MappSent: a Textual Mapping Approach for Question-to-Question Similarity

Amir Hazem, Basma El Amal Boussaha et Nicolas Hernandez

LS2N, Université de Nantes, France
Recent Advances in Natural Language Processing (RANLP 2017)

2-8 September 2017, Varna (Bulgaria)

Outline

Outline

Outline

Outline

Community question answering (CQA)

CQA

- more and more popular (StackExchange, AskUbuntu, etc.)
- useful source of information
- massive resource, full of duplicate posts and similar question variants
- hard to find an answer to a given question

Community question answering (CQA)

CQA

- more and more popular (StackExchange, AskUbuntu, etc.)
- useful source of information
- massive resource, full of duplicate posts and similar question variants
- hard to find an answer to a given question
- emergence of an important area of research: **Community Question Answering**

Community question answering (CQA)

Preliminary step

- identification of similar questions
  - response effectiveness
  - reduce duplicate posts

Community question answering (CQA)

Preliminary step

- identification of similar questions
  - response effectiveness
  - reduce duplicate posts

Question-to-Question similarity task offers a key challenge

- lexical similarity, reformulation, paraphrasing, semantics, etc.

Community question answering (CQA)

Preliminary step

- identification of similar questions
  - response effectiveness
  - reduce duplicate posts

Question-to-Question similarity task offers a key challenge

- lexical similarity, reformulation, paraphrasing, semantics, etc.

SemEval shared task

a subtask is dedicated to question-to-question similarity since 2015

SemEval: Question-to-Question Similarity Task

- consists of reranking the related questions according to their similarity with respect to the original question
- each original question has 10 candidates to rerank
- candidates are labeled as *PerfectMatch*, *Relevant* or *Irrelevant*
- no distinction is made between *PerfectMatch* and *Relevant* labels
- Qatar living corpus
- training dataset consists of 317 original questions and 3,170 related questions
- test sets of 2016 and 2017 respectively consist of 70 original/700 related questions and 88 original/880 related questions

Piece of text = Subject + Body = Context + Question(s)

Piece of text = Subject + Body = Context + Question(s)

Q1: how hard is it for you to get a decent paying job in qatar ?

I had applied in bayt ; monstergulf and gulftalent almost religiously every day and yet I am getting nothing more than having my CV viewed . I have 4 years + experience in Linux and Unix environment and a handful of certifications to boot also . well ; while its back to updating my CV ; I 'd love to hear about your experiences on getting a job here .

Piece of text = Subject + Body = Context + Question(s)

Q1: how hard is it for you to get a decent paying job in qatar ?

I had applied in bayt ; monstergulf and gulftalent almost religiously every day and yet I am getting nothing more than having my CV viewed . I have 4 years + experience in Linux and Unix environment and a handful of certifications to boot also . well ; while its back to updating my CV ; I 'd love to hear about your experiences on getting a job here .

Q11: Why finding a job in Doha so hard ? (Perfect match)

Can somebody tell me how i get a good job in doha ? i have an American passport and holding a husband visas ; with a bachelor degree ; speaking 3 languages ; having a hard time to find a good job ; i post my CV on many website including bayt.com and many more ; but did n't get any respond . Thanks .

Piece of text = Subject + Body = Context + Question(s)

Q1: how hard is it for you to get a decent paying job in qatar ?

I had applied in bayt ; monstergulf and gulftalent almost religiously every day and yet I am getting nothing more than having my CV viewed . I have 4 years + experience in Linux and Unix environment and a handful of certifications to boot also . well ; while its back to updating my CV ; I 'd love to hear about your experiences on getting a job here .

Q11: Why finding a job in Doha so hard ? (Perfect match)

Can somebody tell me how i get a good job in doha ? i have an American passport and holding a husband visas ; with a bachelor degree ; speaking 3 languages ; having a hard time to find a good job ; i post my CV on many website including bayt.com and many more ; but did n't get any respond . Thanks .

Q12: About the job oppourtinites for Women in Doha (Relevant)

hi ; How about the job oppourtinites are there in doha for women who had IT experience ? with regards A.Aravind

Piece of text = Subject + Body = Context + Question(s)

Q1: how hard is it for you to get a decent paying job in qatar ?

