香侬科技
SHANNON.AI

Machine Reading Comprehension (MRC) Framework as Universal Solutions to Various NLP Tasks

2019.11.11

Table of Contents

Table of Contents

The Formulation of MRC Question Answering

- Given a question $Q = \{q_1, \cdots, q_n\}$, context $C = \{c_1, \cdots, c_n\}$, find answer $A = \mathrm{c}_{start:end}$ which is a substring of C, to be the answer to Q.

X = the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

Q = Where do water droplets collide with ice crystals to form precipitation?

A = within a cloud

Transforming Different NLP Tasks to QA

- NER
- Relation Extraction
- Coreference
- Text Classification

Transforming NER to QA

Li et al., A Unified MRC Framework for Named Entity Recognition. ACL 2020

C: at the Chinese embassy in France……

Q: which location is mentioned in the context ?

Transforming NER to QA

Li et al., A Unified MRC Framework for Named Entity Recognition. ACL 2020

C: at the Chinese embassy in France……

Q: find all geographical regions, such as a government, a populace, and a geographic location. (Annotation Guidelines)

Transforming Relation Extraction to QA

Li et al., entity-relation extraction as multi-turn question answering. ACL2019.

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

Transforming Relation Extraction to QA

Li et al., entity-relation extraction as multi-turn question answering. ACL2019.

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

Person	Corp	Time	Position
Musk	SpaceX	2002	CEO
Musk	Tesla	2003	CEO& product architect
Musk	SolarCity	2006	chairman
Musk	Neuralink	2016	CEO
Musk	The Boring Company	2016	-

Transforming Relation Extraction to QA

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

Transforming Relation Extraction to QA

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

- To extract person: Q: who is mentioned in the text? A: Musk

Transforming Relation Extraction to QA

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

- To extract person: Q: who is mentioned in the text? A: Musk
- To extract company: Q: which company did Musk work for? A: SpaceX

Transforming Relation Extraction to QA

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

- To extract person: Q: who is mentioned in the text? A: Musk
- To extract company: Q: which company did Musk work for? A: SpaceX
- To extract position: Q: what was Musk 's position in SpaceX? A: CEO

Transforming Relation Extraction to QA

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

- To extract person: Q: who is mentioned in the text? A: Musk
- To extract company: Q: which company did Musk work for? A: SpaceX
- To extract position: Q: what was Musk 's position in SpaceX? A: CEO
- To extract time: Q: During which period did Musk work for SpaceX as CEO?A : 2002

Transforming Relation Extraction to QA

Li et al., entity-relation extraction as multi-turn question answering. ACL2019.

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

Person	Corp	Time	Position
Musk	SpaceX	2002	CEO
Musk	Tesla	2003	CEO& product architect
Musk	SolarCity	2006	chairman
Musk	Neuralink	2016	CEO
Musk	The Boring Company	2016	-

Transforming Text Classification to QA

C = the text to classify

Q = the description of a category “comp.sys.mac.hardware:The Macintosh (branded simply asMac since 1998) is a family of personal computers designed,manufactured and sold by Apple Inc. since January 1984”

A = binary label yes/no

Template Description
A car (or automobile) is a wheeled motor ... transport people rather than goods.

Extractive Description
the 325is i drove was definitely faster than that

Abstractive Description
the car I drive is fast

Text X
sure sounds like they got a ringer. the 325is i drove was definitely faster than that. if you want to quote numbers, my AW AutoFile shows 0-60 in 7.4, 1/4 mile in 15.9. it quotes Car and Driver's figures of 6.9 and 15.3. … i don't know how the addition of variable valve timing for 1993 affects it. but don't take my word for it. go drive it.

Transforming Co-reference Resolution to QA

Coreference Resolution as Query-based Span Prediction. Wu et al., ACL2020

I was hired to do some Christmas music, and it was just "Jingle Bells" and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.

[Jingle Bells, the song]

Transforming Co-reference Resolution to QA

Coreference Resolution as Query-based Span Prediction. Wu et al., ACL2020

C = I was hired to do some Christmas music, and it was just “Jingle Bells” and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.

Q = it was just “ <mention> Jingle Bells <mention>”

A = the song

Transforming Co-reference Resolution to QA

Coreference Resolution as Query-based Span Prediction. Wu et al., ACL2020

X = I was hired to do some Christmas music, and it was just “Jingle Bells” and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.

Q = it was just “ <mention> to do some<mention>”

A = None

What are the general advantages of this formulation ?

- Questions encode prior knowledge about what we want to extract or classify. (a hard version of attention)

X: at the Chinese embassy in France……

Q: find all geographical regions, such as a government, a populace, and a geographic location. (Annotation Guidelines)

What are the general advantages of this formulation ?

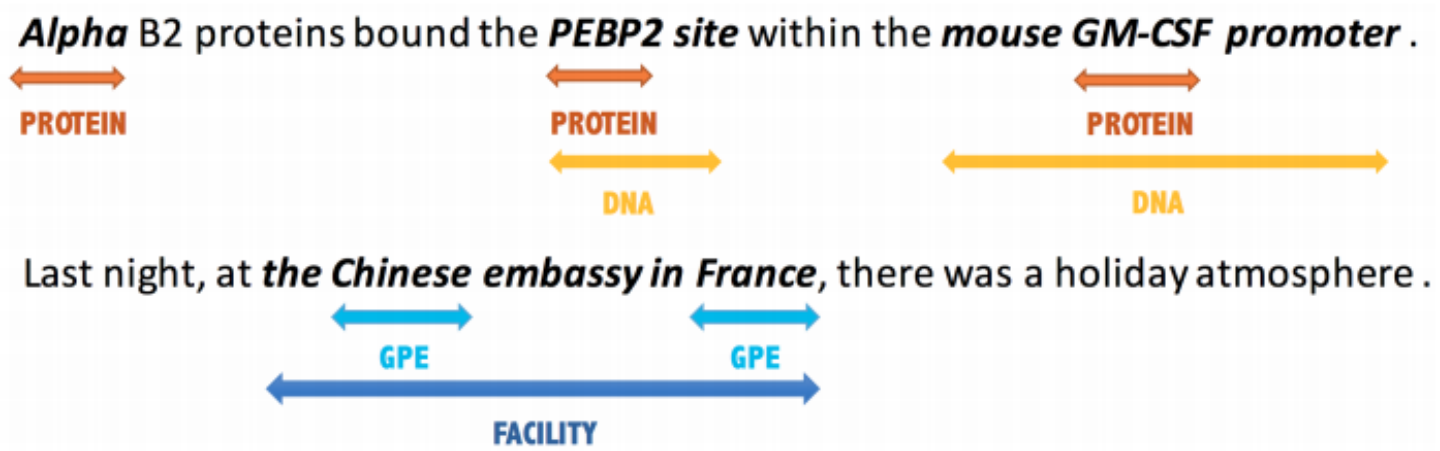
- Questions encode prior knowledge about what we want to extract or clarify. (a hard version of attention)
- Domain adaptation ability
- Few shot learning ability
- Flexibility

MRC for Named Entity Recognition

- In real-world scenarios, entities are often *nested*

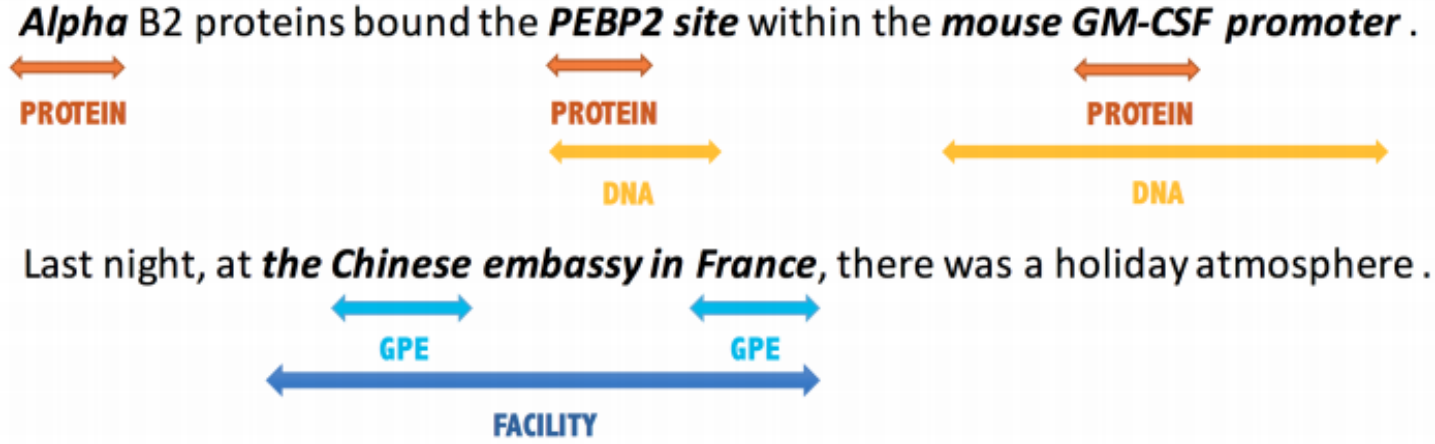
MRC for Named Entity Recognition

- In real-world scenarios, entities are often *nested*



MRC for Named Entity Recognition

- In real-world scenarios, entities are often *nested*



- Sequence labeling model can only predict one label at one time.

Formalization

- Instead of using sequence labeling model, we use an MRC framework to extract entities

Formalization

- Instead of using sequence labeling model, we use an MRC framework to extract entities
- Given a sequence $X = \{x_1, \cdots, x_n\}$, find each entity $x_{start:end}$ in X attached to a specific entity type $y \in Y$, which is encoded as a corresponding question q_y

Formalization

- Instead of using sequence labeling model, we use an MRC framework to extract entities
- Given a sequence $X = \{x_1, \cdots, x_n\}$, find each entity $x_{start:end}$ in X attached to a specific entity type $y \in Y$, which is encoded as a corresponding question q_y
- This question q_y has already encoded significant prior information for extracting entities with label y

Formalization

X: at the Chinese embassy in France……

q_{GPE}: find all geographical regions, such as a government, a populace, and a geographic location.

Query Generation

- We use *annotation guideline notes*, **the guidelines provided to the annotators of the dataset by the dataset builder.**

Query Generation

- We use *annotation guideline notes*, **the guidelines provided to the annotators of the dataset by the dataset builder.**

Entity	Natural Language Question
Location	Find locations in the text, including non-geographical locations, mountain ranges and bodies of water
Facility	Find facilities in the text, including buildings, airports, highways and bridges
Organization	Find organizations in the text, including companies, agencies and institutions

Query Generation

- We use *annotation guideline notes*, **the guidelines provided to the annotators of the dataset by the dataset builder.**
- We also compare several query strategies:
  - Annotation guideline notes
  - Position index of labels
  - Keyword
  - *Rule-based template filling*
  - Wikipedia
  - Synonyms
  - Keyword+synonyms

Model

- BERT
  - Input: $\{[CLS], q_1, \cdots, q_m, [SEP], x_1, \cdots, x_n\}$
  - Output: A matrix $E \in \mathbb{R}^{n \times d}$

Model

- BERT
  - Input: $\{[CLS], q_1, \cdots, q_m, [SEP], x_1, \cdots, x_n\}$
  - Output: A matrix $E \in \mathbb{R}^{n \times d}$
- Predict all possible start indices

Model

- BERT
  - Input: $\{[CLS], q_1, \cdots, q_m, [SEP], x_1, \cdots, x_n\}$
  - Output: A matrix $E \in \mathbb{R}^{n \times d}$
- Predict all possible start indices
  - $\mathbb{R}^{n \times 2} \ni P_{start} = softmax_{row}(ET_{start}),\ \{0,1\}^n \ni I_{start} = argmax_{row}(P_{start})$

Model

- BERT
  - Input: $\{[CLS], q_1, \cdots, q_m, [SEP], x_1, \cdots, x_n\}$
  - Output: A matrix $E \in \mathbb{R}^{n \times d}$
- Predict all possible start indices
  - $\mathbb{R}^{n \times 2} \ni P_{start} = softmax_{row}(ET_{start}),\ \{0,1\}^n \ni I_{start} = argmax_{row}(P_{start})$
- Predict all possible end indices
  - $\mathbb{R}^{n \times 2} \ni P_{end} = softmax_{row}(ET_{end}),\ \{0,1\}^n \ni I_{end} = argmax_{row}(P_{end})$

Model

- BERT
  - Input: $\{[CLS], q_1, \cdots, q_m, [SEP], x_1, \cdots, x_n\}$
  - Output: A matrix $E \in \mathbb{R}^{n \times d}$
- Predict all possible start indices
  - $\mathbb{R}^{n \times 2} \ni P_{start} = softmax_{row}(ET_{start}),\ \{0,1\}^n \ni I_{start} = argmax_{row}(P_{start})$
- Predict all possible end indices
  - $\mathbb{R}^{n \times 2} \ni P_{end} = softmax_{row}(ET_{end}),\ \{0,1\}^n \ni I_{end} = argmax_{row}(P_{end})$
- Start-end matching
  - $p_{ij} = sigmoid(m \cdot [E_i; E_j]);\ \forall (i,j), I^i_{start} = 1, I^j_{end} = 1$

Training and Inference

- Three losses

Training and Inference

- Three losses
  - Start loss: $\mathrm{L}_{\mathrm{start}} = CrossEntropy(P_{start}, Y_{start})$
  - End loss: $\mathrm{L}_{\mathrm{end}} = CrossEntropy(P_{end}, Y_{end})$
  - Span loss: $\mathrm{L}_{\mathrm{span}} = CrossEntropy(P_{ij}, Y_{ij})$

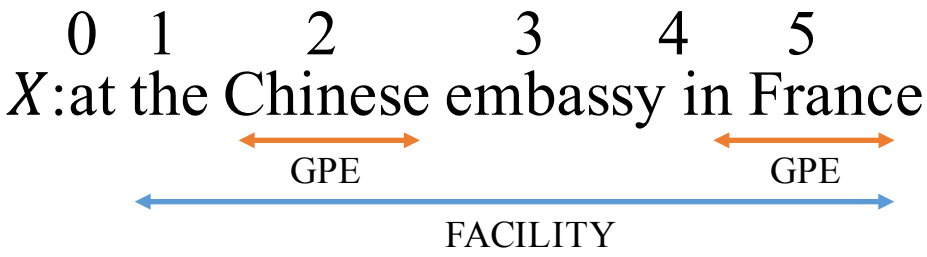
Training and Inference

- Three losses
  - Start loss: $\mathrm{L}_{\mathrm{start}} = CrossEntropy(P_{start}, Y_{start})$
  - End loss: $\mathrm{L}_{\mathrm{end}} = CrossEntropy(P_{end}, Y_{end})$
  - Span loss: $\mathrm{L}_{\mathrm{span}} = CrossEntropy(P_{ij}, Y_{ij})$
- Overall training objective: $L = L_{start} + L_{end} + L_{span}$

Training and Inference

- Three losses
  - Start loss: $\mathrm{L}_{\mathrm{start}} = CrossEntropy(P_{start}, Y_{start})$
  - End loss: $\mathrm{L}_{\mathrm{end}} = CrossEntropy(P_{end}, Y_{end})$
  - Span loss: $\mathrm{L}_{\mathrm{span}} = CrossEntropy(P_{ij}, Y_{ij})$
- Overall training objective: $L = L_{start} + L_{end} + L_{span}$
- For evaluation, we use p_{ij}

