Word Embedding Based Correlation Model for Question/Answer Matching

Yikang Shen, Wenge Rong, Nan Jiang, Baolin Peng, Jie Tang, Zhang Xiong

‘AAAI-17‘

Happy Buzaaba
D1 KDE-LAB
University of Tsukuba

Introduction

- Focus on community based question answering (CQA)
- Example of CQA

- Yahoo! Answers
- Bhaidu Zidao
- Quora
- Zihu

Accumulated
a large scale of Q&A archive

Important
question answer matching

Knowledge reuse

Enhance user experience

Challenge

- Lexical gap between question and candidate answers

For Example: What is the fastest car in the world?

Ans: “The Jaguar XJ220 is the fastest and most sought after car on the planet.”

- No more than 4 words in common

Associated by synonyms, hypernyms or other weaker semantic associations yih et. 2014

Approaches to lexical gap

- Translational model.

- Leverages the Q/A pairs to learn semantically related words ‘Nakov et al, 2016’

Disadvantage
- Curse of dimensionality ‘Bengio et al. 2003’
- Generalization structure is not obvious

- Semantic based model.

- Leveraging the advantages of vector representation of words ‘Zhou et al, 2016’

Disadvantage
- Q&A are heterogenous in many aspects

Contribution

- Proposes a word embedding correlation model (WEC)

  - word level correlation function is designed

$C(q_i, a_j)$

  - sentence level correlation function

$C(q, a)$

- Combined WEC with CNN to integrate both lexical and syntactical

Problem definition

- Given:
  - Question $q = \{q_1, q_2, q_3, ..., q_n\}$, where q_i is the ^{i}th word in the question
  - A set of candidate answers $A = \{a^1, a^2, a^3, ..., a^n\}$

    where $a^j = \{a^j_1, a^j_2, ..., a^j_m\}$, and a^j_k is the ^{k}th word in ^{j}th candidate answer

    Goal: Identify the most relevant answer a^{best}

    This is achieved by:

    calculating matching probabilities between question and each candidate answer

    Rank candidate answer by their matching probabilities

Problem Definition Continuation

- Calculating Matching Probabilities:
- Represent words in questions and answers as vectors in a continuous space
- Calculate word to word correlation score
- Employ phrase level correlation function to obtain question and answer matching

- Incorporate WEC and CNN to achieve better matching precision

Methodology

1. Word Embedding based Correlation Model (**WEC**)

   - Word-level Correlation function

$$C(q_i, a_j) = \cos < v_{qi}, \mathrm{M}\, v_{aj} > = \frac{v^T_{qi}}{||vqi||} \frac{\mathrm{M}v_{aj}}{||\mathrm{M}v_{aj}||}$$

V_{qi} & v_{aj}: d-dimensional word embedding vectors
M: Translation matrix

   - Sentence level Correlation function

$$C(q, a) = \frac{1}{|a|} \sum_j \max C(qi, aj)$$

$|a|$: length of the answer a,
C(qi, aj): correlation score of the i-th word in question & j-th word in answer

Sentence-level correlation score is calculated by averaging selected word level scores

Methodology

2. WEC + CNN

Word-level correlation function

Correlation matrix

- $C(q_i, a_j)$ $\longrightarrow$ $C_{ij} = C(q_{i \bmod |q|}, a_{j \bmod |a|})$

|q| & |a|: Respective length of question and answer

C: n_f x m_f fixed size matrix

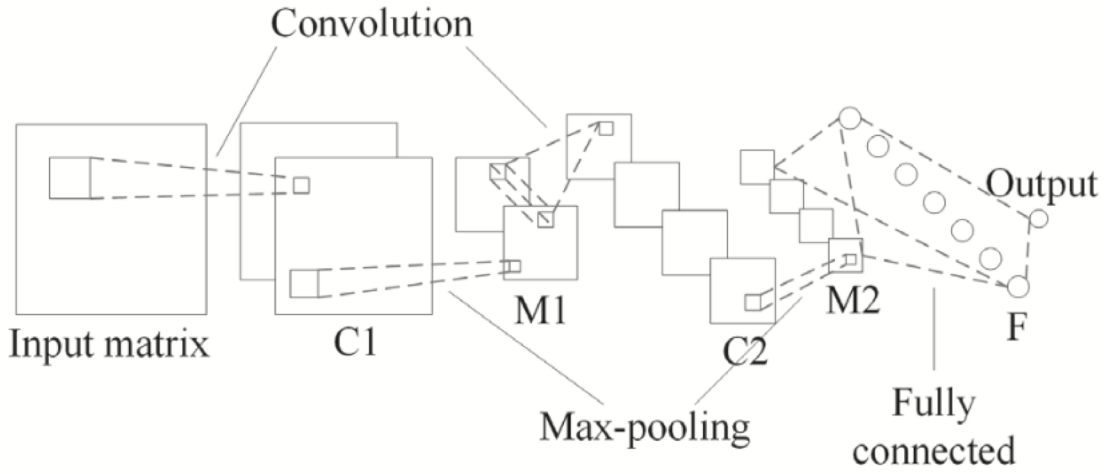


Figure 1. CNN architecture, It has 2 convolutional layers C1 and C2, each followed by a max-pooling layer M1 and M2, and fully connected layer F.

