Subtl.ai at the FinSBD-2 task: Document Structure Identification by Paying Attention

The 2nd Workshop on Financial Technology and Natural Language Processing, 2020

Aman Khullar*, Abhishek Arora*, Sarath Chandra Pakala*, Vishnu Ramesh*, Manish Shrivastava

Subtl.ai, CIE, IIIT Hyderabad

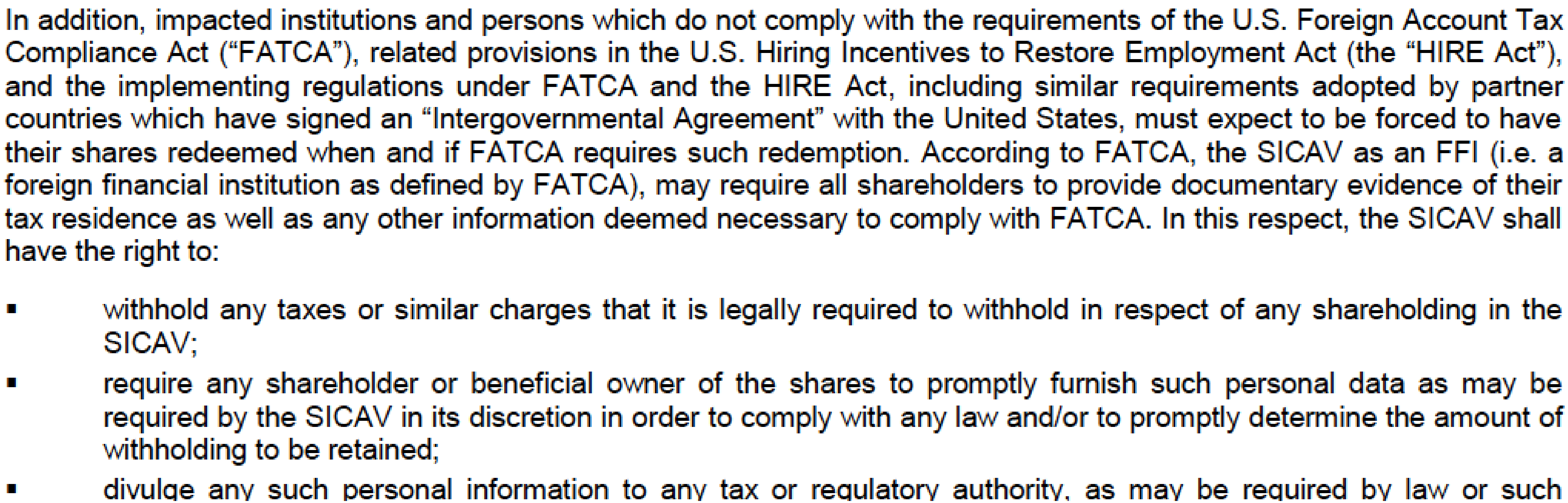
* = equal contribution

What is Sentence Boundary Disambiguation?

Sentence

List

Item



In addition, impacted institutions and persons which do not comply with the requirements of the U.S. Foreign Account Tax Compliance Act ("FATCA"), related provisions in the U.S. Hiring Incentives to Restore Employment Act (the "HIRE Act"), and the implementing regulations under FATCA and the HIRE Act, including similar requirements adopted by partner countries which have signed an "Intergovernmental Agreement" with the United States, must expect to be forced to have their shares redeemed when and if FATCA requires such redemption. According to FATCA, the SICAV as an FFI (i.e. a foreign financial institution as defined by FATCA), may require all shareholders to provide documentary evidence of their tax residence as well as any other information deemed necessary to comply with FATCA. In this respect, the SICAV shall have the right to:

▪ withhold any taxes or similar charges that it is legally required to withhold in respect of any shareholding in the SICAV;

▪ require any shareholder or beneficial owner of the shares to promptly furnish such personal data as may be required by the SICAV in its discretion in order to comply with any law and/or to promptly determine the amount of withholding to be retained;

▪ divulge any such personal information to any tax or regulatory authority, as may be required by law or such