I had applied in bayt ; monstergulf and gulftalent almost religiously every day and yet I am getting nothing more than having my CV viewed . I have 4 years + experience in Linux and Unix environment and a handful of certifications to boot also . well ; while its back to updating my CV ; I 'd love to hear about your experiences on getting a job here .

Q11: Why finding a job in Doha so hard ? (Perfect match)

Can somebody tell me how i get a good job in doha ? i have an American passport and holding a husband visas ; with a bachelor degree ; speaking 3 languages ; having a hard time to find a good job ; i post my CV on many website including bayt.com and many more ; but did n't get any respond . Thanks .

Q12: About the job oppourtinites for Women in Doha (Relevant)

hi ; How about the job oppourtinites are there in doha for women who had IT experience ? with regards A.Aravind

Q13: Reliable recruitment agencies of Doha (Relevant)

Hi all ; Can you please name some of reliable recruitment agencies in Doha you have tried and trust . Do you think it is better to hunt for job through these agent ?

SemEval approaches (2016)

1) *UH-PRHLT* [Franco-Salvador et al., 2016]

- combined lexical and semantic features and representations
- took advantage of distributed representations of words, graph knowledge (BabelNet) and frames extracted from FrameNet

2) *ConvKN* [Barrón-Cedeño et al., 2016]

- used an SVM operating on three kernels and combined convolutional tree kernels with CNN and additional manually extracted features including text similarity and thread specific features

3) *KeLP* [Filice et al., 2016]

- used SVM classifier based on a linear combination of kernel functions
- different features were used such as linguistic similarities, shallow syntactic trees encoding lexical and morpho-syntactic information, feature vectors capturing task specific information, etc.

SemEval approaches (2017) [Nakov et al., 2017]

1) SimBow

- proposed a logistic regression on a combination of different unsupervised textual similarities
- introduced a variant of cosine similarity that uses semantic similarity between words to compute cosine between two bag-of-word vectors

2) LearningToQuestion

- used SVM and logistic regression as integrators of rich features representations (word embeddings, bidirectional LSTMs, gated recurrent unit (GRU), etc.)

3) Kelp

- used SVM classifier based on a linear combination of kernel functions

How it comes to MappSent?

How it comes to MappSent?

- How to measure the similarity between two sentences? paragraphes? pieces of texts...?

How it comes to MappSent?

- How to measure the similarity between two sentences? paragraphes? pieces of texts...?
- How to represent words?
  - distributional approaches
  - distributed approaches

How it comes to MappSent?

- How to measure the similarity between two sentences? paragraphes? pieces of texts...?
- How to represent words?
  - distributional approaches
  - distributed approaches
- How to represent sentences? paragraphes? pieces of texts...?
  - phrase representation [Mikolov et al., 2013a]
  - sentence representation [Wieting et al., 2016, Arora et al., 2017]

How it comes to MappSent?

- How to measure the similarity between two sentences? paragraphes? pieces of texts...?
- How to represent words?
  - distributional approaches
  - distributed approaches
- How to represent sentences? paragraphes? pieces of texts...?
  - phrase representation [Mikolov et al., 2013a]
  - sentence representation [Wieting et al., 2016, Arora et al., 2017]
- How to make pairs of similar sentences closer in the embedding space?
  - bilingual word mapping [Artetxe et al., 2016]
  - paraphrasing [Wieting et al., 2016]