Example



q_{GPE}: find all geographical regions defined by political and/or social groups, such as a government, a populace.

$$P_{start} = \begin{bmatrix} 0.9 & 0.1 \\ 0.85 & 0.1 \\ 0.02 & 0.98 \\ 0.49 & 0.51 \\ 0.99 & 0.01 \\ 0.03 & 0.97 \end{bmatrix} \quad I_{start} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} \qquad P_{end} = \begin{bmatrix} 0.99 & 0.01 \\ 0.89 & 0.11 \\ 0.04 & 0.96 \\ 0.76 & 0.24 \\ 0.99 & 0.01 \\ 0.12 & 0.88 \end{bmatrix} \quad I_{start} = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$p_{22} = 0.96$
$p_{25} = 0.21$
$p_{32} = 0$
$p_{35} = 0.33$
$p_{52} = 0$
$p_{55} = 0.89$

Experiments on Nested NER

English ACE 2004			
Model	P	R	F
Hyper-Graph (Katiyar and Cardie, 2018)	73.6	71.8	72.7
Seg-Graph (Wang and Lu, 2018)	78.0	72.4	75.1
Seq2seq-BERT (Straková et al., 2019)	-	-	84.40
Path-BERT (Shibuya and Hovy, 2019)	83.73	81.91	82.81
DYGIE (Luan et al., 2019)	-	-	84.7
BERT-MRC	**85.05**	**86.32**	**85.98 (+1.28)**
English ACE 2005			
Model	P	R	F
Hyper-Graph (Katiyar and Cardie, 2018)	70.6	70.4	70.5
Seg-Graph (Wang and Lu, 2018)	76.8	72.3	74.5
ARN (Lin et al., 2019a)	76.2	73.6	74.9
Path-BERT (Shibuya and Hovy, 2019)	82.98	82.42	82.70
Merge-BERT (Fisher and Vlachos, 2019)	82.7	82.1	82.4
DYGIE (Luan et al., 2019)	-	-	82.9
Seq2seq-BERT (Straková et al., 2019)	-	-	84.33
BERT-MRC	**87.16**	**86.59**	**86.88 (+2.55)**

English GENIA			
Model	P	R	F
Hyper-Graph (Katiyar and Cardie, 2018)	77.7	71.8	74.6
ARN (Lin et al., 2019a)	75.8	73.9	74.8
Path-BERT (Shibuya and Hovy, 2019)	78.07	76.45	77.25
DYGIE (Luan et al., 2019)	-	-	76.2
Seq2seq-BERT (Straková et al., 2019)	-	-	78.31
BERT-MRC	**85.18**	**81.12**	**83.75 (+5.44)**
English KBP 2017			
Model	P	R	F
KBP17-Best (Ji et al., 2017)	76.2	73.0	72.8
ARN (Lin et al., 2019a)	77.7	71.8	74.6
BERT-MRC	**82.33**	**77.61**	**80.97 (+6.37)**

Experiments on Flat NER

English CoNLL 2003			
Model	P	R	F
BiLSTM-CRF (Ma and Hovy, 2016)	-	-	91.03
ELMo (Peters et al., 2018b)	-	-	92.22
CVT (Clark et al., 2018)	-	-	92.6
BERT-Tagger (Devlin et al., 2018)	-	-	92.8
BERT-MRC	**92.33**	**94.61**	**93.04**
			(+0.24)

English OntoNotes 5.0			
Model	P	R	F
BiLSTM-CRF (Ma and Hovy, 2016)	86.04	86.53	86.28
Strubell et al. (2017)	-	-	86.84
CVT (Clark et al., 2018)	-	-	88.8
BERT-Tagger (Devlin et al., 2018)	90.01	88.35	89.16
BERT-MRC	**92.98**	**89.95**	**91.11**
			(+1.95)

Chinese MSRA			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	93.57	92.79	93.18
BERT-Tagger (Devlin et al., 2018)	94.97	94.62	94.80
Glyce-BERT (Wu et al., 2019)	95.57	95.51	95.54
BERT-MRC	**96.18**	**95.12**	**95.75**
			(+0.21)

Chinese OntoNotes 4.0			
Model	P	R	F
Lattice-LSTM (Zhang and Yang, 2018)	76.35	71.56	73.88
BERT-Tagger (Devlin et al., 2018)	78.01	80.35	79.16
Glyce-BERT (Wu et al., 2019)	81.87	81.40	80.62
BERT-MRC	**82.98**	**81.25**	**82.11**
			(+1.49)

Improvement from BERT or MRC?

- It's not obvious that improvement from BERT or MRC
  - If it's from MRC, then combining MRC with other models except BERT will get improvement as well
  - Compare LSTM-CRF with MRC based models, such as BiDAF and QAnet

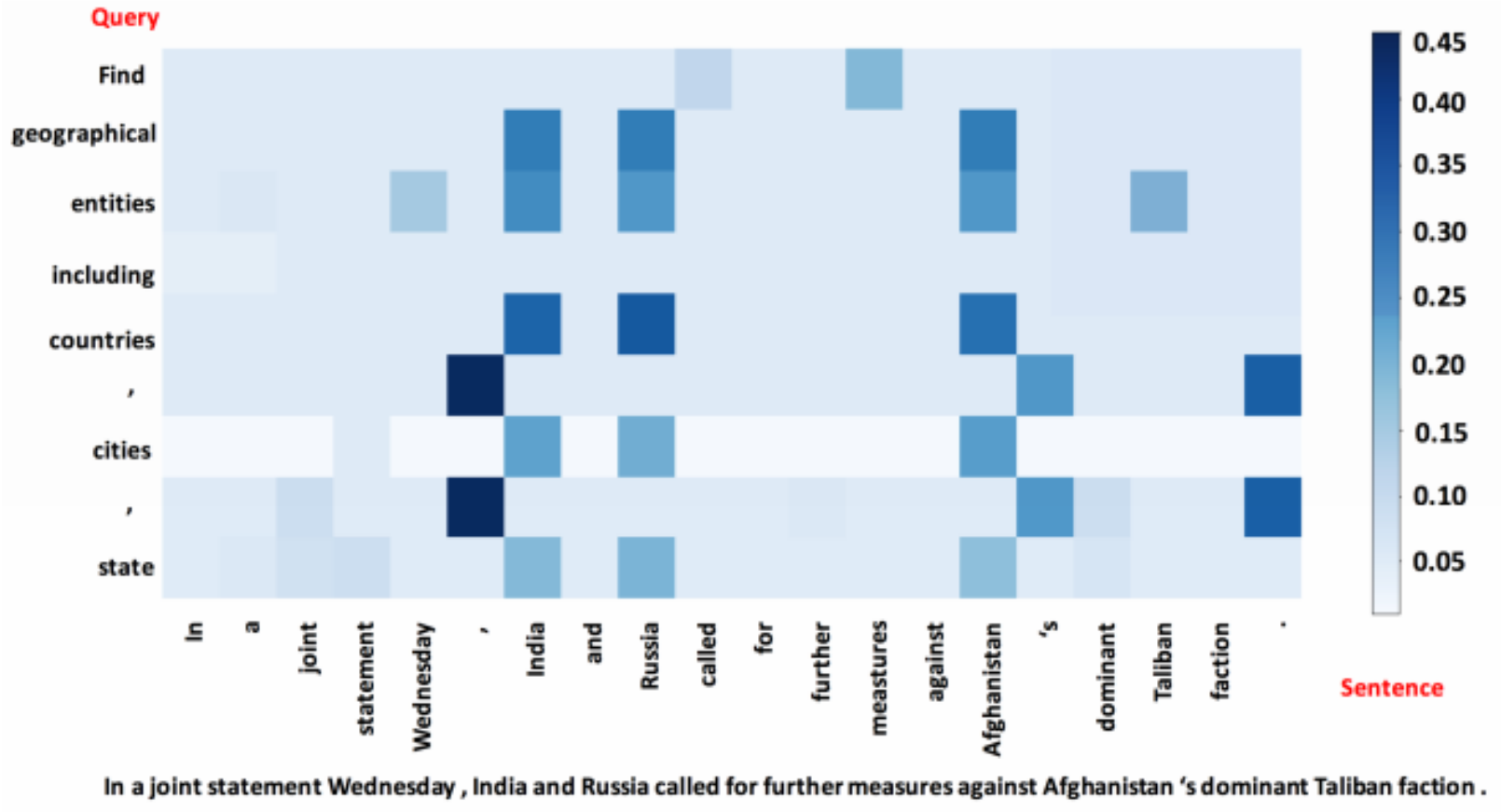
English OntoNotes 5.0	
Model	F
LSTM tagger (Strubell et al., 2017)	86.84
BiDAF (Seo et al., 2017)	87.39 (+0.55)
QAnet (Yu et al., 2018)	87.98 (+1.14)
BERT-Tagger	89.16
BERT-MRC	**91.11** (+1.95)

How to Construct Queries

- How to construct queries has a significant influence
  - Annotation guideline notes
  - Position index of labels: the index of a tag
  - Keyword
  - *Rule-based template filling*
  - Wikipedia: Wikipedia description
  - Synonyms: synonyms from Oxford Dictionary
  - Keyword+synonyms: concatenation

English OntoNotes 5.0	
Model	F1-score
BERT-Tagger	89.16
Position index of labels	88.29 (-0.87)
Keywords	89.74 (+0.58)
Wikipedia	89.66 (+0.59)
Rule-based template filling	89.30 (+0.14)
Synonyms	89.92 (+0.76)
Keywords+Synonyms	90.23 (+1.07)
Annotation guideline notes	91.11 (+1.95)

How to Construct Queries



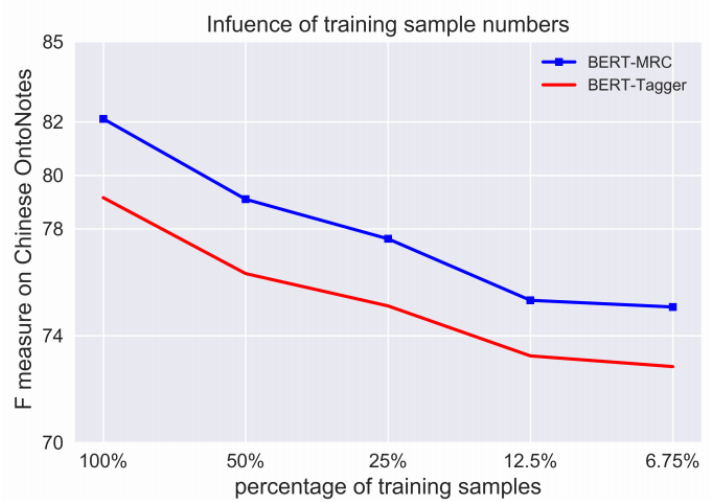
Zero-shot Evaluation on Unseen Labels

- **The question-answering formalization in MRC framework, which predicts the answer to the given query comes with more generalization capability, achieving acceptable results**

Models	Train	Test	F1-score
BERT-tagger	OntoNotes5.0	OntoNotes5.0	89.16
BERT-MRC	OntoNotes5.0	OntoNotes5.0	91.11
BERT-tagger	CoNLL03	OntoNotes5.0	31.87
BERT-MRC	CoNLL03	OntoNotes5.0	72.34

Size of Training Data

- **Since the natural language query encodes significant prior knowledge, we expect that the proposed framework works better with less training data**




香侬科技
SHANNON.AI
Any Questions?
2019.11.11

MRC for Joint Entity and Relation Extraction

- Once we have extracted entities, we are interested in extracting their relations

MRC for Joint Entity and Relation Extraction

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnelconstruction Company.

MRC for Joint Entity and Relation Extraction

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnelconstruction Company.

MRC for Joint Entity and Relation Extraction

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnelconstruction Company.

MRC for Joint Entity and Relation Extraction

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnelconstruction Company.

MRC for Joint Entity and Relation Extraction

In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer. He helped fund Tesla, Inc., an electric vehicle and solar panel manufacturer, in 2003, and became its CEO and product architect. In 2006, he inspired the creation of SolarCity, a solar energy services Company, and operates as its chairman. In 2016, he co-founded Neuralink, a neurotechnology Company focused on developing brain–computer interfaces, and is its CEO. In 2016, Musk founded The Boring Company, an infrastructure and tunnel construction Company.

MRC for Joint Entity and Relation Extraction

- Important entities are usually named
  - Person: Musk
  - Corp: SpaceX, Tesla, SolarCity, Neuralink, The Boring Company
  - Time: 2002, 2003, 2006, 2016
  - Position: CEO, lead designer, product architect, chairman
  - ……

MRC for Joint Entity and Relation Extraction

- Next step: the relations between given two entities?

MRC for Joint Entity and Relation Extraction

- Next step: the relations between given two entities?
  - Musk(Person)____SpaceX(Corp)
  - Musk(Person)____Tesla(Corp)
  - Musk(Person)____CEO(Position)
  - SpaceX(Corp)____2002(Time)

MRC for Joint Entity and Relation Extraction

- Next step: the relations between given two entities?
  - Musk(Person) FOUND SpaceX(Corp)
  - Musk(Person) FOUND Tesla(Corp)
  - Musk(Person) SERVE_AS CEO(Position)
  - SpaceX(Corp) FOUND_TIME 2002(Time)

MRC for Joint Entity and Relation Extraction

- Next step: the relations between given two entities?
  - Musk(Person) FOUND SpaceX(Corp)
  - Musk(Person) FOUND Tesla(Corp)
  - Musk(Person) SERVE_AS CEO(Position)
  - SpaceX(Corp) FOUND_TIME 2002(Time)
- Extracted entities have underlying relations

Person	Corp	Time	Position
Musk	SpaceX	2002	CEO
Musk	Tesla	2003	CEO& product architect
Musk	SolarCity	2006	chairman
Musk	Neuralink	2016	CEO
Musk	The Boring Company	2016	-

Existing Problems

- The style of triple $REL(e_1, e_2)$ can not reflect the structured information due to the widely existing hierarchies in texts

Existing Problems

- The style of triple $REL(e_1, e_2)$ can not reflect the structured information due to the widely existing hierarchies in texts
  - *In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer……*

Existing Problems

- The style of triple $REL(e_1, e_2)$ can not reflect the structured information due to the widely existing hierarchies in texts
  - *In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer……*
  - Extracting Time depends on Position
  - Extracting Position depends on Company

Existing Problems

- The style of triple $REL(e_1, e_2)$ can not reflect the structured information due to the widely existing hierarchies in texts
  - *In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer……*
  - Extracting Time depends on Position
  - Extracting Position depends on Company
- **Entities and relations are often very complicated in real texts**

Existing Problems

- The style of triple $REL(e_1, e_2)$ can not reflect the structured information due to the widely existing hierarchies in texts
  - *In 2002, Musk founded SpaceX, an aerospace manufacturer and space transport services Company, of which he is CEO and lead designer……*
  - Extracting Time depends on Position
  - Extracting Position depends on Company
- **Entities and relations are often very complicated in real texts**
  - Entities can be far away
  - One specific entity can be in more than one triples
  - One sentence contains more same relations
  - Relations can overlap

Solutions

- The task of extracting entities and relations together can be regarded as an ***MRC*** problem, as we do in NER.

