TDNN: A Two-stage Deep Neural Network for Prompt-independent Automated Essay Scoring

Cancan Jin[1] **Ben He**[1,3] **Kai Hui**[2] **Le Sun**[3,4]

[1]School of Computer & Control Engineering,
University of Chinese Academy of Sciences, Beijing, China

[2] SAP SE, Berlin, Germany

[3] Institute of Software, Chinese Academy of Sciences, Beijing, China

[4] Beijing Advanced Innovation Center for Language Resources, Beijing, China

jincancan15@mails.ucas.ac.cn, benhe@ucas.ac.cn
kai.hui@sap.com, sunle@iscas.ac.cn

Outline

- Background
- Method
- Experiments
- Conclusions

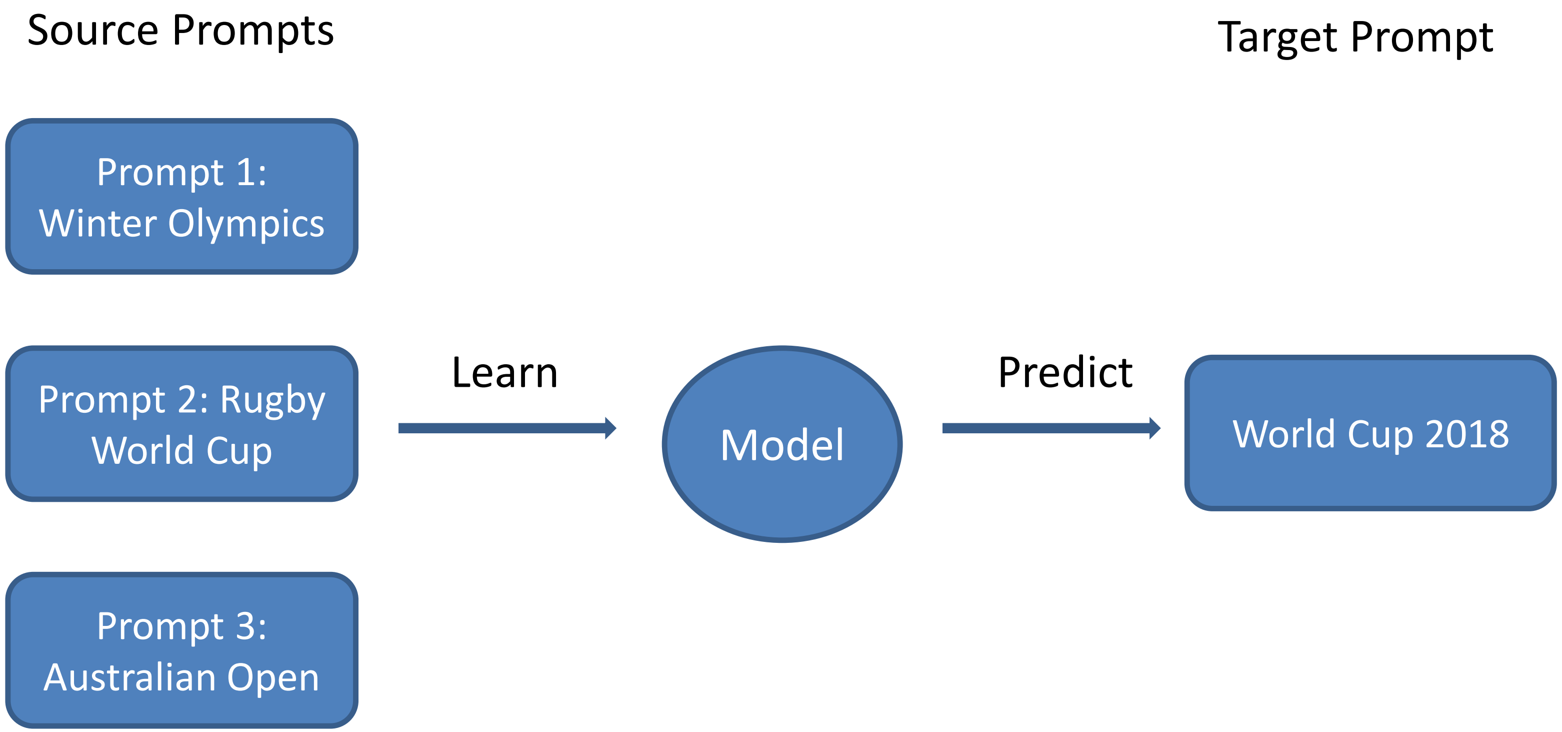
What is Automated Essay Scoring (AES)?

- Computer produces summative assessment for evaluation
- Aim: reduce human workload
- AES has been put into practical use by ETS from 1999

Prompt-specific and -Independent AES

- Most existing AES approaches are prompt-specific
  – Require human labels for each prompt to train
  – Can achieve satisfying human-machine agreement
    - Quadradic weighted kappa (QWK) > 0.75 [Taghipour & Ng, EMNLP 2016]
    - Inter-human agreement: QWK=0.754
- Prompt-independent AES remains a challenge
  – Only non-target human labels are available