Experimental Set up

- Data sets Used:
- 1. Yahoo! Answers

Category	Question#	Training set#	Training triple#	Validation set#	Validation triple#	Test set#
Travel	53,261	35,504	355,040	8,876	88,760	8,881
Relationships	57,576	38,384	383,840	9,596	95,960	9,596
Finance	66,129	44,084	440,840	11,021	110,210	11,024

- 2. Baidu Zhidao

Dataset	Category
Baidu Zhidao	Computer and Internet
	Education and Science
	Games
	Entertainment and Recreation

Evaluation Metrics

- In test set:

  - for each q, $A = \{a_1, a_2, \ldots a_6\}$ { One positive; Five negative

  - Performance of ranking quality is compared with other baselines
  - Discount cumulative gain (DCG) is used to evaluate ranking quality

$$\mathrm{DCG@p} = \sum_{i=1}^{p} \frac{rel_i}{log_2(i+1)}$$ { $rel = 1$, best answer; $rel = 0$, others

DCG@1: Evaluate the precision of choosing best answer

DCG@6: Evaluate the quality of ranking

Experiment Results

- Performance of different approaches on Yahoo! Answers dataset

Approach	Travel		Relationships		Finance	
	DCG@1	DCG@6	DCG@1	DCG@6	DCG@1	DCG@6
WEC + CNN	**0.761**	**0.946**	**0.709**	**0.938**	**0.780**	**0.952**
WEC	0.734	**0.946**	0.698	0.936	0.761	0.949
TRLM	0.727	0.922	0.683	0.910	0.755	0.927
TM	0.698	0.914	0.676	0.912	0.742	0.926
Okapi	0.631	0.875	0.517	0.823	0.646	0.866
LM	0.592	0.848	0.525	0.825	0.595	0.838

Experiment Results

- Performance of different approaches on Baidu Zidhao dataset

Approach	Computers & Internet		Education & Science		Games		Entertainment & Recreation	
	DCG@1	DCG@6	DCG@1	DCG@6	DCG@1	DCG@6	DCG@1	DCG@6
WEC+CNN	**0.826**	**0.970**	**0.870**	**0.980**	**0.703**	**0.941**	**0.780**	**0.963**
WEC	0.821	0.968	0.838	0.975	0.692	0.937	0.778	0.962
TRLM(IBM-1)	0.780	0.937	0.843	0.948	0.654	0.894	0.709	0.918
TM(IBM-1)	0.732	0.925	0.766	0.931	0.598	0.876	0.626	0.892
S+CNN	0.658	0.912	0.734	0.939	0.619	0.894	0.543	0.866
TRLM(cos)	0.601	0.885	0.698	0.924	0.562	0.865	0.492	0.843
TM(cos)	0.596	0.885	0.691	0.922	0.560	0.863	0.486	0.841
Okapi	0.567	0.806	0.702	0.869	0.467	0.747	0.446	0.723
LM	0.624	0.830	0.746	0.881	0.544	0.765	0.488	0.740

Experiment Results

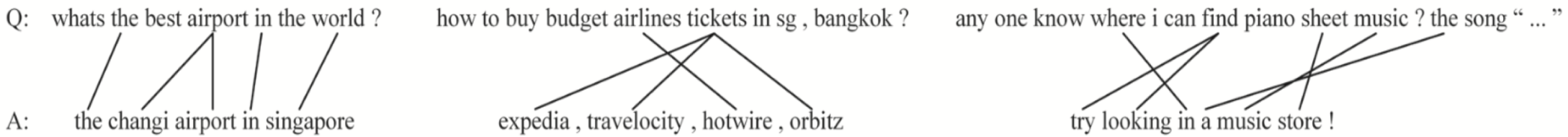
- Word to word translation example, each column show the top 5 related answer words for a given question word

Target	where			when		
TTable	**WEC**	**cos**	**IBM model 1**	**WEC**	**cos**	**IBM model 1**
1	middle	what	hamlets	when	before	visist
2	southern	how	prefecture	after	while	glacial
3	southeastern	which	foxborough	until	once	onward/return
4	situated	tellme	berea	early	because	earthquake
5	burundi	want	unincorporated	planting	if	feb

Target	museum			food		
TTable	**WEC**	**cos**	**IBM model 1**	**WEC**	**cos**	**IBM model 1**
1	exhibits	galleries	rodin	vegetarian	delicious	cassoulet
2	musuem	monuments	montagne	seafood	cuisine	pork
3	planetarium	capitoline	louvre	eat	seafood	cuisine
4	Smithsonian	exhibits	loews	burgers	spicy	prolific
5	archaeology	musicals	chabot	cuisine	vegetarian	delft

Experiment Results

- Example of identified relevance between question and answer using sentence level correlation function



Conclusion

- The paper presents a new approach for Q&A matching in CQA service, The approach uses a word embedding correlation (WEC) model to solve the lexical gap between question and answer.
- The experiments show that WEC and WEC+CNN outperforms the state-of-art models