International Institute of Information Technology, Hyderabad

HinDisSent

Natural Language Processing Project

Ujwal Narayan
20171170
Saujas Vaduguru
20171098
Team: Cookie Monsters

March 22, 2020

Content

Need for Sentence Embedding

We have word embeddings but is that alone enough?

Capturing meaning beyond words

Some tasks that would greatly benefit from sentence embeddings.

- Sentiment Analysis
- Summarization and Paraphrasing
- Natural Language Inference

Introduction

Supervised Learning of Universal Sentence Representations from Natural Language Inference Data - Conneau et al.

Goal
Generate sentence embeddings by trying to classify sentence pairs on the Natural Language Inference task.

Generating Universal Embeddings

Two questions.

1. **What Training Task?**
   Needs to be universal, and capture the underlying meaning.
2. **What Architecture?**
   Needs to be able to perform and generalize well.

For the first question, we need an NLP equivalent for *ImageNet*.

Natural Langauge Inference
High-level understanding task that involves reasoning about the semantic relationships within sentences.

What is Natural Language Inference?

Given two sentence pairs, S_1 and S_2, check whether the hypothesis(S_2) is true given the premise(S_1).

- **True**: Entailment
- **False**: Contradiction
- **Undetermined**: Neutral

Natural Language Inference
Some Examples

Entailment

Premise: A soccer game with multiple males playing.
Hypothesis: Some men are playing a sport.

Contradiction

Premise: Ujwal is drawing dependency trees.
Hypothesis: Ujwal is sleeping.

Neutral

Premise: Saujas is in Bangalore.
Hypothesis: Saujas likes Deep Learning.

Other Sentence Embeddings

Not the first time somebody tried this.

Unsupervised approaches like *SkipThought* (Kiros et al.), *FastSent* (Hill et al.) exist but performance is not good enough for widespread adoption.

Supervised approaches performed even worse .
Learning visually grounded sentence representations - Kiela et al
A structured self-attentive sentence embedding. - Lin et al.

