling over RNN states, ATTN: intra-sentence attention, PRODDIFF: elementwise sentence product and difference features in the final entailment classifier, SNLI: use of the SNLI training set.

Nangia, Nikita, Adina Williams, Angeliki Lazaridou, and Samuel R. Bowman. “The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations.” *arXiv:1707.08172 [Cs]*, July 25, 2017. http://arxiv.org/abs/1707.08172.

Recursive Auto Encoder

- Rede neural recursiva;
- Gera representação não-supervisionada da sentença:
  - Usa a árvore de *parsing* e as palavras para aprender;

- Quatro etapas para aplicação:
  - Codificação (RAE);
  - Comparação (Matriz de similaridade);
  - Repr. em tamanho fixo (Pooling);
  - Classificador final;



Socher, Richard, Eric H. Huang, Jeffrey Pennin, Christopher D Manning, and Andrew Y. Ng. “Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection.” In *Advances in Neural Information Processing Systems 24*, edited by J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, 801–809. Curran Associates, Inc., 2011.

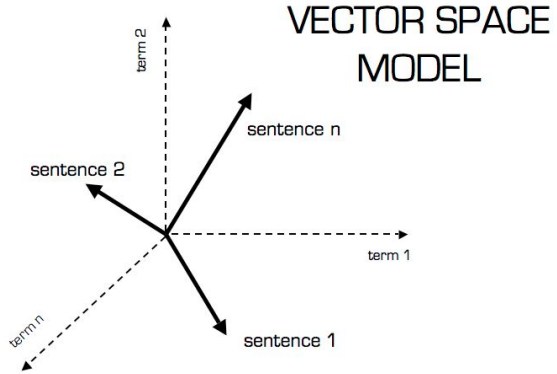
PESC
Programa de Engenharia
de Sistemas e Computação

Experimentos

Modelos Baselines

- Representações vetoriais diferentes:
  - **TF-IDF**:
    - TF-IDF da Premissa vs Hipótese;
  - **Continuous Bag-of-Words** (CBoW):
    - Representação vetorial das frases = soma dos *word embeddings*;
    - Duas variantes:
      - Concatenação dos vetores;
      - Concatenação, multiplicação e diferença absoluta dos vetores;

- **Classificador**: Todos usando regressão logística multinomial (*softmax*);



http://blog.christianperone.com/2013/09/machine-learning-cosine-similarity-for-vector-space-models-part-iii/

Varredura de Hiperparâmetros

Objetivo: Melhorar algoritmos escolhendo os parâmetros *"não-treináveis"*;

- Subconjunto de ~15k do MultiNLI;
- **k-Folds** com k=10;
- Hiperparâmetros:
  - *pool*: tamanho da janela;
  - *reg:* coef. de regularização;
  - *eta:* coef. de aprendizado;
  - *epochs*: num. passos de treinamento;

Modelo	Acurácia
RAE pool=15 reg=1 e-05 eta=0.01 epochs=15	38,48%
RAE pool=15 reg=0.0001 eta=0.01 epochs=10	38,51%
RAE pool=15 reg=1 e-05 eta=0.1 epochs=10	38,52%
RAE pool=15 reg=1 e-05 eta=0.1 epochs=15	38,53%
RAE pool=15 reg=0.0001 eta=0.1 epochs=10	38,54%
RAE pool=15 reg=1 e-05 eta=0.01 epochs=10	38,55%
RAE pool=15 reg=0.0001 eta=0.01 epochs=15	38,56%
RAE pool=15 reg=0.0001 eta=0.1 epochs=15	38,58%
RAE pool=10 reg=0.0001 eta=0.01 epochs=15	40,91%
RAE pool=10 reg=1 e-05 eta=0.01 epochs=15	40,94%
RAE pool=10 reg=1 e-05 eta=0.1 epochs=15	41,05%
RAE pool=10 reg=1 e-05 eta=0.1 epochs=10	41,05%
RAE pool=10 reg=0.0001 eta=0.1 epochs=10	41,07%
RAE pool=10 reg=1 e-05 eta=0.01 epochs=10	41,07%
RAE pool=10 reg=0.0001 eta=0.1 epochs=15	41,15%
RAE pool=10 reg=0.0001 eta=0.01 epochs=10	41,15%
RAE pool=5 reg=1 e-05 eta=0.1 epochs=15	41,95%
RAE pool=5 reg=1 e-05 eta=0.01 epochs=10	42,04%
RAE pool=5 reg=1 e-05 eta=0.1 epochs=10	42,07%
RAE pool=5 reg=0.0001 eta=0.1 epochs=10	42,07%
RAE pool=5 reg=0.0001 eta=0.01 epochs=15	42,09%
RAE pool=5 reg=0.0001 eta=0.01 epochs=10	42,13%
RAE pool=5 reg=1 e-05 eta=0.01 epochs=15	42,14%
RAE pool=5 reg=0.0001 eta=0.1 epochs=15	42,18%
CBoW (Concatenação)	43,76%
TF-IDF	44,19%
CBoW (Concatenação, Subtração e Multiplicação)	48,21%
CBoW (Concatenação) + TF-IDF	50,69%

