A Generative Model for Semantic Question Matching

Michel Deudon
Telecom Paris-Tech

Bruno Taille
Telecom Paris-Tech

Umut Simsekli
Telecom Paris-Tech

Abstract

Measuring semantic similarity between pair of sentences is an important problem in NLP, especially in question answering, information retrieval, information extraction, conversation systems, to name a few. In this paper, we address the problem of automatically detecting duplicate questions with a generative model. Our proposed method is based on a combination of deep generative models (VAE) with sequence-to-sequence models (LSTM) to generate paraphrases, given an input sentence. This factorization allows it to explicitly capture holistic properties of sentences such as style, topic, and high-level syntactic features in an explicit global latent sentence representation for downstream applications.

1 Introduction

Semantic is the study of or relating to meaning in language, used for understanding human expressions. It is concerned with the relationship between signifiers (words, phrases, signs and symbols) and what they stand for, their denotation. Semantics has many applications including question answering, information retrieval, news categorization and clustering, multilingual document matching, machine translation or image captioning evaluation metrics, automated short-answer grading, identifying duplicate posts or any other challenges that arise in building a scalable online knowledge-sharing platform.

Understanding the meaning of text is an abstract goal, difficult to quantify. Thus the Semantic Textual Similarity (STS), Paraphrase detection and Semantic Question Matching tasks propose the more concrete problem to determine the degree of similarity between two sentences or paragraphs. The Natural Language Inference (NLI) task aims to determine whether a Premise is neutral, entails or contradicts a Hypothesis.

2 Background

An important principle used to compute semantic representations of sentences is the principle of compositionality which states that the meaning of a phrase is uniquely determined by the meaning of its parts and the rules that connect those parts.

2.1 Word Embeddings

Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) are semantic embeddings of words based on their context or co-occurrences in a text corpus such as Wikipedia or Google News (unsupervised learning task). Today, Word2Vec or Glove vectors (and their associated latent space) are the building blocks for many NLP tasks such as Machine Translation (Bahdanau et al., 2014), Image captioning and Question Answering.

Word2Vec or Glove vectors can be directly used to measure similarity between pieces of text. A Bag Of Word (BOW) model considers sentences or documents as buckets of words. Such representations could be used to compute weighted sums of corresponding words' embeddings. Weights are either typically uniform (mean, sum) or obtained with an Information Retrieval measure metric (TF-IDF, Okapi BM25 (Beaulieu et al., 1997) to amplify the signal of specific word and reduce noise. Euclidean distance or cosine similarity could then be used to measure the similarity between two sentences representations and classify them as duplicate or not. (Arora et al., 2016) proposed using Smooth Inverse Frequency (SIF) weights and removing the word vectors first principal component to compute a sentence's discourse vector ("what is being talked about"). $v_s = \sum_{w \in s} \frac{a}{a+p(w)} \pi_{c_0^\perp}(v_w)$

ParagraphVector DBOW (Le and Mikolov, 2014) is trained to learn sentence as well as word

embeddings by predicting words contained in sentences, given their vector representation. Sent2Vec is a sentence embedding defined as the average of the source word embeddings (unigrams) and also the present n-grams embeddings.

Distances other than euclidean and cosine similarity could be used to measure BOW document similarity, Word Mover's Distance (Kusner et al., 2015) measures the dissimilarity between two text documents as the minimum amount of distance the embedded words of one document need to "travel" to reach the embedded words of another document. This distance metric is computationally expensive (Transportation problem) and experimentally struggles to finely capture semantics between pieces of text.

Work on non euclidean embedding spaces has been investigated too. In (Nickel and Kiela, 2017), words are embedded in hyperbolic spaces (spaces with constant negative curvature) to simultaneously capture similarity and hierarchy of words (Zipf's power-law distributions in Natural Language).

2.2 Representation learning for semantic textual similarity

Instead of explicitly designing/engineering features (word alignment, n-gram overlap, lexical word overlap, longest common substring, sentences length, euclidean or cosine distance of weighted sum of word embeddings...), one could also learn them using Neural Networks.