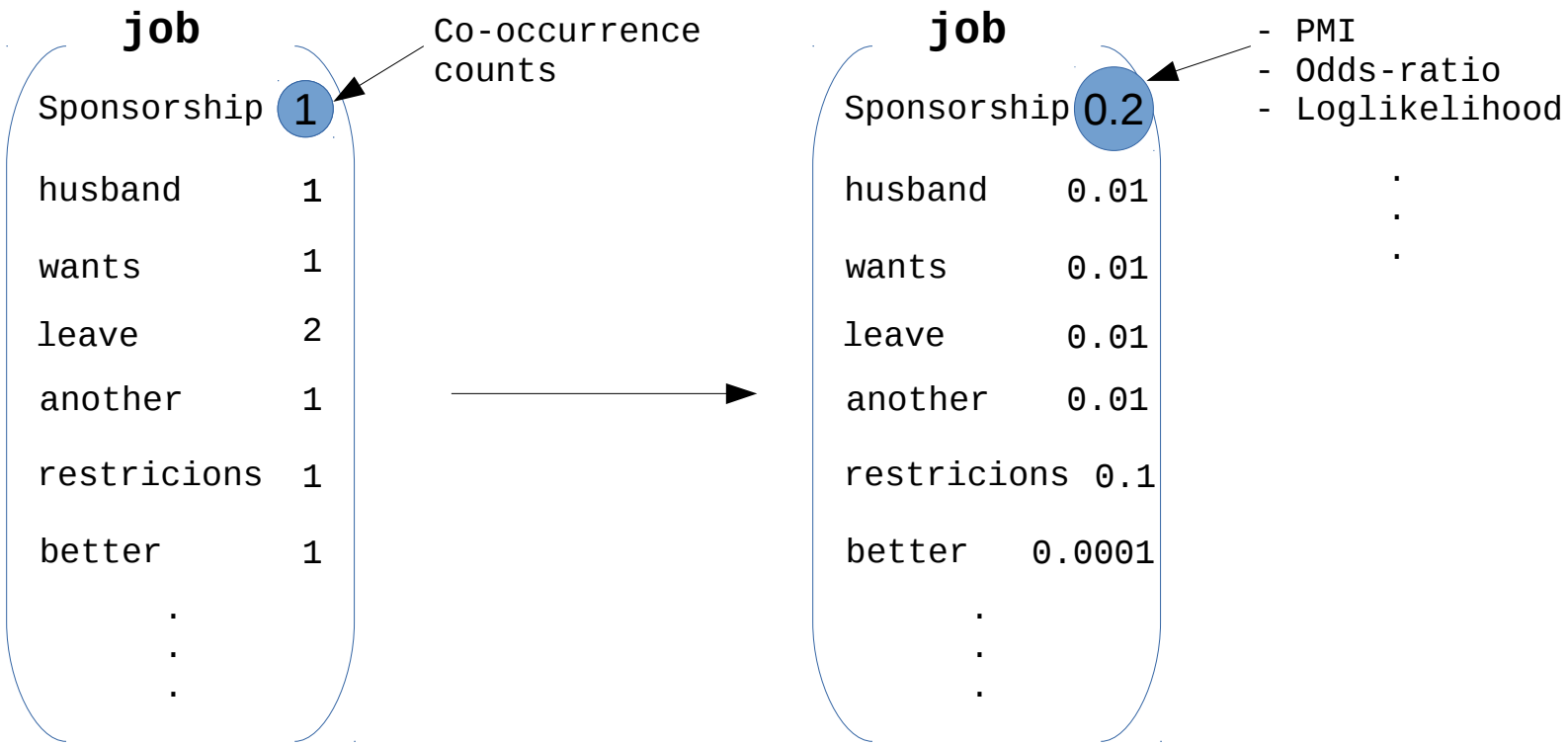
Distributional Hypothesis

Distributional Hypothesis

- one of the most successful ideas of modern statistical NLP
- represents a word by means of its neighbors (Harris, Z. (1954), Firth, J.R (1957))
- co-occurrence can be interpreted as an indicator of semantic proximity of words