Solutions

- The task of extracting entities and relations together can be regarded as an ***MRC*** problem, as we do in NER.
  - To extract person: Q: who is mentioned in the test? A: e_1
  - To extract company: Q: which company did e_1 work for? A: e_2
  - To extract position: Q: what was e_1 's position in e_2? A: e_3
  - To extract time: Q: During which period did e_1 work for e_2 as e_3? A: e_4

Solutions

- The task of extracting entities and relations together can be regarded as an ***MRC*** problem, as we do in NER.
  - To extract person: Q: who is mentioned in the test? A: e_1
  - To extract company: Q: which company did e_1 work for? A: e_2
  - To extract position: Q: what was e_1 's position in e_2? A: e_3
  - To extract time: Q: During which period did e_1 work for e_2 as e_3? A: e_4
- Advantages:
  - Capture hierarchical dependencies between relations and entities

Solutions

- The task of extracting entities and relations together can be regarded as an ***MRC*** problem, as we do in NER.
  - To extract person: Q: who is mentioned in the test? A: e_1
  - To extract company: Q: which company did e_1 work for? A: e_2
  - To extract position: Q: what was e_1 's position in e_2? A: e_3
  - To extract time: Q: During which period did e_1 work for e_2 as e_3? A: e_4
- Advantages:
  - Capture hierarchical dependencies between relations and entities
  - Encode prior information in questions

Solutions

- The task of extracting entities and relations together can be regarded as an ***MRC*** problem, as we do in NER.
  - To extract person: Q: who is mentioned in the test? A: e_1
  - To extract company: Q: which company did e_1 work for? A: e_2
  - To extract position: Q: what was e_1 's position in e_2? A: e_3
  - To extract time: Q: During which period did e_1 work for e_2 as e_3? A: e_4
- Advantages:
  - Capture hierarchical dependencies between relations and entities
  - Encode prior information in questions
  - A natural way to joint extract entities and relations

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?
  - Define question templates

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?
  - Define question templates
  - EntityQuesTemplates: extract entity e_1, which represents the head-entity

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?
  - Define question templates
  - EntityQuesTemplates: extract entity e_1, which represents the head-entity
  - ChainOfRelTemplates: extract relation r and tail-entity e_2

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?
  - Define question templates
  - EntityQuesTemplates: extract entity e_1, which represents the head-entity
  - ChainOfRelTemplates: extract relation r and tail-entity e_2
  - A triple (e_1, r, e_2) has been extracted!

How to Construct Questions?

- How to construct questions for joint entity and relation extraction?
  - Define question templates
  - EntityQuesTemplates: extract entity e_1, which represents the head-entity
  - ChainOfRelTemplates: extract relation r and tail-entity e_2
  - A triple (e_1, r, e_2) has been extracted!
  - A special token *NONE* can be returned meaning there's no answer to the question

Full Algorithm

Input: sentence s, EntityQuesTemplates, ChainOfRelTemplates
Output: a list of list (table) M = []

```
M ← ∅
HeadEntList← ∅
for entity_question in EntityQuesTemplates do
    e_1 = Extract_Answer(entity_question, s)
    if e_1 ≠ NONE do
        HeadEntList = HeadEntList + {e_1}
    endif
end for
for head_entity in HeadEntList do
    ent_list = [head_entity]
    for [rel, rel_temp] in ChainOfRelTemplates do
        for (rel, rel_temp) in List of [rel, rel_temp] do
            q = GenQues(rel_temp, rel, ent_list)
            e = Extract_Answer(rel_question, s)
            if e ≠ NONE
                ent_list = ent_list + e
            endif
        end for
    end for
    if len(ent_list)=len([rel, rel_temp])
        M = M + ent_list
    endif
end for
return M
```

Full Algorithm

Input: sentence s, EntityQuesTemplates, ChainOfRelTemplates
Output: a list of list (table) M = []

```
M ← ∅
HeadEntList← ∅
for entity_question in EntityQuesTemplates do    // extract head entities
    e_1 = Extract_Answer(entity_question, s)
    if e_1 ≠ NONE do
        HeadEntList = HeadEntList + {e_1}
    endif
end for
for head_entity in HeadEntList do
    ent_list = [head_entity]
    for [rel, rel_temp] in ChainOfRelTemplates do
        for (rel, rel_temp) in List of [rel, rel_temp] do
            q = GenQues(rel_temp, rel, ent_list)
            e = Extract_Answer(rel_question, s)
            if e ≠ NONE
                ent_list = ent_list + e
            endif
        end for
    end for
    if len(ent_list)=len([rel, rel_temp])
        M = M + ent_list
    endif
end for
return M
```

Full Algorithm

Input: sentence s, EntityQuesTemplates, ChainOfRelTemplates
Output: a list of list (table) M = []

```
M ← ∅
HeadEntList← ∅
for entity_question in EntityQuesTemplates do    // extract head entities
    e_1 = Extract_Answer(entity_question, s)
    if e_1 ≠ NONE do
        HeadEntList = HeadEntList + {e_1}
    endif
end for                                          // finish extracting head entities
for head_entity in HeadEntList do
    ent_list = [head_entity]
    for [rel, rel_temp] in ChainOfRelTemplates do
        for (rel, rel_temp) in List of [rel, rel_temp] do
            q = GenQues(rel_temp, rel, ent_list)
            e = Extract_Answer(rel_question, s)
            if e ≠ NONE
                ent_list = ent_list + e
            endif
        end for
    end for
    if len(ent_list)=len([rel, rel_temp])
        M = M + ent_list
    endif
end for
return M
```

Full Algorithm

Input: sentence s, EntityQuesTemplates, ChainOfRelTemplates
Output: a list of list (table) M = []

```
M ← ∅
HeadEntList← ∅
for entity_question in EntityQuesTemplates do    // extract head entities
    e_1 = Extract_Answer(entity_question, s)
    if e_1 ≠ NONE do
        HeadEntList = HeadEntList + {e_1}
    endif
end for                                          // finish extracting head entities
for head_entity in HeadEntList do
    ent_list = [head_entity]
    for [rel, rel_temp] in ChainOfRelTemplates do      // ChainOfRelTemplates defines the
        for (rel, rel_temp) in List of [rel, rel_temp] do    // order of relations. Manully.
            q = GenQues(rel_temp, rel, ent_list)
            e = Extract_Answer(rel_question, s)   // joint extraction of the relation and tail-entity
            if e ≠ NONE
                ent_list = ent_list + e
            endif
        end for
    end for
    if len(ent_list)=len([rel, rel_temp])
        M = M + ent_list
    endif
end for
return M
```

Example Templates

Relation Type	head-e	tail-e	Natural Language Question & Template Question
GEN-AFF	FAC	GPE	find a geo-political entity that connects to XXX XXX; has affiliation; geo-political entity
PART-WHOLE	FAC	FAC	find a facility that geographically relates to XXX XXX; part whole; facility
PART-WHOLE	FAC	GPE	find a geo-political entity that geographically relates to XXX XXX; part whole; geo-political entity
PART-WHOLE	FAC	VEH	find a vehicle that belongs to XXX XXX; part whole; vehicle
PHYS	FAC	FAC	find a facility near XXX? XXX; physical; facility
ART	GPE	FAC	find a facility which is made by XXX XXX; agent artifact; facility
ART	GPE	VEH	find a vehicle which is owned or used by XXX XXX; agent artifact; vehicle
ART	GPE	WEA	find a weapon which is owned or used by XXX XXX; agent artifact; weapon
ORG-AFF	GPE	ORG	find an organization which is invested by XXX XXX; organization affiliation; organization
PART-WHOLE	GPE	GPE	find a geo political entity which is controlled by XXX XXX; part whole; geo-political entity
PART-WHOLE	GPE	LOC	find a location geographically related to XXX XXX; part whole; location

Training

- Use the standard MRC framework to extract answer spans

Training

- Use the standard MRC framework to extract answer spans
  - Loss: $L = (1 - \lambda)L(head - entity) + \lambda L(tail - entity, relation)$

Training

- Use the standard MRC framework to extract answer spans
  - Loss: $L = (1 - \lambda)L(head - entity) + \lambda L(tail - entity, relation)$
- **Reinforcement learning to encourage further correctness**

Training

- Use the standard MRC framework to extract answer spans
  - Loss: $L = (1 - \lambda)L(head - entity) + \lambda L(tail - entity, relation)$
- **Reinforcement learning to encourage further correctness**

$$p(y(w_1, w_2, \cdots, w_n) = \text{answer}|\text{question}, s) = p(w_1 = B) \times p(w_n = E) \prod_{2 \le i \le n-1} p(w_i = M)$$

$$\nabla E(\theta) \approx [R(w) - b] \nabla \log \pi(y(w)|\text{question}, s)$$

Experiments

Models	Entity P	Entity R	Entity F	Relation P	Relation R	Relation F
			ACE04			
Li and Ji (2014)	**83.5**	76.2	79.7	**60.8**	36.1	49.3
Miwa and Bansal (2016)	80.8	**82.9**	**81.8**	48.7	**48.1**	**48.4**
Katiyar and Cardie (2017)	81.2	78.1	79.6	46.4	45.3	45.7
Bekoulis et al. (2018)	-	-	81.6	-	-	47.5
Multi-turn QA	**84.4**	82.9	**83.6**	50.1	**48.7**	**49.4**
			ACE05			
Li and Ji (2014)	**85.2**	76.9	80.8	**65.4**	39.8	49.5
Miwa and Bansal (2016)	82.9	**83.9**	83.4	57.2	54.0	55.6
Katiyar and Cardie (2017)	84.0	81.3	82.6	55.5	51.8	53.6
Zhang et al. (2017)	-	-	83.5	-	-	57.5
Sun et al. (2018)	83.9	83.2	**83.6**	64.9	**55.1**	**59.6**
Multi-turn QA	84.7	**84.9**	**84.8**	64.8	**56.2**	**60.2**
			CoNLL-04			
Miwa and Sasaki (2014)	–	–	80.7	–	–	61.0
Zhang et al. (2017)	–	–	85.6	–	–	67.8
Bekoulis et al. (2018)	–	–	83.6	–	–	62.0
Multi-turn QA	**89.0**	**86.6**	**87.8**	**69.2**	**68.2**	**68.9**

Cases

- Example 1
  - **Gold**: [john Scottsdale/PER: PHYS-1] is on the front lines in [Iraq/ GPE:PHYS-1].
  - **MRT**: [john Scottsdale/PER] is on the front lines in [Iraq/GPE].
  - **Multi-QA**: [john Scottsdale/PER: PHYS-1] is on the front lines in [Iraq/ GPE:PHYS-1].
- Example 2
  - **Gold**: The [men/PER: ART-1] held on the sinking [vessel/VEH: ART-1] until the [passenger/PER: ART-2] [ship/VEH: ART-2] was able to reach them.
  - **MRT**: The [men/PER: ART-1] held on the sinking [vessel/VEH: ART-1] until the [passenger/PER] ship was able to reach them.
  - **Multi-QA**: The [men/PER: ART-1] held on the sinking [vessel/VEH: ART-1] until the [passenger/PER: ART-2] [ship/VEH: ART-2] was able to reach them.


香侬科技
SHANNON.AI
Any Questions?
2019.11.11

MRC for Coreference Resolution

- Are MRC framework only useful for entity/relation extraction?

MRC for Coreference Resolution

- Are MRC framework only useful for entity/relation extraction?
- It can be used for much more complicated task

MRC for Coreference Resolution

- Are MRC framework only useful for entity/relation extraction?
- It can be used for much more complicated task
- Such as *coreference resolution*, which considers all text spans in a document

Specific Advantage for Coreference Resolution

- **Standard Method**

  1. Mention Proposal

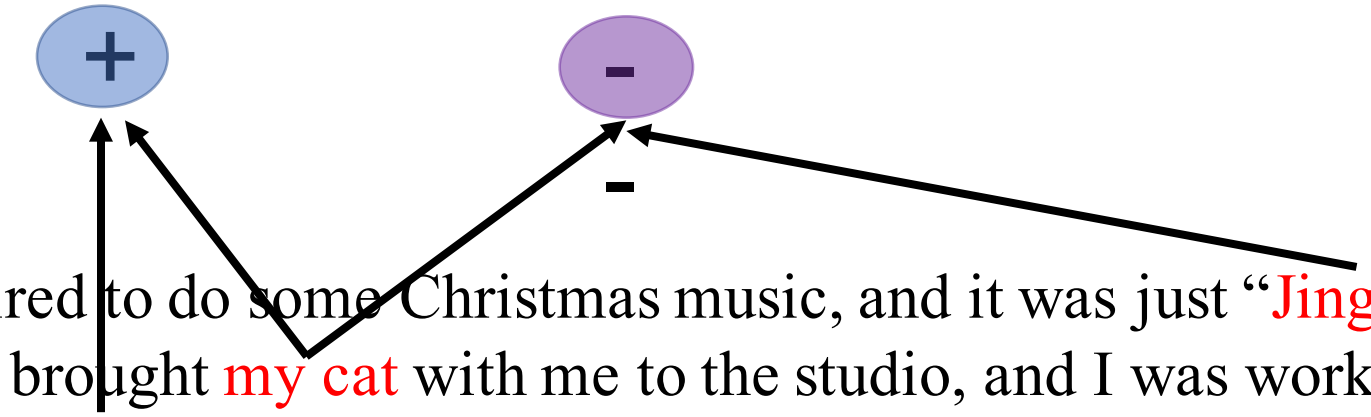
X = I was hired to do some Christmas music, and it was just “Jingle Bells” and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.

Specific Advantage for Coreference Resolution

- **Standard Method**

  1. Mention Proposal

  2. Mention Linking



X = I was hired to do some Christmas music, and it was just "Jingle Bells" and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.

Specific Advantage for Coreference Resolution

- **Standard Method**

  Drawback: Missing mentions at the mention proposal stage can never be retrieved

Specific Advantage for Coreference Resolution

- **Standard Method**

  Drawback: Missing mentions at the mention proposal stage can never be retrieved

- MRC

  X = it was just “Jingle Bells” and I brought my cat with me to the studio, and I was working on the song and the cat jumped up into the record booth and started meowing along, meowing to me.
  Q = it was just “ <mention> Jingle Bells <mention>”
  A = the song

If partial mentions are missed by the mention proposal model, they can still be retrieved in the MRC mention linking stage.

Task Definition

- Given an input sequence $X = \{x_1, \cdots, x_n\}$, there are $N = \frac{n(n+1)}{2}$ spans in total

Task Definition

- Given an input sequence $X = \{x_1, \cdots, x_n\}$, there are $N = \frac{n(n+1)}{2}$ spans in total
- e_i is the i-th span with the start index **FIRSR(i)** and the end index **LAST(i)**

Task Definition

- Given an input sequence $X = \{x_1, \cdots, x_n\}$, there are $N = \frac{n(n+1)}{2}$ spans in total
- e_i is the i-th span with the start index **FIRSR(i)** and the end index **LAST(i)**
- The task of coreference resolution is to *determine the antecedents for all possible spans*

Existing Models and Problems

- Most existing coreference resolution models only consider mentions extracted by *mention proposal module* to find antecedent for each possible mention

Existing Models and Problems

- Most existing coreference resolution models only consider mentions extracted by *mention proposal module* to find antecedent for each possible mention
  - Two issues: (1) task formalization; (2) the algorithm

Existing Models and Problems

- Most existing coreference resolution models only consider mentions extracted by *mention proposal module* to find antecedent for each possible mention
  - Two issues: (1) task formalization; (2) the algorithm
  - **Task formalization**: mentions left out at the *mention proposal stage* can never be recovered since the mention-ranking model only operates on the proposed mentions.

