Siamese RNN for Learning Text Similarity

2017-12-13

CONTENT

●Learning text / sentence / sequence similarity
●Siamese network
●LSTM
●Character-level & word-level

◆***Reference***

1. 《Learning Text Similarity with Siamese Recurrent Networks》,2016
2. 《Siamese Recurrent Architectures for Learning Sentence Similarity》,2016

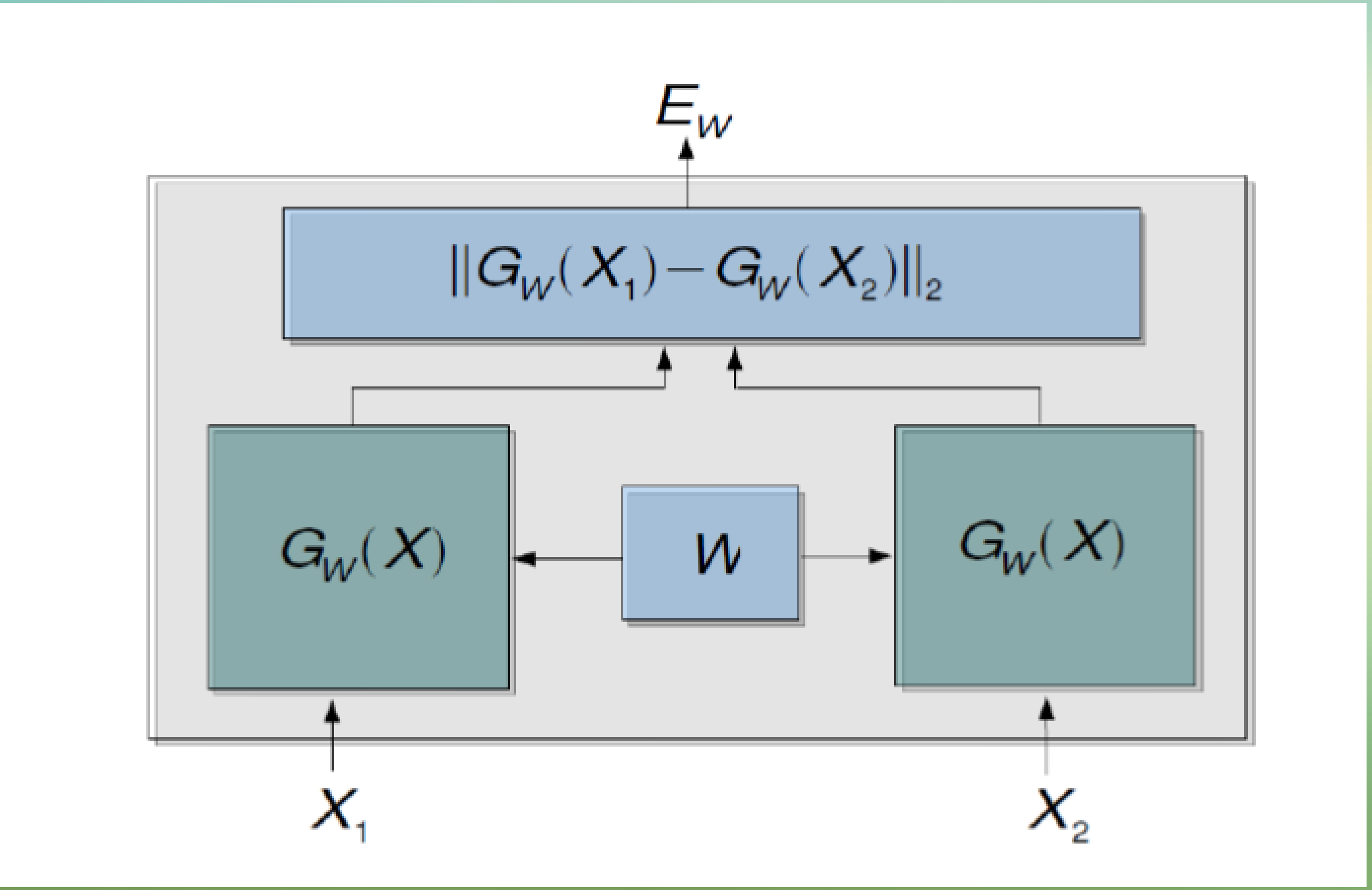
CONTENT

- Background
- Model
- Experiment
- Conclusion

PART ONE

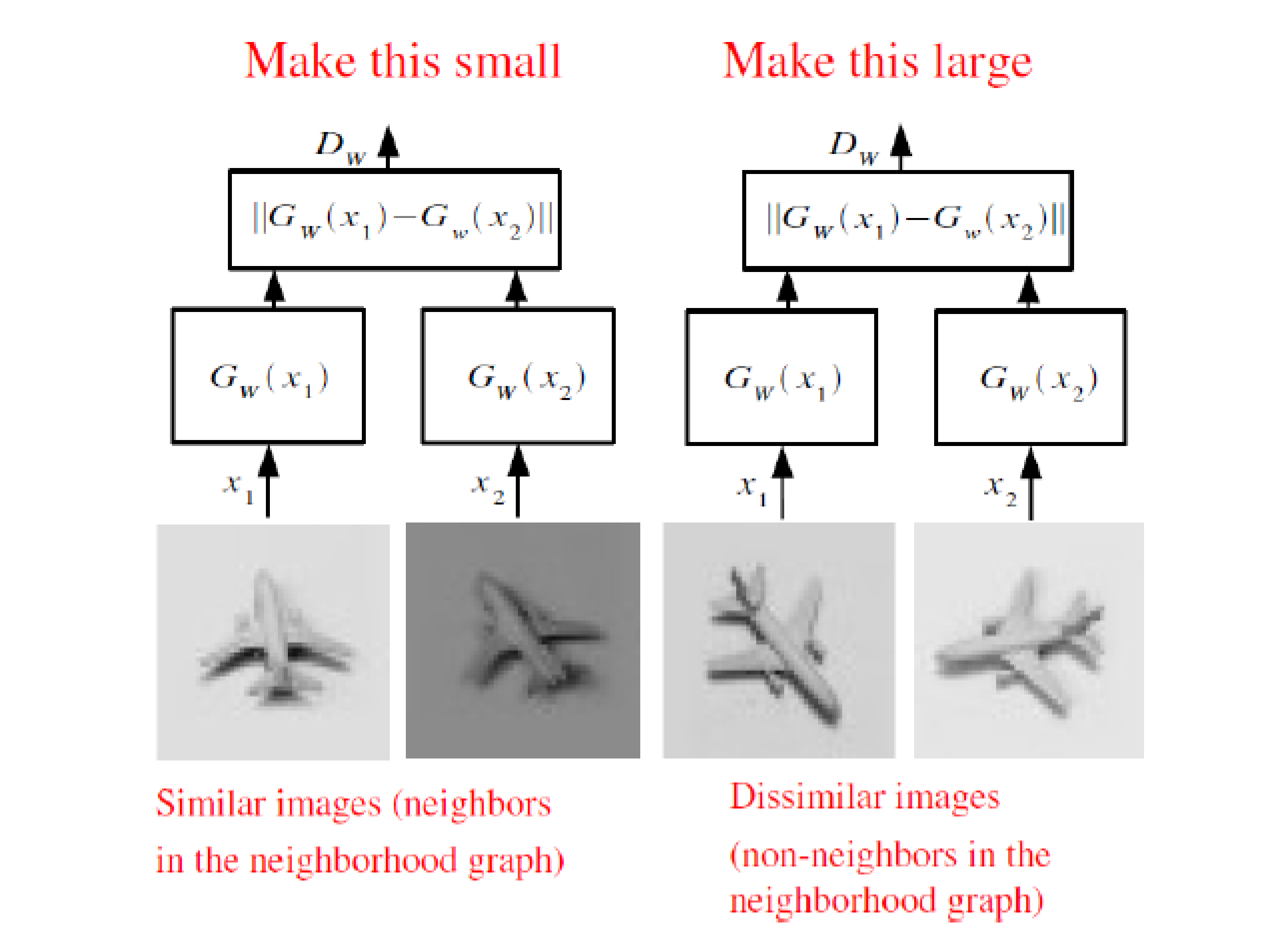
Background