Motivation

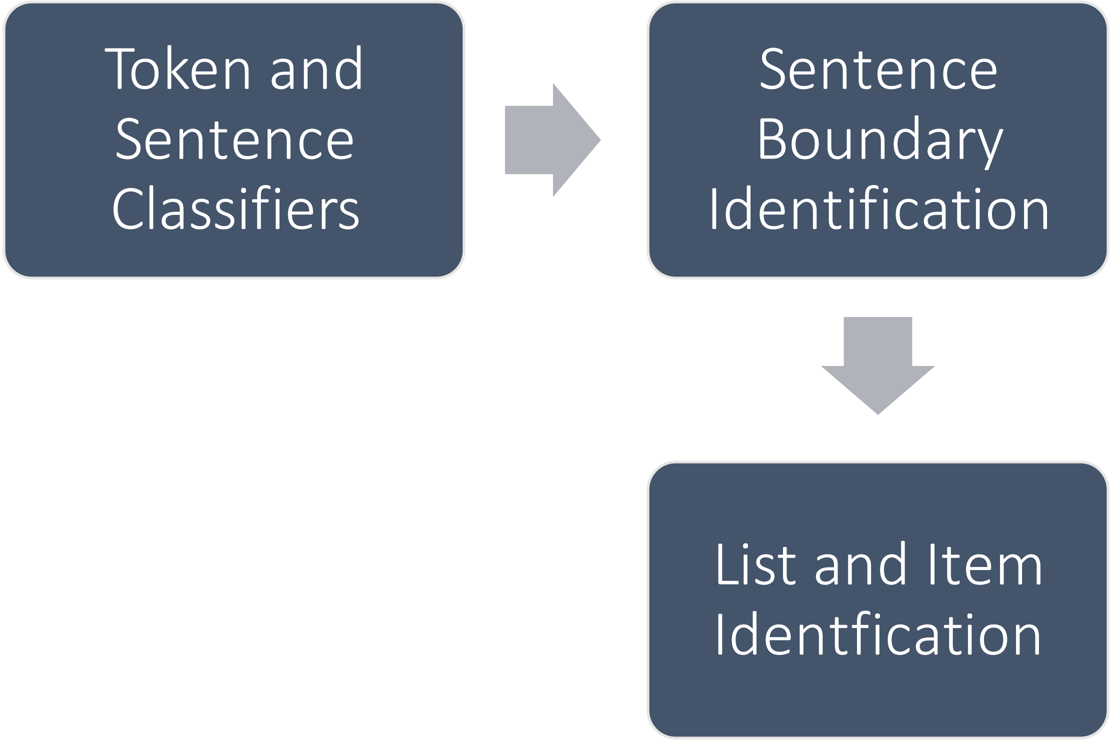
- Under-studied problem for complex PDF documents
- 1st step of Subtl's question-answering pipeline
- Evaluating DL + heuristics approach

Dataset

Provided as part of FinSBD-2 2020 Challenge

	TRAINING SET	VALIDATION SET	TEST SET
No. Of Documents	5	1	2
No. Of Characters	1,322,767	169,546	558,611
No. Of Tokens	290,092	39,816	141,138
No. Of Sentences	7,282	788	2,450
No. Of Lists	207	42	69
No. Of Items	843	268	332
No. Of OOV tokens	1,079	248	443