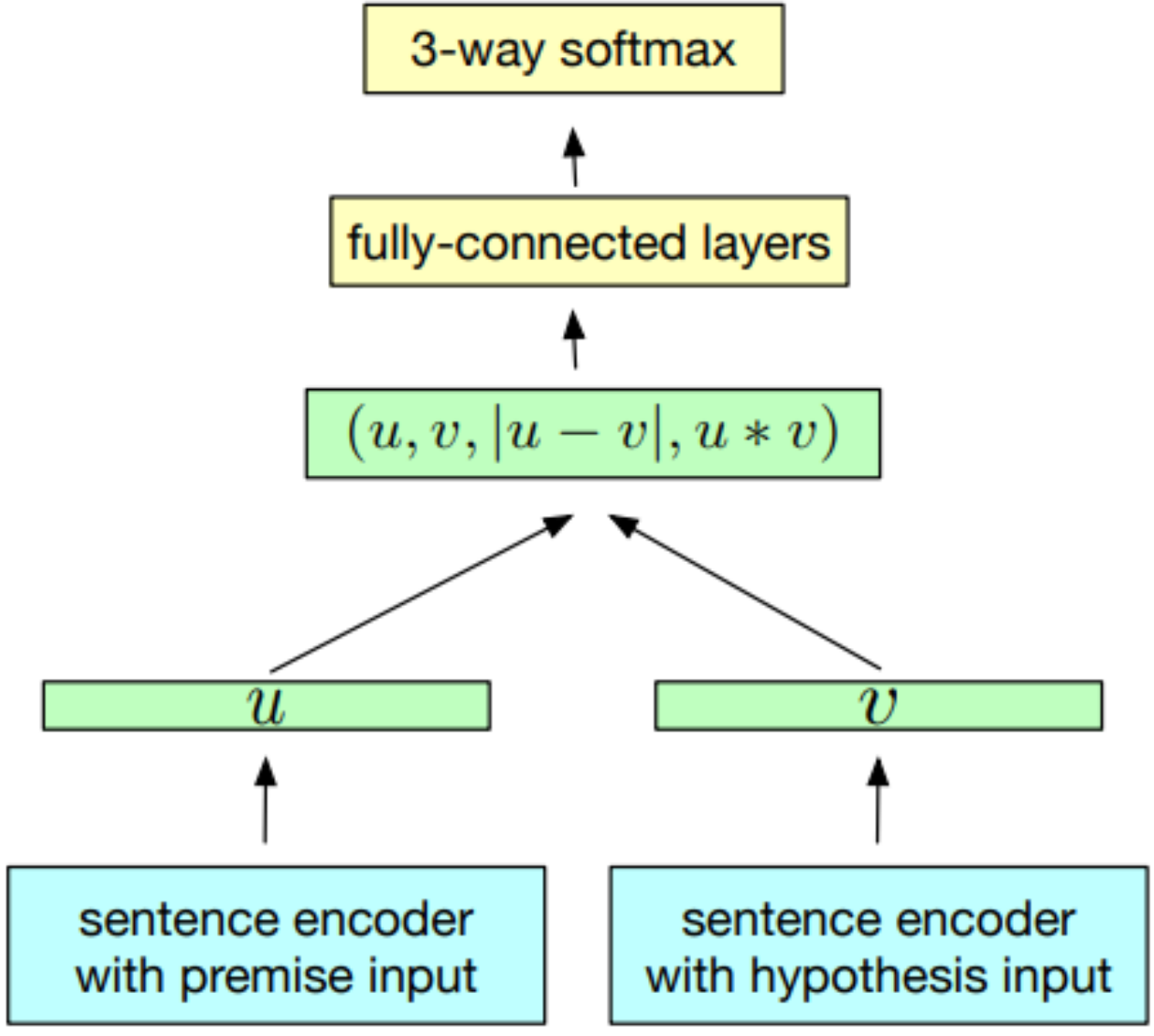


Figure: Task Setup

Architectures for Sentence Encoding

1. Standard LSTM.
2. Standard GRU.
3. Concatenation of last hidden states of forward and backward GRU.
4. Bi-LSTM with mean polling.
5. Bi-LSTM with max polling.
6. Bi-LSTM with Attention.
7. Hierarchical Conv Nets.

Evaluation Tasks

- Sentiment Analysis
- Entailment and Semantic Relatedness
- Semantic Textual Similarity
- Paraphrase Detection
- Caption Image Retrieval
- Question Type
- Opinion Polarity
- Subjectivity/ Objectivity

Results
Architecture

Bi-LSTM with Max polling performs the best.
Other candidates like Bi-LSTM with attention and Bi-LSTM with mean polling does not perform as well due to the incorporation of task specific biases and the lack of discriminatory capabilities respectively.

Unsupervised representation training (unordered sentences)										
Unigram-TFIDF	73.7	**79.2**	90.3	82.4	-	85.0	73.6/81.7	-	-	.58/.57
ParagraphVec (DBOW)	60.2	66.9	76.3	70.7	-	59.4	72.9/81.1	-	-	.42/.43
SDAE	74.6	78.0	90.8	86.9	-	78.4	**73.7**/80.7	-	-	.37/.38
SIF (GloVe + WR)	-	-	-	-	82.2	-	-	-	**84.6**	**.69**/ -
word2vec BOW†	77.7	79.8	90.9	88.3	79.7	83.6	72.5/81.4	0.803	78.7	.65/.64
fastText BOW†	78.3	81.0	**92.4**	87.8	**81.9**	84.8	**73.9/82.0**	0.815	78.3	.63/.62
GloVe BOW†	**78.7**	78.5	91.6	87.6	79.8	83.6	72.1/80.9	0.800	78.6	.54/.56