Matriz de confusão

RAE pool=5 reg=0.0001 eta=0.1 epochs=15			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	11,69%	11,53%	10,11%
Neutro	8,77%	15,91%	8,65%
Inferência	9,11%	9,64%	14,58%

RAE pool=10 reg=0.0001 eta=0.1 epochs=15			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	11,39%	11,76%	10,19%
Neutro	9,37%	15,51%	8,46%
Inferência	9,74%	9,34%	14,25%

RAE pool=15 reg=0.0001 eta=0.1 epochs=15			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	15,67%	10,95%	6,71%
Neutro	12,59%	14,18%	6,56%
Inferência	14,42%	10,19%	8,73%

CBoW (Concatenação)			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	15,34%	8,67%	9,32%
Neutro	9,03%	14,37%	9,93%
Inferência	9,14%	10,15%	14,05%

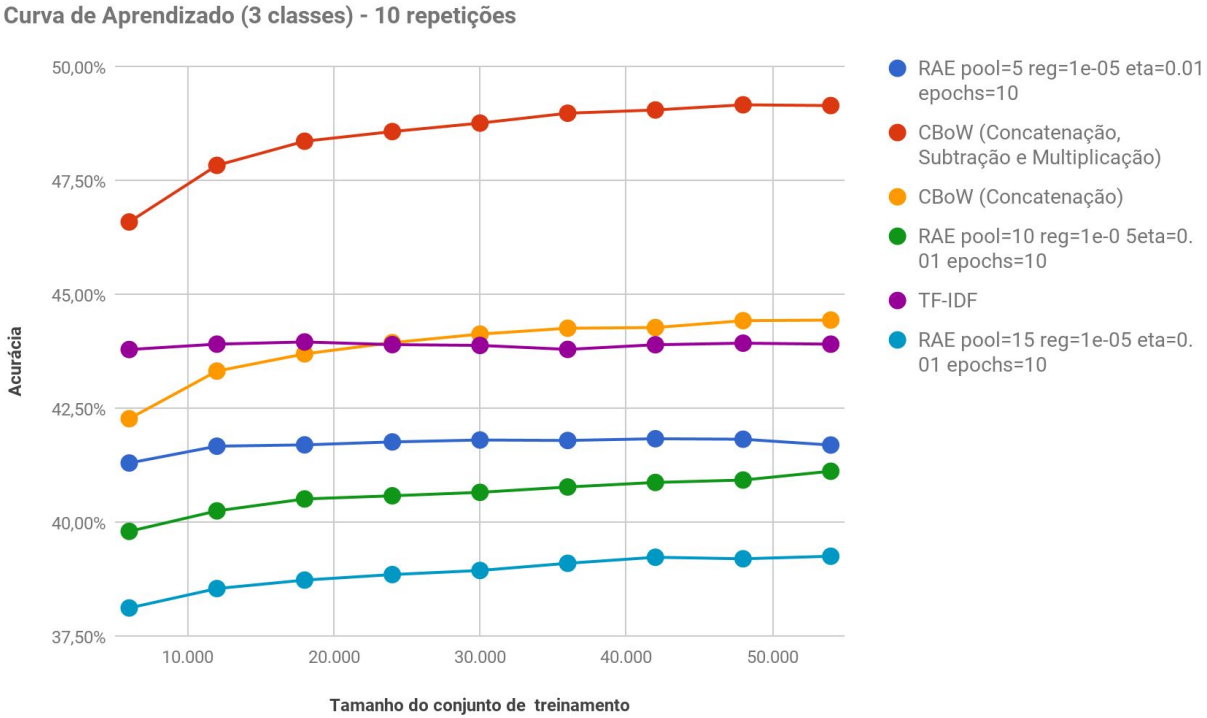
CBoW (Concatenação, Subtração e Multiplicação)			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	13,99%	9,74%	9,61%
Neutro	8,35%	17,01%	7,98%
Inferência	8,13%	7,99%	17,21%