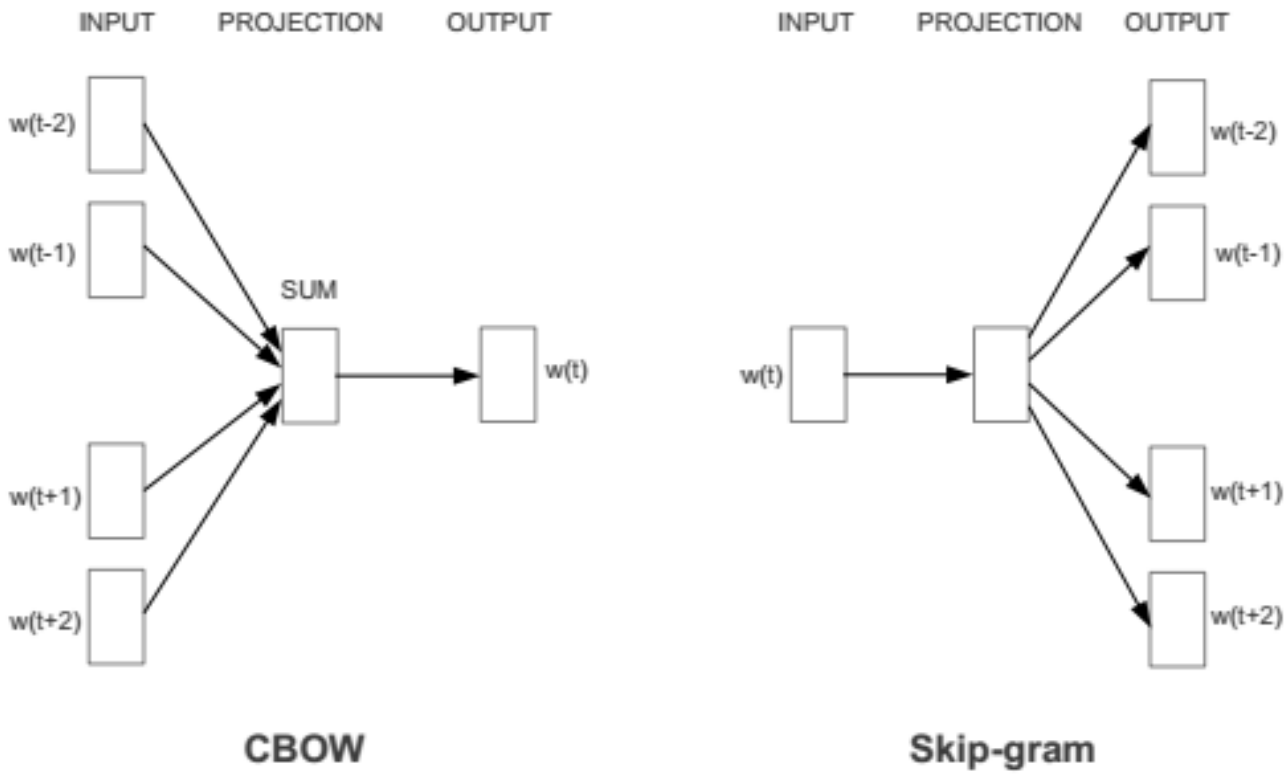
Example

If a wife that arrived on husband 's sponsorship gets a **job** ; and wants to leave that **job** for another better paying **job** ; can she do it without any restricions ?