GloVe Positional Encoding†	78.3	77.4	91.1	87.1	80.6	83.3	72.5/81.2	0.799	77.9	.51/.54
BiLSTM-Max (untrained)†	77.5	**81.3**	89.6	**88.7**	80.7	**85.8**	73.2/81.6	**0.860**	83.4	.39/.48
Unsupervised representation training (ordered sentences)										
FastSent	70.8	78.4	88.7	80.6	-	76.8	72.2/80.3	-	-	**.63/.64**
FastSent+AE	71.8	76.7	88.8	81.5	-	80.4	71.2/79.1	-	-	.62/.62
SkipThought	76.5	80.1	93.6	87.1	82.0	**92.2**	**73.0/82.0**	**0.858**	82.3	.29/.35
SkipThought-LN	**79.4**	**83.1**	**93.7**	**89.3**	82.9	88.4	-	**0.858**	79.5	.44/.45
Supervised representation training										
CaptionRep (bow)	61.9	69.3	77.4	70.8	-	72.2	73.6/81.9	-	-	.46/.42
DictRep (bow)	76.7	78.7	90.7	87.2	-	81.0	68.4/76.8	-	-	**.67/.70**
NMT En-to-Fr	64.7	70.1	84.9	81.5	-	82.8	69.1/77.1	-		.43/.42
Paragram-phrase	-	-	-	-	79.7	-	-	0.849	83.1	**.71**/ -
BiLSTM-Max (on SST)†	(*)	83.7	90.2	89.5	(*)	86.0	72.7/80.9	0.863	83.1	.55/.54
BiLSTM-Max (on SNLI)†	79.9	84.6	92.1	**89.8**	83.3	**88.7**	75.1/82.3	**0.885**	**86.3**	.68/.65
BiLSTM-Max (on AllNLI)†	**81.1**	**86.3**	**92.4**	**90.2**	**84.6**	88.2	**76.2/83.1**	**0.884**	**86.3**	**.70/.67**
Supervised methods (directly trained for each task – no transfer)										
Naive Bayes - SVM	79.4	81.8	93.2	86.3	83.1	-	-	-	-	-
AdaSent	83.1	86.3	95.5	93.3	-	92.4	-	-	-	-
TF-KLD	-	-	-	-	-	-	80.4/85.9	-	-	-
Illinois-LH	-	-	-	-	-	-	-	-	84.5	-
Dependency Tree-LSTM	-	-	-	-	-	-	-	0.868	-	-

