Learning Multi-level Dependencies for Robust Word Recognition

Zhiwei Wang, Michigan State University
Hui Liu, Michigan State University
Jiliang Tang, Michigan State University
Songfan Yang, TAL AI Lab
Gale Yan Huang, TAL AI Lab
Zitao Liu, TAL AI Lab

Language Processing Mechanism of Humans Are Very Robust

- Humans can read the following Cambridge sentence easily

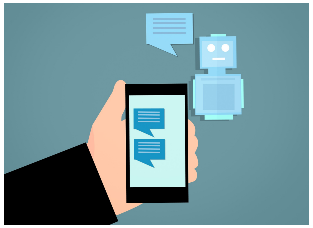
Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae. The rset can be a toatl mses and you can sitll raed it wouthit porbelm. Tihs is bcuseae the huamn mnid deos not raed ervey lteter by istlef, but the wrod as a wlohe.

- In contrast, modern natural language processing systems are very brittle

TAL 好未来 每个人的好未来
MICHIGAN STATE UNIVERSITY

Robust Natural Language Processing Systems Are Important

- Severe consequences that can be made by adversarial samples
  - Carefully misspelled spam emails that fool spam detection systems
  - Deliberately designed input sentences that force chatbot to emit offensive language



- We focus on building a word recognition framework which can denoise the misspellings
  - It can be straightforwardly used to correct misspellings

rhgit → Framework → right

  - It could serves as a denoising component of other natural language processing systems

rhgit → Framework / Translation → 对

Intuition Behind Human Word Recognition Mechanism

- Consider the following sentence from the "Cambridge" example

… but the wrod as a wlohe.

- To recognize the correct form of "wlohe", two types of information is extremely important
  - Character-level sequential dependencies

wlohe → whole

wlohe → whelo

  - Word-level sequential dependencies

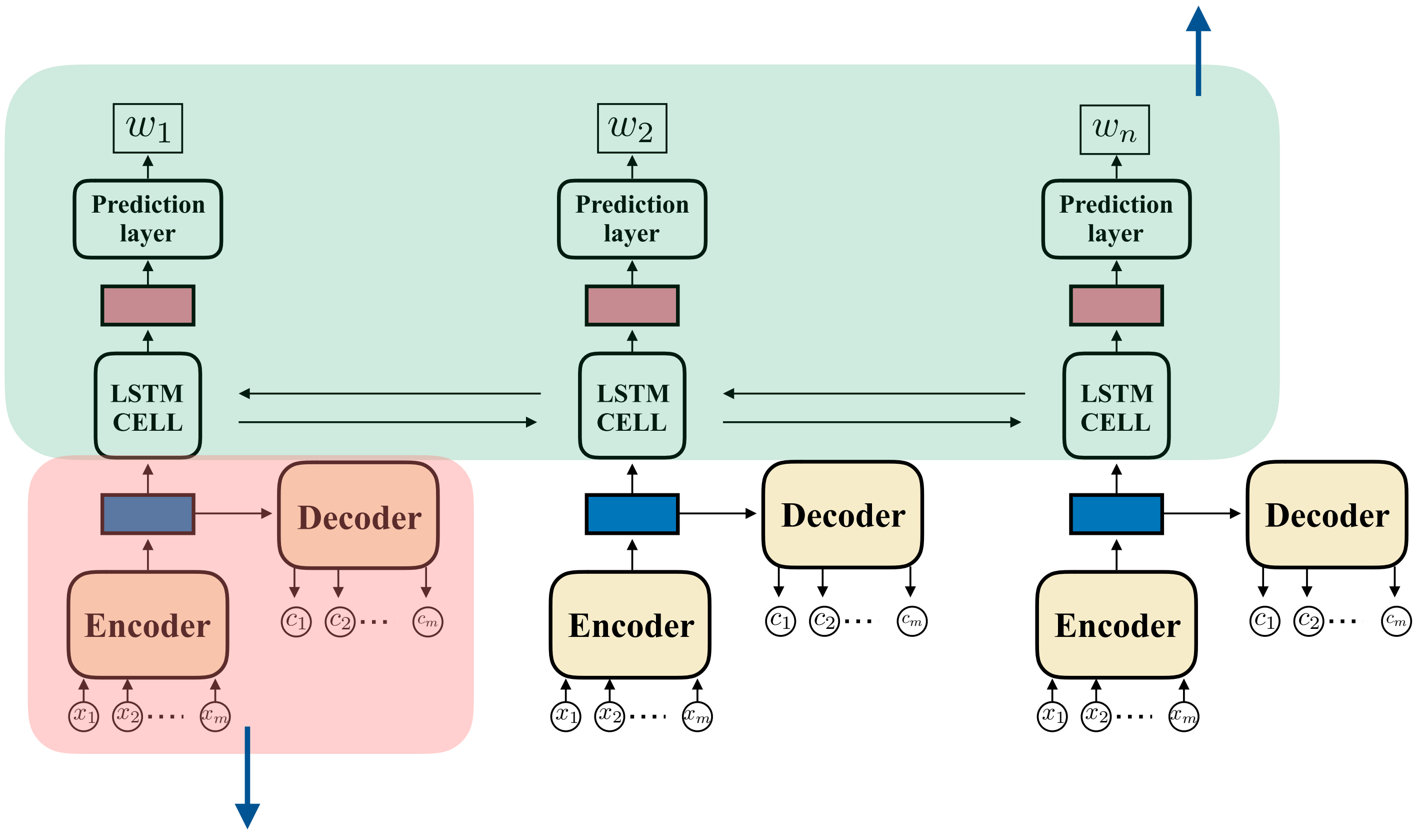
… but the wrod as a whelo .