Siamese architecture



a network architecture & a similarity metric

Siamese architecture



Application:

- *Dimensionality reduction*
- *Signature Verification*
- *Face identification & Face verification*

Siamese network loss function

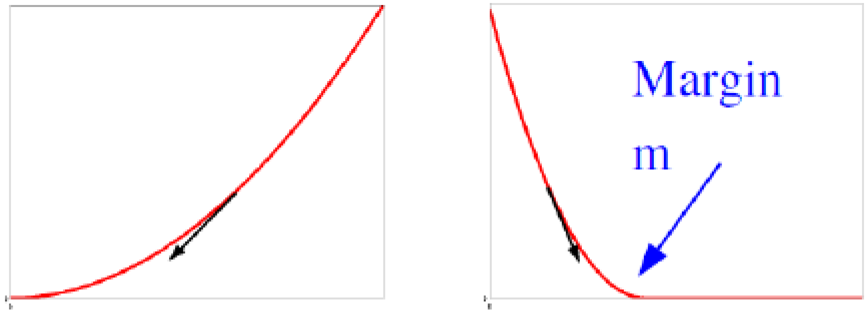
- 两个相同结构、权值共享的子网络
- 三元组（X1,X2,Y），X1和X2相同时y=1，不同时y=0
- 损失函数设计：

　　1. 输入样本对不相似（y=0）时，距离越大（相似度越小），损失越小，即是关于距离的单调递减函数（相似度的单调递增函数）

　　2. 输入样本对相似（y=1）时，距离越大（相似度越小），损失越大，即是关于距离的单调递增函数（相似度的单调递减函数）

Loss function:

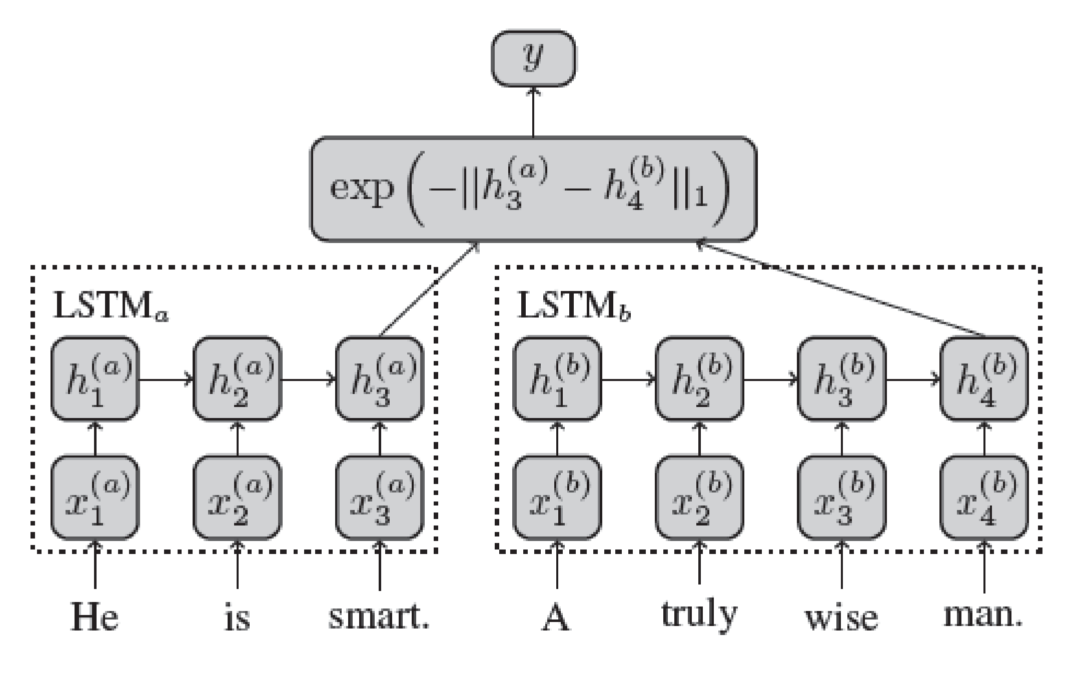
$$L_{similar}=\frac{1}{2}D_w^2 \qquad L_{dissimilar}=\frac{1}{2}\{max(0,m-D_W)\}^2$$



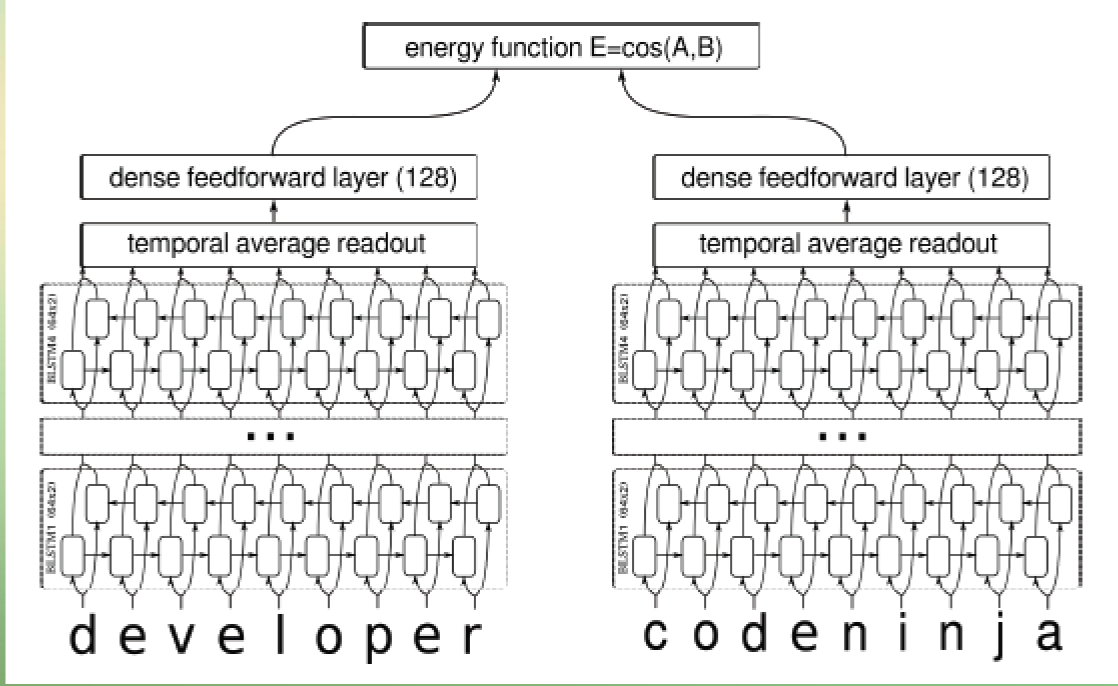
PART TWO

Model

Model architectures



Word-level



Character-level

Model ~ character-level

INPUT Phrase:

International Business Machines = I.B.M

Synergy Telecom = SynTel

Beam inc = Beam Incorporate

Sir J J Smith = Johnson Smith

Alex, Julia = J Alex

James B. D. Joshi = James Joshi

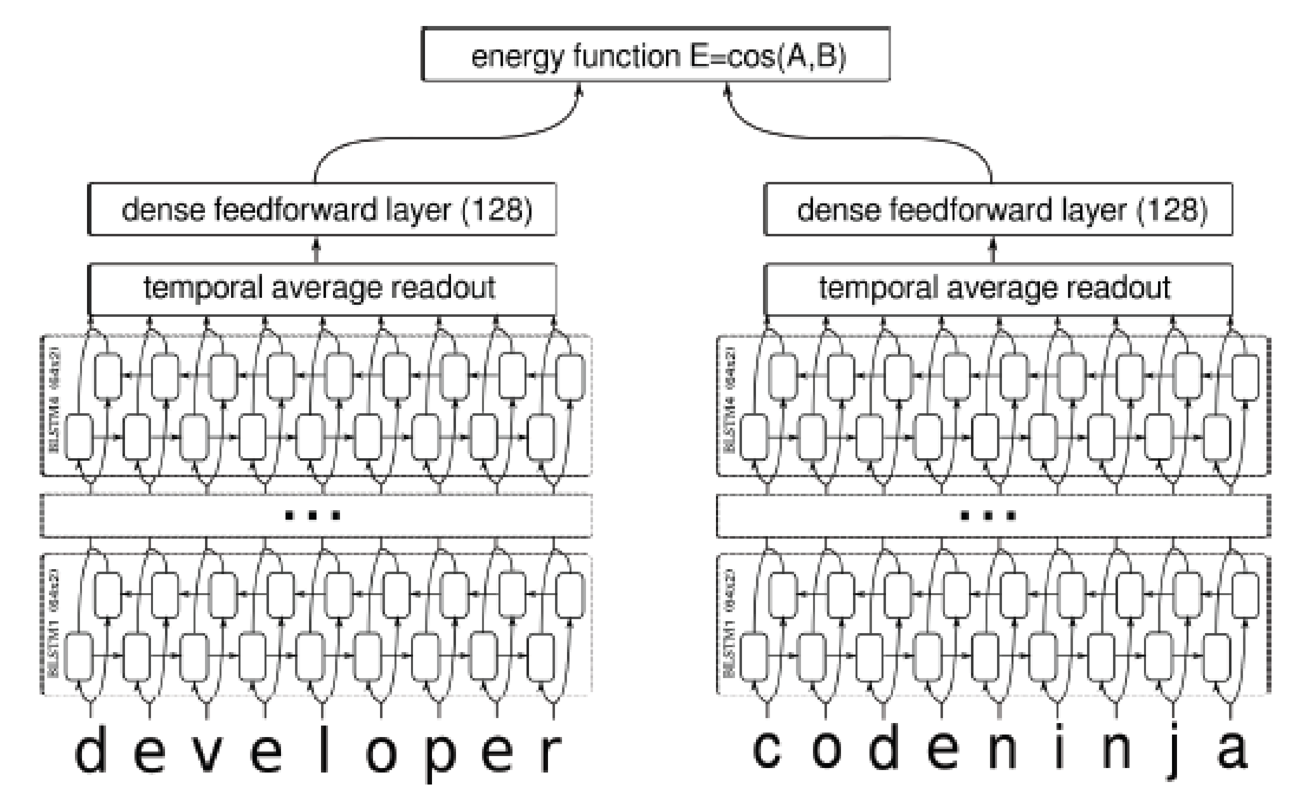
James Beaty.Jr. = Beaty

Job title :