Figure: Comparison on various tasks and models

Results
Comparison with encoding size

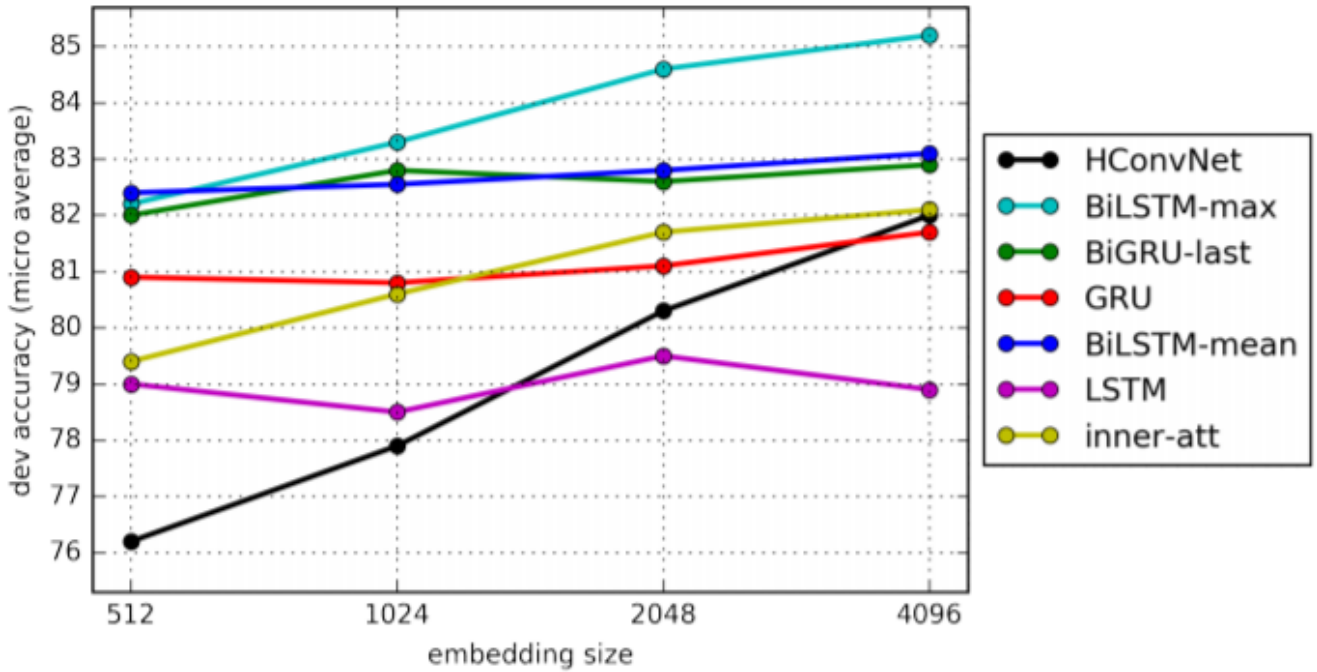


Figure: Embedding Sizes and the accuracy

Results
Extensions

Trained with MultiGenre-NLI dataset.
Performance on the Sentiment Analysis of Product Reviews task increased massively, bringing it up to SOTA standards.
Did not degrade the performance on the other tasks

HinDisSent

Inspired by *DisSent: Sentence Representation Learning from Explicit Discourse Relations* - Nie et al.