Challenges in Prompt-independent AES



Challenges in Prompt-independent AES

Source Prompts

Target Prompt

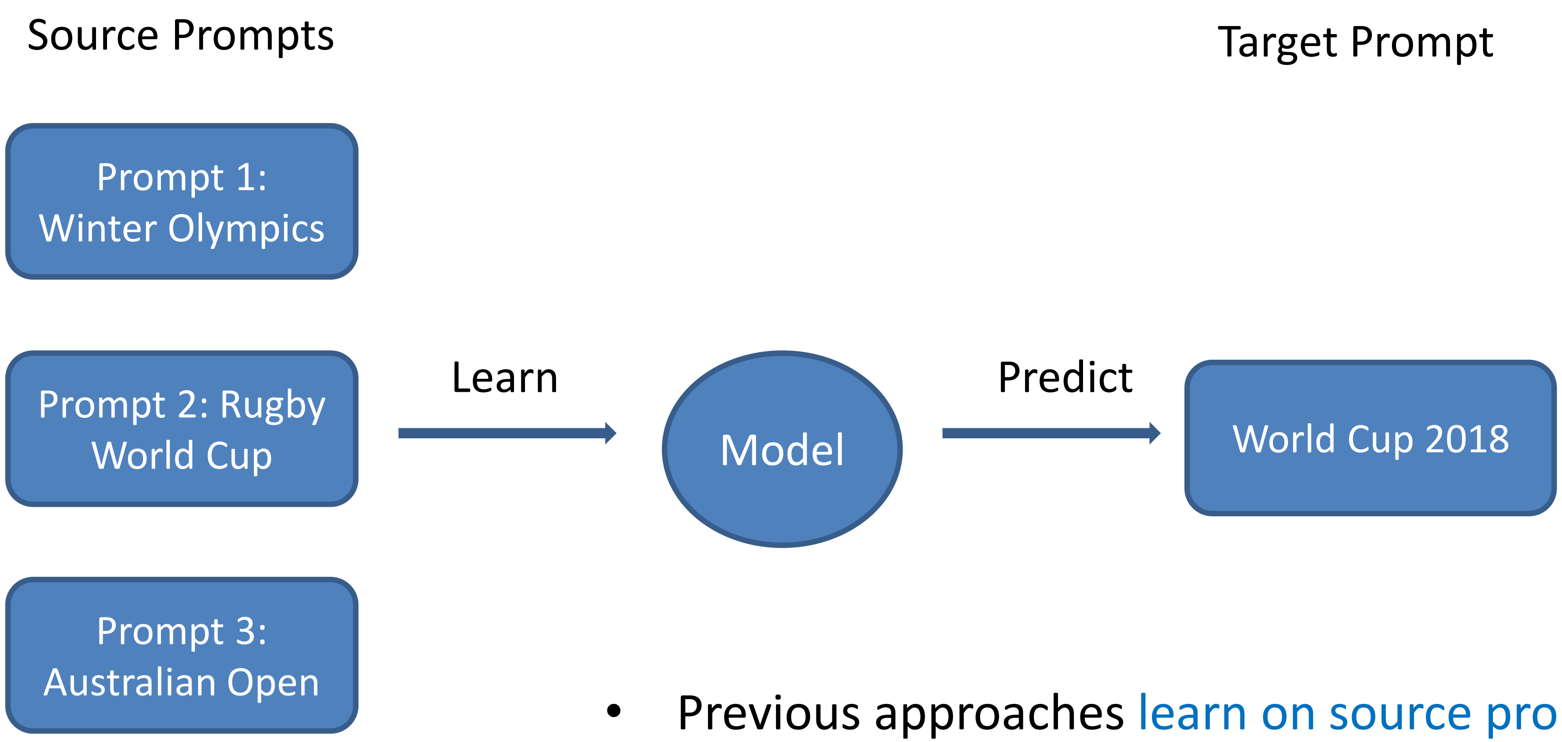
Prompt 1:
Winter Olympics

Prompt 2: Ru
World Cup

Prompt 3:
Australian Open

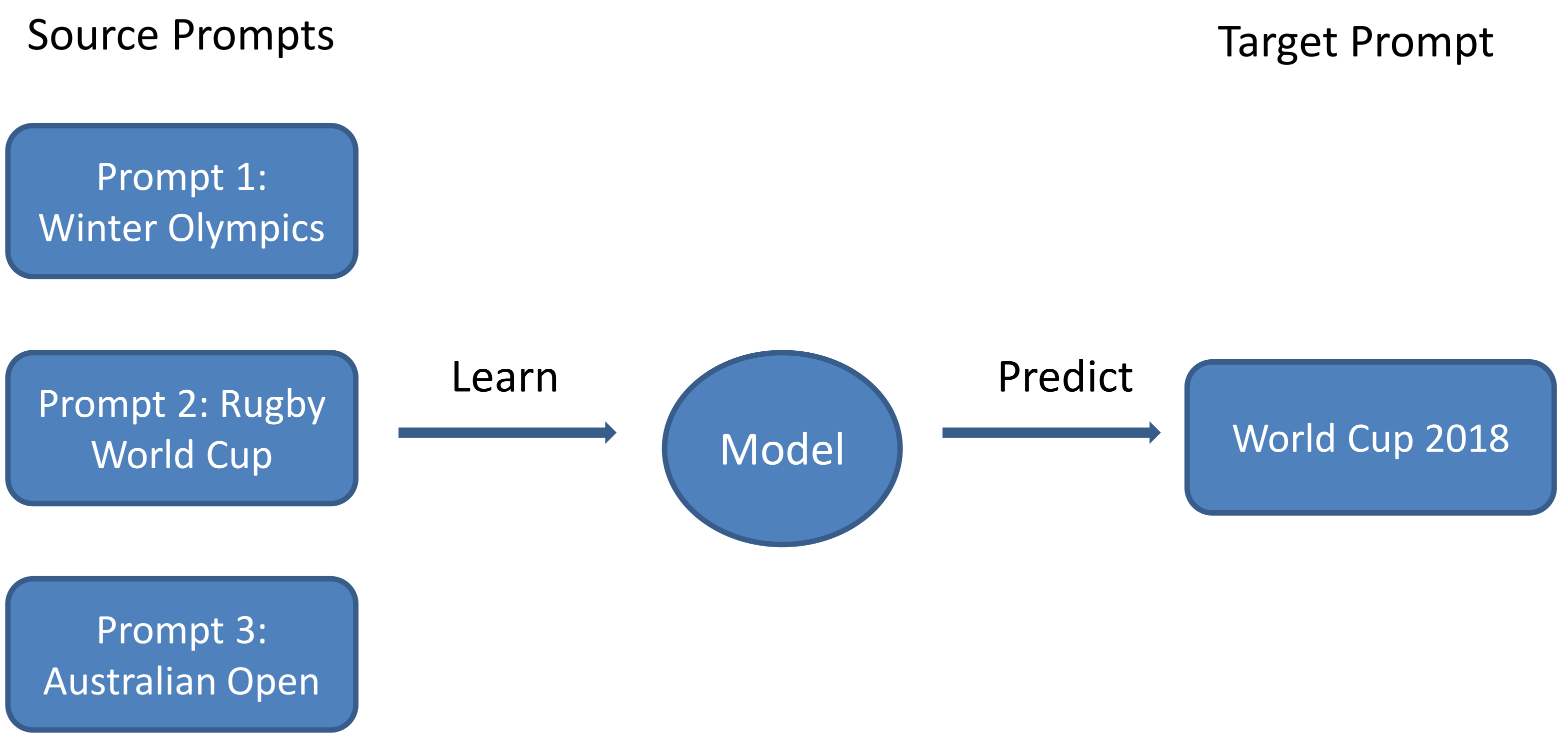
Unavailability of rated essays written for the target prompt

Challenges in Prompt-independent AES



- Previous approaches learn on source prompts
  - Domain adaption [Phandi et al. EMNLP 2015]
  - Cross-domain learning [Dong & Zhang, EMNLP 2016]
  - Achieved Avg. QWK = 0.6395 at best with up to 100 labeled target essays

