“You Should Probably Read This”: Hedge Detection in Text

Denys Katerenchuk, Rivka Levitan
The Graduate Center, CUNY
Brooklyn College, CUNY

Hedge Detection

Problem:

Understanding the certainty level of a claim is crucial in areas such as medicine, finance, engineering, and many others where errors can lead to disastrous results.

What if a doctor says:

“*I think* you need surgery *immediately*!”

Hedges are linguistic devices that are used to indicate uncertainty and mitigate orders.

Hedge phrase identifiers:

- modal verbs (“could”, “might”, etc.)
- peacock expressions (“very likely”, “everyone”, “I think”, etc.)
- weasel words (“some believe”, “clearly”, etc.)

Data and Contributions

The CoNLL-2010 Wikipedia dataset

Wikipedia discussions pages are manually annotated as curtain or uncertain.

Corpus:

- 11110 - train sentences
- (10% is used as validation set)
- 9634 - test sentences
- F1-score

Challenges:

- The dataset is very small and unbalanced
- The data has been around for over 10 years
- The results hasn’t been improved for over 4 years.

Related Work and Motivation

Bag-of-Words

Result: F1 - 60.17

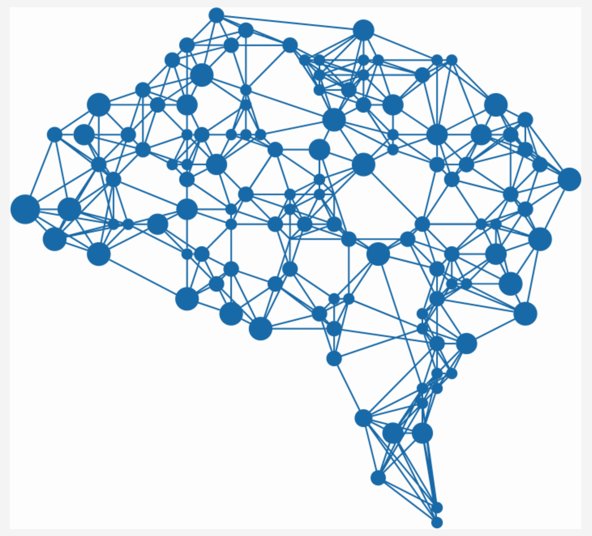
M. Georgescul. A hedgehop over a max-margin framework using hedge cues. 2012.

Probabilistic

Result: F1 - 62.8

P. Jean, et al. "Uncertainty detection in natural language: A probabilistic model.” 2016.

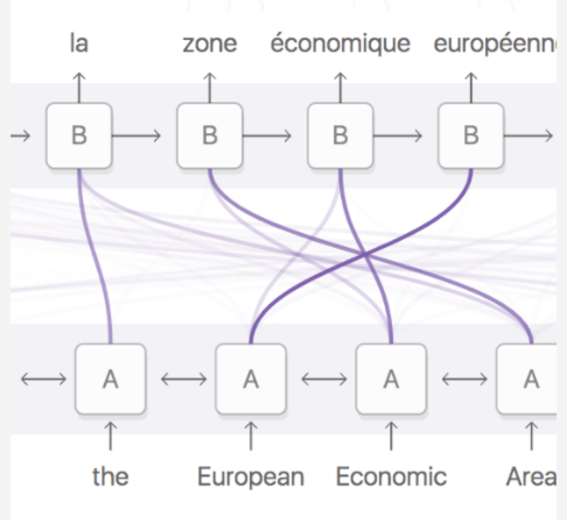
Neural Networks



Result: F1 - 67.52

H. Adel, et al., “Exploring Different Dimensions of Attention for Uncertainty Detection” 2017.

Transformers



Result: None
(BioScope: F1 - 85)

M. Sinha, et al., “Relation Aware Attention Model for Uncertainty Detection in Text” 2020.

Methods

1. Word Embeddings

Language Model	F1 Score
GoogleNews 300d	60.34
GloVe 100d	55.74
GloVe 300d 6B	**63.12**
GloVe 300d 840B	63.09
FastText 1M	62.08
FastText 2M	**63.57**
Custom Wiki 1Gb	61.99