Existing Models and Problems

- Most existing coreference resolution models only consider mentions extracted by *mention proposal module* to find antecedent for each possible mention
  - Two issues: (1) task formalization; (2) the algorithm
  - **Task formalization**: mentions left out at the *mention proposal stage* can never be recovered since the mention-ranking model only operates on the proposed mentions.
  - **The algorithm**: models *score each pair of mentions* at the output layer level, which means that the model lacks explicit emphasis on the mentions and their contexts

MRC as Solution

- To alleviate these two issues, we propose to formulate the coreference resolution problem as a span prediction task, akin to machine reading comprehension (MRC).

MRC as Solution

- Reformulate the coreference resolution problem as a span prediction task, akin to machine reading comprehension (MRC).
- A query is generated for each candidate mention using its surrounding context

MRC as Solution

- Reformulate the coreference resolution problem as a span prediction task, akin to machine reading comprehension (MRC).
- A query is generated for each candidate mention using its surrounding context
- A span prediction module is employed to extract the text spans of the coreferences within the document using the generated query.

MRC as Solution

- Reformulate the coreference resolution problem as a span prediction task, akin to machine reading comprehension (MRC).
- A query is generated for each candidate mention using its surrounding context (***Mention Proposal Module***)
- A span prediction module is employed to extract the text spans of the coreferences within the document using the generated query (***Mention Linking Module***)

Mention Proposal Module

- SpanBERT as backbone to induce representation x_i

Mention Proposal Module

- SpanBERT as backbone to induce representation x_i
- We only keep λT (usually set to 0.2 and T is the document length) spans with the highest mention scores by calculating each mention score using three feedforward layer:

$$S_m(x_{FIRST(i)}) = FFN([x_{FIRST(i)}]), S_m(x_{LAST(i)}) = FFN([x_{LAST(i)}])$$
$$s_m(x_{FIRST(i)}, x_{LAST(i)}) = FFN_m([x_{FIRST(i)}, x_{LAST(i)}])$$

$$S_m(i) = \frac{[S_m(x_{FIRST(i)}) + S_m(x_{LAST(i)}) + s_m(x_{FIRST(i)}, x_{LAST(i)})]}{3}$$

Mention Proposal Module

$$S_m(x_{FIRST(i)}) = FFN([x_{FIRST(i)}]), S_m(x_{LAST(i)}) = FFN([x_{LAST(i)}])$$
$$s_m(x_{FIRST(i)}, x_{LAST(i)}) = FFN_m([x_{FIRST(i)}, x_{LAST(i)}])$$

$$S_m(i) = \frac{[S_m(x_{FIRST(i)}) + S_m(x_{LAST(i)}) + s_m(x_{FIRST(i)}, x_{LAST(i)})]}{3}$$

- The mention proposal module is pretrained via three binary classification losses:

$$\begin{aligned}\text{Loss(m)} = &\text{sigmoid}(s_m(\boldsymbol{x}_{\text{FIRST}(i)})) \\ &+\text{sigmoid}(s_m(\boldsymbol{x}_{\text{END}(i)})) \\ &+\text{sigmoid}(s_m(\boldsymbol{x}_{\text{FIRST}(i)}, \boldsymbol{x}_{\text{LAST}(i)}))\end{aligned}$$

Mention Linking Module

- We use a mention linking network to give a score $s_a(i, j)$ to **any** mention pair e_i, e_j indicating whether they are coreference

Mention Linking Module

- We use a mention linking network to give a score $s_a(i, j)$ to **any** mention pair e_i, e_j indicating whether they are coreference
- This is done by an ***MRC*** framework

Mention Linking Module

- We use a mention linking network to give a score $s_a(i, j)$ to **any** mention pair e_i, e_j indicating whether they are coreference
- This is done by an ***MRC*** framework
  - Triple: {context(X), Query(q), answer(A)}

Mention Linking Module

- We use a mention linking network to give a score $s_a(i, j)$ to **any** mention pair e_i, e_j indicating whether they are coreference
- This is done by an ***MRC*** framework
  - Triple: {context(X), Query(q), answer(A)} (blue denotes input)
  - X is the full document, and q is the sentence e_i residing in and e_i is encapsulated with two special tokens <mention> and </mention>

Mention Linking Module

- The score of mention e_j being a coreference of e_i is given by:

$$s_a(j \mid i) = FFN_{j|i}[x_{FIRST(j)|i}, x_{LAST(j)|i}]$$

Mention Linking Module

- The score of mention e_j being a coreference of e_i is given by:

$$s_a(j \mid i) = FFN_{j|i}[x_{FIRST(j)|i}, x_{LAST(j)|i}]$$

- Which is further modified for bidirectional relations:

$$s_a(i,j) = \frac{1}{2}(s_a(j|i) + s_a(i|j))$$

Mention Linking Module

- The score of mention e_j being a coreference of e_i is given by:

$$s_a(j \mid i) = FFN_{j|i}[x_{FIRST(j)|i}, x_{LAST(j)|i}]$$

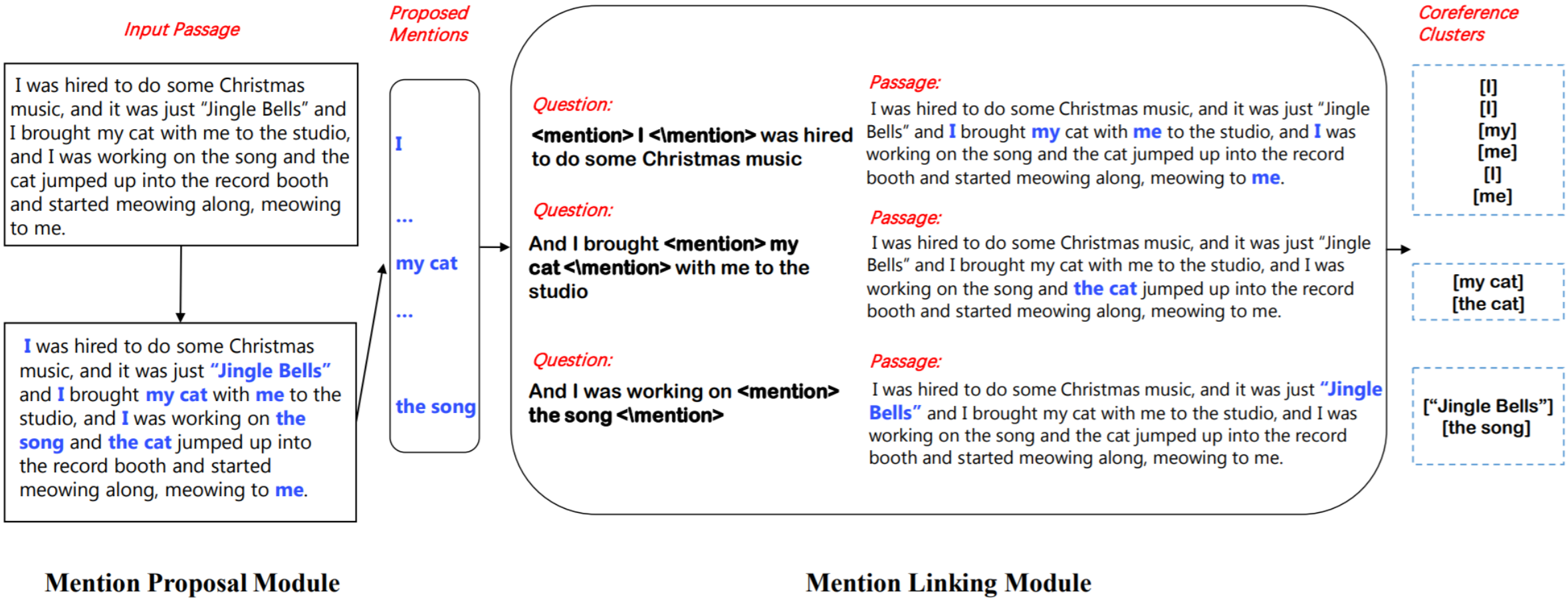
- Which is further modified for bidirectional relations:

$$s_a(i,j) = \frac{1}{2}(s_a(j|i) + s_a(i|j))$$

- The final score is:

$$s(i,j) = \lambda[s_{m(i)} + s_{m(j)}] + (1-\lambda)s_a(i,j)$$

Overall Structure



Advantages

- Has the flexibility of retrieving mentions left out at the mention proposal stage

Advantages

- Has the flexibility of retrieving mentions left out at the mention proposal stage
  - However, if all the mentions within the cluster are missed by the mention proposal module, none of them can be used for query construction, which means they all will be irreversibly left out

Advantages

- Has the flexibility of retrieving mentions left out at the mention proposal stage
  - However, if all the mentions within the cluster are missed by the mention proposal module, none of them can be used for query construction, which means they all will be irreversibly left out
  - The chance that mentions within a mention cluster are all missed is relatively low

Advantages

- Has the flexibility of retrieving mentions left out at the mention proposal stage
  - However, if all the mentions within the cluster are missed by the mention proposal module, none of them can be used for query construction, which means they all will be irreversibly left out
  - The chance that mentions within a mention cluster are all missed is relatively low
- Explicitly emphasizes the surrounding context of the mentions of interest

Training and Inference

- For each mention e_i proposed, it's associated with C potential spans based on $s(j|i)$

Training and Inference

- For each mention e_i proposed, it's associated with C potential spans based on $s(j|i)$
- Optimize the marginal log-likelihood of all correct antecedents

Training and Inference

- For each mention e_i proposed, it's associated with C potential spans based on $s(j|i)$
- Optimize the marginal log-likelihood of all correct antecedents
  - A dummy token ϵ is appended to the C candidates

Training and Inference

- For each mention e_i proposed, it's associated with C potential spans based on $s(j|i)$
- Optimize the marginal log-likelihood of all correct antecedents
  - A dummy token ϵ is appended to the C candidates
  - The model will output it if none of the C candidates is coreferential with e_i

Training and Inference

- For each mention e_i proposed, it's associated with C potential spans based on $s(j|i)$
- Optimize the marginal log-likelihood of all correct antecedents
  - A dummy token ϵ is appended to the C candidates
  - The model will output it if none of the C candidates is coreferential with e_i
  - A distribution is learned through softmax:

$$P(e_j) = \frac{e^{s(i,j)}}{\sum_{j' \in C} e^{s(i,j')}}$$

Training and Inference

$$P(e_j) = \frac{e^{s(i,j)}}{\sum_{j' \in C} e^{s(i,j')}}$$

- Given an input document, we can obtain an undirected graph using the overall score, each node of which represents a mention.
- We prune the graph by keeping the edge whose weight is the largest for each node based on $p(e_j)$. Nodes whose closest neighbor is ϵ are abandoned.
- Therefore, the mention clusters can be decoded from the graph.

Pretraining

- Further pretrain the mention linking network on the Queref (Dasigi et al., 2019b) and the SQuAD dataset (Rajpurkar et al., 2016b)

Experiments

	MUC			B^3			$CEAF_{\phi_4}$			
	P	R	F1	P	R	F1	P	R	F1	Avg. F1
e2e-coref(Lee et al., 2017)	78.4	73.4	75.8	68.6	61.8	65.0	62.7	59.0	60.8	67.2
c2f-coref + ELMo (Lee et al., 2018)	81.4	79.5	80.4	72.2	69.5	70.8	68.2	67.1	67.6	73.0
EE + BERT-large (Kantor and Globerson, 2019)	82.6	84.1	83.4	73.3	76.2	74.7	72.4	71.1	71.8	76.6
c2f-coref + BERT-large (Joshi et al., 2019b)	84.7	82.4	83.5	76.5	74.0	75.3	74.1	69.8	71.9	76.9
c2f-coref + SpanBERT-large (Joshi et al., 2019a)	85.8	84.8	85.3	78.3	77.9	78.1	76.4	74.2	75.3	79.6
CorefQA + SpanBERT-base	85.2	87.4	86.3	78.7	76.5	77.6	76.0	75.6	75.8	79.9 (+0.3)
CorefQA + SpanBERT-large	**88.6**	**87.4**	**88.0**	**82.4**	**82.0**	**82.2**	**79.9**	**78.3**	**79.1**	**83.1** (+3.5)

Experiments

- Results on GAP dataset (Webster et al. 2018)

Model	M	F	B	O
e2e-coref	67.2	62.2	0.92	64.7
c2f-coref + ELMo	75.8	71.1	0.94	73.5
c2f-coref + BERT-large	86.9	83.0	0.95	85.0
c2f-coref + SpanBERT-large	88.8	84.9	0.96	86.8
CorefQA + SpanBERT-large	**88.9**	**86.1**	**0.97**	**87.5**

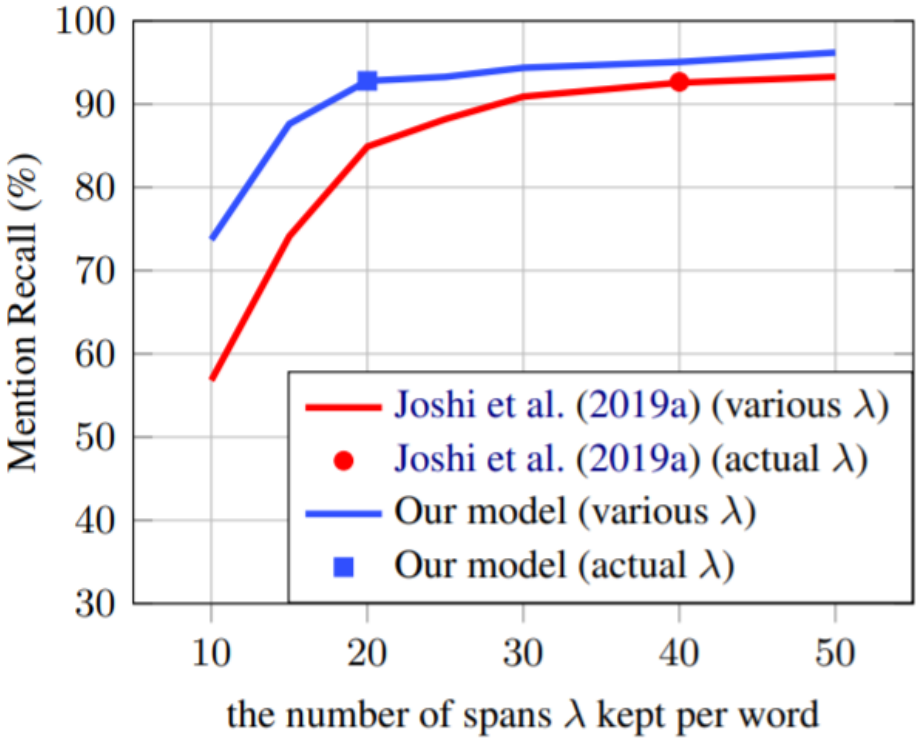
Table 2: CorefQA achieves the state-of-the-art performance on all metrics including F1 scores on **M**asculine and **F**eminine examples, a **B**ias factor (F / M) and the **O**verall F1 score.

