Principle-to-Program: Neural Methods for Similar Question Retrieval in Online Communities

Muthusamy Chelliah[1(✉)], Manish Shrivastava[2], and Jaidam Ram Tej[1]

[1] Flipkart, Bangalore, India
chelgeetha@yahoo.com
[2] IIIT Hyderabad, Hyderabad, India

Abstract. Similar question retrieval is a challenge due to lexical gap between query and candidates in archive and is very different from traditional IR methods for duplicate detection, paraphrase identification and semantic equivalence. This tutorial covers recent deep learning techniques which overcome feature engineering issues with existing approaches based on translation models and latent topics. Hands-on proposal thus will introduce each concept from end user (e.g., question-answer pairs) and technique (e.g., attention) perspectives, present state of the art methods and a walkthrough of programs executed on Jupyter notebook using real-world datasets demonstrating principles introduced.

Keywords: Question answering · Semantic similarity · Neural networks · Paraphrase identification · Duplicate detection

1 Introduction

Time lag between user posting a question and receiving its answer could be reduced by retrieving similar, historic questions from community question answering (cQA) archives. Two seemingly different questions may refer implicitly to a common problem with the same answer. Identifying semantic equivalence is thus critical for retrieving similar questions and to automate reusing answers available for such previous questions. This is a difficult task because different users may formulate the same question in a variety of ways, using different vocabulary (e.g., watersports vs. snorkeling) and structure. Similar questions hence vary in style, length and content quality (e.g., blood pressure vs. hypertension).

Lexical gap - different but related words between queried (e.g., knot) and existing (e.g., tangle) questions - rules out traditional IR models (e.g., BM25) as a solution. Text fragments in questions (e.g., disk full) could lead to correlated content (e.g., format) in answers. Similarity - different from relatedness (e.g., synonymy, antonymy) - has been addressed with strategies like machine translation, knowledge graphs and topic models though. Treating question-answer pairs

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 663–668, 2020.
https://doi.org/10.1007/978-3-030-45442-5_88

as parallel text, relationships can be established through translation probability (e.g., word-to-word) - asymmetry being a handicap. Latent topics aligned across question-answer pairs is another option - heterogeneity being an issue.

2 Background/Motivation

Title/body of a question is concatenated into problem definition and IR task is to decide if 2 such text fragments are similar. Detecting (almost) exact copies of the same document in corpora from Web crawling/search systems is not enough as equivalent questions may have very little (or no) overlap. Considering only surface form of a question without factoring in semantics makes it hard to identify duplicates. Keyword-based retrieval methods (e.g., language/vector-space models) hence predict questions with same meaning as different. Traditional similarity measures based on word overlap - even those on a graded scale from 0 to 5 - have thus proven to be inadequate for capturing semantic equivalence.

Paraphrase identification which helps determine if 2 sentences have the same meaning is not sufficient as well for similar question retrieval. Polysemy and word order are other challenges which similar question retrieval has to tackle like any other NLP tasks. Also, submitted questions have extraneous details - which obscures key information buried in the noise - in the body irrelevant to main question being asked. Title alone on the other hand lacks crucial detail present in question body. Building a large amount of training data with similar questions is expensive and careful feature engineering is time consuming. Deep learning techniques recently have been effective for sentence-level analysis of short texts in a variety of IR tasks. Size of a question in an online community however varies from a single sentence to detailed problem description with many sentences.

3 Convolutional Neural Networks (CNNs)/Question-Question Pairs

CNNs transform first words into embeddings with unlabeled data and then build distributed, vector representations for pairs of original and related questions. Questions are scored next with a metric (e.g., cosine similarity) and those pairs above a threshold based on a held-out set are considered equivalent. During training, CNN is induced to produce similar vector representations for equivalent questions. We discuss:

- [1] evaluates in-domain word embeddings vs. one trained with Wikipedia, estimates impact of training set size and evaluates aspects of domain adaptation,
- [2] combines bag of words (BoW) to retrieve equivalent questions while learning to rank them according to similarity with a loss function,

– [4] integrates sentence modeling and semantic matching into a single framework without syntactic analysis and prior knowledge (e.g., wordnet). Word tokens are converted into vectors by a lookup layer and useful information is captured with convolutional/pooling layers; finally, matching metric is learnt - better than traditional ones (e.g., inner-product, Euclidean distance) - between question/answer capturing their interaction with a tensor layer.

4 Recurrent Neural Networks (RNNs)/Similarity Features

Available annotations on similar questions however are noisy and fragmented. An encoder maps title/body combination of questions - treated as word sequences - into vector representation with a recurrent model. Complementary decoder is trained to reproduce title from noisy question body. We discuss:

– [6] incorporates adaptive gating in non-consecutive CNNs to focus temporal averaging on key pieces of questions. Training paradigm utilizes entire corpus of unannotated questions in a semi-supervised manner and fine tunes learning model discriminatively with limited annotations,
– [14,15] applies LSTM with attention to select entire sentences and subparts (word/chunk) from shallow syntactic trees towards question retrieval and tree kernels to filtered text representations exploiting implicit features of subtree space for learning question reranking.