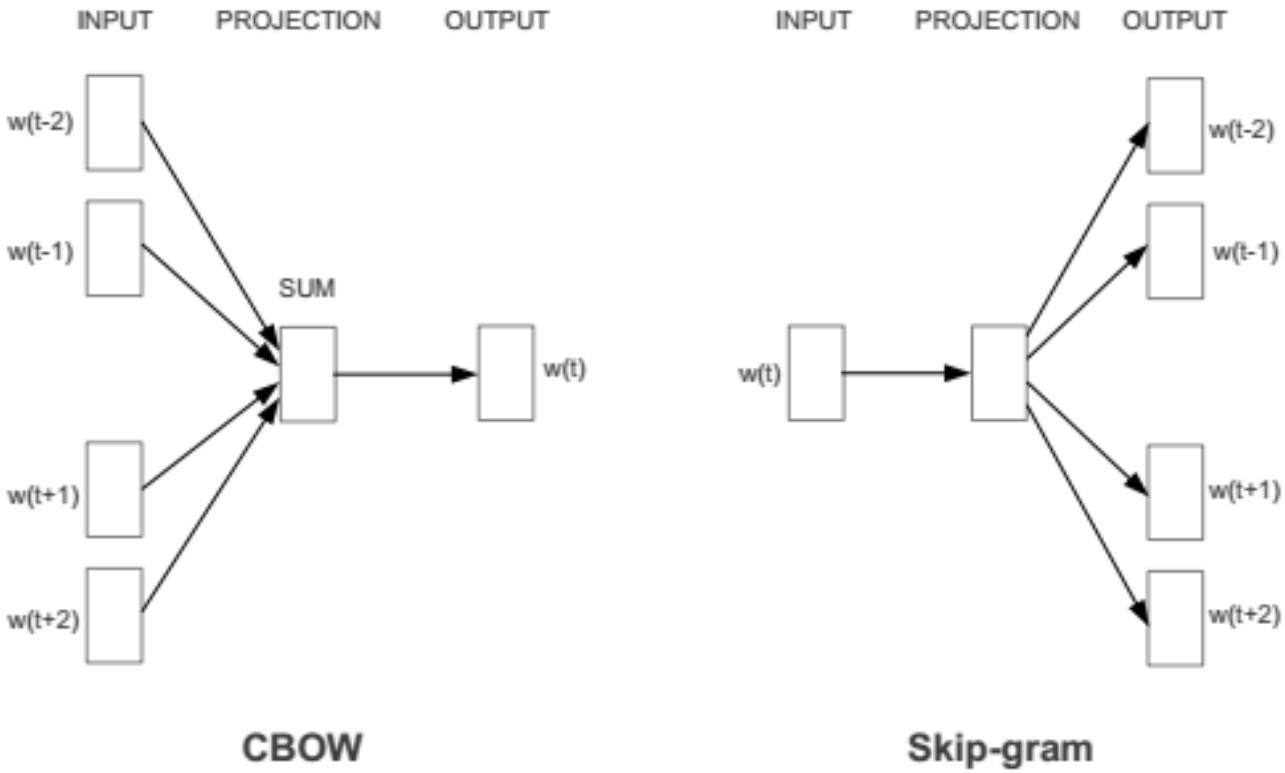


Distributed approach: word embeddings

- capture lexical and semantic word's properties by representing words in a low continuous dimensional space
  [Bengio et al., 2003, Collobert and Weston, 2008,
  Mikolov et al., 2013a, Mikolov et al., 2013b, Pennington et al., 2014]



Distributed approach: word embeddings



[Mikolov et al., 2013a]

- Skip-gram > CBOW when data is SMALL
- Cbow is faster
- Skip-gram is preferable for infrequent words

Longer textual embedding methods

- use operations on vectors and matrices like addition or multiplication to represent phrases, sentences or paragraphs
  [Mitchell and Lapata, 2010, Mikolov et al., 2013a, Socher et al., 2011, Mikolov et al., 2013a, Le and Mikolov, 2014, Kalchbrenner et al., 2014, Kiros et al., 2015, Wieting et al., 2016, Arora et al., 2017]

Word embeddings: learning phrases [Mikolov et al., 2013a]

- many phrases have a meaning that is not a simple composition of the meaning of its individual words
- Skip-gram model exhibits a linear structure for analogical reasoning using simple vector arithmetics
- Skip-gram model exhibits another kind of linear structure for meaningfully combine words by an element-wise addition of their vector representations
  - Russia + river = Volga river
  - German + airlines = airline Lufthansa

Sentence embedding representation [Arora et al., 2017]

- first compute a weighted average sum of the word embedding vectors of sentences
- then remove the projections of the average vectors on their first principal components
- Like [Mikolov et al., 2013a] and [Wieting et al., 2016], their approach is based on word embedding sum, but the difference is remarkable on the weighted schema and on the use of principal component analysis (PCA) method to remove the correlation of sentence vectors dimensions
- significantly achieved better performance than the unweighted average on a variety of textual similarity tasks
- outperformed sophisticated supervised methods such as RNN's and LSTM's