Effect of Different Modules

	Avg. F1	Δ
CorefQA	83.4	
–– SpanBERT	79.6	-3.8
–– Mention Proposal Pre-training	75.9	-7.5
–– Question Answering	75.0	-8.4
–– Quoref Pre-training	82.7	-0.7
–– Squad Pre-training	83.1	-0.3

Effect of the Overall Mention Recall

- Since the proposed framework has the potential to retrieve *mentions missed at the mention proposal stage*, we expect it to have higher overall mention recall rate



Case

- Example mention clusters that were correctly predicted by our model, but *wrongly* predicted by c2fcoref + SpanBERT-large

1	[**Freddie Mac**] is giving golden parachutes to two of its ousted executives. . . . Yesterday Federal Prosecutions announced a criminal probe into [**the company**].
2	[**A traveling reporter**] now on leave and joins us to tell [**her**] story. Thank [**you**] for coming in to share this with us.
3	*Paula Zahn:* [**Thelma Gutierrez**] went inside the forensic laboratory where scientists are trying to solve this mystery. *Thelma Gutierrez:* In this laboratory alone [**I**] 'm surrounded by the remains of at least twenty different service members who are in the process of being identified so that they too can go home.


香侬科技
SHANNON.AI
Any Questions?
2019.11.11

Text Classification

- Text classification is the task to assign one or multiple category label(s) to a sequence of text tokens.
  - Sentiment analysis

    “I really like the movie!” —> Positive (Negative)
  - Topic classification

    “iPhone sales drop 20 percent in China”—> Economy (Politics, Sport)
  - Multi-aspect sentiment analysis

    “Clean updated room, friendly efficient staff, but rate was too high”

    —> cleanliness: positive; service: positive; price: negative

Text Classification

- More formally, given a sequence of text x, text classification aims at mapping it to category label(s) y:

$$P(y|x)$$

- Standardly, category labels are merely vocabulary indices, i.e. 1, 2, 3, …
- Two drawbacks then arise:
  - Labels also contain rich information, which is missing under the standard formalization
  - Different categories may entangle in the text (e.g. multi-aspect sentiment analysis)

MRC for Text Classification

- Solution: explicitly fuse label semantics into the process of text classification, i.e., using *label descriptions*, akin to machine reading comprehension

"I really like the movie!"—> Positive / Negative

"I really like the movie! [SEP]$\boldsymbol{q}_{\text{positive}}$"—> Probability

"I really like the movie! [SEP]$\boldsymbol{q}_{\text{negative}}$"—> Probability

MRC for Text Classification

- Source text $x = \{x_1, \cdots, x_T\}$
- Target label $y \in \{1, \cdots, N\} = \mathbf{Y}$
- Label Description q_y
- Two approaches
  - N binary classifiers: $\{[\text{CLS}];\ q_y;\ [\text{SEP}];\ x\}, \forall y \in \mathbf{Y}$
  - One N-class classifier: $\{[\text{CLS}];\ q_1;\ [\text{CLS}];\ q_2; \cdots;\ [\text{CLS}];\ q_N;\ [\text{SEP}];\ x\}$

- One main question: **how to generate q_y?**

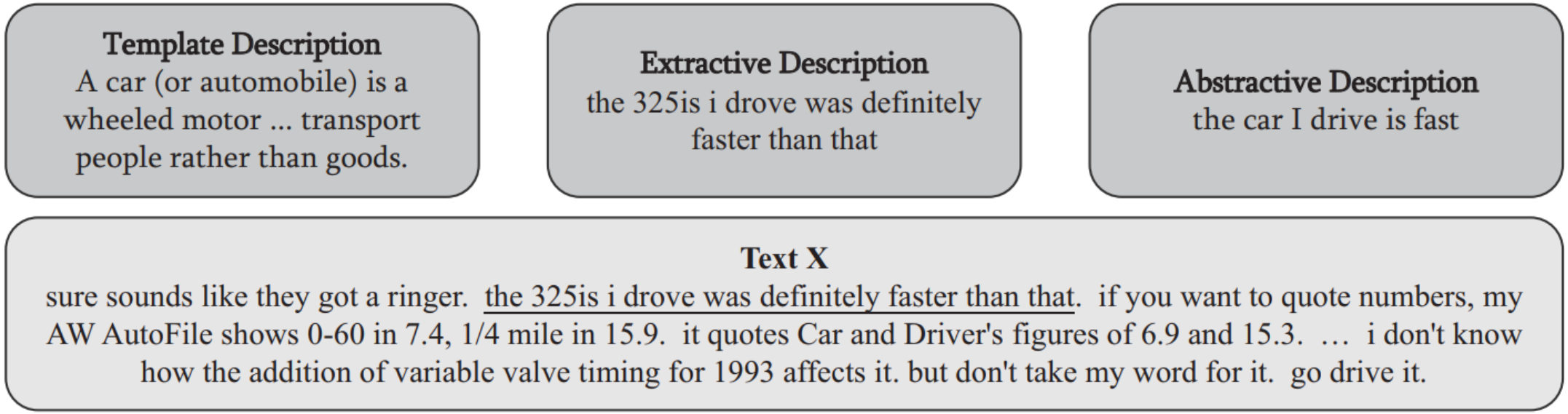
Generating Descriptions

- The simplest way is to use fixed template descriptions for category labels (called *Template Strategy*)

Label	Description
COMP.SYS.MAC.HARDWARE	The Macintosh is a family of personal computers designed ... since January 1984.
REC.AUTOS	A car (or automobile) is a wheeled motor ... transport people rather than goods.
TALK.POLITICS.MISC	Politics is a set of activities ... making decisions that apply to groups of members.

- But template descriptions can be not appropriate for all categories
  - Automatic descriptions according to the current text
  - Two strategies can be exploited: *extractive* and *abstractive*

Generating Descriptions



- Use *reinforcement learning* for both extractive and abstractive strategies!

Extractive Strategy

- Source text $x = \{x_1, \cdots, x_T\}$, Target label y, Label Description q_y
- Extracting a piece of text as description= selecting the start index i_s and the end index i_e

- Action: a_{i_s,i_e} as a text span

- Policy: $P_{\text{start}}(y,k) = \frac{\exp(W^{ys}h_k)}{\sum_{t=1}^{t=T}\exp(W^{ys}h_t)}$ $P_{\text{end}}(y,k) = \frac{\exp(W^{ye}h_k)}{\sum_{t=1}^{t=T}\exp(W^{ye}h_t)}$ $P_{\text{span}}(y, a_{i_s,i_e}) = P_{\text{start}}(y,i_s) \times P_{\text{end}}(y,i_e)$

- Reward: The probability of assigning the correct label(s) to x
- Update by REINFORCE(Williams, 1992)

Extractive Strategy

- For classes that should not be assigned, there might be no corresponding span in x that can be used as descriptions
  - Solution: append N dummy tokens to x
  - If the extractive model selects a dummy token, it means the text should not be assigned the category label

Input Text	Category	Extracted Span	Concatenation
I really like the movie! [T1] [T2]	Positive	really like	[CLS] really like [SEP] I really like the movie!
I really like the movie! [T1] [T2]	Negative	[T2]	[CLS] [T2] [SEP] I really like the movie!

- The extractive model is initialized to pick dummy tokens

Abstractive Strategy

- Source text $x = \{x_1, \cdots, x_T\}$, Target label y, Label Description q_y
- Use seq2seq models to generate descriptions for each label
- Action and Policy:

$$P_{\text{SEQ2SEQ}}(\boldsymbol{q}_y|x) = \prod_{i=1}^{L} p_\theta(q_i|q_{<i}, x, y)$$

  - Append a unique class embedding when generating
- Reward:
  - REGS (Reward for Every Generation Step): trains a discriminator that is able to assign rewards to partially decoded sequences rather than to all tokens within one sequence

$$\nabla \mathcal{L} \approx -\sum_{i=1}^{L} \nabla \log \pi(q_i|q_{<i}, h_y)[R(q_{<i}) - b(q_{<i})]$$

- Initialized using a pretrained seq2seq model with input being x and output being template descriptions

Experiments

- Three tasks
  - **Single-label Classification**: AGNews, 20newsgroups, DBPedia, Yahoo! Answers, YelpP, IMDB
  - **Multi-label Classification**: Reuters, AAPD
  - **Multi-aspect Sentiment Analysis**: BeerAdvocate, TripAdvisor
- Baselines
  - LSTM (Zhang et al., 2015)
  - Hierarchical Attention (Yang et al., 2016)
  - Label Embedding (Wang et al., 2018)
  - BERT (Devlin et al., 2019; Adhikari et al., 2019)

Results

Table 2. Test error rates on the AGNews, 20news, DBPedia, Yahoo, Yelp P and IMDB datasets for single-label classification.

Model	AGNews	20news	DBPedia	Yahoo	YelpP	IMDB
Char-level CNN (Zhang et al., 2015)	8.5	–	1.4	28.8	4.4	–
VDCNN (Conneau et al., 2016)	8.7	–	1.3	26.6	4.3	–
DPCNN (Johnson & Zhang, 2017)	6.9	–	0.91	23.9	2.6	–
Label Embedding (Wang et al., 2018b)	7.5	–	1.0	22.6	4.7	–
LSTMs (Zhang et al., 2015)	13.9	22.5	1.4	29.2	5.3	9.6
Hierarchical Attention (Yang et al., 2016)	11.8	19.6	1.6	24.2	5.0	8.0
D-LSTM(Yogatama et al., 2017)	7.9	–	1.3	26.3	7.4	–
Skim-LSTM (Seo et al., 2018)	6.4	–	–	–	–	8.8
BERT (Devlin et al., 2018)	5.9	16.9	0.72	22.7	2.4	6.8
Description (Tem.)	5.2	15.8	0.65	22.1	2.2	5.8
Description (Ext.)	**5.0**	15.6	**0.63**	22.0	2.1	**5.5**
Description (Abs.)	5.1	**15.4**	0.62	**21.8**	**2.0**	**5.5**

Table 3. Test error rates on the Reuters and AAPD datasets for multi-label classification.

Model	Reuters	AAPD
LSTMs (Zhang et al., 2015)	16.8	33.5
Hi-Attention (Yang et al., 2016)	13.9	30.3
Label-Emb (Wang et al., 2018b)	13.6	29.9
LSTM$_{reg}$ (Adhikari et al., 2019a)	13.0	29.5
BERT (Adhikari et al., 2019b)	11.0	26.6
Description (Tem.)	10.3	25.9
Description (Ext.)	10.1	26.0
Description (Abs.)	**10.0**	**25.7**

Table 4. Test error rates on the BeerAdvocate (Beer), TripAdvisor (Trip) for multi-aspect sentiment classification.

Model	Beer	Trip
LSTMs (Zhang et al., 2015)	34.9	47.6
Hi-Attention (Yang et al., 2016)	33.3	42.2
Label-Emb (Wang et al., 2018b)	32.0	43.5
BERT (Devlin et al., 2018)	27.8	35.6
Description (Tem.)	17.4	18.1
Description (Ext.)	16.0	**17.0**
Description (Abs.)	**15.6**	17.6

Results

Table 2. Test error rates on the AGNews, 20news, DBPedia, Yahoo, Yelp P and IMDB datasets for single-label classification.

Model	AGNews	20news	DBPedia	Yahoo	YelpP	IMDB
Char-level CNN (Zhang et al., 2015)	8.5	–	1.4	28.8	4.4	–
VDCNN (Conneau et al., 2016)	8.7	–	1.3	26.6	4.3	–
DPCNN (Johnson & Zhang, 2017)	6.9	–	0.91	23.9	2.6	–
Label Embedding (Wang et al., 2018b)	7.5	–	1.0	22.6	4.7	–
LSTMs (Zhang et al., 2015)	13.9	22.5	1.4	29.2	5.3	9.6
Hierarchical Attention (Yang et al., 2016)	11.8	19.6	1.6	24.2	5.0	8.0
D-LSTM(Yogatama et al., 2017)	7.9	–	1.3	26.3	7.4	–
Skim-LSTM (Seo et al., 2018)	6.4	–	–	–	–	8.8
BERT (Devlin et al., 2018)	5.9	16.9	0.72	22.7	2.4	6.8
Description (Tem.)	5.2	15.8	0.65	22.1	2.2	5.8
Description (Ext.)	**5.0**	15.6	**0.63**	22.0	2.1	**5.5**
Description (Abs.)	5.1	**15.4**	0.62	**21.8**	**2.0**	**5.5**

Table 3. Test error rates on the Reuters and AAPD datasets for multi-label classification.

Model	Reuters	AAPD
LSTMs (Zhang et al., 2015)	16.8	33.5
Hi-Attention (Yang et al., 2016)	13.9	30.3
Label-Emb (Wang et al., 2018b)	13.6	29.9
LSTM$_{reg}$ (Adhikari et al., 2019a)	13.0	29.5
BERT (Adhikari et al., 2019b)	11.0	26.6
Description (Tem.)	10.3	25.9
Description (Ext.)	10.1	26.0
Description (Abs.)	**10.0**	**25.7**

Table 4. Test error rates on the BeerAdvocate (Beer), TripAdvisor (Trip) for multi-aspect sentiment classification.

Model	Beer	Trip
LSTMs (Zhang et al., 2015)	34.9	47.6
Hi-Attention (Yang et al., 2016)	33.3	42.2
Label-Emb (Wang et al., 2018b)	32.0	43.5
BERT (Devlin et al., 2018)	27.8	35.6
Description (Tem.)	17.4	18.1
Description (Ext.)	16.0	**17.0**
Description (Abs.)	**15.6**	17.6

Ability to disentangle multiple aspects and categories!

Effect of Templates

- Comparison of different templates
  - **Label Index**: the description is the index of a class, *i.e.* “one”, “two”, “three”.
  - **Keyword**: the description is the keyword extension of each category.
  - **Keyword Expansion**: we use Wordnet to retrieve the synonyms of keywords and the description is their concatenation.
  - **Wikipedia**: definitions drawn from Wikipedia

Model	Error Rate
BERT	16.9
Template Description (Label Index)	16.8 (-0.1)
Template Description (Keyword)	16.4 (-0.5)
Template Description (Key Expansion)	16.2 (-0.7)
Template Description (Wiki)	15.8 (-1.1)

Discussion

- Several questions needed to be answered
  - Are text lengths influential on performances?
  - Do the description-based models converge more slowly than standard models?
  - Is the training size influential on performances?