TF-IDF			
Alvo \ Predito	Contradição	Neutro	Inferência
Contradição	6,10%	15,34%	11,89%
Neutro	5,41%	17,18%	10,75%
Inferência	5,00%	7,42%	20,91%

Curva de Aprendizado

Objetivo: Avaliar melhorias nos algoritmos conforme crescimento do *dataset*;

- Subconjunto de ~ 60k pares;
- Treino com x% e teste com restante;



Varredura de Hiperparâmetros (2)

Objetivo: Melhorar algoritmos escolhendo os parâmetros *"não-treináveis"*;

- Subconjunto de ~15k do MultiNLI;
- Hiperparâmetros:
  - *pool*: tamanho da janela;
  - *reg:* coef. de regularização;
  - *eta:* coef. de aprendizado;
  - *epochs*: num. passos de treinamento;
- **Caso simplificado:**
  - **Inferência** vs demais (contradição e neutro);
- *Downsample* para balancear as classes;

Modelo	Acurácia
CBoW (Concatenação, Subtração e Multiplicação)	51,22%
RAE pool=15 reg=0.0001 eta=0.1 epochs=10	52,67%
RAE pool=15 reg=0.0001 eta=0.01 epochs=10	52,85%
RAE pool=15 reg=1 e-05 eta=0.01 epochs=15	52,89%
RAE pool=15 reg=1 e-05 eta=0.01 epochs=10	52,90%
RAE pool=15 reg=0.0001 eta=0.1 epochs=15	53,08%
RAE pool=15 reg=1 e-05 eta=0.1 epochs=10	53,09%
RAE pool=15 reg=0.0001 eta=0.01 epochs=15	53,22%
RAE pool=15 reg=1 e-05 eta=0.1 epochs=15	53,68%
CBoW (Concatenação)	55,05%
RAE pool=5 reg=0.0001 eta=0.1 epochs=10	58,21%
RAE pool=5 reg=1 e-05 eta=0.1 epochs=10	58,23%
RAE pool=5 reg=1 e-05 eta=0.1 epochs=15	58,27%
RAE pool=10 reg=0.0001 eta=0.1 epochs=15	58,30%
RAE pool=10 reg=1 e-05 eta=0.01 epochs=15	58,38%
RAE pool=10 reg=0.0001 eta=0.01 epochs=10	58,40%
RAE pool=10 reg=0.0001 eta=0.01 epochs=15	58,41%
RAE pool=5 reg=1 e-05 eta=0.01 epochs=10	58,42%
RAE pool=10 reg=1 e-05 eta=0.01 epochs=10	58,43%
RAE pool=10 reg=0.0001 eta=0.1 epochs=10	58,45%
RAE pool=5 reg=0.0001 eta=0.01 epochs=15	58,50%
RAE pool=5 reg=1 e-05 eta=0.01 epochs=15	58,52%
RAE pool=5 reg=0.0001 eta=0.01 epochs=10	58,54%
RAE pool=10 reg=1 e-05 eta=0.1 epochs=15	58,57%
RAE pool=10 reg=1 e-05 eta=0.1 epochs=10	58,69%
RAE pool=5 reg=0.0001 eta=0.1 epochs=15	58,70%
TF-IDF	59,78%
CBoW (Concatenação) + TF-IDF	65,93%

Matriz de confusão (2)

RAE pool=5 reg=0.0001 eta=0.1 epochs=15		
Alvo \ Predito	Outros	Inferência
Outros	27,74%	22,26%
Inferência	19,04%	30,96%