Introduction

NLI needs expensive data.
Such quality data may not be available for low resource languages .

Can we use some other kind of inter-sentence relations to get better universal embeddings?

Discourse Markers

What are discourse markers

A discourse marker is a word or a phrase that plays a role in managing the flow and structure of discourse.
Example: *because*, *since* etc.

Why choose this?

- Small and closed subset.
- Explicitly marked.

Task

इस क्षेत्र की व्यवस्था पहले बहुत अच्छी नहीं थीं __ यह पहले कभी किसी राज्य का अभिन्न अंग नहीं रहा था

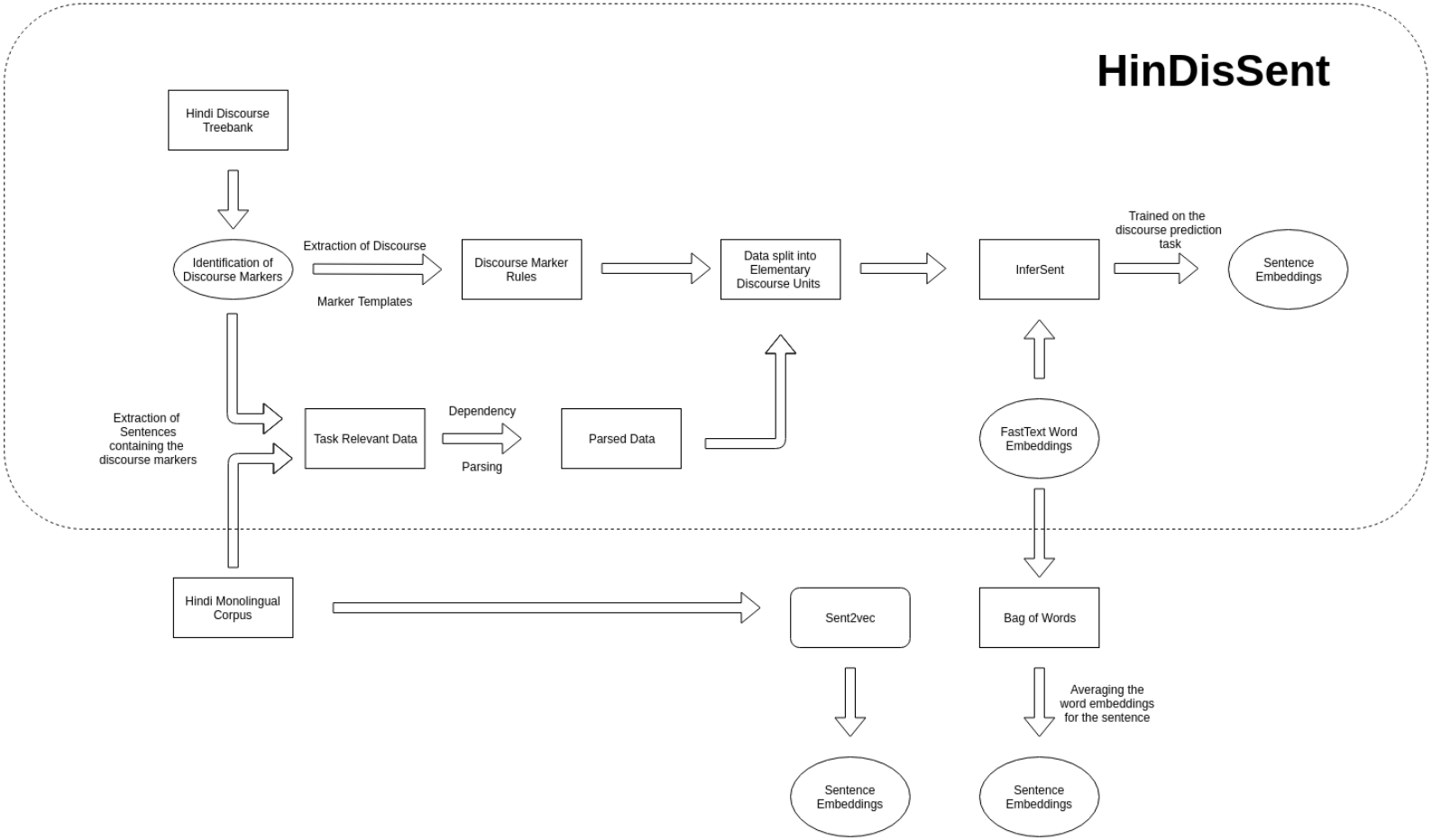
we can complete the sentence with क्योंकि.