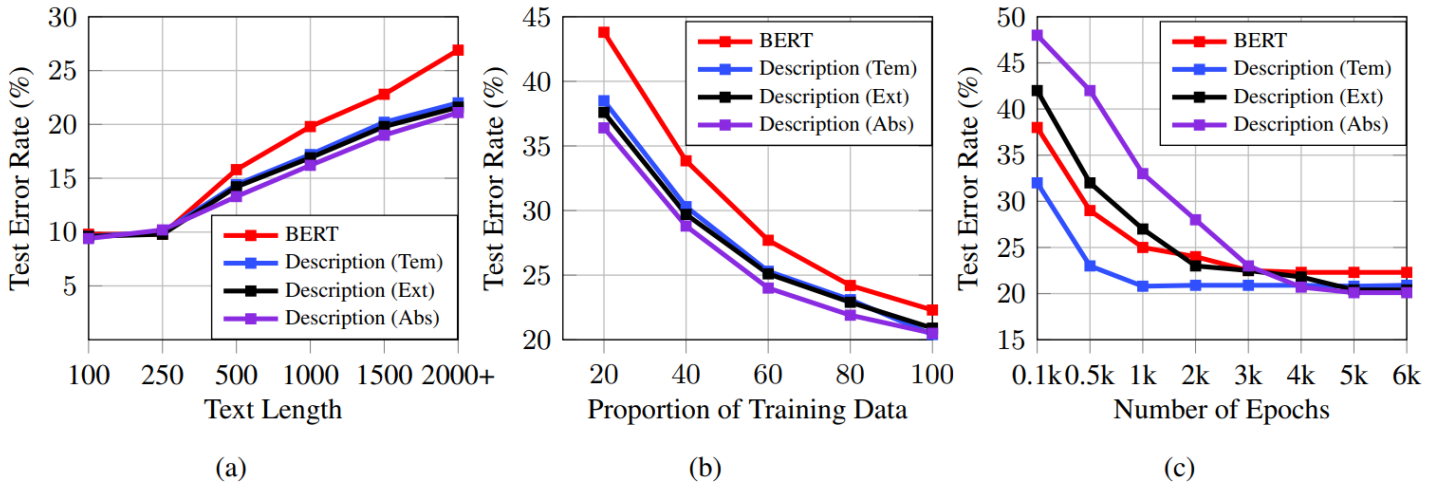


Figure 2. (a) Test error rate vs text length (b) Test error rate vs proportion of training data (c) Test error rate vs the number of epochs.

Impact of Initializations

- To what extent different initialization methods impact downstream performances?
  - Recap: we use *dummy Init* for extractive models and *template Init* for abstractive models

Table 6. Error rates for different RL initialization strategies.

	Yahoo Answer	AAPD
Template	22.4	25.9
Ext (dummy Init)	22.2	26.0
Ext (ROUGE-L Init)	25.3	27.2
Ext (random Init)	28.0	30.1
Abs (template Init)	22.0	25.7
Abs (random Init)	87.9	78.4


香侬科技
SHANNON.AI
Any Questions?
2019.11.11

Table of Contents

Code: https://github.com/ShannonAI/glyce

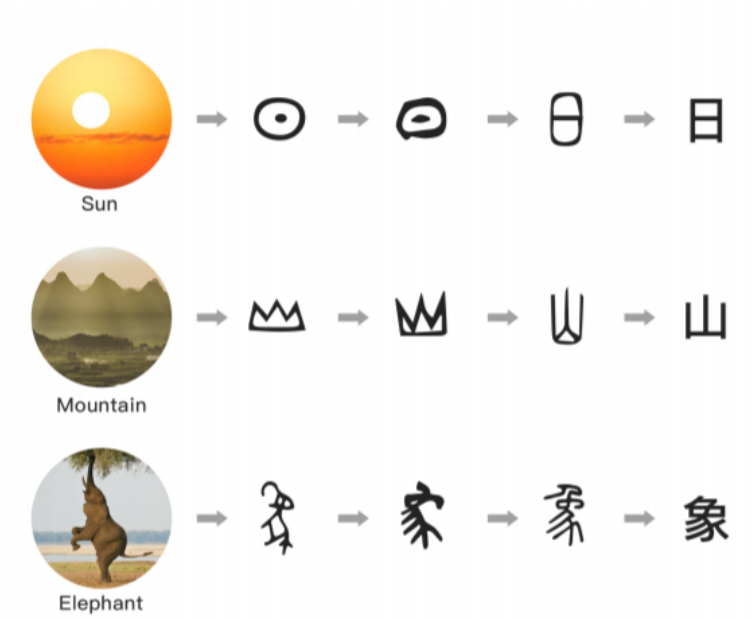
Link: https://arxiv.org/pdf/1901.10125.pdf

Chinese Character

- **The history of Chinese characters has witnessed the change of language semantics**

Chinese Character

- **The history of Chinese characters has witnessed the change of language semantics**



Chinese Character

- **The simplified characters still keep some common glyph properties, which can be used to extract similarities among characters.**

Chinese Character

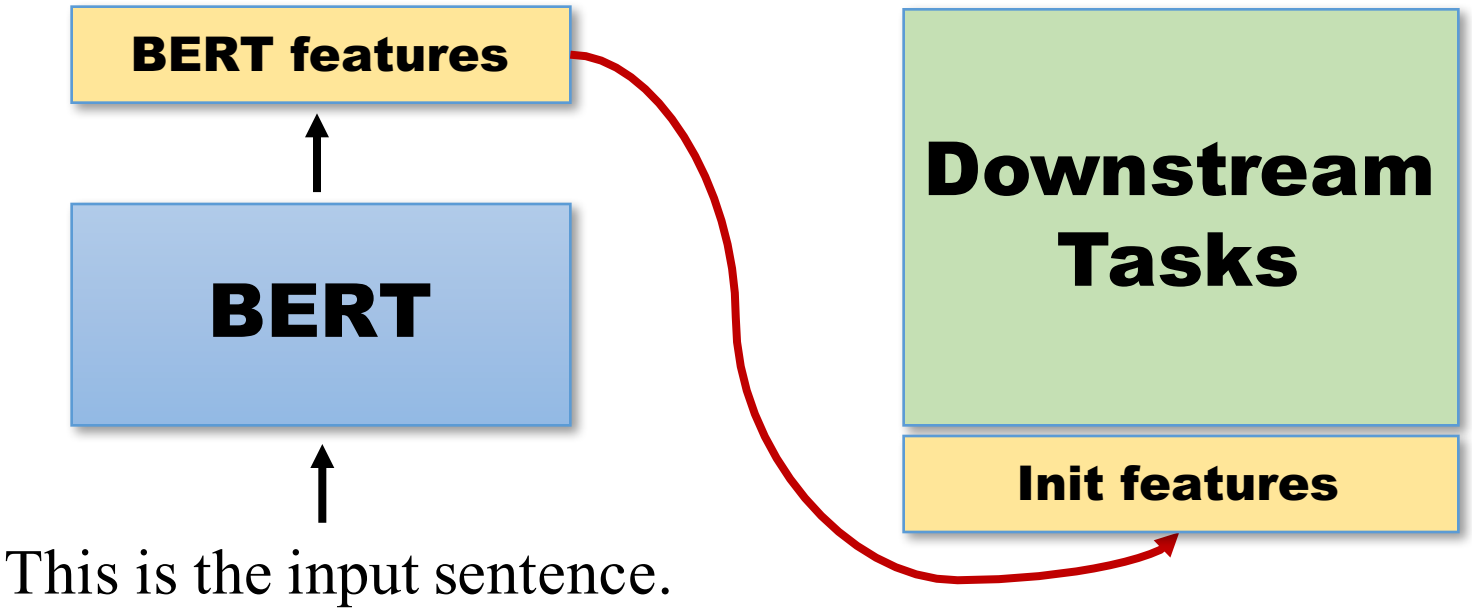
- **The simplified characters still keep some common glyph properties, which can be used to extract similarities among characters.**
  - 打、把、提……
  - 木、林、森、棵……
  - 烧、燃、炒、灯、炖……

Chinese Character

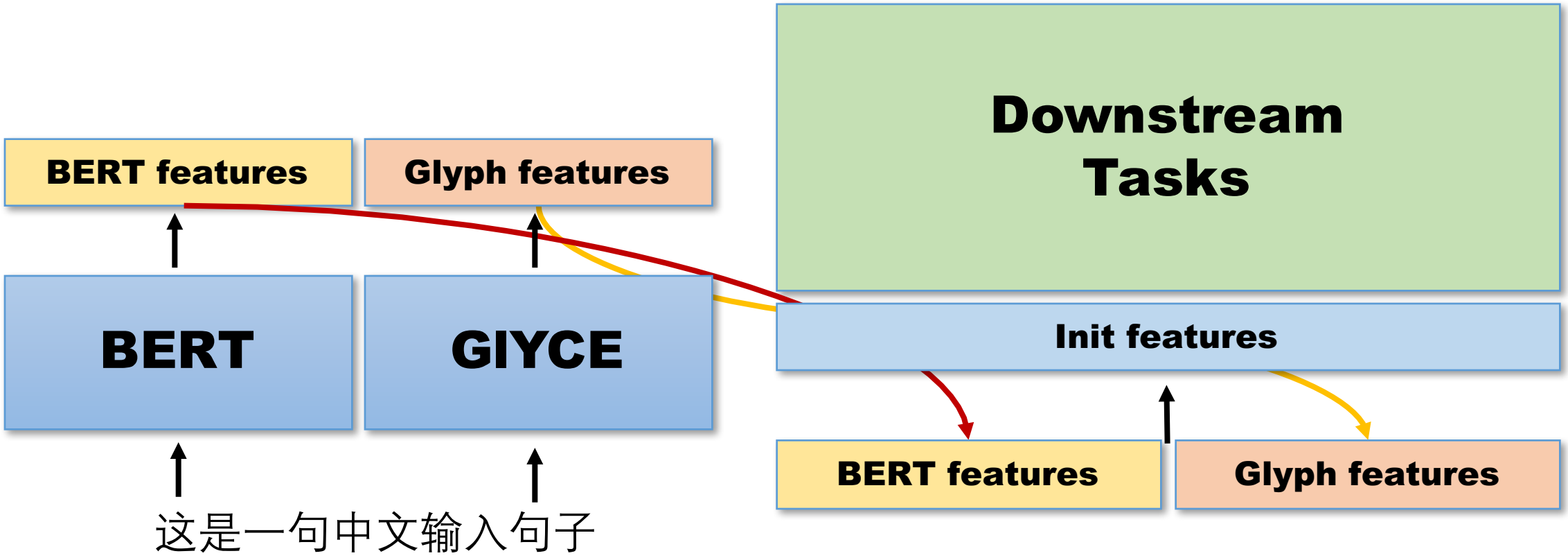
- **The simplified characters still keep some common glyph properties, which can be used to extract similarities among characters.**
  - 打、把、提……
  - 木、林、森、棵……
  - 烧、燃、炒、灯、炖……
- **Enhance the feature representation under Deep Learning!**

BERT Features

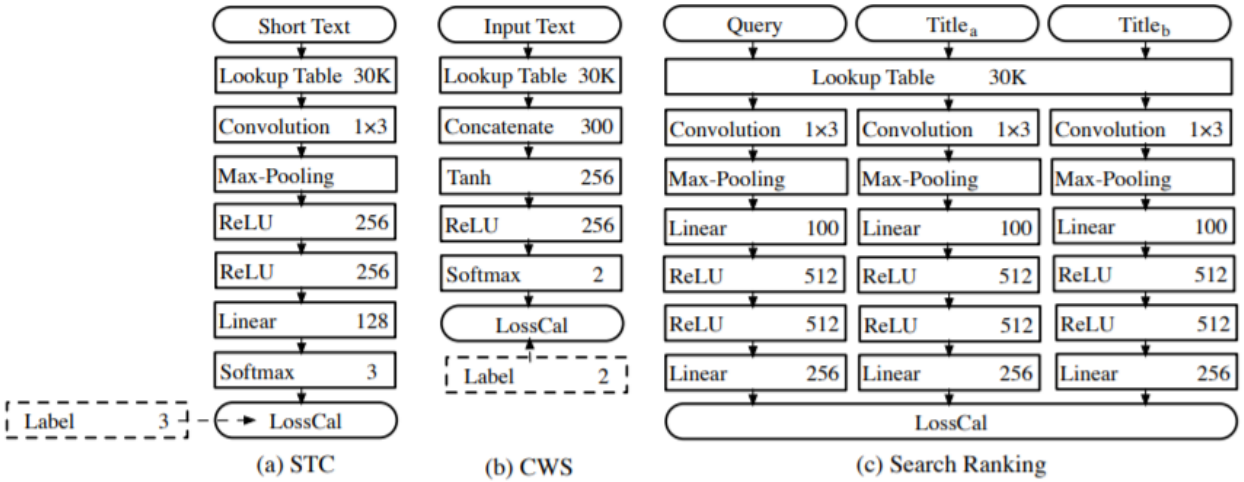
- **Recent works have widely used BERT as initialized word/char features.**



Enhance Chinese Token Features by Character Glyphs



Using Radical Information

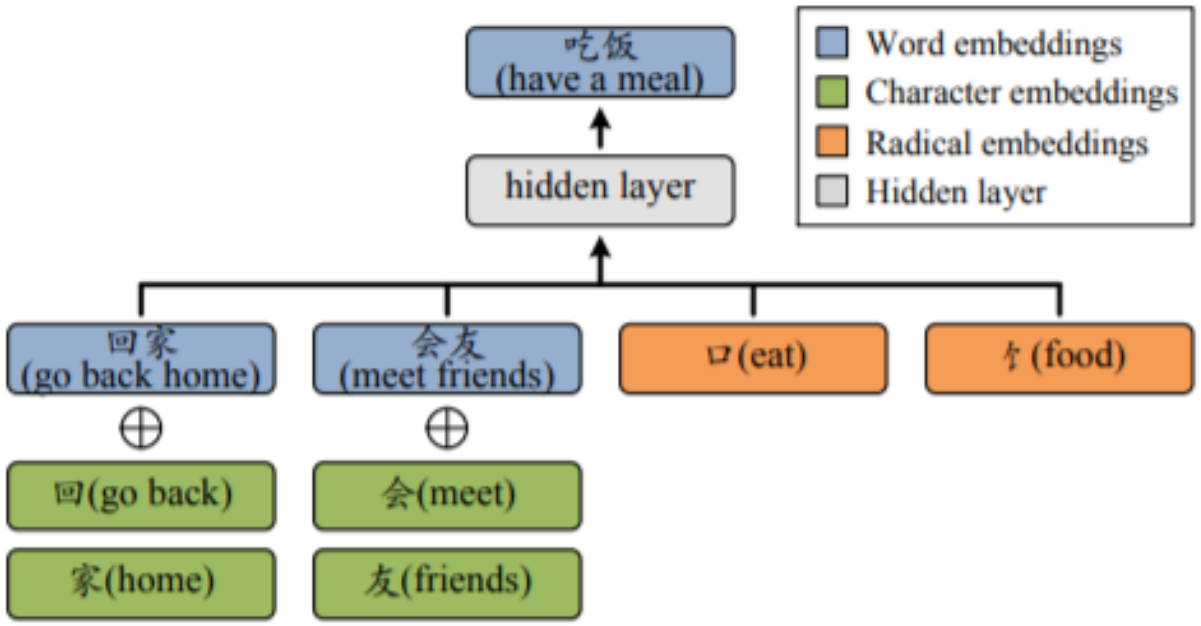


Data	Approach	Precision	Recall	F1
PKU	CRF	88.1	86.2	87.1
	FNLM	87.1	87.9	87.5
	PSA	**92.8**	92.0	**92.4**
	RdE	92.6	**92.1**	92.3
MSR	CRF	89.3	87.5	88.4
	FNLM	92.3	92.2	92.2
	PSA	92.9	**93.6**	93.3
	RdE	**93.4**	93.3	**93.3**

Table 3: CWS Result Comparison

Shi et al. 2015

Using Radical Information

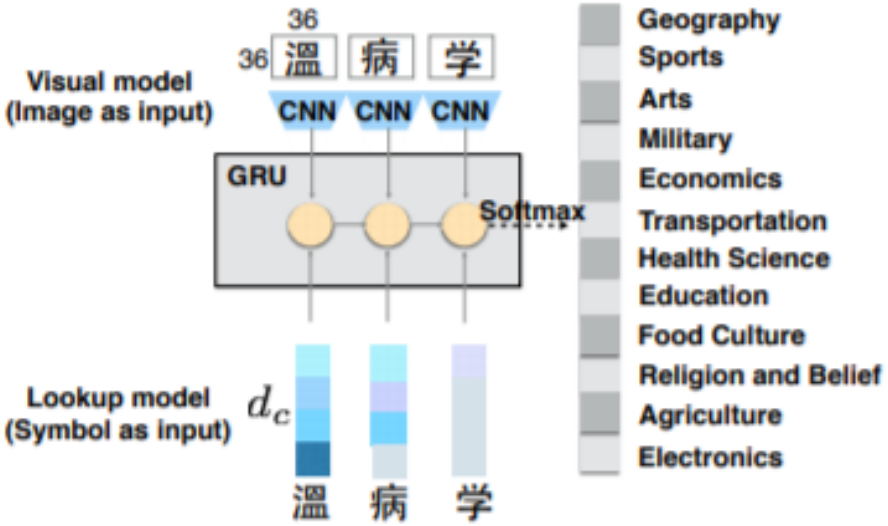


Method	WordSim-239		WordSim-293	
	$k=100$	$k=200$	$k=100$	$k=200$
CBOW	0.4917	0.4971	**0.5667**	**0.5723**
CWE	0.5121	0.5197	0.5511	0.5655
CWE+P	0.4989	0.5026	0.5427	0.5545
MGE	**0.5670**	**0.5769**	0.5555	0.5659
MGE+P	0.5511	0.5572	0.5530	0.5692

Table 1: Results on word similarity computation.