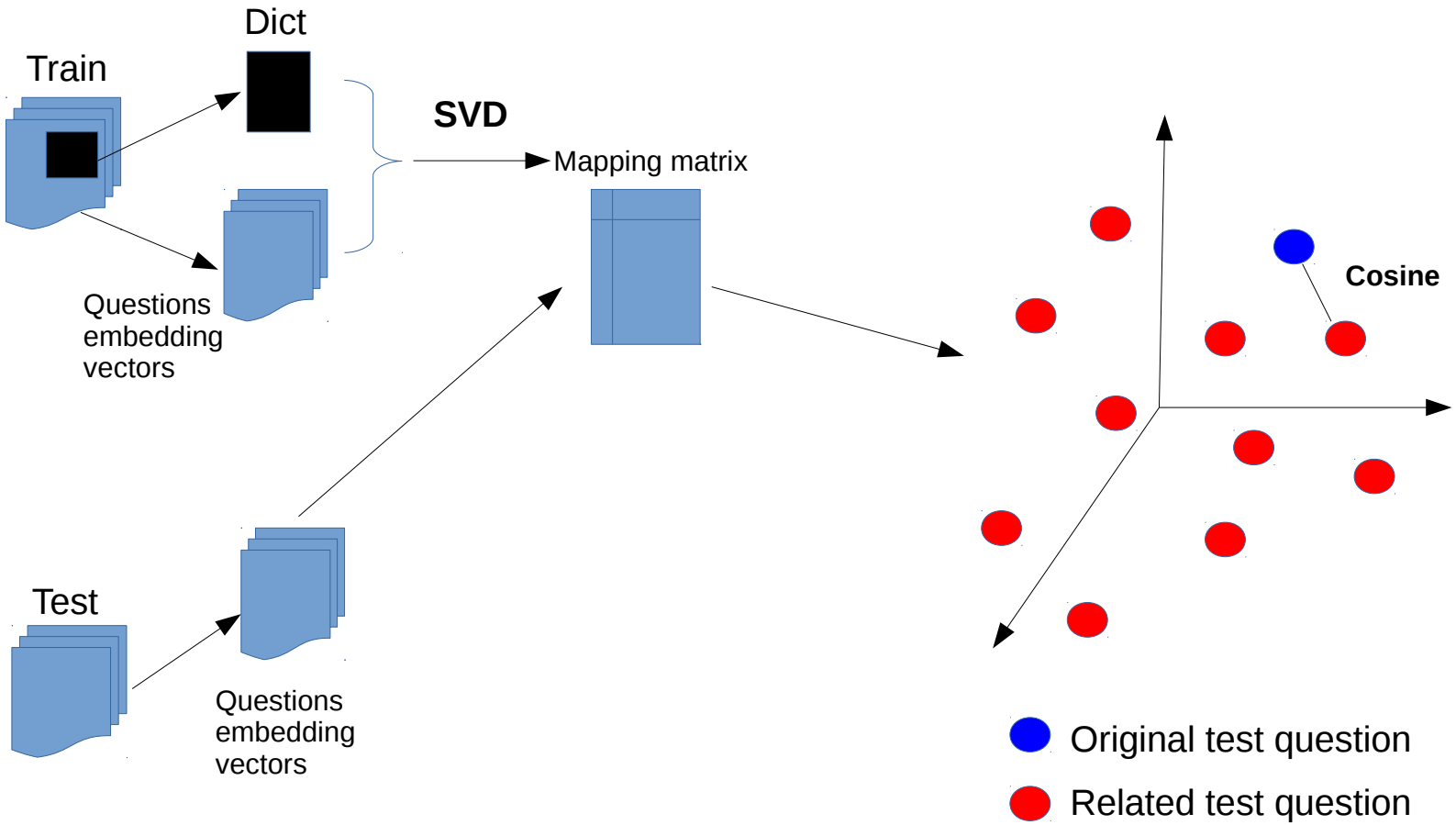
More sophisticated approaches

- recurrent neural networks (RNN)
  [Socher et al., 2014, Kiros et al., 2015],
- long short-term memory (LSTM) [Tai et al., 2015]
- convolutional neural networks (CNNs) [Kalchbrenner et al., 2014]
- remarkable improvements in a wide range of applications
- computational cost and the need of large amount of training data
- inefficient on small and specific datasets