Developer = software architect

J2EE = Java developer

Model ~ character-level

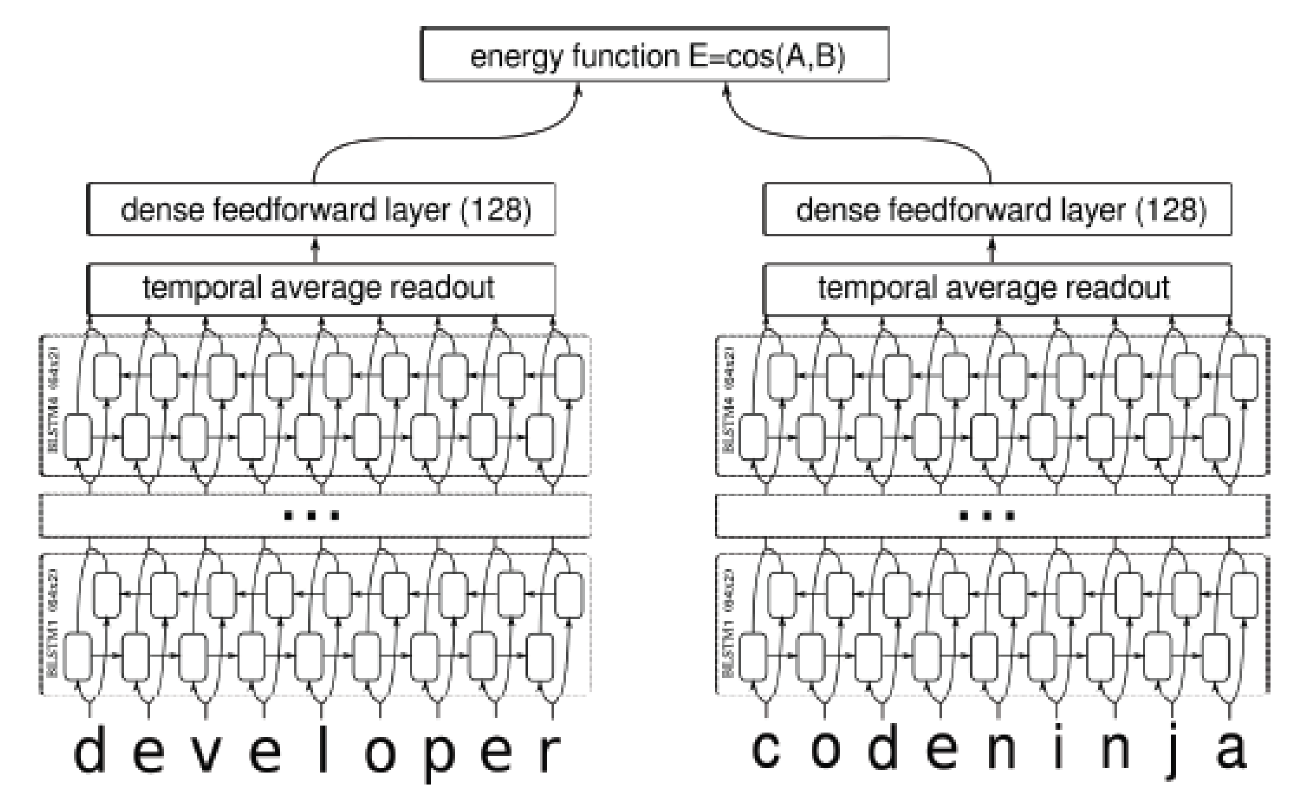


INPUT:
(x1, x2 ,y)
y=1 similar
y=0 dissimilar

OUTPUT:
Sequence similarity

Sentence embedding:
Output of feedforward layer

Model ~ character-level



$$E_{\mathrm{W}}(x_1, x_2) = \frac{\langle f_{\mathrm{W}}(x_1), f_{\mathrm{W}}(x_2) \rangle}{\| f_{\mathrm{W}}(x_1) \| \| f_{\mathrm{W}}(x_2) \|}$$

$$\mathcal{L}_{\mathrm{W}}(X) = \sum_{i=1}^{N} L_{\mathrm{W}}^{(i)}(x_1^{(i)}, x_2^{(i)}, y^{(i)})$$

$$\begin{aligned} L_{\mathrm{W}}^{(i)} = {} & y^{(i)} L_+(x_1^{(i)}, x_2^{(i)}) + \\ & (1 - y^{(i)}) L_-(x_1^{(i)}, x_2^{(i)}) \end{aligned}$$

$$L_+(x_1, x_2) = \frac{1}{4}(1 - E_{\mathrm{W}})^2$$

$$L_-(x_1, x_2) = \begin{cases} E_{\mathrm{W}}^2 & \text{if } E_{\mathrm{W}} < m \\ 0 & \text{otherwise} \end{cases}$$

Model ~ word-level

Sentence sample:

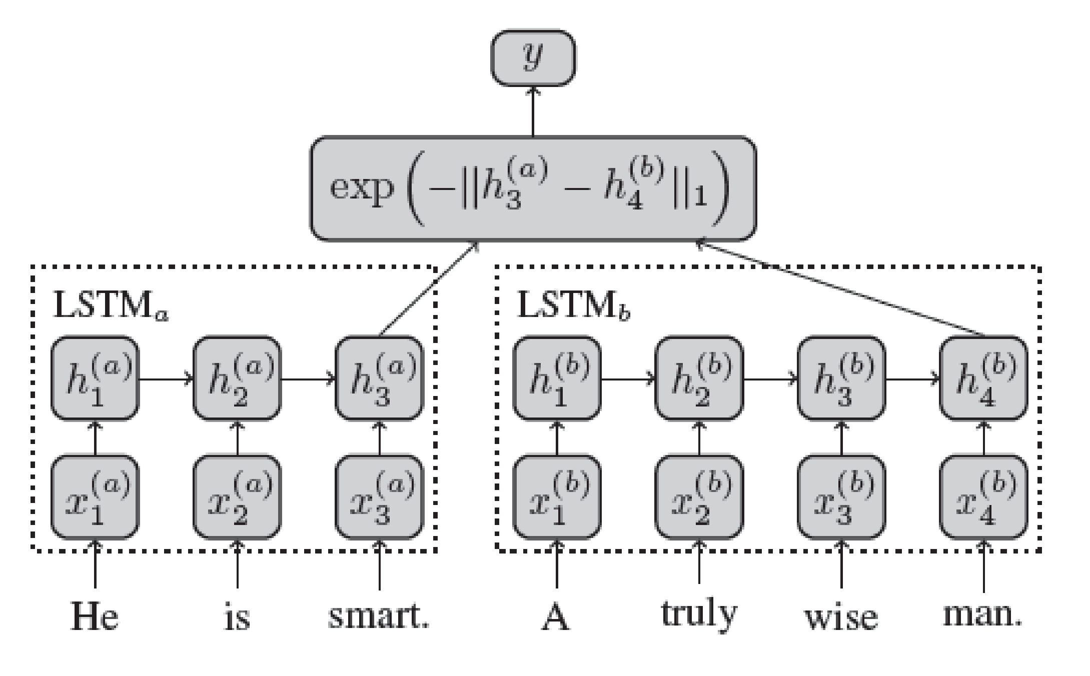
He is smart = He is a wise man.