Different neural architectures have been investigated to encode sentences into fixed-size representation. (Bi-directional) RNN encoders with LSTM or GRU cells (with either mean or max pooling) treat sentences sequentially. Convolutional networks (Ma et al., 2015) treat them hierarchically. (Self-)attentive embeddings (Lin et al., 2017) (Vaswani et al., 2017) treat them as a set or sequence of (query,references).

When designing a neural network for a text-pair task, probably the most important decision is whether the meanings of the texts should be represented independently, or jointly. An independent representation means that the network can read a text in isolation, and produce a vector representation for it. This is great if lots of comparisons over the same texts have to be done, for instance to find their pairwise-similarities. However, reading the sentences independently makes the text-pair task more difficult. Models which read the sentences together before reducing them to vectors have an accuracy advantage.

How to capture the relationships among multiple words and phrases in a single vector still remains a question to be solved and depends on the neural architecture and the (supervised or unsupervised) training task.

Supervised discriminative models

Discriminative models learn the (hard or soft) boundary between classes. In the supervised setting, a labeled dataset is available during training.

The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015a) or the Multi-Genre NLI Corpus (Williams et al., 2017) consist of sentence pairs annotated with textual entailment information ("entails", "neutral", "contradicts") for Natural Language Inference (NLI) aka Recognizing Textual Entailment (RTE). The Paraphrase Database (Ganitkevitch et al., 2013) and the Microsoft Research Paraphrase Corpus (Dolan et al., 2005) contain labeled paraphrases. The SemEval2015 dataset (Agirre et al., 2014) consists of sentence pairs annotated with their degree of similarity (0 to 5). The Quora Question Pairs Dataset `https://data.quora.com/` consists of question pairs and a binary value that indicates whether the pair is a duplicate pair (same content) or not.

One could frame the above tasks as classification tasks. A traditional Deep Learning approach to solve these tasks consist in building a "siamese" network. Each sentence is separately encoded into a fixed length hidden vector (sentence representation). The encoder module typically is a Recursive Neural Network, a 1D Convolutional Neural Network, LSTM based RNN or a biLSTM architecture with max pooling that reads the input sentence as a sequence of word embeddings. The concatenation, Hadamard product and/or squared difference of these sentence representations [concat(u,v), u*v, $|u - v|$] is then used as input layer for a deep neural network trained to predict the degree of similarity or entailment between two sentences as in STS (Sanborn and Skryzalin, 2015), SNLI (Conneau et al., 2017) (Munkhdalai and Yu, 2016) (Liu et al., 2016) (Shen et al., 2017b), Quora (Dadashov et al.,). Instead of considering the final layer as a classification layer, one could also consider an energy based model to learn a similarity metric discriminatively, as in (Chopra et al.,

2005).

An attention-based approach was proposed in (Parikh et al., 2016) (soft) aligns words (intra + inter attention), compare them and aggregate computations to predict the degree of similarity between two sentences. (Tomar et al., 2017) modified the model to use sums of character n-gram embeddings instead of word embeddings.

The bilateral multi-perspective matching model (BIMPM) (Wang et al., 2017) uses a character-based LSTM, a layer of bi-LSTMs for context, four different types of multi-perspective matching layers, an additional bi-LSTM aggregation layer, followed by a 2-layer FFN for prediction.

Interactive Inference Network (IIN) (Gong et al., 2017) achieves state-of-the-art results on SNLI, Multi-Genre NLI and Quora dataset. Similarly to a siamese network, IIN first embeds and encodes separately two sentences. Instead of matching vectorial representations of sentences into a matrix (as for the siamese network), IIN matches matrix representation of sentence (word-by-word) in an interactive layer (dense 3D tensor). A Convolutional layer with 2-D kernels (AlexNet, VGG, Inception, ResNet and DenseNet) extracts the semantic features from the interaction tensor and an output layer predicts classification.