Challenges in Prompt-independent AES



Off-topic: essays written for source prompts are mostly irrelevant

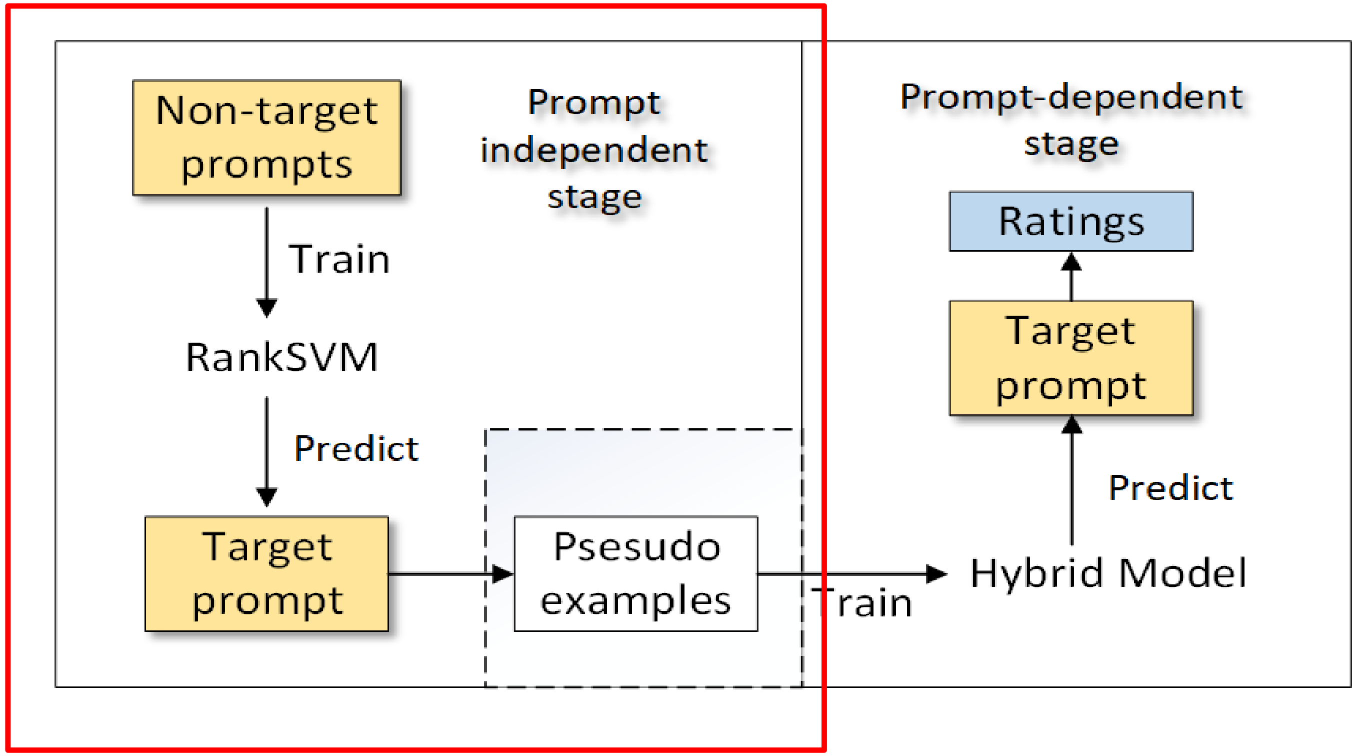
Outline

- Background
- Method
- Experiments
- Conclusions

TDNN: A Two-stage Deep Neural Network for Prompt-independent AES

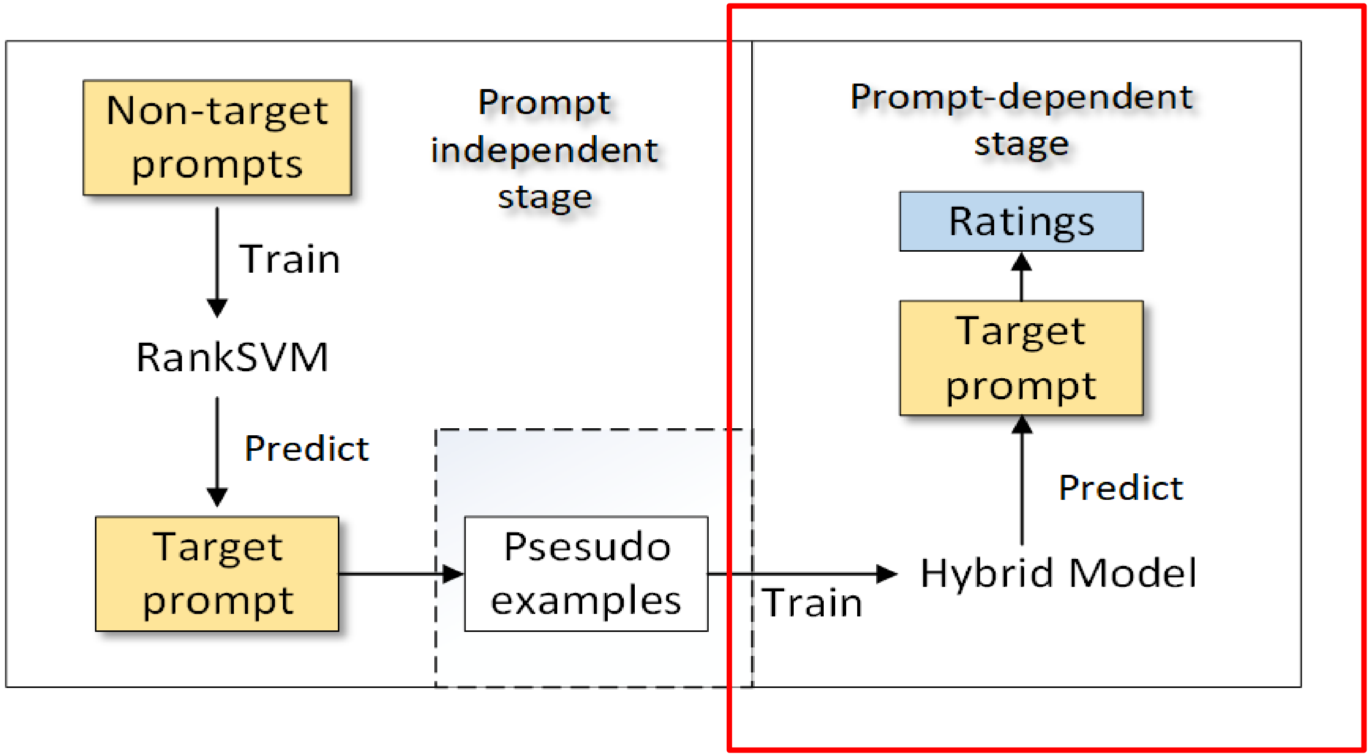
- Based on the idea of transductive transfer learning
- Learn on target essays
- Utilize the content of target essays to rate

The Two-stage Architecture



- Prompt-independent stage: train a shallow model to create pseudo labels on the target prompt

The Two-stage Architecture



- Prompt-dependent stage: learn an end-to-end model to predict essay ratings for the target prompts

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM for AES
  - Pre-defined prompt-independent features

- Select confident essays written for the target prompt

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM
  - Pre-defined prompt-independent features

- Select confident essays written for the target prompt

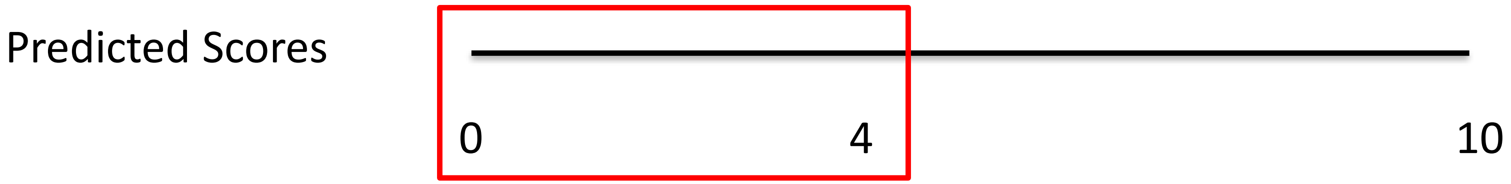
Predicted Scores

0 10

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM
  - Pre-defined prompt-independent features

- Select confident essays written for the target prompt



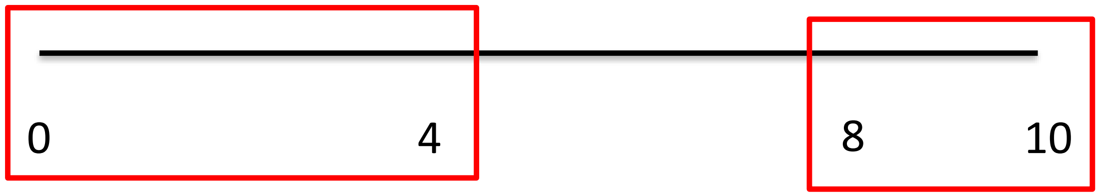
Predicted ratings in [0, 4] as negative examples

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM
  - Pre-defined prompt-independent features

- Select confident essays written for the target prompt

Predicted Scores

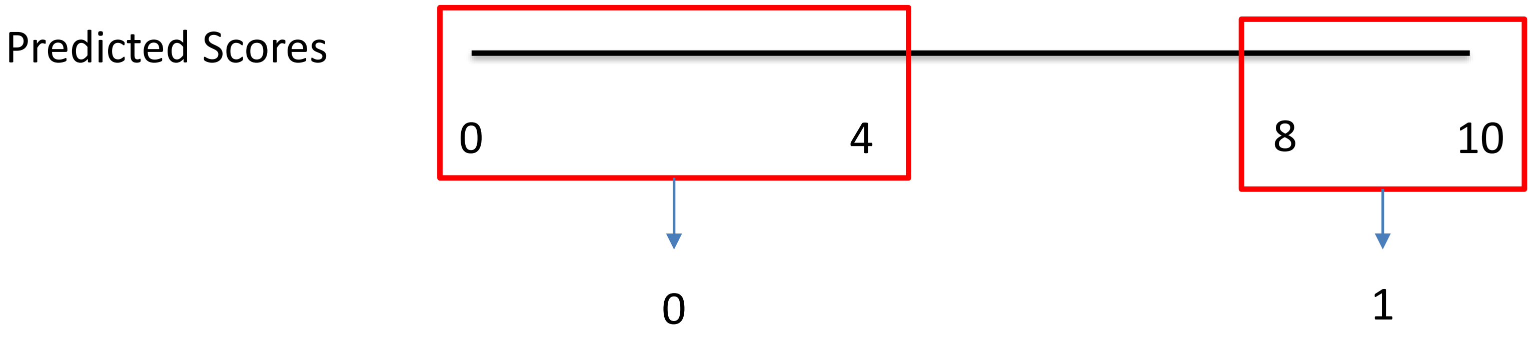


Predicted ratings in [8, 10] as positive examples

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM
  - Pre-defined prompt-independent features