Someone is travelling countryside = He is travelling to a village.

She is cooking a dessert = Pudding is being cooked.

Microsoft to acquire Linkedin ≠ Linkedin to acquire microsoft

Model ~ word-level



$$g(h_{T_a}^{(a)}, h_{T_b}^{(b)}) = \exp(-\|h_{T_a}^{(a)} - h_{T_b}^{(b)}\|_1) \in [0,1].$$

INPUT:
(x1,x2,y)
y=1 similar
y=0 dissimilar

OUTPUT:
y

Sentence embedding:
last hidden state of the model

PART THREE

Experiment

Data and data augmentation

Data augmentation:

- Sysnonyms:

  Replace random word with one of their synonyms found in Wordnet
- Typo and spelling invariance

  characters randomly substituted and deleted
- Extra words

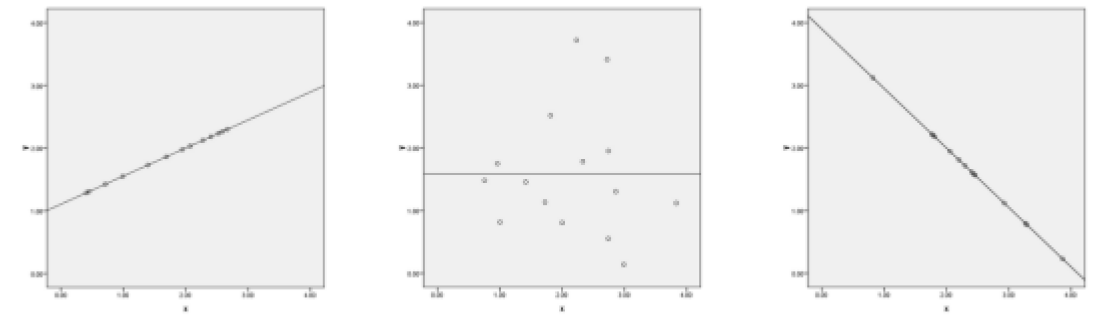
  invariant to superfluous words

Data:

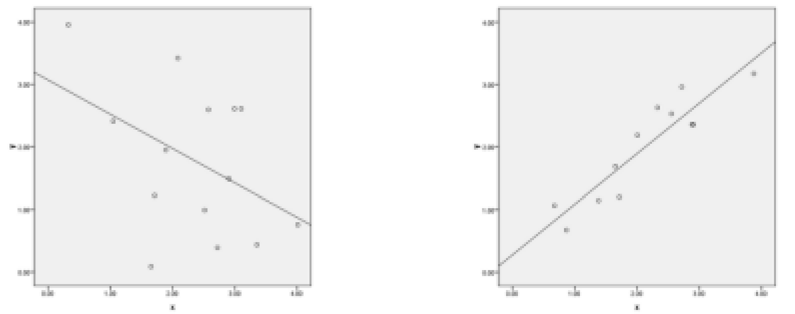
- Train data : SemEval 2014 train data ,5000
- Test data: SemEval 2014 test data ,4927
- Sample : [“Nobody is pouring ingredients into a pot.”, ”Someone is pouring ingredients into a pot.”,3.5]

experiment

Pearson's correlation coefficient is a statistical measure of the strength of a **linear** relationship between paired data, r belongs to [-1,1]



$r = -1$ perfect -ve correlation | $r = 0$ no correlation | $r = 1$ perfect +ve correlation

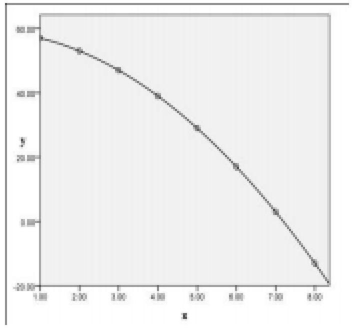


$r = -.45$ moderate -ve correlation | $r = .92$ very strong +ve correlation

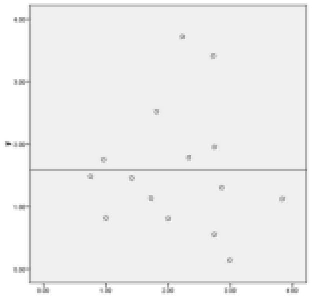
$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

experiment

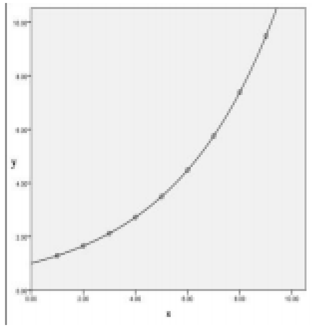
Spearman's correlation coefficient is a statistical measure of the strength of a **monotonic** relationship between paired data , rho belongs to [-1,1]