Unsupervised Auto Encoders

Neural autoencoders are models where output units are identical to input units. Inputs D are compressed with neural encoders into a representation e_D, which is then used to reconstruct it back.

Sequential Denoising Autoencoder SDAE introduces noise in the input to predict (as target) the original source sentence S given a corrupted version of it N(S—po, px). The noise function N(S—po, px), is determined by free parameters po, px. For each word w in S, N deletes w with (independent) probability po. Then, for each non-overlapping bigram wiwi+1 in S, N swaps wi and wi+1 with probability px. The model then uses an LSTM-based architecture to recover the original sentence from the corrupted version. As a result of this process, SDAEs learn to represent the data in terms of features that explain its important factors of variation. The model can then be used to encode new sentences into vector representations. SDAEs are the top performer on paraphrase identification, among unsupervised model (Hill et al., 2016).

Skip Thought (ST) vectors (Kiros et al., 2015) adapt the skip-gram model for words to the sentence level, by encoding a sentence to predict (decode) the sentences around it. ST vectors require a consequent training corpus of ordered sentences, with a coherent narrative. On the (unsupervised) SICK sentence relatedness benchmark (10,000 sentence pairs) (Marelli et al., 2014), FastSent, a BOW variant of SkipThought, performs best among unsupervised model (Hill et al., 2016).

Hierarchical LSTMs with attention (Li et al., 2015) can operate at the token level and document-level representations, in a hierarchical structure. To improve reconstruction, one could inject context information (topics associated with the sentence, sentence length feature e.g.) as low dimensional vectors in a deep autoencoder (Amiri et al., 2016).

2.3 Generative models

Generative models model the distribution of individual classes. They suffer less from overfitting or missing data and are perfectly suited for semi-supervised learning (clustering).

Generative models with latent variables have been successfully used for topic modeling. In Latent Dirichlet Allocation (LDA) (Blei et al., 2003), the document distribution over topics θ has a Dirichlet prior. In probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999), θ is simply a multinomial parameter.

The latent distributions grant the ability to sum over all the possibilities in terms of semantics. From the perspective of optimization, Bayesian learning guards against overfitting.

Variational Auto Encoders

In contrast to the standard autoencoder which learns, for any input x, a deterministic latent code z via a deterministic encoder function q_ϕ, the VAE (Kingma and Welling, 2013)(Rezende et al., 2014), a deep generative model, learns a posterior distribution $q_\phi(z|x)$ over the latent code $z \in R^k$ (sentence representation), given an input $x \in R^d$. The posterior $q_\phi(z|x)$ is usually assumed to be a Gaussian distribution $z \sim \mu(x) + \sigma(x)N(0;1)$, and the functions $\mu(x), \sigma(x)$ are nonlinear transformations of the input x.

The VAE also consists of a decoder model, which is another distribution $p_\theta(x|z)$ that takes as input a random latent code z and produces an observation x. The parameters defining the VAE are learned by maximizing the following lower

bound on the model evidence $p(x|\theta, \phi)$:
$L(\theta; \phi; x) = E_{q_\phi(z|x)}[\log p_\theta(x|z)] - KL(q_\phi(z|x)||p(z))$

The VAE learns to reconstruct the original input x from a latent code z and encourages its posterior distribution $q_\phi(z|x)$ to be close to a prior p(z), typically a standard normal distribution $N(0; I)$.

In (Miao et al., 2016), a deep neural inference network conditioned on the discrete text input (BOW model) approximates the intractable distributions over the latent variables and provides the variational distribution. NVI learns to model the posterior probability, thus endowing the model with strong generalisation abilities. By using the reparameterisation method, the inference network is trained through back-propagating unbiased and low variance gradients w.r.t. the latent continuous variables.

(Bowman et al., 2015b) presented a text generation model in which the encoder and decoder were modeled by long short-term memory (LSTM) networks. Moreover, training tricks such as KL-term annealing and dropout of inputs of the decoder were employed to circumvent the problems encountered when using the standard VAE for the task of modeling text data.