2. NN Architectures

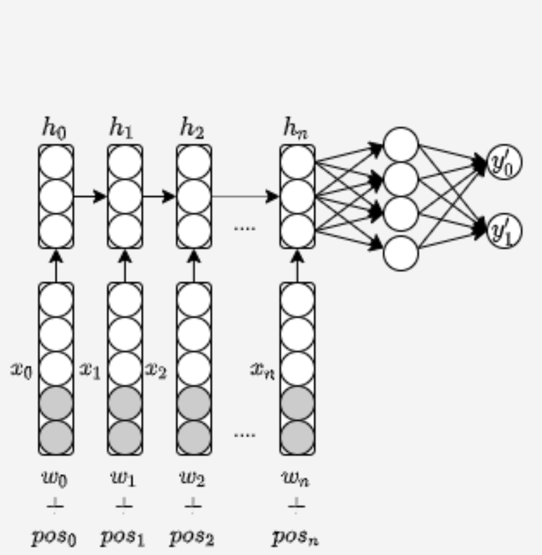
NN Model	Wiki 1G	GloVe	FastText	*Mean*	*STD*
CNN	57.38	58.34	59.36	*58.36*	*0.99*
GRU	62.07	64.23	65.04	***63.78***	*1.54*
LSTM	64.35	62.14	63.91	*63.47*	*1.17*
Transformer	57.18	57.91	54.74	*56.61*	*1.66*
CNN+At	56.86	58.22	62.79	*59.29*	*3.11*
GRU+At	62.13	63.33	65.40	*63.62*	*1.65*
LSTM+At	59.37	60.11	64.26	*61.25*	*2.64*
Mean	*59.91*	*60.61*	*62.21*		
STD	*2.97*	*2.62*	*3.86*		

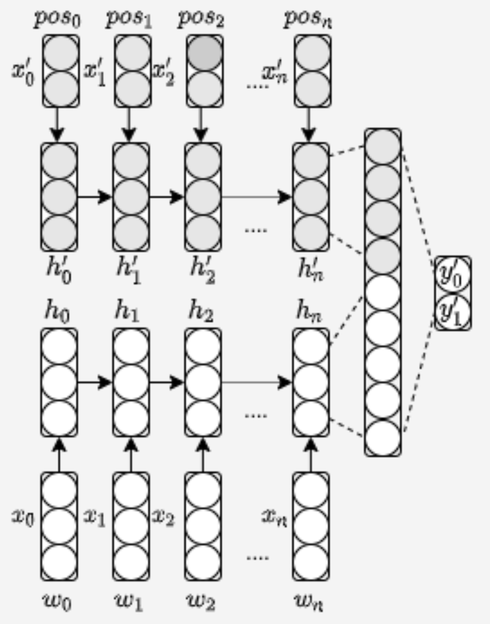
3. POS Tags

NN Model	F1 score
GRU	47.54
LSTM	48.47
GRU+At	48.27
LSTM+At	**48.9**

4. POS and Word Joint Models

4.1. Model Architectures





4.2. Training Results

	Wiki 1G	FastText	*Mean*	*STD*
Joint Input				
GRU	65.81	66.08	***65.95***	*0.19*
LSTM	61.69	66.14	*63.92*	*3.15*
GRU+At	65.46	65.45	***65.46***	*0.01*
LSTM+At	66.57	64.26	*65.42*	*1.63*
Joint Model				
GRU+At	64.42	64.54	*64.48*	*0.08*
LSTM+At	64.22	62.47	*63.35*	*1.24*
GRU&LSTM+At	64.82	66.09	***65.46***	*0.90*
Mean	*64.71*	*65.00*		
STD	*1.56*	*1.36*		

Results

	Wiki 1G	FastText	*Mean*	*STD*
Joint Input				
GRU	68.25	67.69	*67.97*	*0.40*
GRU+At	68.97	66.32	*67.65*	*1.87*
Joint Model				
GRU&LSTM+At	**69.21**	**69.74**	***69.48***	***0.37***
Mean	*68.81*	*67.92*		
STD	*0.50*	*1.72*		

	Wiki 1G	GloVe	FastText	*Mean*	*STD*
CNN	60.00	62.92	64.28	*62.40*	*2.19*
GRU	**70.24**	67.54	68.97	***68.92***	*1.35*
LSTM	69.14	65.22	68.43	*67.60*	*2.09*
Transformer	63.27	65.22	68.43	*65.64*	*2.61*
CNN+At	58.44	65.77	66.83	*63.68*	*4.57*
GRU+At	68.22	66.81	67.11	*67.38*	*0.74*
LSTM+At	68.27	65.26	67.94	*67.16*	*1.65*
Joint Input					
GRU	68.25	66.74	67.69	*67.56*	*0.76*
LSTM	68.91	67.92	66.45	*67.76*	*1.24*
GRU+At	68.97	66.35	66.32	*67.21*	*1.52*
LSTM+At	69.00	64.54	64.15	*65.90*	*2.69*
Joint Model					
GRU+At	68.29	63.21	66.85	*66.12*	*2.62*
LSTM+At	68.27	62.31	65.74	*65.44*	*2.99*
GRU&LSTM+At	69.21	68.35	69.74	***69.10***	*0.70*
Mean	*67.03*	*65.58*	*67.07*		
STD	*3.68*	*1.87*	*1.63*		

Conclusion

- We find that joint GRU & LSTM attention model overall produces high scores across three language models.
- However, the top score of 70.24 is achieved with a domain specific language model with a GRU based neural network.

The main contributions of this work are:

1. a comprehensive analysis of various neural network architectures and their performance on this task
2. a model formulation for including part-of-speech information in the input
3. a new top score on the CoNLL-2010 Wikipedia dataset

Email your questions to: dkaterenchuk@gradcenter.cuny.edu
levitan@sci.brooklyn.cuny.edu