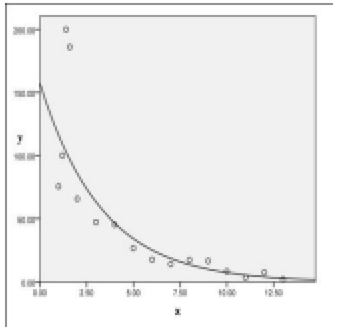
$r_s = -1$
perfect –ve
monotonic correlation



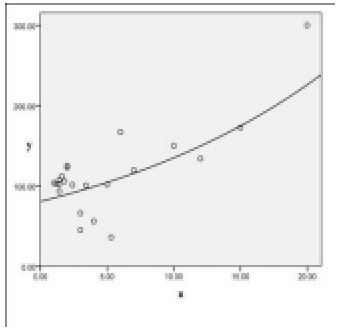
$r_s = 0$
no correlation



$r_s = 1$
perfect +ve
monotonic correlation

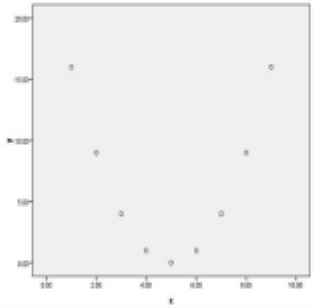


$r_s = -.941$
very strong -ve
monotonic correlation



$r_s = .372$
weak +ve
monotonic correlation

$$\rho = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$



$r_s = 0$
perfect quadratic relationship

experiment

- Semantic relatedness scoring
- Sentence representation:

  1、 sentence similarity between sentences by aggregating differences

  2、 semantic information
- Entailment classification：

  textual entailment task

Word-level experiment

Sentence pair	G	S	M
A little girl is looking at a woman in costume. A young girl is looking at a woman in costume.	4.7	4.5	4.8
A person is performing tricks on a motorcycle. The performer is tricking a person on a motorcycle.	2.6	4.4	2.9
Someone is pouring ingredients into a pot. A man is removing vegetables from a pot.	2.4	3.6	2.5
Nobody is pouring ingredients into a pot. Someone is pouring ingredients into a pot.	3.5	4.2	3.7

Example sentence pairs from the SICK test data. G denotes ground truth relatedness $\in$ *[1, 5], S = skip-thought* predictions, and M = MaLSTM predictions.

Word-level experiment

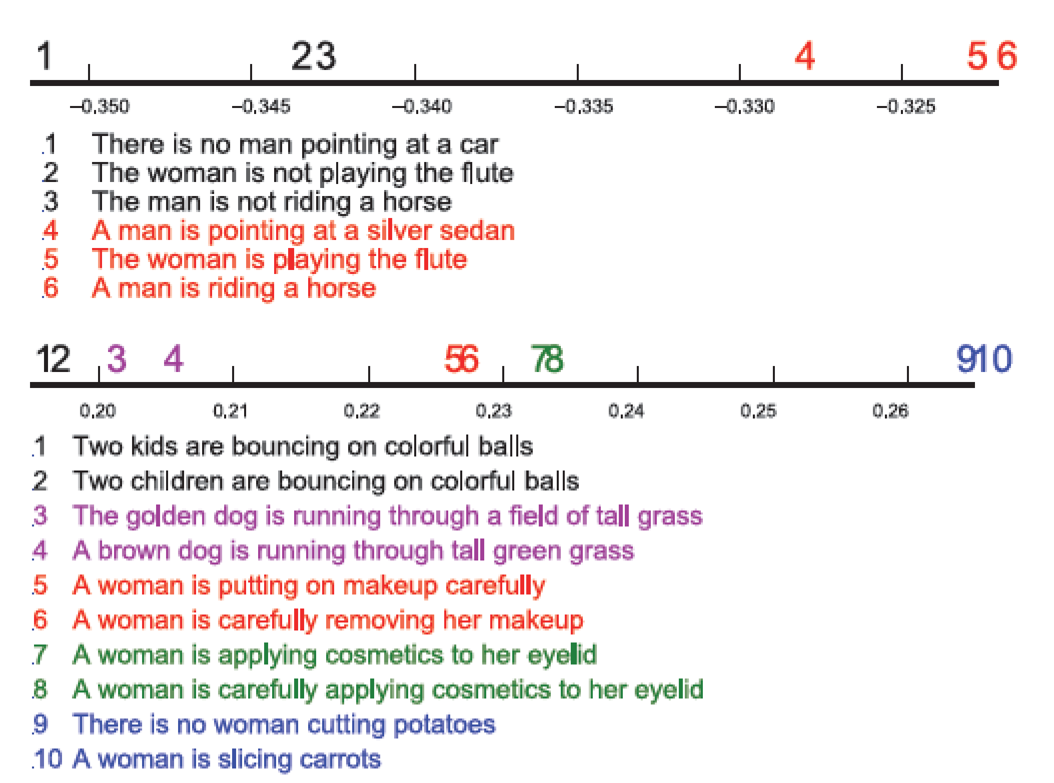
Method	r	ρ	MSE
Illinois-LH (Lai and Hockenmaier 2014)	0.7993	0.7538	0.3692
UNAL-NLP (Jimenez et al. 2014)	0.8070	0.7489	0.3550
Meaning Factory (Bjerva et al. 2014)	0.8268	0.7721	0.3224
ECNU (Zhao, Zhu, and Lan 2014)	0.8414	–	–
Skip-thought+COCO (Kiros et al. 2015)	0.8655	0.7995	0.2561
Dependency Tree-LSTM (Tai, Socher, and Manning 2015)	0.8676	0.8083	0.2532
ConvNet (He, Gimpel, and Lin 2015)	0.8686	0.8047	0.2606
MaLSTM	**0.8822**	**0.8345**	**0.2286**