RAE pool=10 reg=1 e-05 eta=0.1 epochs=10		
Alvo \ Predito	Outros	Inferência
Outros	29,25%	20,75%
Inferência	20,56%	29,44%

RAE pool=15 reg=1 e-05 eta=0.1 epochs=15		
Alvo \ Predito	Outros	Inferência
Outros	17,22%	32,78%
Inferência	13,54%	36,46%

CBoW (Concatenação, Subtração e Multiplicação)		
Alvo \ Predito	Outros	Inferência
Outros	2,58%	47,42%
Inferência	1,36%	48,64%

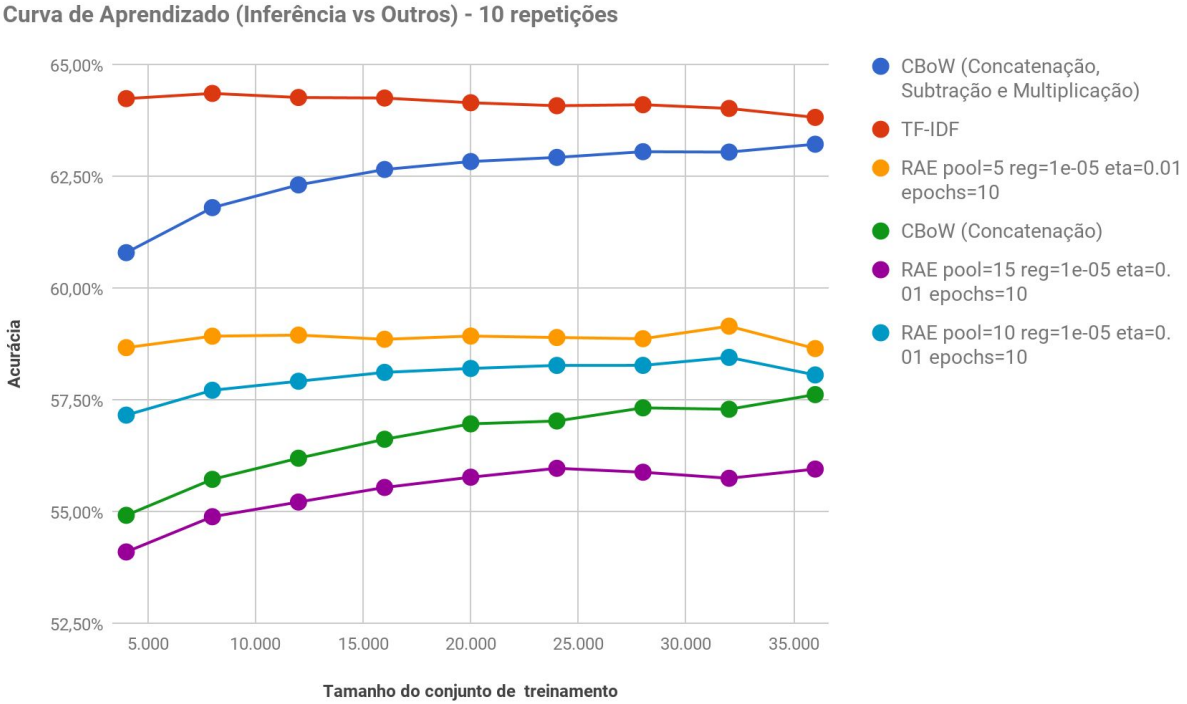
CBoW (Concatenação)		
Alvo \ Predito	Outros	Inferência
Outros	15,39%	34,61%
Inferência	10,34%	39,66%

TF-IDF		
Alvo \ Predito	Outros	Inferência
Outros	44,75%	5,25%
Inferência	34,97%	15,03%