MappSent

- build a sentence embedding space
- sentences are represented by the sum of their word embedding vectors
- similar sentences are moved closer thanks to a mapping matrix [Artetxe et al., 2016]
- mapping matrix is learned from a seed training dataset containing annotated similar sentences
- optimal mapping is computed by minimizing the distance between the seed sentence pairs

MappSent


Dict
Train
SVD
Mapping matrix
Questions
embedding
vectors
Cosine
Test
Questions
embedding
vectors
Original test question
Related test question

MappSent steps:

1. train a Skip-Gram model using Gensim [Řehůřek and Sojka, 2010] on the *Qatar Living* lemmatized training dataset
2. training and test sentences were pre-processed (remove stopwords and only keep nouns, verbs and adjectives while computing sentence embedding vectors and the mapping matrix)
3. compute each sentence embedding vector (element-wise addition of its words embedding vectors [Mikolov et al., 2013a]
4. build a mapping matrix (by adapting [Artetxe et al., 2016] approach in a monolingual scenario)
5. project test sentences in the new subspace using the mapping matrix
6. compute Cosine between the projected test sentences

MappSent: How to build the mapping matrix?

- need a mapping dictionary which contains similar sentence pairs
  - consider pairs of sentences that are labeled as *PerfectMatch* and *Relevant* in the Qatar Living training dataset
- learn a linear transformation which minimizes the sum of squared Euclidean distances for the dictionary entries
  - use orthogonality constraint
  - preserve length normalization
  - use mean centering
- in the bilingual scenario, source words are projected in the target space using the bilingual mapping matrix
- in our case, original and related questions are both projected in a similar subspace using the monolingual sentence mapping matrix

Results

Method	MAP(%)
UH-PRHLT	76.70
ConvKN	76.02
KeLP	75.83
Arora	77.87
$Arora_{pca}$	78.81
$MappSent^{-}$	78.56
$MappSent^{-}_{pca}$	78.66
MappSent	**79.18**
$MappSent_{pca}$	79.09

Table: Results on SemEval-2016 Task3 Subtask B

Results

Method	MAP(%)
Simbow	47.22
LearningToQuestion	46.93
KeLP	46.66
Arora	46.93
$Arora_{pca}$	46.66
$MappSent^{-}$	46.90
$MappSent^{-}_{pca}$	46.53
MappSent	47.50
$MappSent_{pca}$	**49.29**

Table: Results on SemEval-2017 Task3 Subtask B

Results

# PCA	*Arora*	*MappSent*$^{-}$	*MappSent*
0	76.47	77.45	**79.18**
1	**78.81**	78.66	78.39
2	77.46	77.80	77.66
3	77.20	78.35	77.63
4	77.91	**78.82**	78.02
5	78.20	78.01	77.13
6	78.59	78.14	77.34
7	78.33	78.09	77.60
8	77.64	77.69	77.51
9	77.64	77.72	78.13
10	77.16	77.14	78.19
20	76.51	75.86	77.08

Table: Comparison of *Arora* and *MappSent* on SemEval 2016 while removing different numbers of principal components (w=20 and dim= 300)

Results

# PCA	*Arora*	*MappSent*$^{-}$	*MappSent*
0	44.90	45.83	47.36
1	46.66	46.53	46.77
2	**47.40**	46.81	48.57
3	46.86	46.52	49.07
4	46.50	46.70	**49.29**
5	45.60	46.79	48.69
6	45.72	46.52	47.55
7	47.19	**47.21**	47.77
8	46.97	46.53	47.24
9	45.51	46.48	47.41
10	45.35	46.15	46.84
20	46.53	47.07	46.70

Table: Comparison of *Arora* and *MappSent* on SemEval 2017 while removing different numbers of principal components (w=10 and dim=500)

Conclusion

- MappSent, a novel approach for textual similarity
- Map sentences in a joint more representative sub-space
- Thanks to questions mapping matrix, similar questions are pushed closer suggesting that the new sub-space is more discriminant
- Experimental results confirm our intuition while MappSent and its PCA-based variant obtain the best results on SemEval (2016/2017) question-to-question similarity task
- One remarkable advantage is its simplicity (neither intensive computation nor external resources or metadata are needed)
- Can be applied to pieces of text of any length as long as a training set of similar texts is available