Feeding the decoder with ground-truth words during training has two potential issues: (i) given the powerful recursive and autoregressive nature of LSTM decoders, the latent-variable model tends to ignore the latent vector altogether, thus reducing to a pure language model. (ii) the learned latent vector does not necessarily encode all the information needed to reconstruct the entire sequence, since additional guidance is provided while generating every word, i.e., exposure bias. (Shen et al., 2017a) proposed deconvolutional networks for the decoder (generator), in a latent-variable model, for matching natural language sentences. They jointly infer sentence representations by optimizing generative and discriminative objective: $L = -L_{VAE}(\theta, \phi, u) - L_{VAE}(\theta, \phi, v) + \alpha L_{task}(\psi, z_u, z_v, label(u, v))$.

Paraphrase Generation

(Gupta et al., 2017) uses a variational autoencoder (VAE) as a generative model for paraphrase generation. Their VAE-LSTM architecture is customized for the paraphrase generation task, where the training examples are given in form of pairs of sentences (original sentence and its paraphrased version). Both the encoder and decoder of their VAE-LSTM are conditioned on the original sentence which amounts to 4 LSTMs!

3 Model

3.1 Intuition

In Community Question Answering (cQA), there should be a single canonical question page for each logically distinct question/query. Behind duplicate questions, the intent is identical!

Our generative model considers that two duplicate questions or paraphrases were generated from a same latent intent (hidden variable). We thus take the VAE approach but instead of only recovering question q from q (REPEAT), we also let the model learn to recover q from q' (REFORMULATE). Our intuition is that such a model should learn to capture a question's intent, beyond its formulation.

Our work is closely related to SDAE. Instead of corrupting the input by removing words or swapping n-grams, we corrupt inputs by replacing them by paraphrases. Our model makes use of Neural Variational Inference to learn the posterior probability over intents as in (Bowman et al., 2015b).

For transfer to Semantic Question Matching, we infer latent intents of both questions and learn a classifier on the paired intents as in the siamese network framework.

3.2 Probabilistic Graph

Two questions are independent given their intent (markov chain).

3.3 Architecture details

We embed words in a 300 dimensional space, using Glove vectors as initialization. A single layer LSTM encodes sentences in fixed length vectors that are transformed in a hidden mean μ and diagonal variance σ vectors. The intent is sampled from the posterior $N(\mu, \sigma) = \mu + \sigma N(0, 1)$, using the reparameterization trick. The sampled intent is then used to condition a LSTM decoder network that generates a paraphrase, token after token with teacher forcing. The model is trained to minimize the regularized loss (= cross_entropy + KLD) by gradient descent (ADAM).

4 Results

Testing accuracy on the Quora duplicate question dataset,
* pLSA + MLP = 75% (ours)

* Siamese bi-LSTMs with max pooling and concatenation = 87% (ours)
* Siamese LSTM: 83.2% (Dadashov et al.,)
* Decomposable Attention: 87% (2016)
* Decomposable Attention (variant): 88% (2017)
* Multi-perspective matching model: 88% (2017)
* Interactive Inference Network (IIN): 89% (2017)

Acknowledgments

The acknowledgments should go immediately before the references. Do not number the acknowledgments section. Do not include this section when submitting your paper for review.

References

[Agirre et al.2014] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel M Cer, Mona T Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *SemEval@ COLING*, pages 81–91.

[Amiri et al.2016] Hadi Amiri, Philip Resnik, Jordan Boyd-Graber, and Hal Daumé III. 2016. Learning text pair similarity with context-sensitive autoencoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1882–1892.

[Arora et al.2016] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

[Bahdanau et al.2014] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

[Beaulieu et al.1997] M Beaulieu, M Gatford, Xiangji Huang, S Robertson, S Walker, and P Williams. 1997. Okapi at trec-5. *NIST SPECIAL PUBLICATION SP*, pages 143–166.