5 Latent Space/Meta-data

User asking a question in cQA sites is required to choose a label from a predefined hierarchy of categories. This meta-data encodes attributes/properties of words from which similar words can be grouped according to categories. Language models represent words and question categories in a vector space and calculate question-question similarity with linear combinations of dot products of vectors - thus being heuristic on data or difficult to scale up. Each question is thus defined as a distribution which generates each word (embedding) independently and subsequently a kernel is used to assess question similarities. This design will require representation of words that belong to the same category to be close to each other thus benefiting embedding learning. We discuss:

– [3] learns variable-length word embeddings with category information and aggregates them into fixed-size vectors,
– [8] optimizes an objective which in turn applies a non-linear transformation considering only local-relatedness of words (i.e., category and small window in a question/associated answers),
– [12] outperforms text-based methods in misflagged duplicate detection with features like user authority, question quality and relational data between questions.

6 Representation Learning/Question-Answer Pairs

Due to relatively short text, question-question pairs have insignificant information to determine their relationship. To combat scarcity of similar question pairs for training, question-answer pairs from archives can be leveraged in a weakly supervised fashion without manual labeling. An added advantage of this approach is mapping simple terms used by novice askers (e.g., short sighted) to technical terms (e.g., myopia) and concepts (e.g., lasik/laser surgery, contact lens) used by expert answerers. We discuss:

- [5] learns shared parameters and similarity metric minimizing contrastive-loss energy function connecting twin networks,
- [7] preserves local neighborhood structure of and mirrors semantic similarity among question and answer spaces,
- [9] represents hierarchical structures of word and concept information with layer-by-layer composition and pooling leading to question embedding that captures semantics/syntax.

7 Attention/Constituent Matching

Essential constituents (e.g., destination) are those - name and value - important to meaning of the question (e.g., route). Units in a semantic parse can be leveraged to alleviate defining/labeling them in open domain. We discuss:

- [11] combines FrameNet with neural networks through ensemble and embedding approaches for question retrieval with constituent matching,
- [13] integrates shallow lexical mismatching information with initial rank by an external search engine to generate deep question representation with attention autoencoder,
- [10] leverages semantic information in paired answers while alleviating noise caused by adding answers with three heterogeneous attention mechanisms for modeling temporal interaction in a long sentence, capturing relevance between questions and relevance between answers and extracting knowledge from answers.

8 Conclusion

Distributed representations help tackle lexical gap in question retrieval as features based on word embeddings that enable similarity calculation through neural networks; gated convolutions map key question information from lengthy detail to semantic representations and LSTM with attention weights alleviates noise in syntactic structure selecting most significant parse tree fragments from question text.

Simultaneously embedding categories of questions into vector space helps model local relatedness of words in learning. Misflagging duplicate detection through user authority and question quality is more indicative of behavior problems (e.g., posting questions). Local linear embedding is leveraged to use collective corpus-level information for embedding historical question-answer pairs in a latent space without lexical correlation and separate topic/translation models. Attention encoders contain context information with focus on current word of input sequence thus avoiding bias towards sentence end.

Incorporating user ratings/reputation still remains unexplored. Semantic parsing techniques like abstract meaning representation is a future direction for essential constituent matching.

References

1. Bogdanova, D., dos Santos, C., Barbosa, L., Zadrozny, B.: Detecting semantically equivalent questions in online user forums. In: Proceedings of the Nineteenth Conference on Computational Natural Language Learning, pp. 123–131 (2015)
2. Dos Santos, C., Barbosa, L., Bogdanova, D., Zadrozny, B.: Learning hybrid representations to retrieve semantically equivalent questions. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing, pp. 694–699 (2015)
3. Zhou, G., He, T., Zhao, J., Hu, P.: Learning continuous word embedding with metadata for question retrieval in community question answering. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 250–259 (2015)
4. Qiu, X., Huang, X.: Convolutional neural tensor network architecture for community-based question answering. In: Twenty-Fourth IJCAI (2015)
5. Das, A., Yenala, H., Chinnakotla, M., Shrivastava, M.: Together we stand: siamese networks for similar question retrieval. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 378–387 (2016)
6. Lei, T., et al.: Semi-supervised question retrieval with gated convolutions. arXiv preprint arXiv:1512.05726 (2015)
7. Deepak, P., Garg, D., Shevade, S.: Latent space embedding for retrieval in question-answer archives. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 855–865 (2017)
8. Zhang, K., Wu, W., Wang, F., Zhou, M., Li, Z.: Learning distributed representations of data in community question answering for question retrieval. In: Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pp. 533–542. ACM (2016)
9. Wang, P., Zhang, Y., Ji, L., Yan, J., Jin, L.: Concept embedded convolutional semantic model for question retrieval. In: WSDM 2017 (2017)
10. Liang, D., et al.: Adaptive multi-attention network incorporating answer information for duplicate question detection (2019)
11. Zhang, X., Sun, X., Wang, H.: Duplicate question identification by integrating FrameNet with neural networks. In: 32nd AAAI Conference on Artificial Intelligence (2018)

12. Hoogeveen, D., Bennett, A., Li, Y., Verspoor, K.M., Baldwin, T.: Detecting misflagged duplicate questions in community question-answering archives. In: Twelfth International AAAI Conference on Web and Social Media (2018)
13. Zhang, M., Wu, Y.: An unsupervised model with attention autoencoders for question retrieval. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
14. Romeo, S., et al.: Neural attention for learning to rank questions in community question answering. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, pp. 1734–1745 (2016)
15. Barrórn-Cedeno, A., Da San Martino, G., Romeo, S., Moschitti, A.: Selecting sentences versus selecting tree constituents for automatic question ranking. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, pp. 2515–2525 (2016)