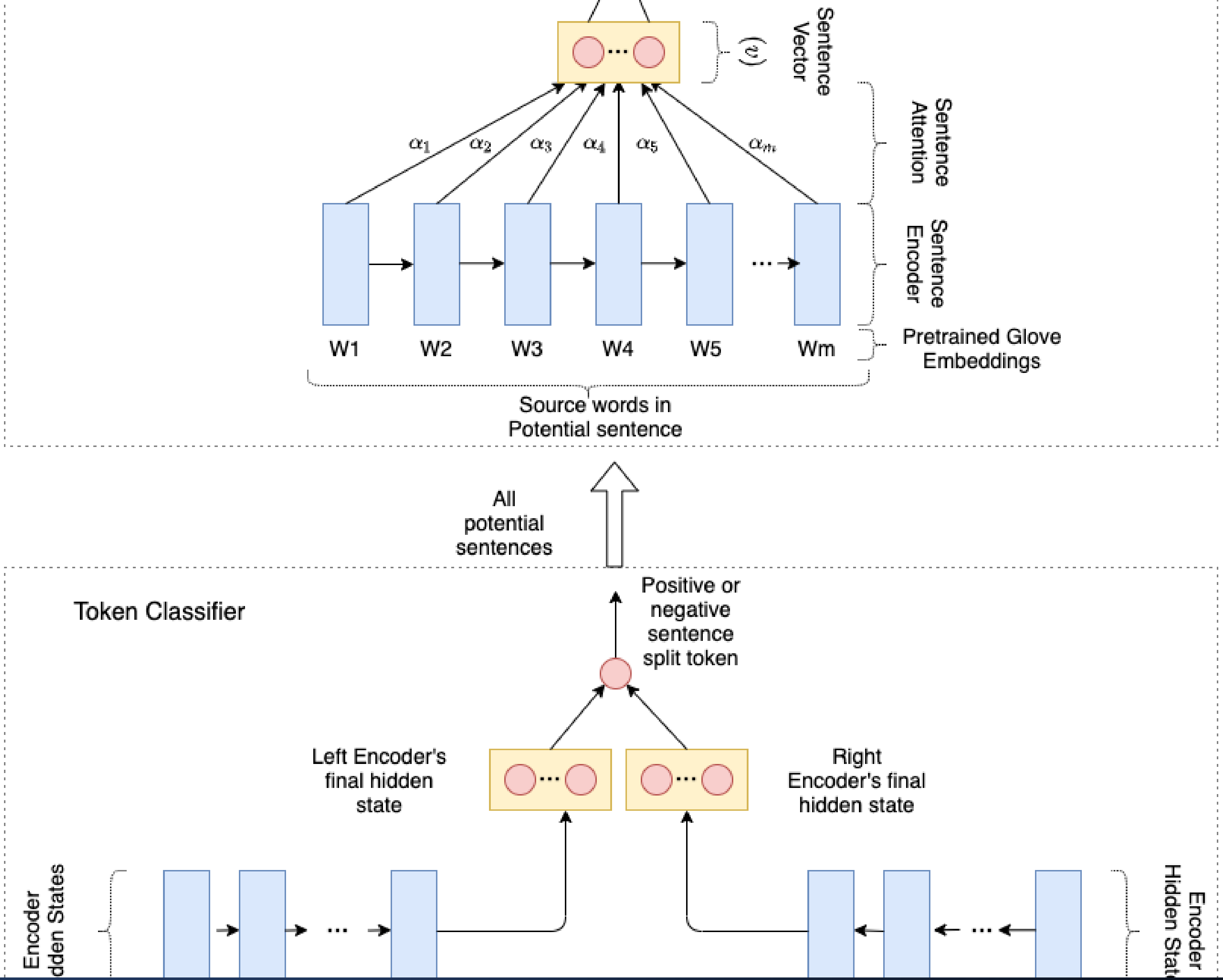
Methodology



Token and Sentence Classifiers

Sentence Boundary Identification

List and Item Identfication

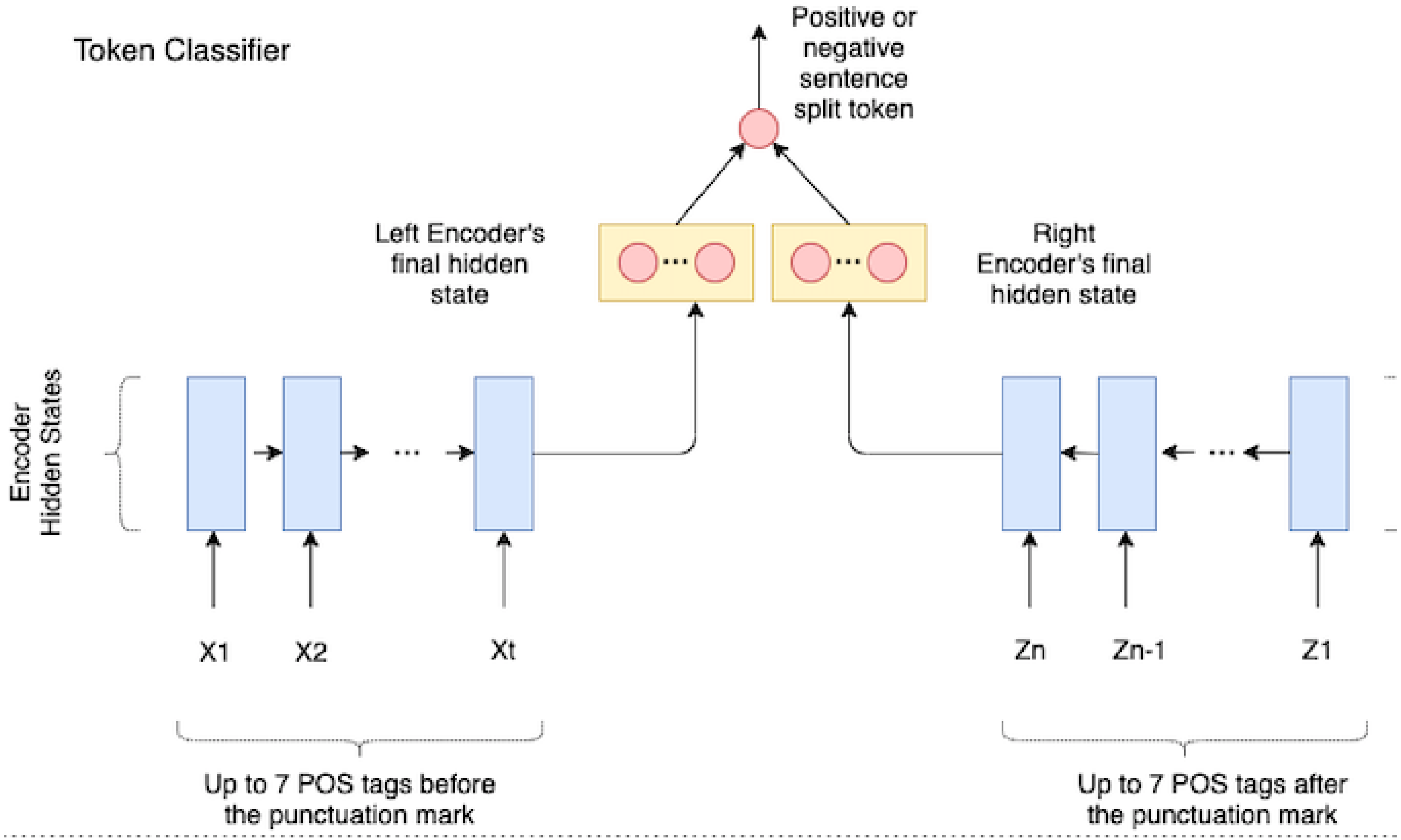


Our Models

- Token Classifier
- Sentence Classifier

Token Classifier

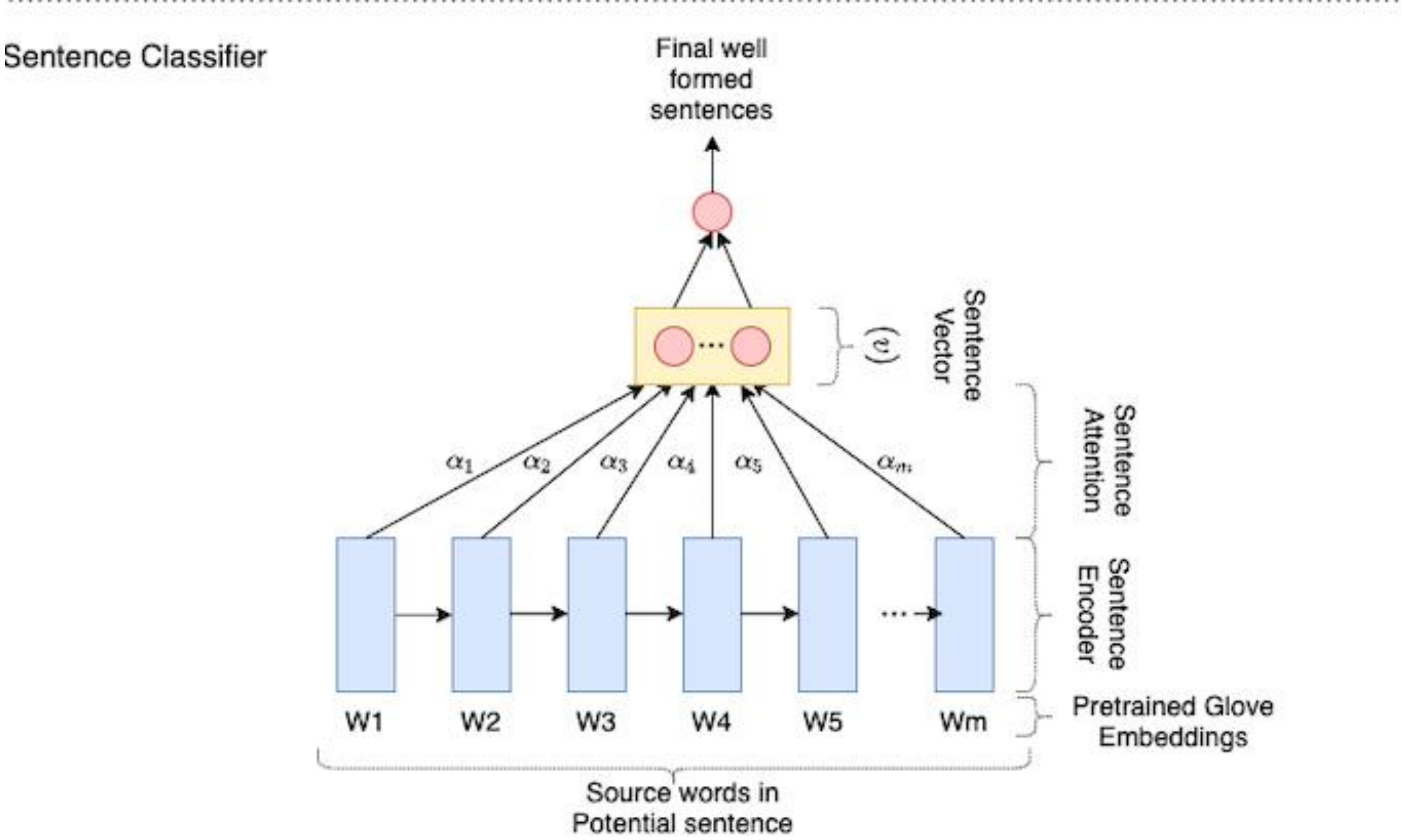
- Potential split points - '\n', ':', '.'
- Convert the previous 7 tokens and the next 7 tokens into their corresponding POS tags
- Convert these POS tags into their one hot vectors
- Pass the previous 7 one hot POS encodings into a forward directional LSTM
- Pass the next 7 one hot POS encodings into a backward directional LSTM
- Concatenate the final hidden states of these two LSTM encoders and pass a linear classifier layer



Token Classifier
Positive or negative sentence split token
Left Encoder's final hidden state
Right Encoder's final hidden state
Encoder Hidden States
X1
X2
Xt
Zn
Zn-1
Z1
Up to 7 POS tags before the punctuation mark