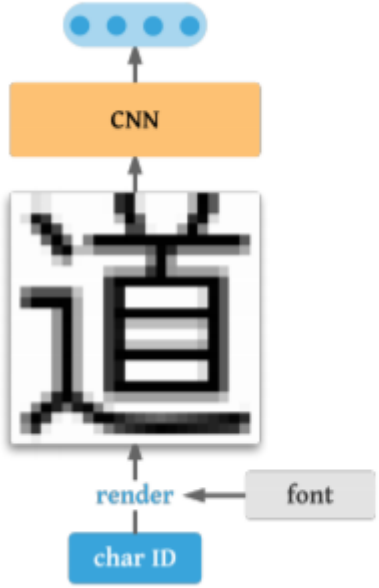
Yin et al. 2016

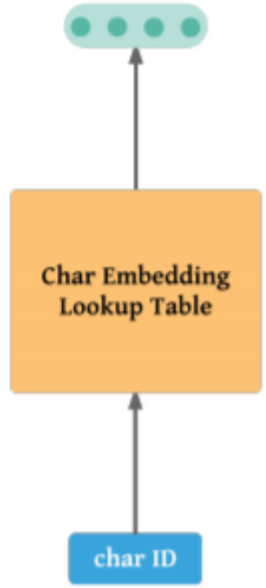
Using Glyph Information



Layers	Description
1-2	Conv 64x3x3
3	Pool 2
4-5	Conv 128x3x3
6	Pool 2
7-8	Conv 256x3x3
9	Pool 2
10	Full 1024
11	Full 256

Table 3: Large GlyphNet encoder

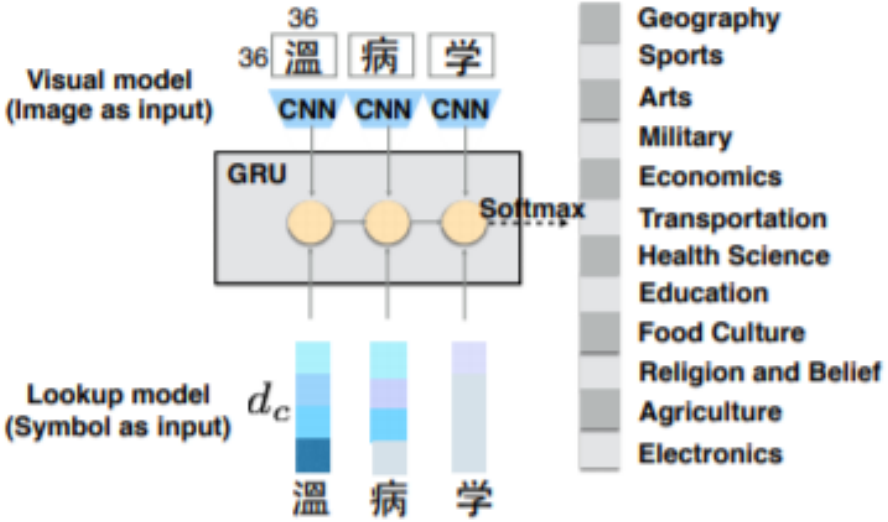




Liu et al., 2017
Zhang and LeCun, 2017a
Dai and Cai, 2017

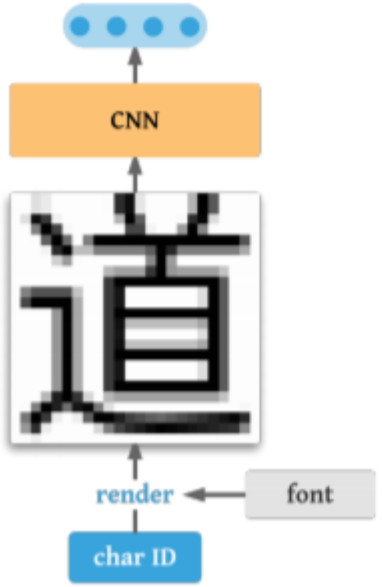
Using Glyph Information

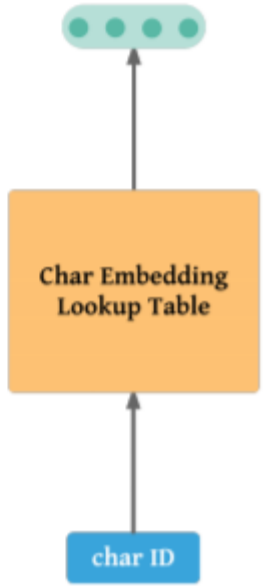
- **Improvements are not significant or even negative.**



Layers	Description
1-2	Conv 64x3x3
3	Pool 2
4-5	Conv 128x3x3
6	Pool 2
7-8	Conv 256x3x3
9	Pool 2
10	Full 1024
11	Full 256

Table 3: Large GlyphNet encoder





Liu et al., 2017
Zhang and LeCun, 2017a
Dai and Cai, 2017

Existing Problems

- **Only one single script is used, mostly simplified or complicated scripts. Chinese characters progress amid history.**

Existing Problems

- **Only one single script is used, mostly simplified or complicated scripts. Chinese characters progress amid history.**

- **Improper CNN structures. Character glyphs are pretty small(12*12) compared to common images(512*512/256*256).**

Existing Problems

- **Only one single script is used, mostly simplified or complicated scripts. Chinese characters progress amid history.**
- **Improper CNN structures. Character glyphs are pretty small(12*12) compared to common images(512*512/256*256).**
- **Incorrect loss functions. The number of Chinese characters is small.**

Solutions

- **Only one single script is used, mostly simplified or complicated script. Chinese characters progress amid history. Use multiple scripts.**
- **Improper CNN structures. Character glyphs are pretty small(12*12) compared to common images(512*512/256*256).**

- **Incorrect loss functions. The number of Chinese characters is small.**

Solutions

- **Only one single script is used, mostly simplified or complicated script. Chinese characters progress amid history. Use multiple scripts.**
- **Improper CNN structures. Character glyphs are pretty small(12*12) compared to common images(512*512/256*256). Design a dedicated CNN structure.**
- **Incorrect loss functions. The number of Chinese characters is small.**

Solutions

- **Only one single script is used, mostly simplified or complicated script. Chinese characters progress amid history. Use multiple scripts.**
- **Improper CNN structures. Character glyphs are pretty small(12*12) compared to common images(512*512/256*256). Design a dedicated CNN structure.**
- **Incorrect loss functions. The number of Chinese characters is small. Add a image classification task as regularization.**

The Usage of Historical Scripts

- **We use the following 8 scripts in history.**

Chinese	English	Time Period
金文	Bronzeware script	Shang dynasty and Zhou dynasty (2000 BC – 300 BC)
隶书	Clerical script	Han dynasty (200BC-200AD)
篆书	Seal script	Han dynasty and Wei-Jin period (100BC - 420 AD)
魏碑	Tablet script	Northern and Southern dynasties 420AD - 588AD
繁体中文	Traditional Chinese	600AD - 1950AD (mainland China). still currently used in HongKong and Taiwan
简体中文(宋体)	Simplified Chinese - Song	1950-now
简体中文(仿宋体)	Simplified Chinese - FangSong	1950-now
草书	Cursive script	Jin Dynasty to now

Tianzige–CNN Structure

- **We design the Tianzige-CNN to better capture the glyph features.**
- **Each glyph is of shape 12*12, followed by a 5*5 conv layer, a max pooling layer and a 1*1 conv layer, which results in a 2*2 structure similar to Tianzige.**
- **Last, the 2*2*256 tensor goes through a group conv layer to get the final 1024d feature vector, which is passed to the softmax layer for image classification.**

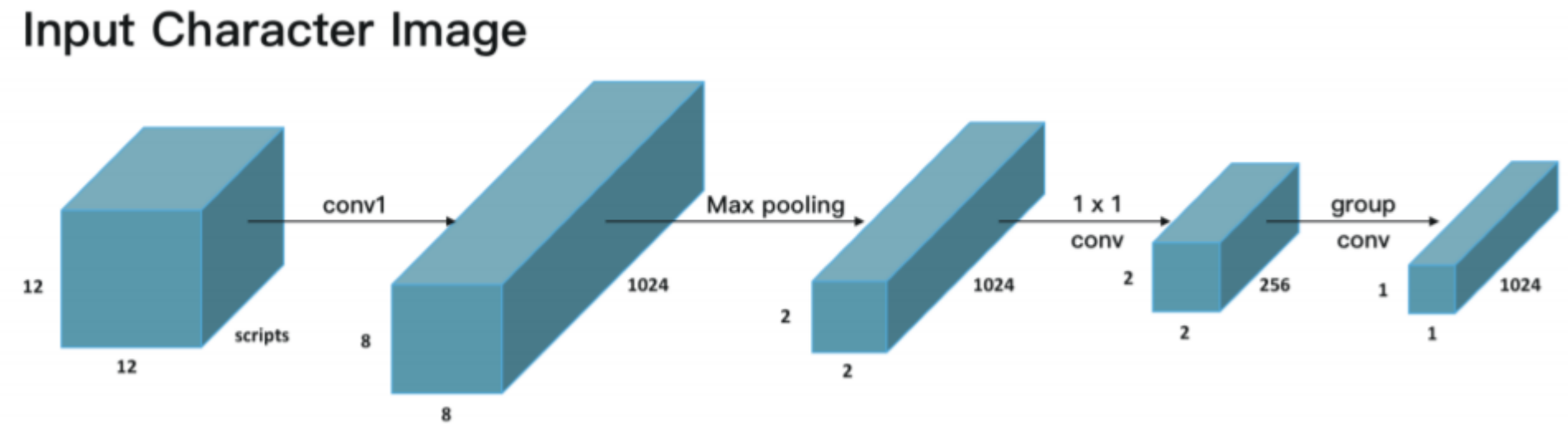


Image Classification Task

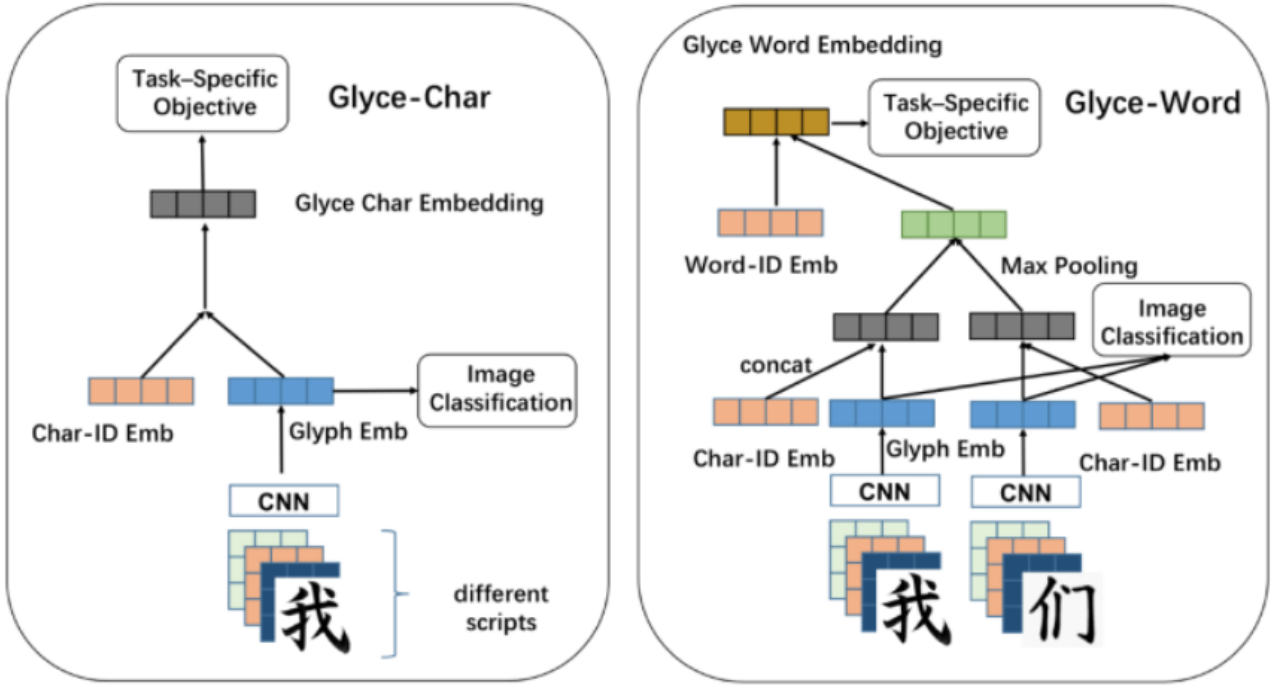
- **Image classification is used as a regularization method.**

$$\mathcal{L}(\text{cls}) = -\log p(z|x) = -\log \text{softmax}(W \times h_{\text{image}})$$

$$\mathcal{L} = (1 - \lambda(t))\mathcal{L}(\text{task}) + \lambda(t)\mathcal{L}(\text{cls})$$