Dataset Stats

- Most frequent discourse markers are chosen from the Hindi Discourse Treebank
- Data taken from IIT-B Hindi monolingual corpus.
- Due to lack of compute, we chose to take only 10% giving nearly 47,00,000 sentences, spanning across BBC and MonoCorp domains
- Extracting the sentences with the relevant discourse markers lead to 3,64,076 sentence pairs

Count of Discourse Markers	Discourse Marker
650	इसलिए
282486	और
8413	क्योंकि
6401	जबकि
22781	तथा
13210	तो
29920	लेकिन
215	हालांकि

HinDisSent



HinDisSent Architecture

Template Rules

Rules were manually created to extract the relevant clauses based on the discourse marker present.

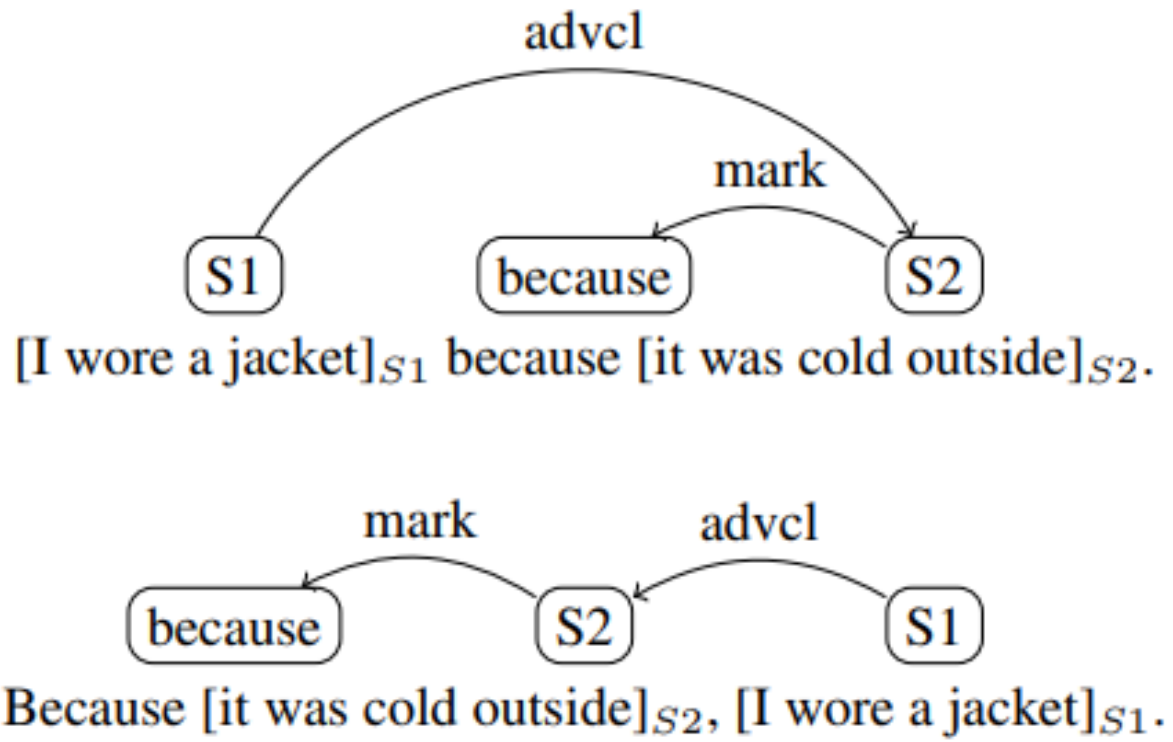


Figure: Template Rules

Hindi Templates


```
(और,cc,conj,,,)
(तथा,cc,conj,,,)
(इसके,obl,conj,,,,अलावा,इसके,case)
(इसके,obl,advcl,,,,अलावा,इसके,case)
(इसके,obl,IPS,,,,बाद,इसके,case)
(इसलिए,obl,acl,,,)
(इसलिए,,,obl,,nmod)
(इसलिए,,,obl,,acl)
(इसलिए,obl,conj,,,)
(इसलिए,obl,obj,,,)
(बाद,,,case,obl,)
(क्योंकि,mark,advcl,,,)
(क्योंकि,mark,obl,,,)
(क्योंकि,mark,nmod,,,)
(जबकि,cc,conj,,,)
(तो,mark,acl,,,)
(तो,mark,obj,,,)
(तो,mark,conj,,,)
(बल्कि,,advcl,mark,,)
(बल्कि,,acl:relcl,mark,,)
(लेकिन,cc,conj,,,)
(वहीं,,,obl,,acl:relcl)
(साथ,,,mark,advcl,,ही,साथ,dep)
(साथ,obl,advcl,,,,ही,साथ,dep)
(साथ,,,mark,obl,,ही,साथ,dep)
(हालांकि,,advcl,mark,,)
(हालांकि,,nmod,mark,,)
(हालांकि,mark,advcl,,,)
```

Figure: Hindi Templates

Examples of extracted sentence pairs

S_1	DM	S_2
अंग्रेज तो वह साहस नही कर पाए	लेकिन	मालवीय जी की आशंका स्वतंत्र भारत में सही साबित हुई
प्रेम को हम आदर्शवाद के रूप में चुनते	जबकि	नफरत हमारे अन्दर एक प्रवृति के रूप में होती है
जब केवल पंचगंगा घाट में ही देव दीपावली मनाई जाती थी	और	बगल के दुर्गाघाट में दुर्गाघाटी

Experiments

Models Trained

InferSent encoder with Tied Weights

- Using only the top 5 discourse marker
- Using the top 8 discourse markers

Baselines

- Sent2vec
- Mean of FastText BOW

Downstream Evaluation Task

- *Aspect Based Sentiment Analysis in Hindi: Resource Creation and Evaluation* - Akhtar et al.

Sentiment Analysis

We train a multitude of classifiers for the Sentiment Analysis Task.

- KNN
- Linear SVM
- Kernel SVM
- LR
- Gaussian Process
- Decision Tree
- Random Forests
- Multi Layer Perceptrons
- AdaBoost
- QDA