Up to 7 POS tags after the punctuation mark

Sentence Classifier

- Pretrained Glove word embeddings
- Attention based LSTM encoder
- Linear classifier layer



Sentence Boundary Identification

- Sentence classifier identifies if the sentence is well-formed
- If the sentence is well-formed, keep it
- Else merge previous and next sentence and pass through the sentence classifier

Lists and Item Identification

- Analysis of training data yielded 90% of items with alphanumeric pattern or non-ASCII characters like bullet points
- Items identified using the heuristics based on observations
- Items aggregated to identify the circumscribing list
- Window size of 7 sentences to identify starting sentence
- Sentence classifier used to identify the list end point

Metrics: Sentence Coverage

$$\text{Sentence Overlap \%} = \frac{\text{Len(Common substring)}}{\text{Len(Ground truth sentence)}}$$

$$\text{Average Overlap \%} = \frac{\Sigma\ \text{(Sentence Overlap)}}{\text{Total No. Of predicted sentences}}$$

- F1 Score becomes a strict metric for industry applications
- Sentence Coverage is a softer evaluation metric

Document	Class	Precision	Recall	F1 score
Document 1	Sentence	0.67	0.62	0.64
	List	0.00	0.00	0.00
	Items	0.00	0.00	0.00
	Average	0.22	0.20	0.21
Document 2	Sentence	0.71	0.62	0.66
	List	0.00	0.00	0.00
	Items	0.00	0.00	0.00
	Average	0.24	0.21	0.22

Results

Document	Class	Coverage
Document 1	Sentence	0.82
	List	0.20
	Items	0.16
	Average	0.39
Document 2	Sentence	0.91
	List	0.16
	Items	0.15
	Average	0.41

- Deep learning models outperform rule-based approach
- Encouraging results for industry applications

Conclusion

- Promising SBD results using Token + Sentence classifier
- Heuristics provide a starting point for list and item identification
- Softer sentence coverage evaluation metric for industry applications

THANK YOU