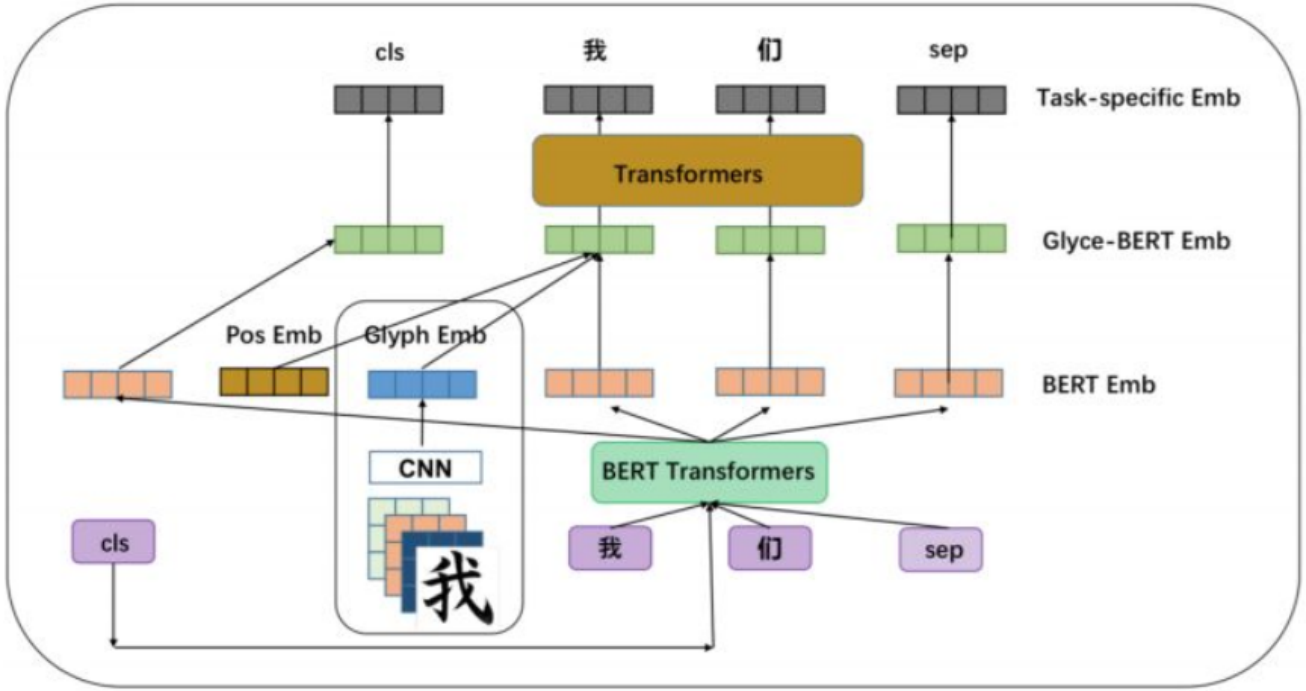
GLYCE

- **Glyph features can be simple combined with token features, such as word embeddings.**



GLYCE with BERT

- **Moreover, Glyce can further complement BERT features.**



Glyce-BERT for Different Tasks

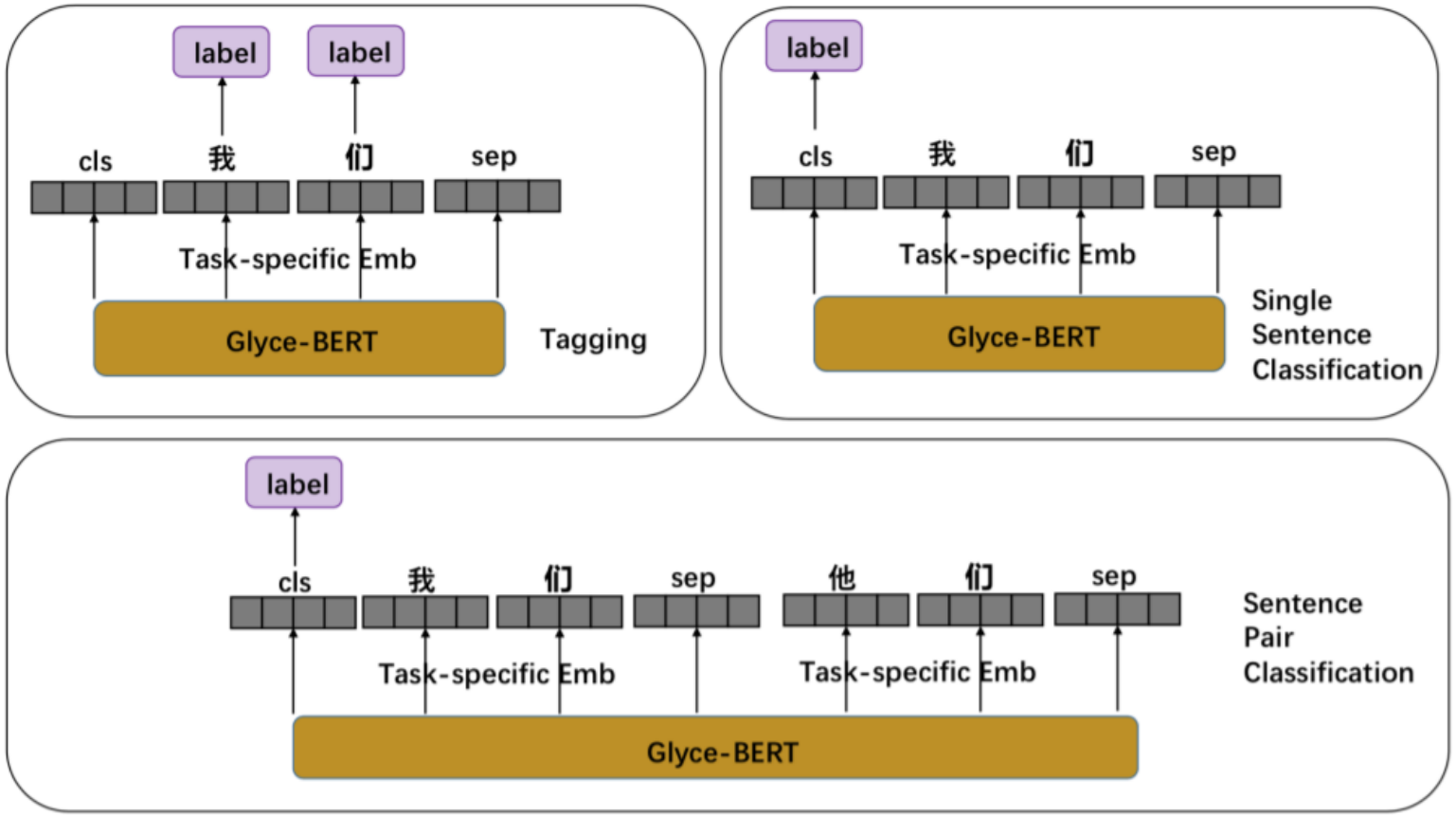


Figure 4: Using Glyce-BERT model for different tasks.

Experiments

- **Experiments are conducted on the following tasks:**
  - **Sequence labeling**
    - **NER、CWS、POS**
  - **Single sentence classification**
    - **ChnSentiCorp、FudanCorpus、Ifeng**
  - **Sentence pair classification**
    - **BQ、LCQMC、XNLI、DBQA**

NER Results

OntoNotes			
Model	P	R	F
CRF-LSTM	74.36	69.43	71.81
Lattice-LSTM	76.35	71.56	73.88
Glyce+Lattice-LSTM	82.06	68.74	74.81
			(+ 0.93)
BERT	78.01	80.35	79.16
Glyce+BERT	81.87	81.40	80.62
			(+1.46)

Weibo			
Model	P	R	F
CRF-LSTM	51.16	51.07	50.95
Lattice-LSTM	52.71	53.92	53.13
Lattice-LSTM+Glyce	53.69	55.30	54.32
			(+1.19)
BERT	67.12	66.88	67.33
Glyce+BERT	67.68	67.71	67.60
			(+0.27)

MSRA			
Model	P	R	F
CRF-LSTM	92.97	90.80	91.87
Lattice-LSTM	93.57	92.79	93.18
Lattice-LSTM+Glyce	93.86	93.92	93.89
			(+0.71)
BERT	94.97	94.62	94.80
Glyce+BERT	95.57	95.51	95.54
			(+0.74)

resume			
Model	P	R	F
CRF-LSTM	94.53	94.29	94.41
Lattice-LSTM	94.81	94.11	94.46
Lattice-LSTM+Glyce	95.72	95.63	95.67
			(+1.21)
BERT	96.12	95.45	95.78
Glyce+BERT	96.62	96.48	96.54
			(+0.76)

CWS Results

PKU			
Model	P	R	F
Yang et al. [2017]	-	-	96.3
Ma et al. [2018b]	-	-	96.1
Huang et al. [2019]	-	-	96.6
BERT	96.8	96.3	96.5
Glyce+BERT	97.1	96.4	96.7 **(+0.2)**

CITYU			
Model	P	R	F
Yang et al. [2017]	-	-	96.9
Ma et al. [2018b]	-	-	97.2
Huang et al. [2019]	-	-	97.6
BERT	97.5	97.7	97.6
Glyce+BERT	97.9	98.0	97.9 **(+0.3)**

MSR			
Model	P	R	F
Yang et al. [2017]	-	-	97.5
Ma et al. [2018b]	-	-	98.1
Huang et al. [2019]	-	-	97.9
BERT	98.1	98.2	98.1
Glyce+BERT	98.2	98.3	98.3 **(+0.2)**

AS			
Model	P	R	F
Yang et al. [2017]	-	-	95.7
Ma et al. [2018b]	-	-	96.2
Huang et al. [2019]	-	-	96.6
BERT	96.7	96.4	96.5
Glyce+BERT	96.6	96.8	96.7 **(+0.2)**

POS Results

CTB5			
Model	P	R	F
Shao et al. [2017] (Sig)	93.68	94.47	94.07
Shao et al. [2017] (Ens)	93.95	94.81	94.38
Lattice-LSTM	94.77	95.51	95.14
Glyce+Lattice-LSTM	95.49	95.72	95.61 **(+0.47)**
BERT	95.86	96.26	96.06
Glyce+BERT	96.50	96.74	96.61 **(+0.55)**

CTB6			
Model	P	R	F
Shao et al. [2017] (Sig)	-	-	90.81
Lattice-LSTM	92.00	90.86	91.43
Glyce+Lattice-LSTM	92.72	91.14	91.92 **(+0.49)**
BERT	94.91	94.63	94.77
Glyce+BERT	95.56	95.26	95.41 **(+0.64)**

CTB9			
Model	P	R	F
Shao et al. [2017] (Sig)	91.81	94.47	91.89
Shao et al. [2017] (Ens)	92.28	92.40	92.34
Lattice-LSTM	92.53	91.73	92.13
Lattice-LSTM+Glyce	92.28	92.85	92.38 **(+0.25)**
BERT	92.43	92.15	92.29
Glyce+BERT	93.49	92.84	93.15 **(+0.86)**

UD1			
Model	P	R	F
Shao et al. [2017] (Sig)	89.28	89.54	89.41
Shao et al. [2017] (Ens)	89.67	89.86	89.75
Lattice-LSTM	90.47	89.70	90.09
Lattice-LSTM+Glyce	91.57	90.19	90.87 **(+0.78)**
BERT	95.42	94.17	94.79
Glyce+BERT	96.19	96.10	96.14 **(+1.35)**

Sentence Pair Results

BQ				
Model	P	R	F	A
BiMPM	82.3	81.2	81.7	81.9
Glyce+BiMPM	81.9	85.5	83.7 **(+2.0)**	83.3 **(+1.4)**
BERT	83.5	85.7	84.6	84.8
Glyce+BERT	84.2	86.9	85.5 **(+0.9)**	85.8 **(+1.0)**

LCQMC				
Model	P	R	F	A
BiMPM	77.6	93.9	85.0	83.4
Glyce+BiMPM	80.4	93.4	86.4 **(+1.4)**	85.3 **(+1.9)**
BERT	83.2	94.2	88.2	87.5
Glyce+BERT	86.8	91.2	88.8 **(+0.6)**	88.7 **(+1.2)**

XNLI				
Model	P	R	F	A
BiMPM	-	-	-	67.5
Glyce+BiMPM	-	-	-	67.7 **(+0.2)**
BERT	-	-	-	78.4
Glyce+BERT	-	-	-	79.2 **(+0.8)**

NLPCC-DBQA				
Model	P	R	F	A
BiMPM	78.8	56.5	65.8	-
Glyce+BiMPM	76.3	59.9	67.1 **(+1.3)**	-
BERT	79.6	86.0	82.7	-
Glyce+BERT	81.1	85.8	83.4 **(+0.7)**	-

Single Sentence Results

Model	ChnSentiCorp	the Fudan corpus	iFeng
LSTM	91.7	95.8	84.9
LSTM + Glyce	93.1	96.3	85.8
	(+ 1.4)	**(+0.5)**	**(+0.9)**
BERT	95.4	99.5	87.1
Glyce+BERT	95.9	99.8	87.5
	(+0.5)	**(+0.3)**	**(+0.4)**

Training Strategies

- **How to combine Glyce and BERT?**
  - **BERT-Glyce joint training**
  - **Glyce-joint training.**
  - **Joint training**
  - **Only BERT?**
- **Results on LCQMC**

Strategy	Precision	Recall	F1	Accuracy
BERT-glyce-joint	86.8	91.2	88.8	88.7
Glyph-Joint	82.5	94.0	87.9	87.1
joint	81.5	95.1	87.8	86.8
only BERT	83.2	94.2	88.2	87.5

The Effect of Image Classification Task

Strategy	Precision	Recall	F1	Accuracy
W image-cls	86.8	91.2	88.8	88.7
WO image-cls	83.9	93.6	88.4	87.9

Table 8: Impact of the auxilliary image-classification training objective.

The Effect of CNNs

- **Are Tianzige-CNN more effective than other CNNs?**
  - **Vanillia-CNN**
  - **ResNet**
  - **Tianzige-CNN**

	Precision	Recall	F1
Vanilla-CNN	85.3	89.8	87.4
He et al. [2016]	84.5	90.8	87.5
Tianzige-CNN	86.8	91.2	88.8

Table 10: Impact of CNN structures.


香侬科技
SHANNON.AI
Any Questions?
2019.11.11

Summary

- Glyce encodes information from glyphs, leading to further performance boost beyond BERT

Summary

- Glyce encodes information from glyphs, leading to further performance boost beyond BERT
- Model structures based on MRC framework encodes prior knowledge, and gains consistent performance improvement over various NLP tasks

Summary

- Glyce encodes information from glyphs, leading to further performance boost beyond BERT
- Model structures based on MRC framework encodes prior knowledge, and gains consistent performance improvement over various NLP tasks

Unify NLP tasks with…?

T5?

Summary

- Glyce encodes information from glyphs, leading to further performance boost beyond BERT
- Model structures based on MRC framework encodes prior knowledge, and gains consistent performance improvement over various NLP tasks

Unify NLP tasks with…?

T5? Maybe MRC is another answer.

Thanks for watching

Glyce: Glyph-vectors for Chinese Character Representations
https://arxiv.org/pdf/1901.10125.pdf
Entity-Relation Extraction as Multi-turn Question Answering
https://arxiv.org/pdf/1905.05529.pdf
A Unified MRC Framework for Named Entity Recognition
https://arxiv.org/pdf/1910.11476.pdf
Coreference Resolution as Query-based Span Prediction
https://arxiv.org/pdf/1911.01746.pdf
Description Based Text Classification with Reinforcement Learning
https://arxiv.org/pdf/2002.03067.pdf