- Select confident essays written for the target prompt

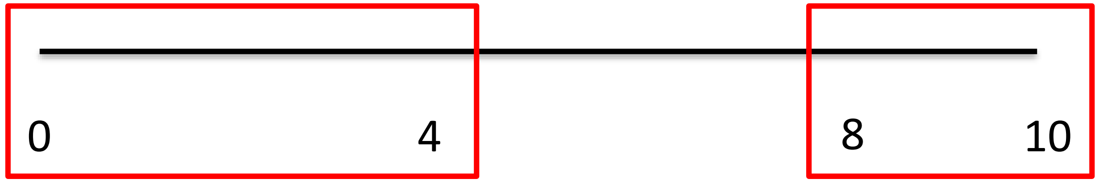


Converted to 0/1 labels

Prompt-independent stage

- Train a robust prompt-independent AES model
  - Using Non-target prompts
  - Learning algorithm: RankSVM
  - Pre-defined prompt-independent features

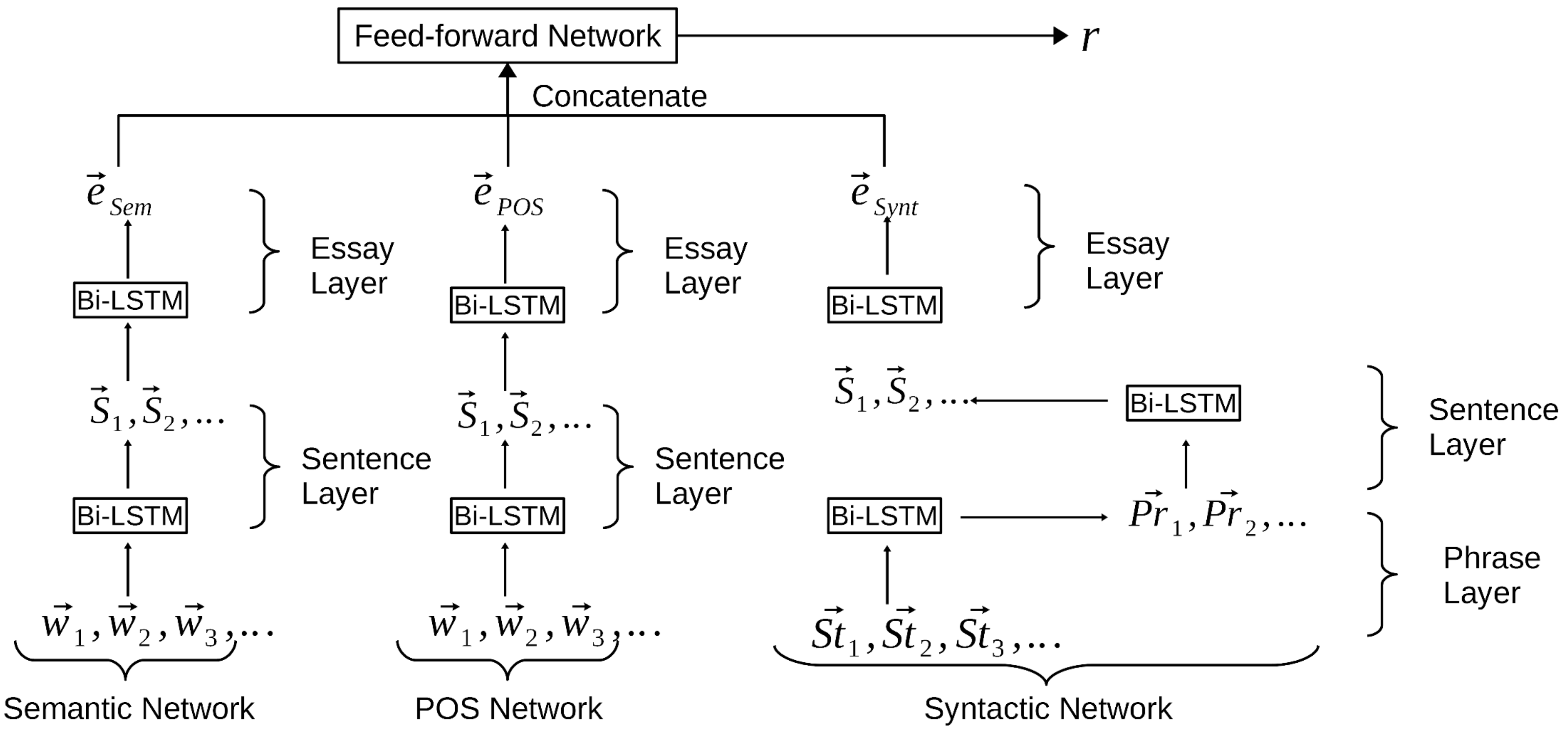
- Select confident essays written for the target prompt
  - Common sense: ≥8 is good, <5 is bad
  - Enlarge sample size



Prompt-dependent stage

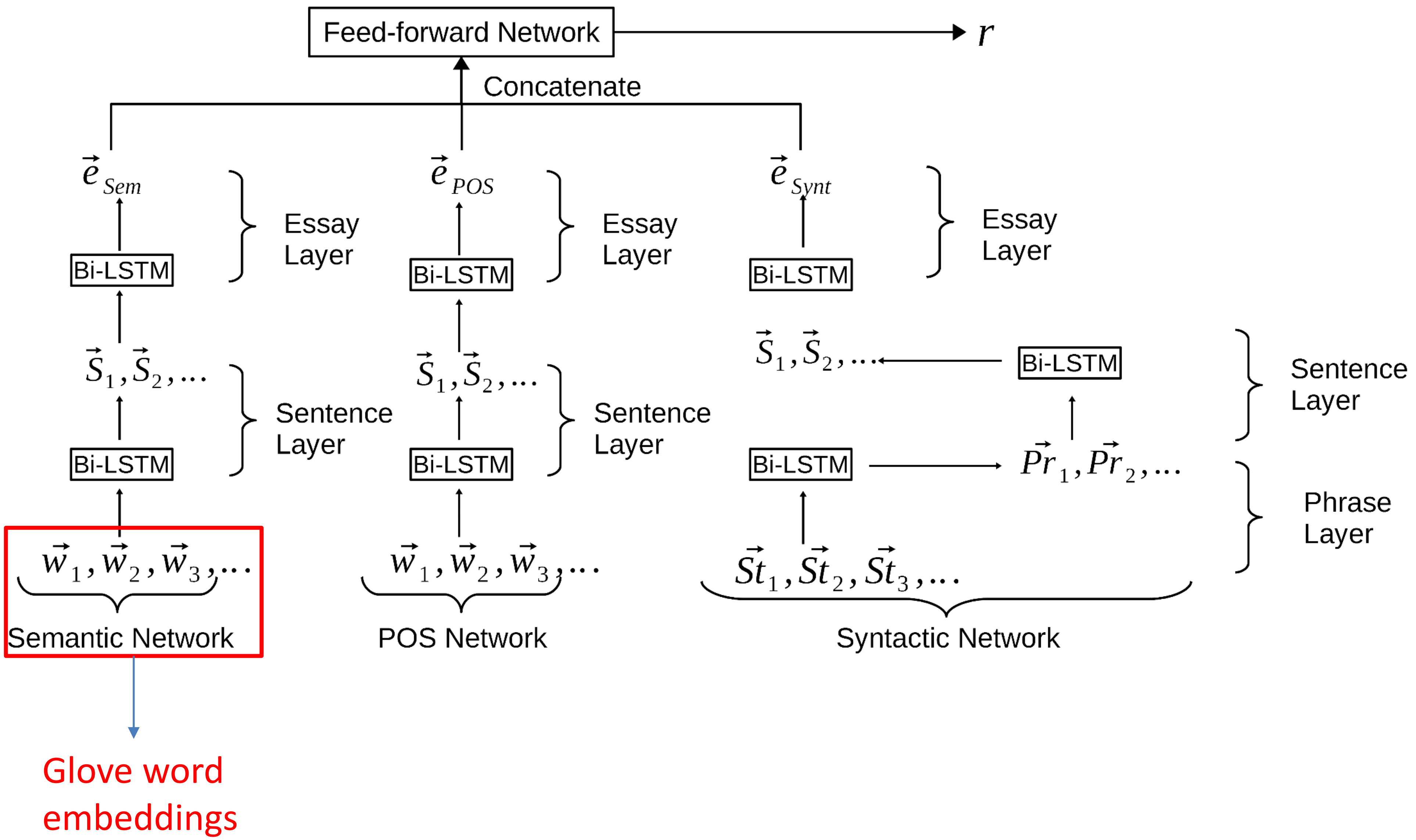
- Train a hybrid deep model for a prompt-dependent assessment
- An end-to-end neural network with three parts of inputs:
  - Word semantic embeddings
  - Part-of-speech (POS) taggings
  - Syntactic taggings