Results
Mean FastText BoW

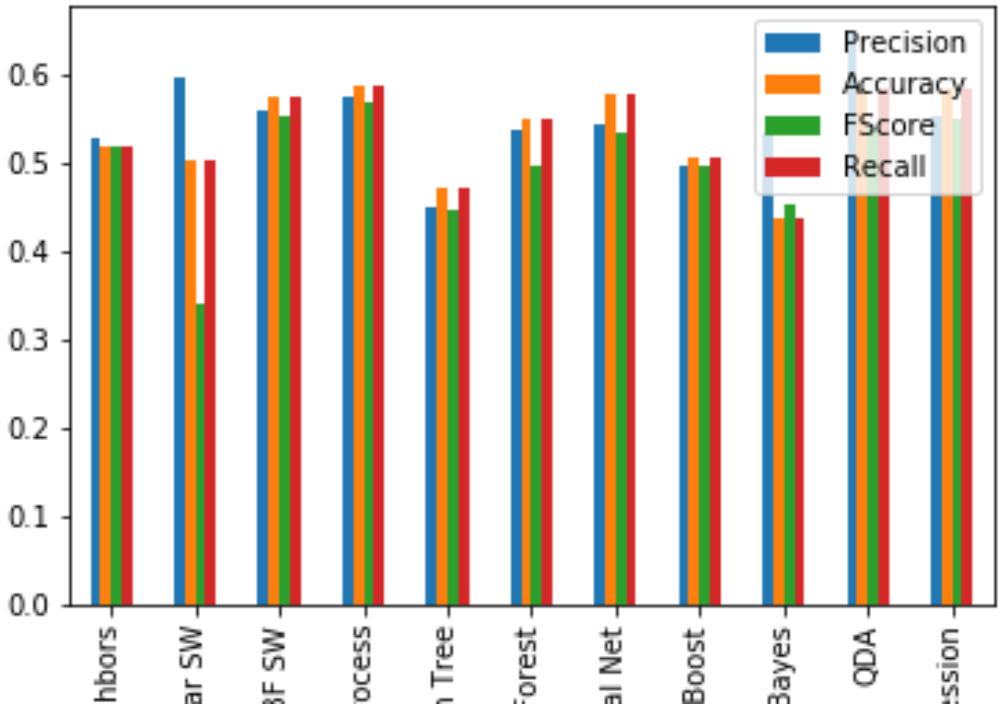


Figure: Accuracy on Mean FastText BoW for Sentiment Analysis

Results
Sent2Vec

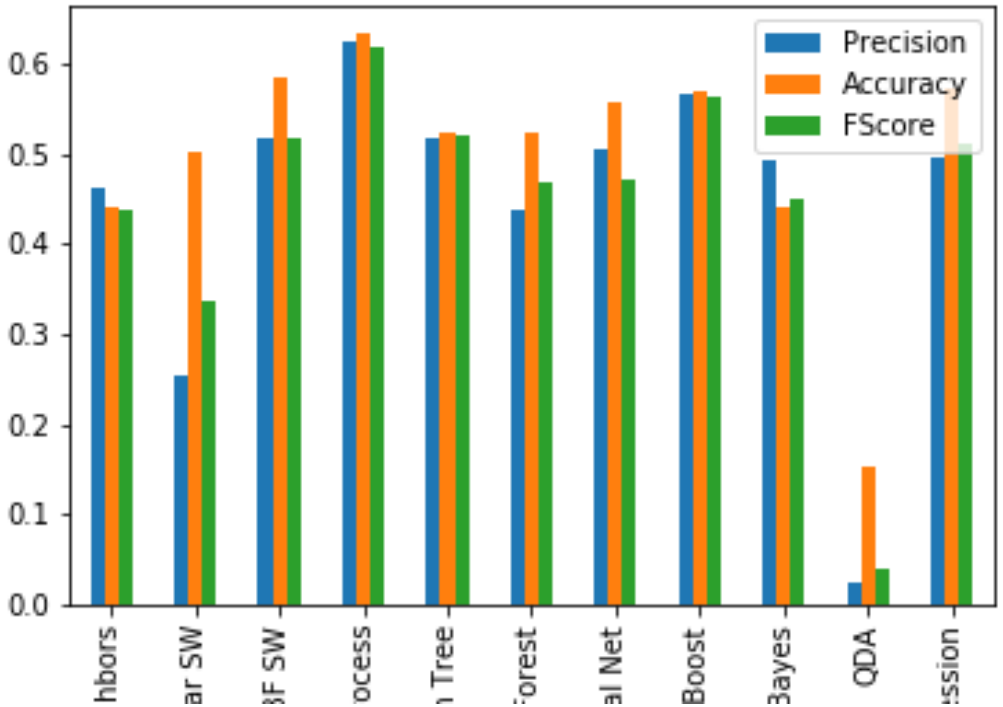


Figure: Accuracy on Sent2Vec for Sentiment Analysis

Results
HinDisSent: Top 5

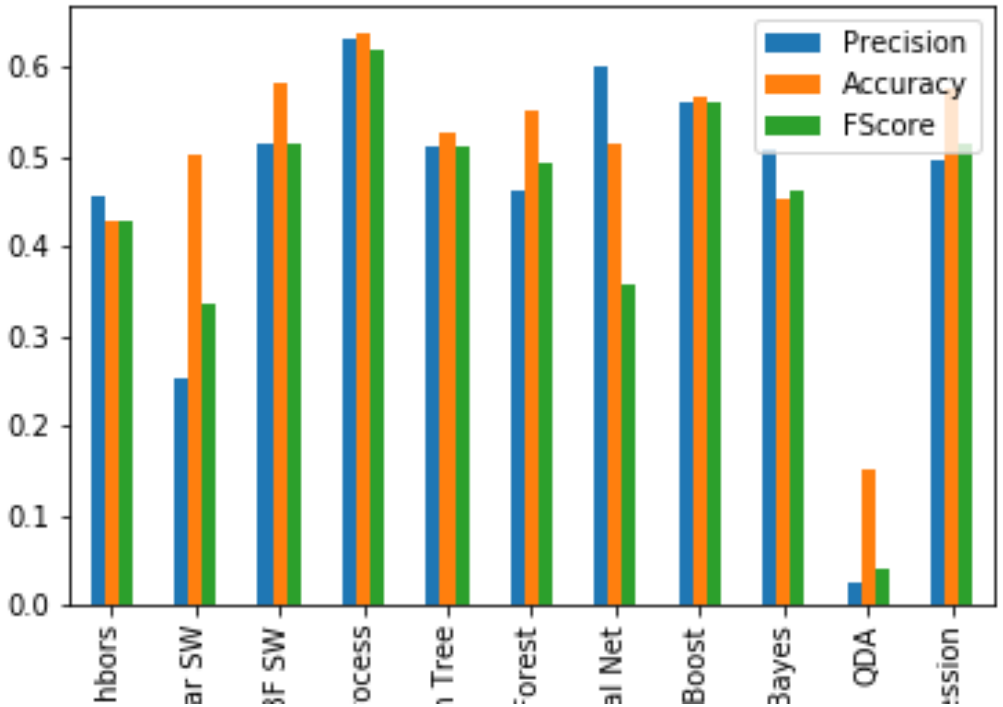


Figure: Accuracy on *HindDisSent*$_5$ for Sentiment Analysis

Results
HinDisSent: Top 8

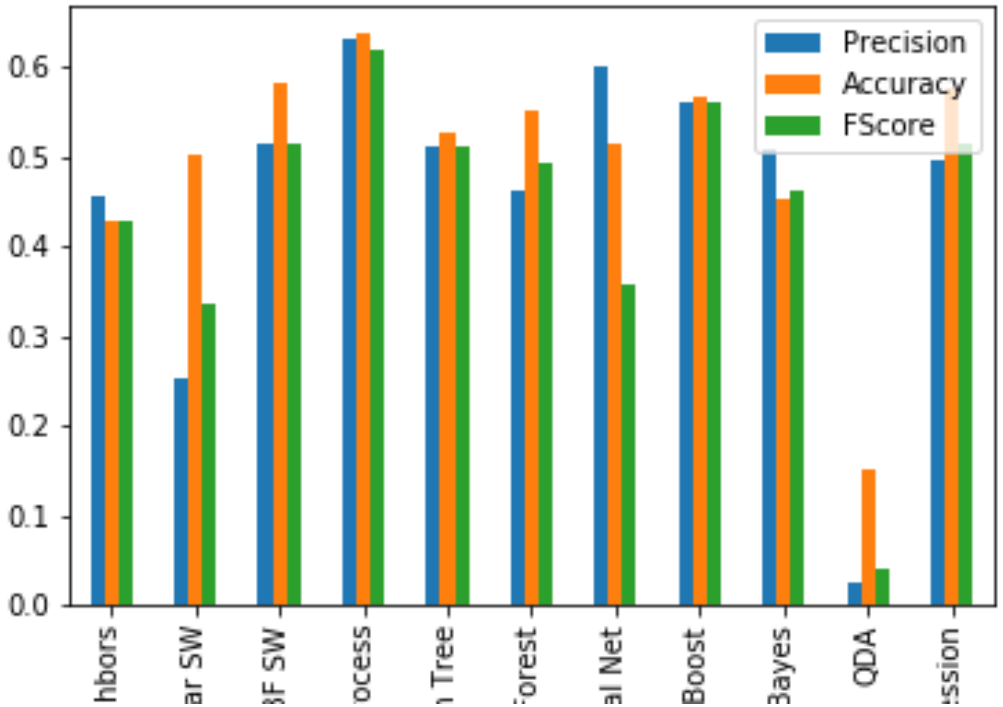


Figure: Accuracy on *HindDisSent*$_8$ for Sentiment Analysis

Conclusions
Sentence size and Representation

- Highest accuracy for smaller sentences.
- Accuracy for LR model takes a steep dive dropping from 70 to 60 from the first bucket to the second, stabilising at
- Gaussian Process is fairly consistent, with accuracy showing a very small decrease(1%) every bucket.

Conclusions
OP Baselines

A simple but tough-to-beat baseline for sentence embeddings - Arora et al.

Sent2vec had access to nearly 13 times the amount of data of HinDisSent

Conclusions

`Less data, but better representations`

Future Work

- Use more data.
- Use more discourse markers.
- Evaluate against other SentEval tasks, and compare performances
- Evaluate with other types of encoders.
- More experimentation with hyperparameters.
- Better filtering out of non discourse entities.

Questions?