Unified sentence embedding for question answering

Abstract

In Huang et al. 2013 Deep Structured Semantic Model was proposed. General idea is finding a deep vector representation for logically binded objects. In this study the idea is used for learning representative question and answer embeddings.

Afterwards, this idea is being combined with sequence-to-sequence modeling to develop more informative vector representations.

DSSM

Each answer and question is a sequence of word-level tokens. First, a word2vec transition is applied. For each unique word a corresponding vector is being matched.

The resulting dataset D is a set of pairs (q, a) where $q \in Q$ is a sequence of vectors corresponding to a single question and $a \in A$ is the same for an answer.

Final goal is to learn two mappings: ϕ and ψ. Assume there is a latent semantic space $V = \mathbb{R}^n$.

$$\phi(q) : Q \to V \quad \psi(a) : A \to V \tag{1}$$

There is some similarity measure ρ on V.

$$\rho(v_1, v_2) : V \times V \to \mathbb{R} \tag{2}$$

ρ is supposed to show how good a given A fits to a given Q.

The last question is how to measure how good a given ϕ and ψ solve our task?

Triplet loss was proposed by Xiao et al. 2016. We try to make related question and answer closer in the semantic space, but unrelated objects are supposed to be farther. The criteria of significant difference in those distances is denoted as *margin*.

More formally, the goal is to minimize the following expectation. Assume $p(q, a_+, a_-)$ is a distribution over $Q \times A \times A$. χ_D stands for identity function of D, $p(q, a_+, a_-) \propto \chi_D(q, a_+)(1 - \chi_D(q, a_-))$

$$\mathop{\mathbb{E}}_{(q,a_+,a_-)\sim p(q,a_+,a_-)} \left[\rho(\phi(q), \psi(a_+)) - \rho(\phi(q), \psi(a-)) + margin\right]_+ \tag{3}$$

Practical

DNN methods were used for an implementation of this model. ϕ and ψ would be LSTM layers. Last state of LSTM is cut and passed to one (or more) dense layer.

In original paper ρ is a cosine similarity. But, in this particular case, euclidean metric worked well enough.

$$\rho(v_1, v_2) = ||v_1 - v_2||_2 \tag{4}$$

The data was collected from https://thequestion.ru. It contains near 100 thousands questions and 200 thousands answers.

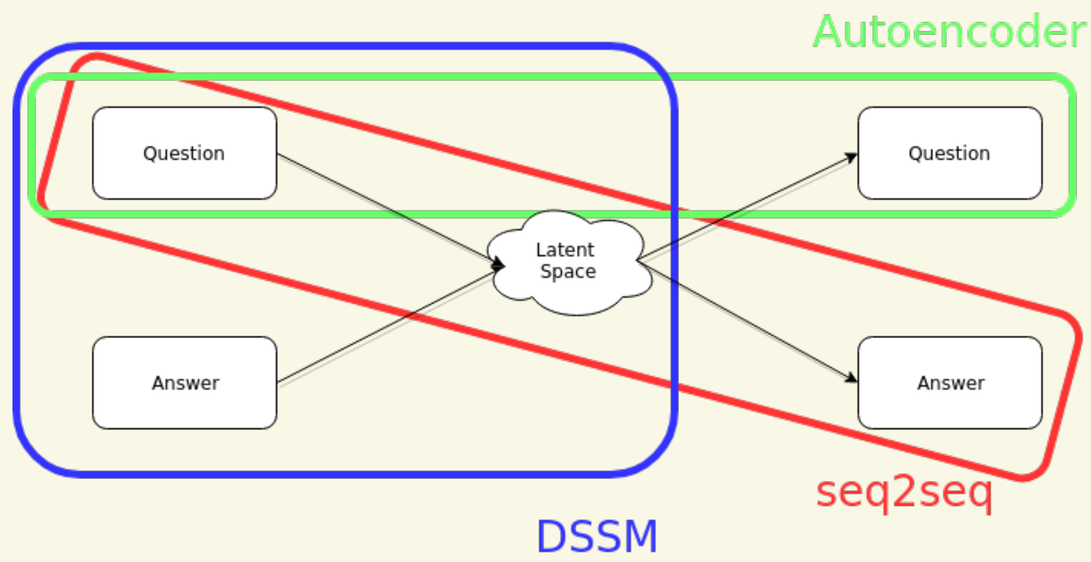
Theano and lasagne were used as main frameworks. Due to huge vocabulary size, a hierarchical softmax were used to reduce memory usage.

Word2vec size was empirically chosen equal to 512, latent semantic space dimensionality is 256.

After a long time of trainings loss became equal to 1.5 on validation sample, with $margin = 10$. This means, that expected $\rho(\phi(q), \psi(a_+))$ and $\rho(\phi(q), \psi(a_-))$ remarkably differ.

Generative model

The idea of sentence embeddings is not unique for DSSM. A brief review over known models will show at least three types of models, pictured bellow.



Embeddings, built by different models, would be different. There might exist a common optimum for those models: a single set of mappings which minimizes all loss functions.

A DSSM model was combined with sequence-to-sequence model proposed in Sutskever, Vinyals, and Le 2014. A new mapping is defined.

$$\alpha(v) : V \to A \tag{5}$$

And a model $\alpha(\phi(q)) : Q \to A$ is being trained to minimize negative log-likelihood simultaneously with DSSM, sharing the same parameters for ϕ.

Examples

The answer is $\arg\min_a \rho(\phi(q), \psi(a))$, while q was absent in D.

Q: Куда съездить в путешествие?
A: Суздаль, Екатеринбург, Калининград, Казань, Набережные Челны.Города красивые, есть что посмотреть, есть чего попробовать. Рекомендую в общем.

Conclusion

For evaluation Mean Reciprocal Rank metric and First Hit Success were used

$$MRR = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{1}{rank_{a_i} \{\rho(a, q_i) | a \in A\}} \tag{6}$$

FHS is simply a part of cases when first proposed answer was correct.

As a baseline a Doc2Vec (Le and Mikolov 2014) were taken.

	MMR	FHS	Triplet loss
Doc2vec	0.0059	0.003	-
DSSM	0.0821	0.029	2.86
DSSM & Generative	0.0995	0.049	2.99

Deep end-to-end models are not applicable to general question-answering yet. On the other hand, it might work well in vertical cases, like FAQ and product support optimization.

Having a unified sentence embedding would allow to choose online between two options: pick a nearest answer from data bank or generate a new one.

References

Huang, Po S. et al. (2013). "Learning Deep Structured Semantic Models for Web Search Using Clickthrough Data". In: *Proceedings of the 22Nd ACM International Conference on Information & Knowledge Management*. CIKM '13. San Francisco, California, USA: ACM, pp. 2333–2338. URL: http://dx.doi.org/10.1145/2505515.2505665.

Xiao, Qiqi et al. (2016). *Cross Domain Knowledge Transfer for Person Re-identification*. arXiv: 1611.06026.pdf. URL: http://arxiv.org/abs/1611.06026.pdf.

Sutskever, Ilya, Oriol Vinyals, and Quoc V. Le (2014). *Sequence to Sequence Learning with Neural Networks*. arXiv: 1409.3215. URL: http://arxiv.org/abs/1409.3215.

Le, Quoc V. and Tomas Mikolov (2014). *Distributed Representations of Sentences and Documents*. arXiv: 1405.4053. URL: http://arxiv.org/abs/1405.4053.

Oleg Vasilev
omrigann@gmail.com
NRU HSE

Ilia Yakubovskiy
ilia.yakubovskiy@skolkovotech.ru
Skoltech