Architecture of the hybrid deep model

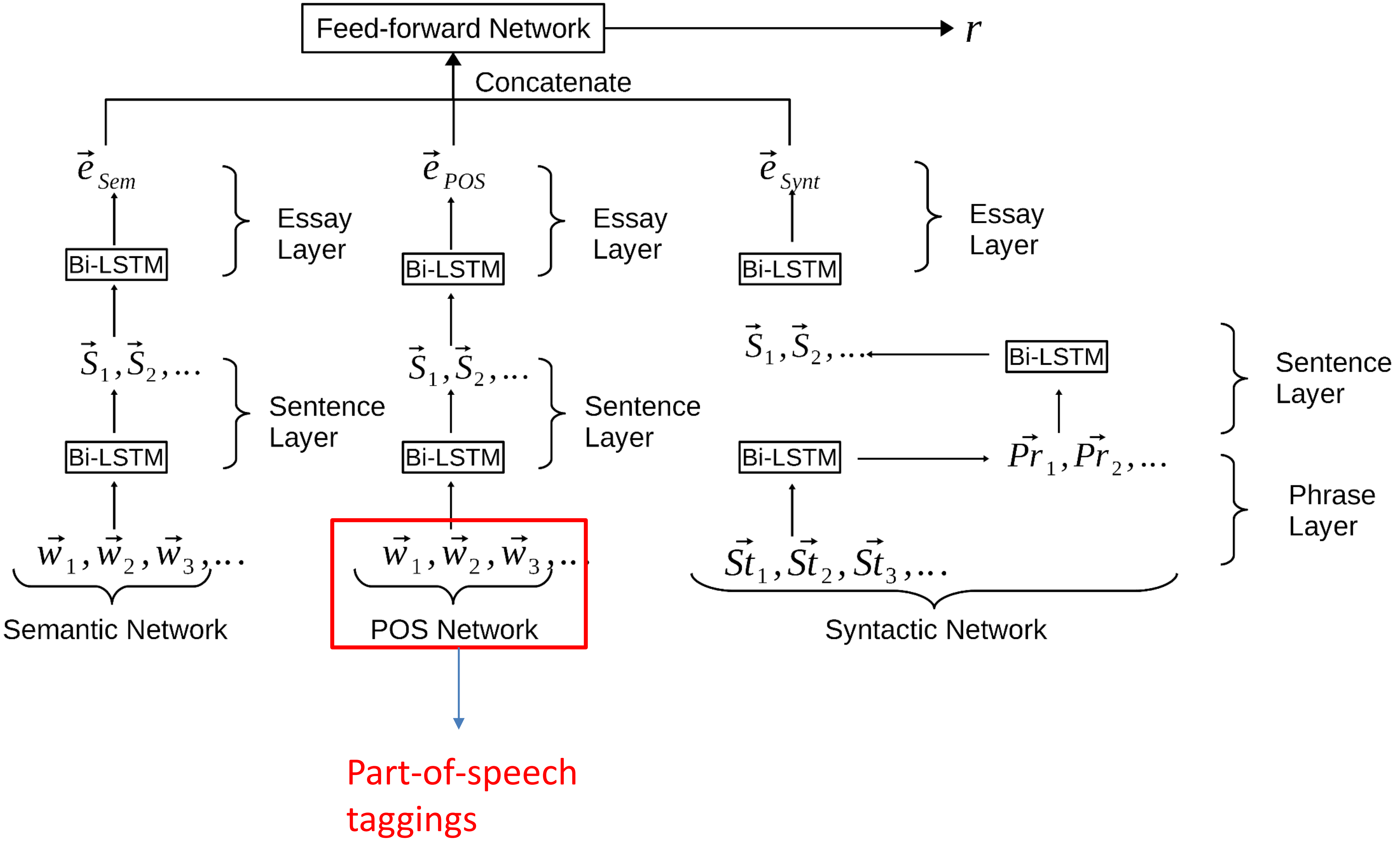


Multi-layer structure: Words – (phrases) - Sentences – Essay

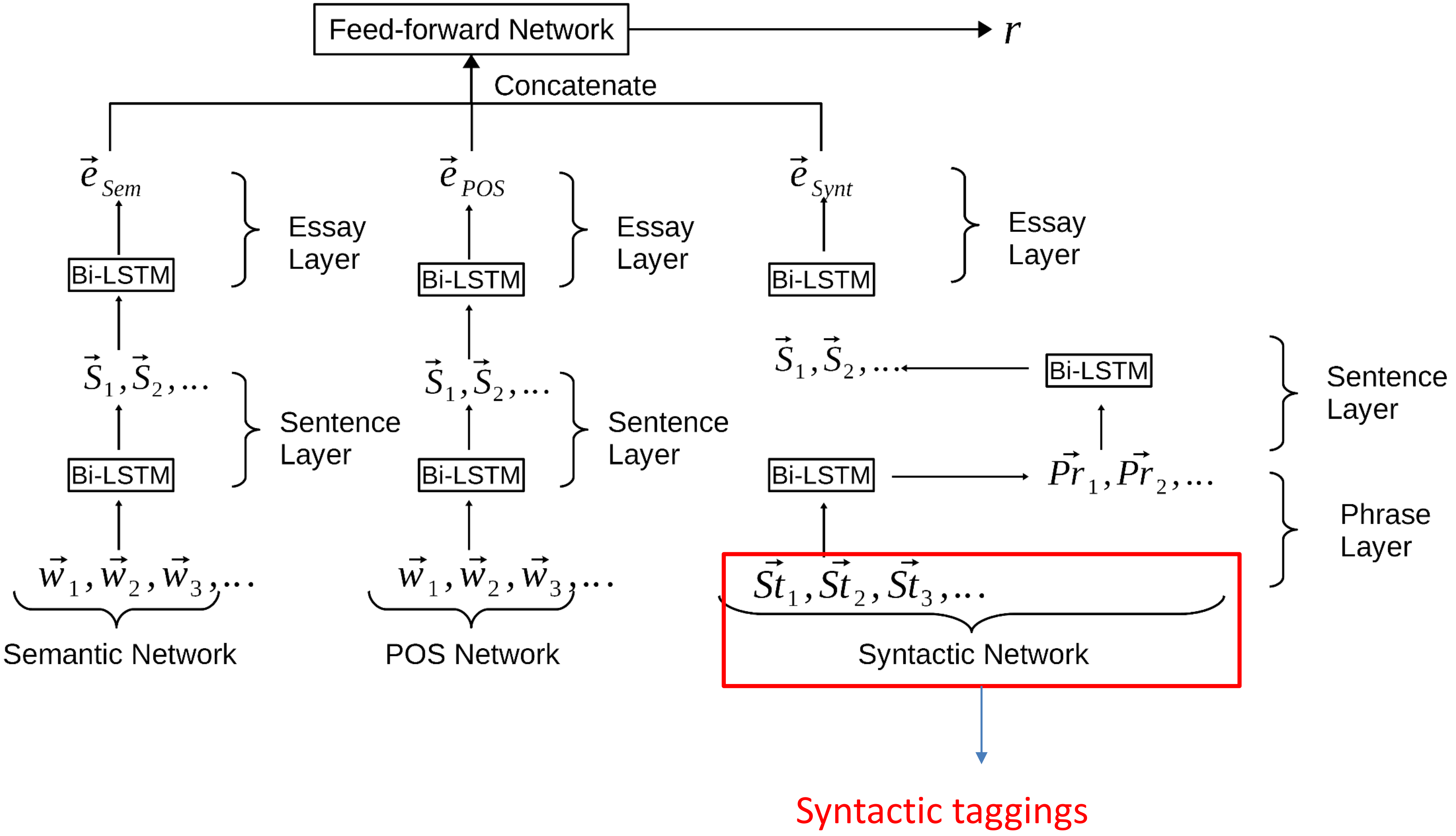
Architecture of the hybrid deep model



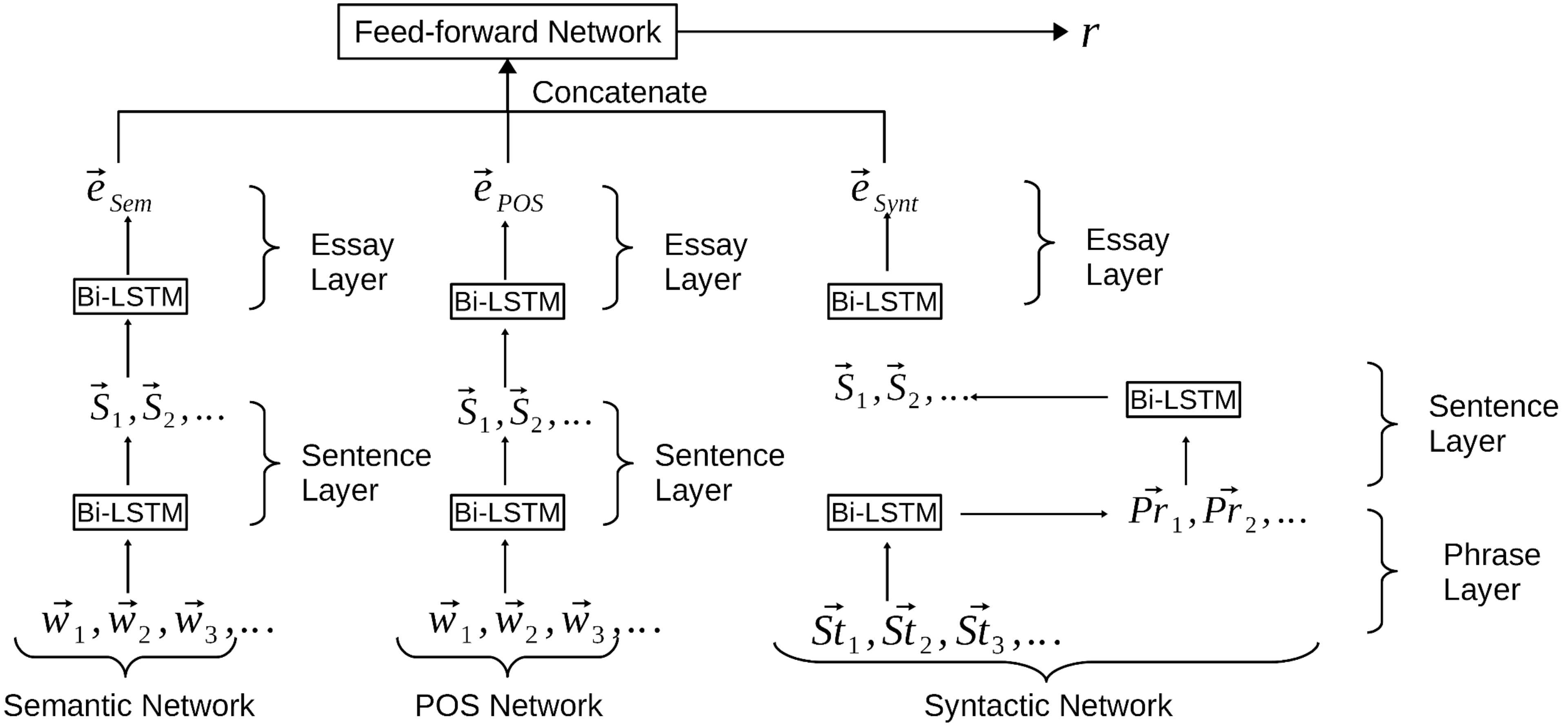
Architecture of the hybrid deep model



Architecture of the hybrid deep model

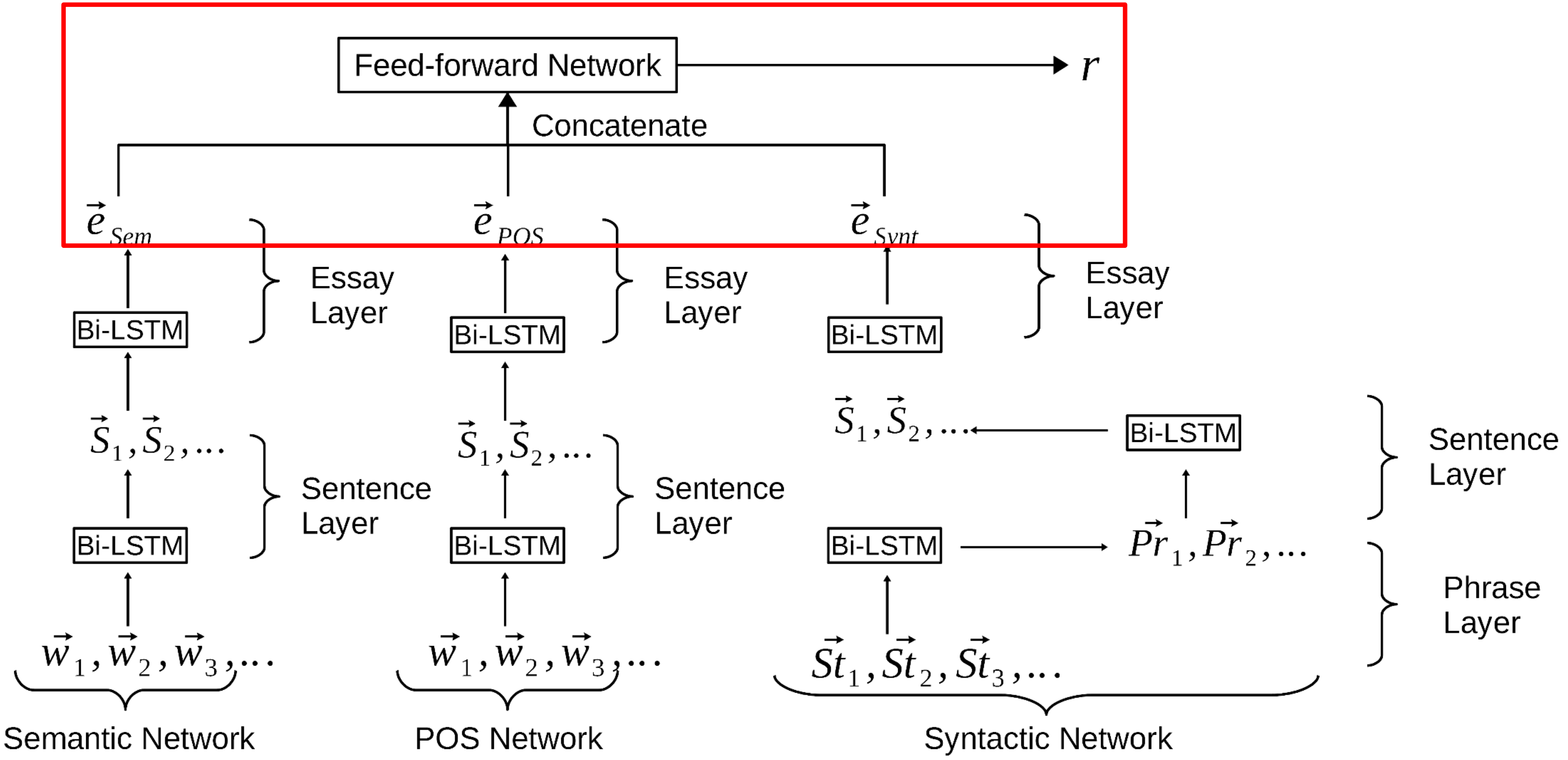


Architecture of the hybrid deep model



Multi-layer structure: Words – (phrases) - Sentences – Essay