[Blei et al.2003] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

[Bowman et al.2015a] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015a. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*.

[Bowman et al.2015b] Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015b. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

[Chopra et al.2005] Sumit Chopra, Raia Hadsell, and Yann LeCun. 2005. Learning a similarity metric discriminatively, with application to face verification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 539–546. IEEE.

[Conneau et al.2017] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364*.

[Dadashov et al.] Elkhan Dadashov, Sukolsak Sakshuwong, and Katherine Yu. Quora question duplication.

[Dolan et al.2005] Bill Dolan, Chris Brockett, and Chris Quirk. 2005. Microsoft research paraphrase corpus. *Retrieved March*, 29:2008.

[Ganitkevitch et al.2013] Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. Ppdb: The paraphrase database. In *HLT-NAACL*, pages 758–764.

[Gong et al.2017] Yichen Gong, Heng Luo, and Jian Zhang. 2017. Natural language inference over interaction space. *arXiv preprint arXiv:1709.04348*.

[Gupta et al.2017] Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation. *arXiv preprint arXiv:1709.05074*.

[Hill et al.2016] Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016. Learning distributed representations of sentences from unlabelled data. *arXiv preprint arXiv:1602.03483*.

[Hofmann1999] Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

[Kingma and Welling2013] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

[Kiros et al.2015] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302.

[Kusner et al.2015] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966.

[Le and Mikolov2014] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 1188–1196.

[Li et al.2015] Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*.

[Lin et al.2017] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

[Liu et al.2016] Yang Liu, Chengjie Sun, Lei Lin, and Xiaolong Wang. 2016. Learning natural language inference using bidirectional lstm model and inner-attention. *arXiv preprint arXiv:1605.09090*.

[Ma et al.2015] Mingbo Ma, Liang Huang, Bing Xiang, and Bowen Zhou. 2015. Dependency-based convolutional neural networks for sentence embedding. *arXiv preprint arXiv:1507.01839*.

[Marelli et al.2014] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *LREC*, pages 216–223.

[Miao et al.2016] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International Conference on Machine Learning*, pages 1727–1736.

[Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[Munkhdalai and Yu2016] Tsendsuren Munkhdalai and Hong Yu. 2016. Neural semantic encoders. *arXiv preprint arXiv:1607.04315*.

[Nickel and Kiela2017] Maximilian Nickel and Douwe Kiela. 2017. Poincar\'e embeddings for learning hierarchical representations. *arXiv preprint arXiv:1705.08039*.

[Parikh et al.2016] Ankur P Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *arXiv preprint arXiv:1606.01933*.

[Pennington et al.2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

[Rezende et al.2014] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

[Sanborn and Skryzalin2015] Adrian Sanborn and Jacek Skryzalin. 2015. Deep learning for semantic similarity. *CS224d: Deep Learning for Natural Language Processing. Stanford, CA, USA: Stanford University*.

[Shen et al.2017a] Dinghan Shen, Yizhe Zhang, Ricardo Henao, Qinliang Su, and Lawrence Carin. 2017a. Deconvolutional latent-variable model for text sequence matching. *arXiv preprint arXiv:1709.07109*.

[Shen et al.2017b] Tao Shen, Tianyi Zhou, Guodong Long, Jing Jiang, Shirui Pan, and Chengqi Zhang. 2017b. Disan: Directional self-attention network for rnn/cnn-free language understanding. *arXiv preprint arXiv:1709.04696*.

[Tomar et al.2017] Gaurav Singh Tomar, Thyago Duque, Oscar Täckström, Jakob Uszkoreit, and Dipanjan Das. 2017. Neural paraphrase identification of questions with noisy pretraining. *arXiv preprint arXiv:1704.04565*.

[Vaswani et al.2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

[Wang et al.2017] Zhiguo Wang, Wael Hamza, and Radu Florian. 2017. Bilateral multi-perspective matching for natural language sentences. *arXiv preprint arXiv:1702.03814*.

[Williams et al.2017] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.