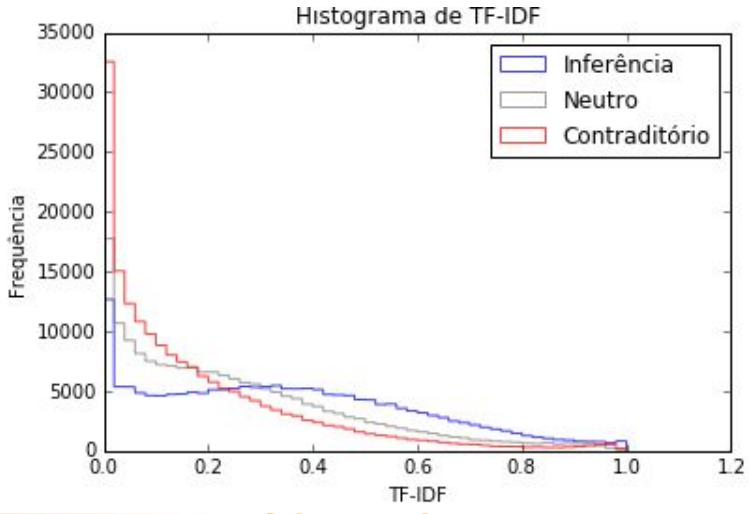
Curva de Aprendizado (2)

Objetivo: Avaliar melhorias nos algoritmos conforme crescimento do *dataset*;

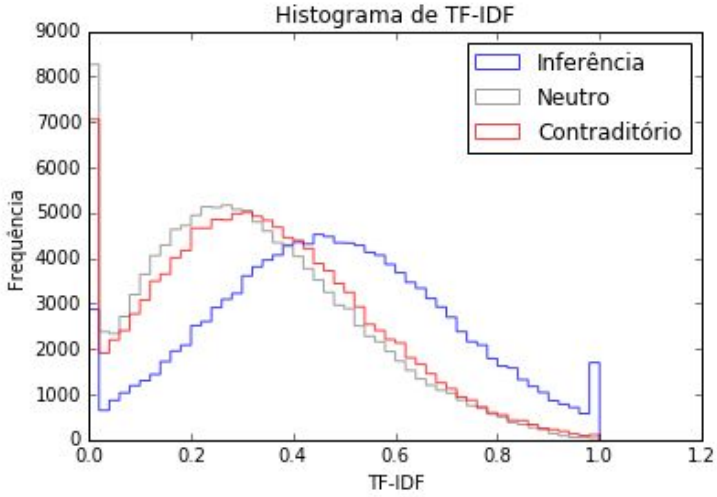
- Subconjunto de ~ 60k pares;
- Treino com x% e teste com restante;
- **Caso simplificado:**
  - **Inferência** vs demais (contradição e neutro);
- *Downsample* para balancear as classes;



Exploração dos datasets



SNLI



MultiNLI

Análise de erros

- MultiNLI tem 900 pares classificados com tags especiais:

  - **#CONDITIONAL**: Alguma sentença tem algum condicional (*if*);
  - **#ACTIVE/PASSIVE**: Existe uma troca entre voz ativa ou passiva no par;
  - **#PARAPHRASE**: As sentenças são paráfrases;
  - **#COREF**: Precisa de resolução de correferência da hipótese com relação a premissa;
  - **#QUANTIFIER**: Alguma sentença tem algum quantificador (*much*, *enough*...);
  - **#MODAL**: Alguma sentença tem algum verbo modal (can, could, may ...);
  - **#BELIEF**: Alguma sentença tem algum verbo como *agree, disagree, deny, promise* ...
  - **#NEGATION**: Alguma sentença tem alguma palavra de negação;
  - **#ANTO**: As sentenças têm pares de antônimos;
  - **#TENSE_DIFFERENCE**: As sentenças usam diferentes tempos verbais;
  - **#QUANTITY/TIME_REASONING**: A classificação do par requer raciocínio com quantidades ou tempo;
  - **#WORD_OVERLAP**: Sentenças têm mais do que 70% das palavras em comum;
  - **#LONG_SENTENCE**: Alguma sentença é maior do que 30 (premissa) ou 16 palavras (hipótese);

Análise dos erros: Baselines

CBoW (Concatenação, Subtração e Multiplicação)			
Tag	**Corretos**	**Total**	**Acurácia**
#ACTIVE/PASSIVE	17	25	68,00%
#ANTO	12	37	32,43%
#BELIEF	57	124	45,97%
#CONDITIONAL	25	49	51,02%
#COREF	34	59	57,63%
#LONG_SENTENCE	102	208	49,04%
#MODAL	152	270	56,30%
#NEGATION	97	233	41,63%
#PARAPHRASE	35	62	56,45%
#QUANTIFIER	139	265	52,45%
#QUANTITY / TIME_REASONING	24	54	44,44%
#TENSE_DIFFERENCE	33	69	47,83%
#WORD_OVERLAP	35	65	53,85%