Architecture of the hybrid deep model



Model Training

- Training loss: MSE on 0/1 pseudo labels

- Validation metric: Kappa on 30% non-target essays
  - Select the model that can best rate

Outline

- Background
- Method
- Experiments
- Conclusions

Dataset & Metrics

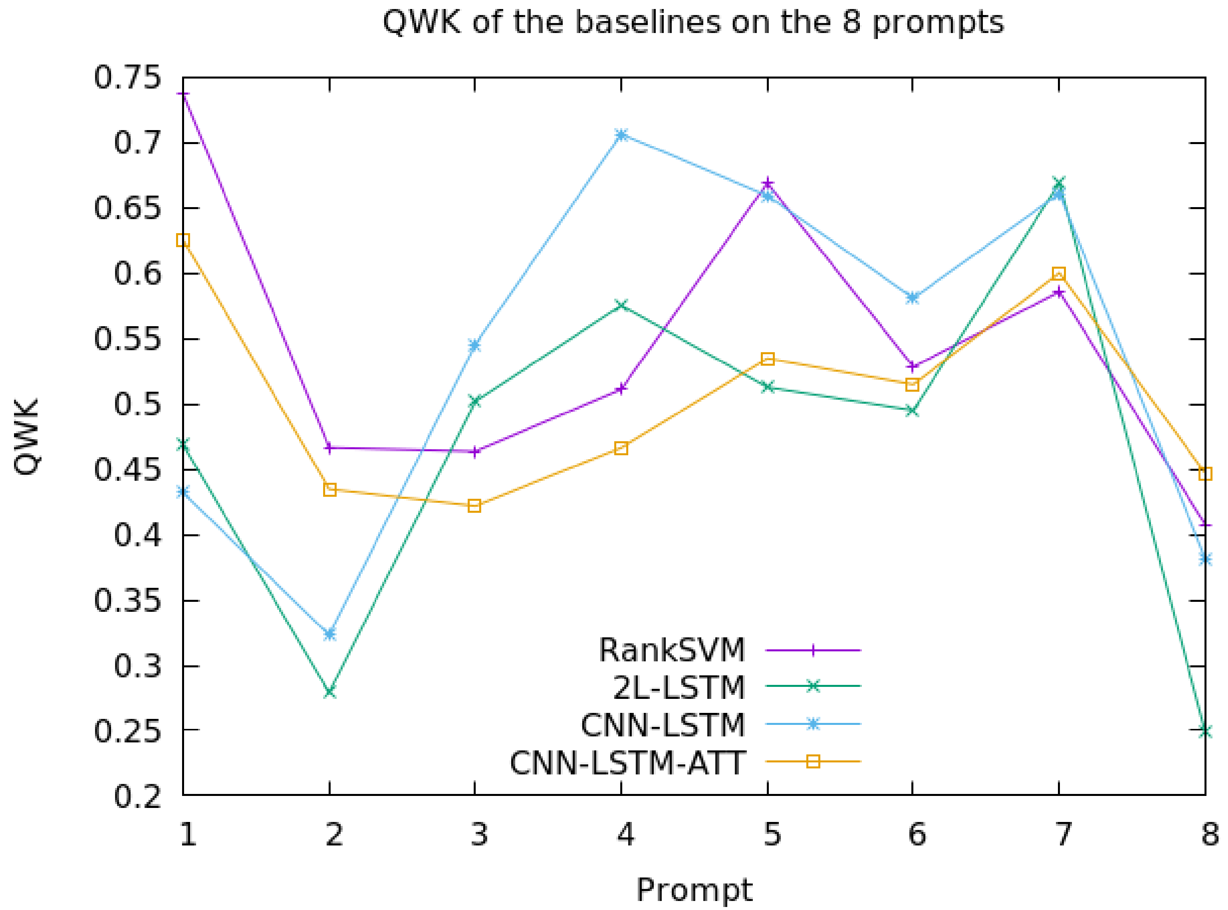
- We use the standard ASAP corpus
  - 8 prompts with >10K essays in total
- Prompt-independent AES: 7 prompts are used for training, 1 for testing

- Report on common human-machine agreement metrics