Perspectives

- explore linguistic features and sentence structure
- exploiting the context of a question and the question itself differently
- apply our approach to questions and answers
- use metadata
- paraphrasing
- RNN, LSTM, CNN...
- pragmatic relations...

https://github.com/hazemAmir/MappSent.git

SOON AVAILABLE

https://github.com/hazemAmir/MappSent.git

SOON AVAILABLE

Arora, S., Yingyu, L., and Tengyu, M. (2017).
A simple but tough to beat baseline for sentence embeddings.
In *Proceedings of the 17th International Conference on Learning Representations (ICLR'17)*, pages 1–11.

Artetxe, M., Labaka, G., and Agirre, E. (2016).
Learning principled bilingual mappings of word embeddings while preserving monolingual invariance.
In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP'16)*, pages 2289–2294, Austin, TX, USA.

Barrón-Cedeño, A., Da San Martino, G., Joty, S., Moschitti, A., Al-Obaidli, F., Romeo, S., Tymoshenko, K., and Uva, A. (2016).
Convkn at semeval-2016 task 3: Answer and question selection for question answering on arabic and english fora.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 896–903, San Diego, California. Association for Computational Linguistics.

Bengio, Y., Ducharme, R., Vincent, P., and Jauvin, C. (2003).
A neural probabilistic language model.
JOURNAL OF MACHINE LEARNING RESEARCH, 3:1137–1155.

Collobert, R. and Weston, J. (2008).
A unified architecture for natural language processing: Deep neural networks with multitask learning.
In *International Conference on Machine Learning, ICML*.

Filice, S., Croce, D., Moschitti, A., and Basili, R. (2016).
Kelp at semeval-2016 task 3: Learning semantic relations between questions and answers.
In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1116–1123, San Diego, California. Association for Computational Linguistics.

Franco-Salvador, M., Kar, S., Solorio, T., and Rosso, P. (2016).
UH-PRHLT at semeval-2016 task 3: Combining lexical and semantic-based features for community question answering.

In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 814–821.


Kalchbrenner, N., Grefenstette, E., and Blunsom, P. (2014).
A convolutional neural network for modelling sentences.
CoRR, abs/1404.2188.


Kiros, R., Zhu, Y., Salakhutdinov, R. R., Zemel, R., Urtasun, R., Torralba, A., and Fidler, S. (2015).
Skip-thought vectors.
In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc.


Le, Q. V. and Mikolov, T. (2014).
Distributed representations of sentences and documents.
CoRR, abs/1405.4053.


Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013a).

Distributed representations of words and phrases and their compositionality.
In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.


Mikolov, T., Yih, S. W.-t., and Zweig, G. (2013b).
Linguistic regularities in continuous space word representations.
In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.


Mitchell, J. and Lapata, M. (2010).
Composition in distributional models of semantics.
Cognitive Science, 34(8):1388–1439.


Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017).
SemEval-2017 task 3: Community question answering.

In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada. Association for Computational Linguistics.


Pennington, J., Socher, R., and Manning, C. D. (2014).
Glove: Global vectors for word representation.
In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.


Řehůřek, R. and Sojka, P. (2010).
Software Framework for Topic Modelling with Large Corpora.
In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.
http://is.muni.cz/publication/884893/en.


Socher, R., Huang, E. H., Pennington, J., Ng, A. Y., and Manning, C. D. (2011).
Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection.
In *Advances in Neural Information Processing Systems 24*.

Socher, R., Karpathy, A., Le, Q. V., Manning, C. D., and Ng, A. Y. (2014).
Grounded compositional semantics for finding and describing images with sentences.
TACL, 2:207–218.

Tai, K. S., Socher, R., and Manning, C. D. (2015).
Improved semantic representations from tree-structured long short-term memory networks.
CoRR, abs/1503.00075.

Wieting, J., Bansal, M., Gimpel, K., and Livescu, K. (2016).
Towards universal paraphrastic sentence embeddings.
Internationa Conference on Learning Representations, CoRR, abs/1511.08198.