… but the wrod as a whole . ✓

The Proposed Framework: MUDE

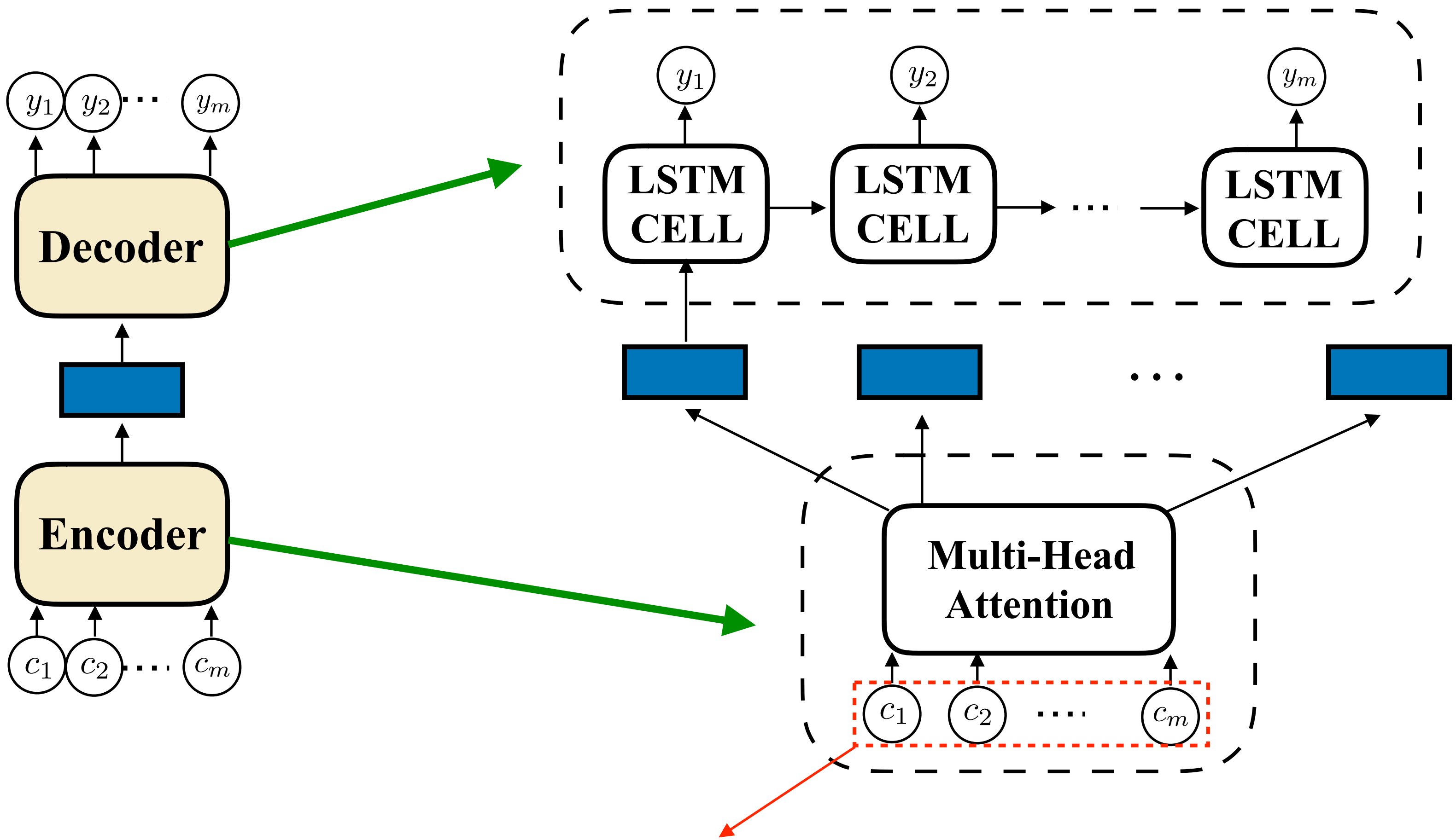
Overview of the framework

Word-level dependencies



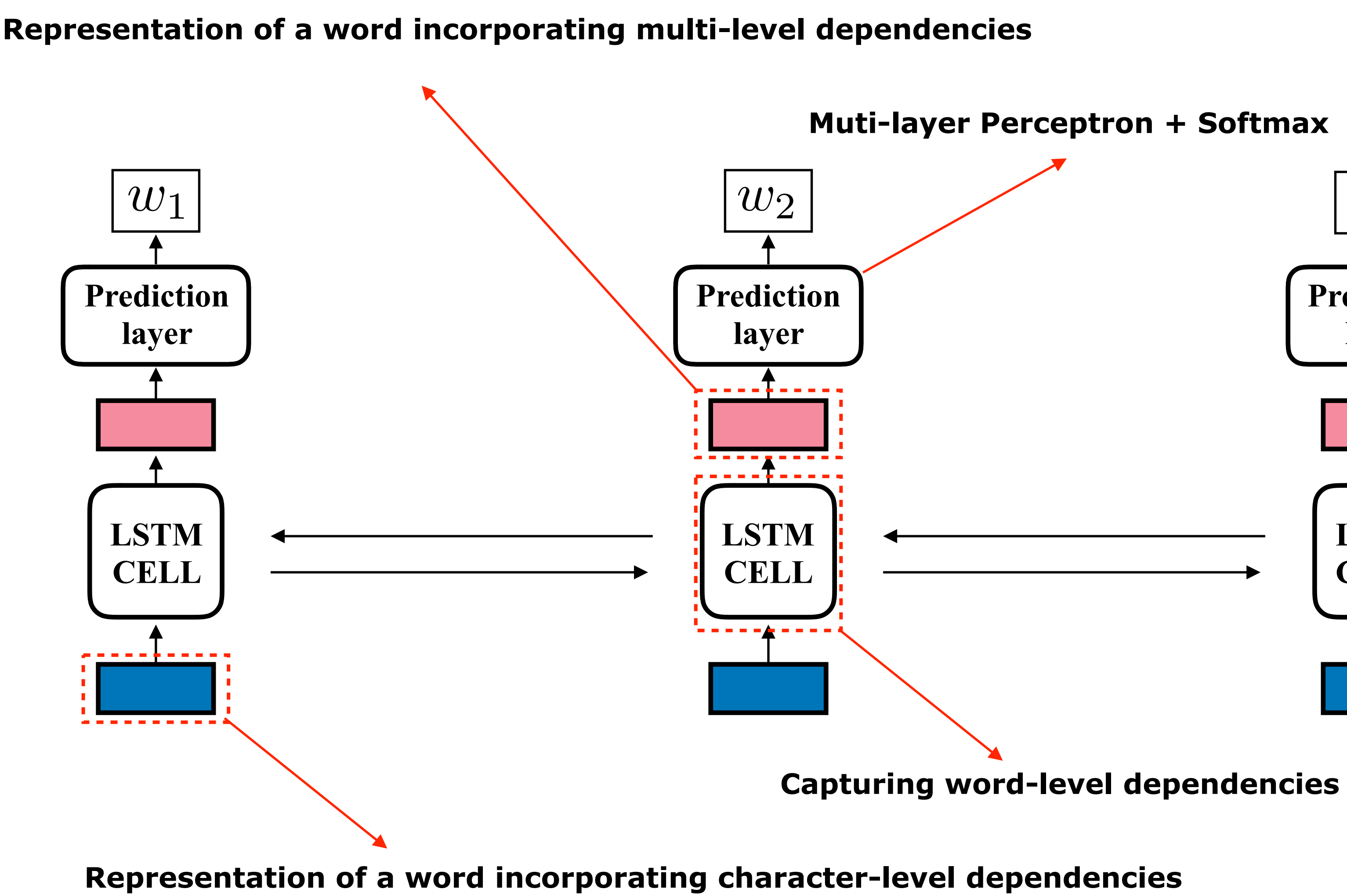
Character-level dependencies

Seq2Seq Model Captures Character-level Dependencies



The order of characters can be misleading

BiLSTM Captures Word-level Dependencies



Experiment Settings

- Dataset: Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993)
  - Training: 39,832 sentences
  - Validation: 1,700 sentences
  - Testing: 2,416 sentences
  - Four types of easy noise: Permutation (PER), Deletion (DEL), Insertion (INS), and Substitution (SUB)

Noise Type	Sentence
Correct	An example of noised text
PER	An epaxmle of nsieod txet
DEL	An examle of nosed tet
INS	An edxample of nmoised texut
SUB	An exsmple of npised test

  - Four types of difficult noise: W-PER, W-DEL, W-INS, and W-SUB

- Baselines:
  - Enchant: a widely used dictionary-based method
  - scRNN: a state-of-the-art word recognition method

Experimental Results

- Easy Cases

Method	INT	DEL	INS	SUB
Enchant	72.33	71.23	93.93	79.77
scRNN	98.23	91.55	95.95	87.09