CBoW (Concatenação)			
Tag	**Corretos**	**Total**	**Acurácia**
#ACTIVE/PASSIVE	16	25	64,00%
#ANTO	15	37	40,54%
#BELIEF	60	124	48,39%
#CONDITIONAL	20	49	40,82%
#COREF	31	59	52,54%
#LONG_SENTENCE	101	208	48,56%
#MODAL	134	270	49,63%
#NEGATION	112	233	48,07%
#PARAPHRASE	19	62	30,65%
#QUANTIFIER	133	265	50,19%
#QUANTITY / TIME_REASONING	23	54	42,59%
#TENSE_DIFFERENCE	28	69	40,58%
#WORD_OVERLAP	28	65	43,08%

TF-IDF			
Tag	**Corretos**	**Total**	**Acurácia**
#ACTIVE/PASSIVE	21	25	84,00%
#ANTO	11	37	29,73%
#BELIEF	57	124	45,97%
#CONDITIONAL	23	49	46,94%
#COREF	29	59	49,15%
#LONG_SENTENCE	91	208	43,75%
#MODAL	121	270	44,81%
#NEGATION	84	233	36,05%
#PARAPHRASE	45	62	72,58%
#QUANTIFIER	107	265	40,38%
#QUANTITY / TIME_REASONING	18	54	33,33%
#TENSE_DIFFERENCE	28	69	40,58%
#WORD_OVERLAP	36	65	55,38%

Análise dos erros: RAE

RAE pool 5			
Tag	Corretos	Total	Acurácia
#ACTIVE/PASSIVE	13	25	52,00%
#ANTO	13	37	35,14%
#BELIEF	61	124	49,19%
#CONDITIONAL	22	49	44,90%
#COREF	20	59	33,90%
#LONG_SENTENCE	95	208	45,67%
#MODAL	115	270	42,59%
#NEGATION	88	233	37,77%
#PARAPHRASE	25	62	40,32%
#QUANTIFIER	119	265	44,91%
#QUANTITY / TIME_REASONING	24	54	44,44%
#TENSE_DIFFERENCE	32	69	46,38%
#WORD_OVERLAP	32	65	49,23%

RAE pool 10			
Tag	Corretos	Total	Acurácia
#ACTIVE/PASSIVE	8	25	32,00%
#ANTO	16	37	43,24%
#BELIEF	54	124	43,55%
#CONDITIONAL	25	49	51,02%
#COREF	25	59	42,37%
#LONG_SENTENCE	102	208	49,04%
#MODAL	122	270	45,19%
#NEGATION	91	233	39,06%
#PARAPHRASE	29	62	46,77%
#QUANTIFIER	120	265	45,28%
#QUANTITY / TIME_REASONING	17	54	31,48%
#TENSE_DIFFERENCE	29	69	42,03%
#WORD_OVERLAP	35	65	53,85%

RAE pool 15			
Tag	Corretos	Total	Acurácia
#ACTIVE/PASSIVE	8	25	32,00%
#ANTO	18	37	48,65%
#BELIEF	53	124	42,74%
#CONDITIONAL	20	49	40,82%
#COREF	18	59	30,51%
#LONG_SENTENCE	84	208	40,38%
#MODAL	107	270	39,63%
#NEGATION	85	233	36,48%
#PARAPHRASE	18	62	29,03%
#QUANTIFIER	111	265	41,89%
#QUANTITY / TIME_REASONING	19	54	35,19%
#TENSE_DIFFERENCE	25	69	36,23%
#WORD_OVERLAP	30	65	46,15%

Ideias de passos futuros

- Testar outros modelos:
  - *RepEval;*
  - *Leaderboard* do SNLI;
  - Combinações dos baselines;
  - ...
- Esquema hierárquico:
  - Aproveitar viés apresentado pelo TF-IDF;
  - Especialista em remover casos neutros:
    - Expansão aleatória do dataset (neutro);
  - Especialista em diferenciar *inferência* vs *contradição*;