  - Pearson's correlation coefficient (PCC)
  - Spearman's correlation coefficient (SCC)
  - Quadratic weighted Kappa (QWK)

Baselines

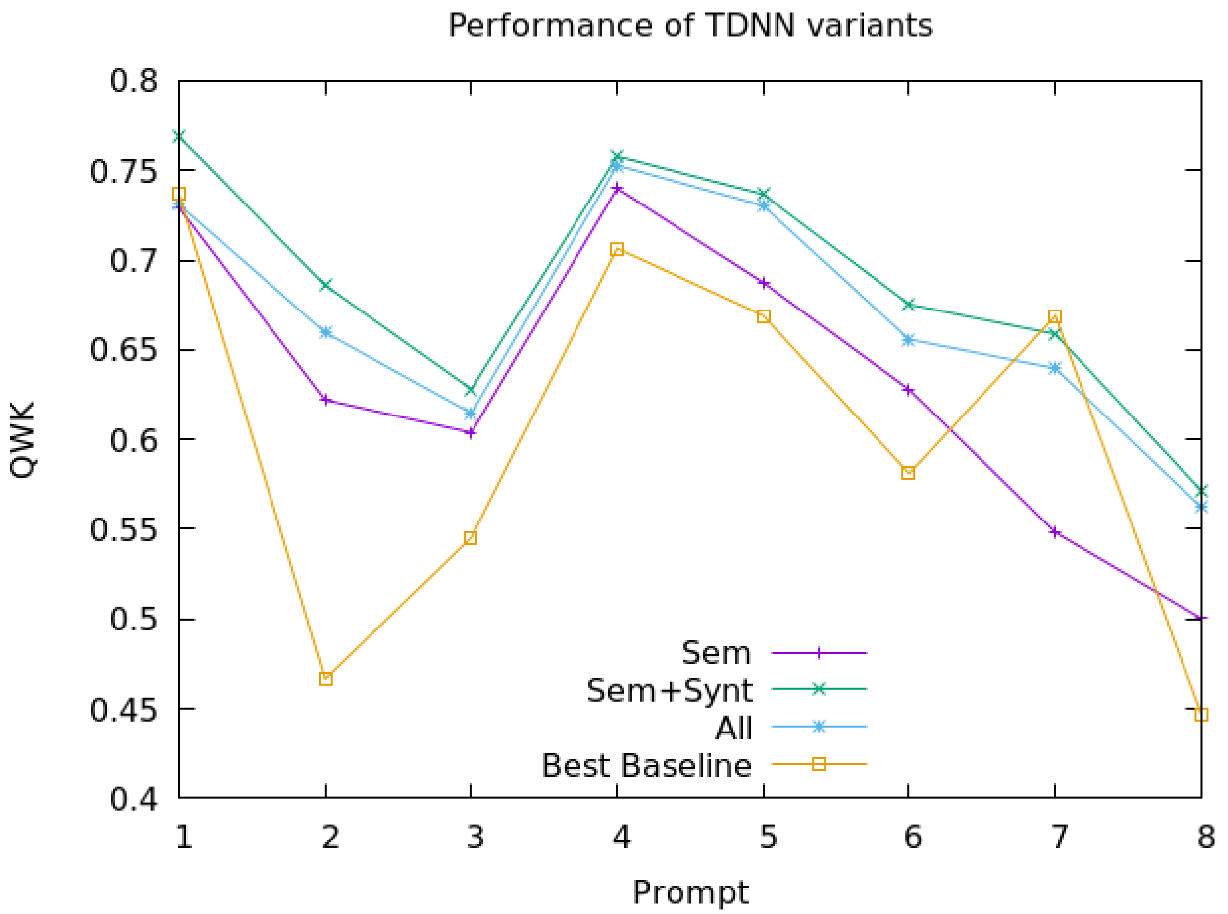
- RankSVM based on prompt-independent handcrafted features
  - Also used in the prompt-independent stage in TDNN
- 2L-LSTM [Alikaniotis et al. , ACL 2016]
  - Two LSTM layer + linear layer
- CNN-LSTM [Taghipour & Ng, EMNLP 2016]
  - CNN + LSTM + linear layer
- CNN-LSTM-ATT [Dong et al. , CoNLL 2017]
  - CNN-LSTM + attention