MUDE	**98.81**	**95.86**	**97.16**	**90.52**

- Difficult Cases:

Method	W-PER	W-DEL	W-INS	W-SUB
Enchant	59.08	69.84	93.77	77.23
scRNN	97.36	89.99	95.96	81.12
MUDE	**98.75**	**93.29**	**97.10**	**85.17**

- Observations:
  - Our proposed framework MUDE significantly outperforms the baselines
  - In more difficult cases, the performance gain of MUDE is more obvious

Case Study: Permutations

- Model Performance

	W-PER
Noised	iodrcAngc ot a reeachsr at meigaCdbr srtiinUyve , it seod tn' amrtte in wtah rerdo het tserelt in a rdwo rae , the onyl onmtiaptr ingth si tath hte itfrs dan stla treelt be ta het tgrhi place . hTe rset nca be a aotlt mess dan ouy anc lsilt drae ti tthwuoi lorbmpe . hTsi is aubeecs the huamn dmni edos nto erad evrye lteter by etfisl , but het rdwo sa a eholw .
scRNN	According to a **research** at Cambridge University , it does n't matter in what order the letters in a word are , the only important thing is that the first and last letter be at the right place . The rest can be a total mess and you can still read it without problem . This is because the human mind does not read **very** letter by itself , but the word as a whole .
MUDE	According to a **research** at Cambridge University , it does n't matter in what order the letters in a word are , the only important thing is that the first and last letter be at the right place . The rest can be a total mess and you can still read it without problem . This is because the human mind does not read every letter by itself , but the word as a whole .

- Observations:
  - When permuting all characters in a word, it becomes difficult to read even for humans
  - MUDE is able to capture the character-level dependencies while scRNN can not

Summary and Future Work

- Summary
  - We design a novel framework MUDE to recognize correct words from their noised version
  - It can capture both character-level and word-level dependencies in natural languages
  - Through extensive experiments, we demonstrate the effectiveness of the proposed framework
  - It has great potential to improve robustness of other natural language systems straightforwardly

- Future Works
  - Verify its effectiveness in different languages
  - Extend it to deal with sentence-level noises, i.e., grammar noises

Thanks you

Questions ?