- Three metrics:
  r: Pearson correlation
  p: Spearman's p
  MSE: mean-squared-error

- only an evaluation of complete relatedness-scoring systems

- [0,1] ~ [1,5] by an additional nonparametric regression

Word-level experiment



- **Semantic similarity** between sentences by simply aggregating their differences in various characteristics.
- Different dimensions of h (h1,h2): negation, categorization

Word-level experiment

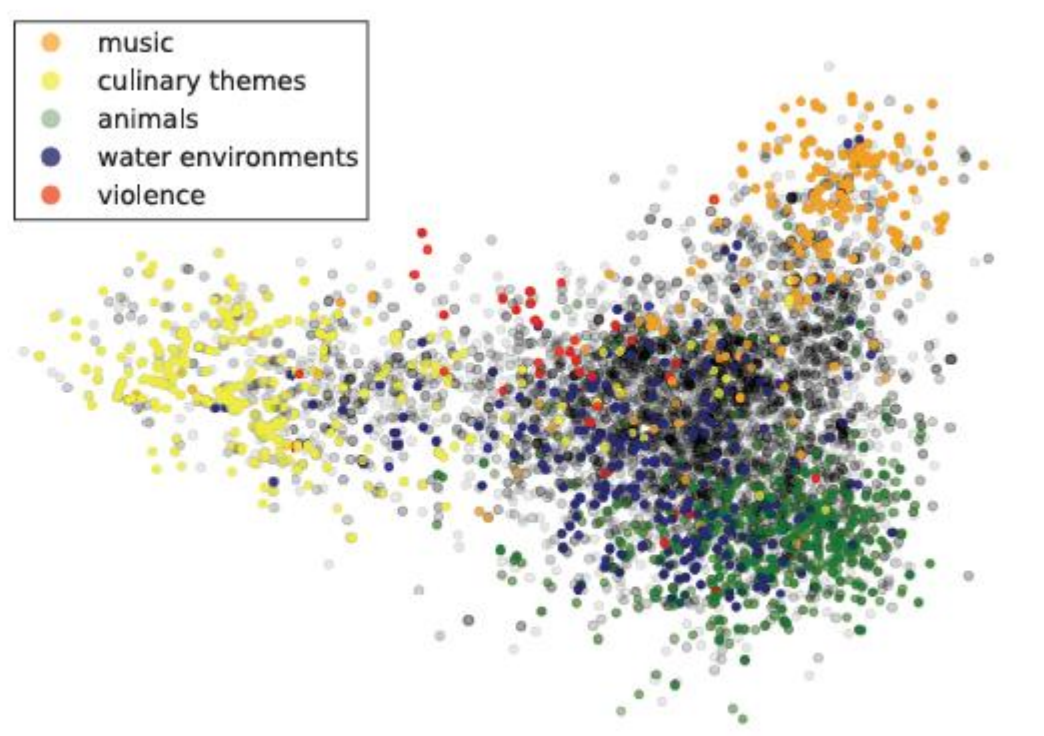


Figure 3: MaLSTM representations for all sentences from the SICK test set, projected onto two principal components.

- Sentence representation space
- PCA

Word-level experiment

Method	Accuracy
Illinois-LH (Lai and Hockenmaier 2014)	**84.6**
ECNU (Zhao, Zhu, and Lan 2014)	83.6
UNAL-NLP (Jimenez et al. 2014)	83.1
Meaning Factory (Bjerva et al. 2014)	81.6
Reasoning-based n-best (Lien and Kouylekov 2015)	80.4
LangPro Hybrid-800 (Abzianidze 2015)	81.4
SNLI-transfer 3-class LSTM (Bowman et al. 2015)	80.8
MaLSTM features + SVM	84.2

- **Evaluate sentence representation**
- Feature:
  element-wise (abs) differences &
  element-wise product
- SVM
- Data: SemEval 2014 textual entailment task
  pair label: entailment, contradiction or neutral
- LH features specially constructed for this task

PART FOUR

Conclusion

Conclusion

- Map from a general space of variable length sequence into an structured metric space of fixed dimensionality
- Simple metrics capture sentence similarity
- Sentence embedding
  - Last hidden state of the model
  - Add a BP net based last hidden state
- Sentence distance
  - Cosine distance
  - Manhattan distance

Discussion

- Different weight W:
  asymmetric domains such as IR(query and document)

- sentence represent:
  RNN、CNN

- Triplet network

THANKS