RankSVM is the most robust baseline



- High variance of DNN models’ performance on all 8 prompts
  - Possibly caused by learning on non-target prompts
- RankSVM appears to be the most stable baseline
  - Justifies the use of RankSVM in the first stage of TDNN

Comparison to the best baseline



- TDNN outperforms the best baseline on 7 out of 8 prompts
- Performance improvements gained by learning on the target prompt

Average performance on 8 prompts

	Method	QWK	PCC	SCC
Baselines	RankSVM	.5462	.6072	.5976
	2L-LSTM	.4687	.6548	.6214
	CNN-LSTM	.5362	.6569	.6139
	CNN-LSTM-ATT	.5057	.6535	.6368
TDNN	TDNN(Sem)	.5875	.6779	.6795
	TDNN(Sem+POS)	.6582	.7103	.7130
	TDNN(Sem+Synt)	.6856	.7244	.7365
	TDNN(POS+Synt)	.6784	.7189	.7322
	TDNN(ALL)	.6682	.7176	.7258

Average performance on 8 prompts

	Method	QWK	PCC	SCC
Baselines	RankSVM	.5462	.6072	.5976
	2L-LSTM	.4687	.6548	.6214
	CNN-LSTM	.5362	.6569	.6139
	CNN-LSTM-ATT	.5057	.6535	.6368
TDNN	TDNN(Sem)	.5875	.6779	.6795
	TDNN(Sem+POS)	.6582	.7103	.7130
	TDNN(Sem+Synt)	.6856	.7244	.7365
	TDNN(POS+Synt)	.6784	.7189	.7322
	TDNN(ALL)	.6682	.7176	.7258

Average performance on 8 prompts

	Method	QWK	PCC	SCC
Baselines	RankSVM	.5462	.6072	.5976
	2L-LSTM	.4687	.6548	.6214
	CNN-LSTM	.5362	.6569	.6139
	CNN-LSTM-ATT	.5057	.6535	.6368
TDNN	TDNN(Sem)	.5875	.6779	.6795
	TDNN(Sem+POS)	.6582	.7103	.7130
	TDNN(Sem+Synt)	.6856	.7244	.7365
	TDNN(POS+Synt)	.6784	.7189	.7322
	TDNN(ALL)	.6682	.7176	.7258

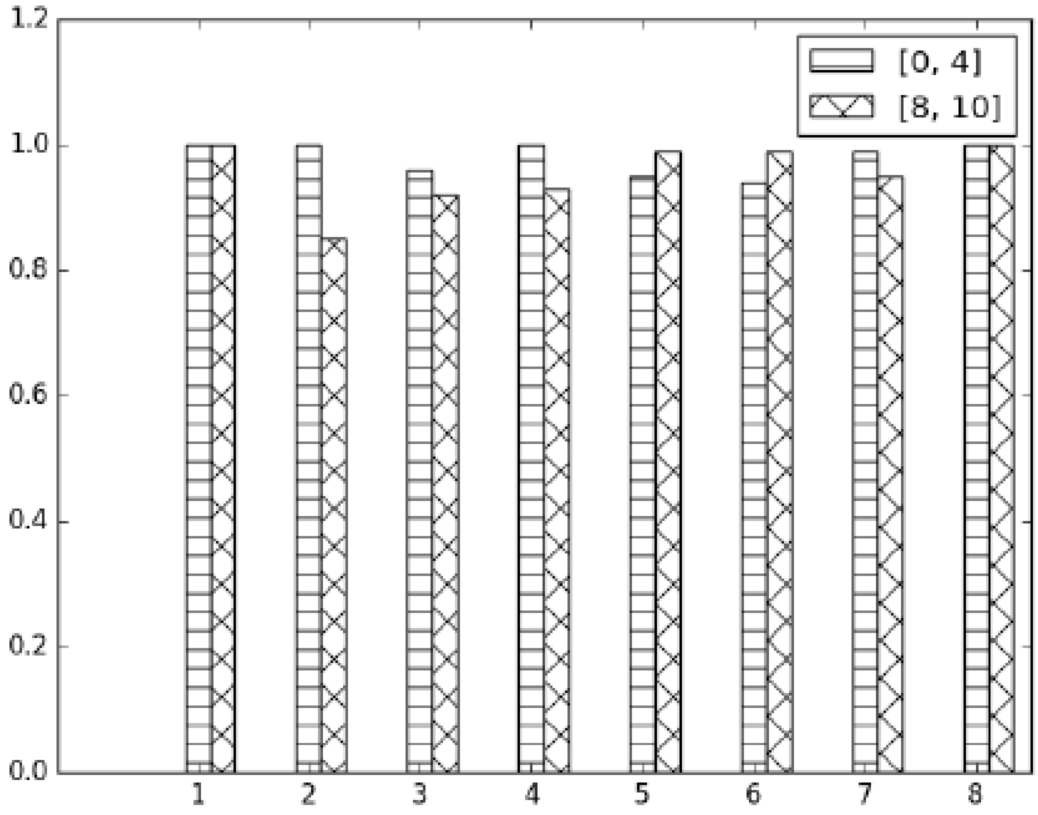
Sanity Check: Relative Precision

How the quality of pseudo examples affects the performance of TDNN?

- The sanctity of the selected essays, namely, the number of positive (negative) essays that are better (worse) than all negative (positive) essays.

- Such relative precision is at least 80% and mostly beyond 90% on different prompts

- TDNN can at least learn from correct 0/1 labels



Conclusions

- It is beneficial to learn an AES model on the target prompt
- Syntactic features are useful addition to the widely used Word2Vec embeddings
- Sanity check: small overlap between pos/neg examples
- Prompt-independent AES remains an open problem
  - ETS wants Kappa>0.70
  - TDNN can achieve 0.68 at best

Thank you!

INSTITUTE OF SOFTWARE CHINESE ACADEMY OF SCIENCES

中国科学院大学
University of Chinese Academy of Sciences