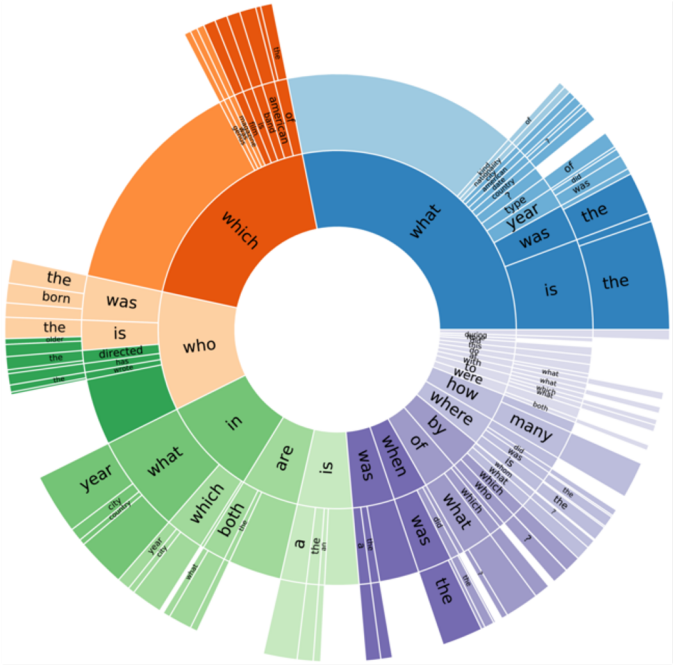
WassersteinQA

An Exploration of Generalization using Wasserstein distance

Matteo Donati Amir Ghomeshi Neil Leiser Zacharie Rozenberg



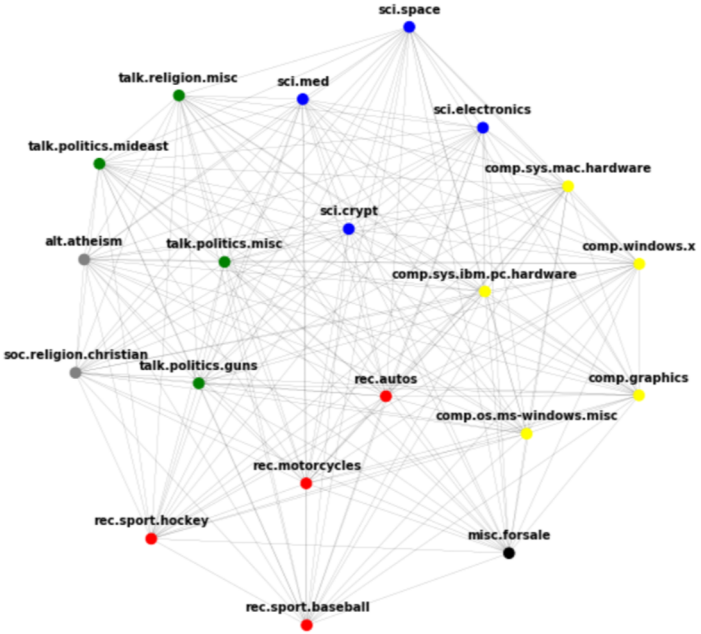
Background:

- Popularity of Optimal Transport (OT) in Machine Learning (Vision)
- Efficiency and potential of OT distances in NLP
- Alvarez-Melis and Fusi 2020, ...

Motivation :

- Applicability of Wasserstein distance to language (in particular : Q&A datasets)
- Use for improving generalization/transfer learning

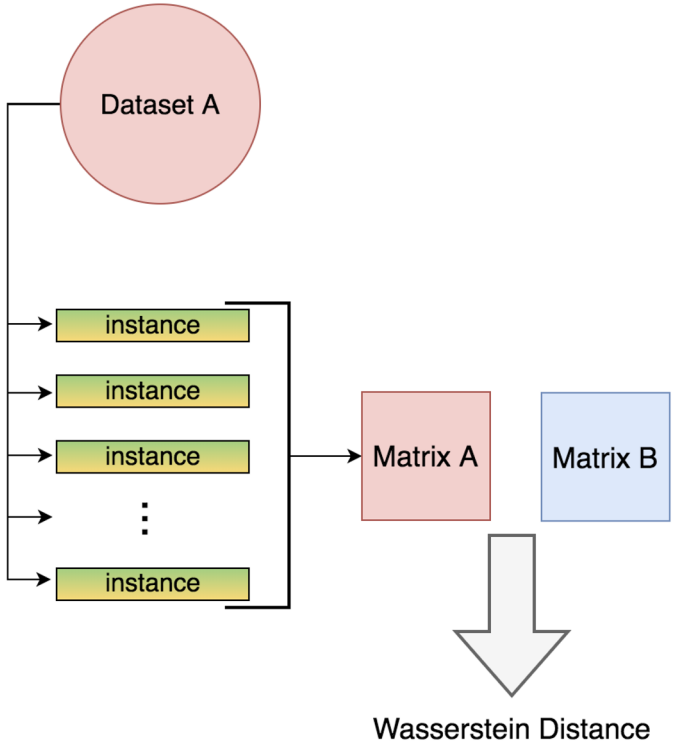
$$W_\epsilon(\mathbb{P}_x, \mathbb{P}_y) = \min_{\pi \in \Pi(\mathbb{P}_x, \mathbb{P}_y)} \mathbb{E}_{(x,y)\sim\pi} c(x,y) - \epsilon h(\pi)$$



Wasserstein distance - 20 NewsGroups Dataset

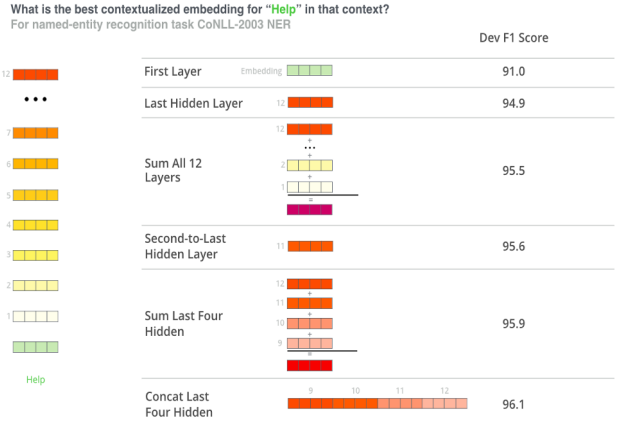
Dataset to distribution

- Objective : converting dataset to matrix
  - Embedding instances to vectors
- Issue : Q&A instance = context + question
- 2 different approaches :
  - Averaging contextual word embeddings (based on BERT)
  - Sentence embeddings (based on SBERT)



Approach 1 : Averaging contextual word embeddings

- Averaging word embeddings to represent document
- Embedding context and question together using BertForQuestionAnswering
- BERT provides good **contextual** embeddings
- Each instance (question + context) to 1024-vector

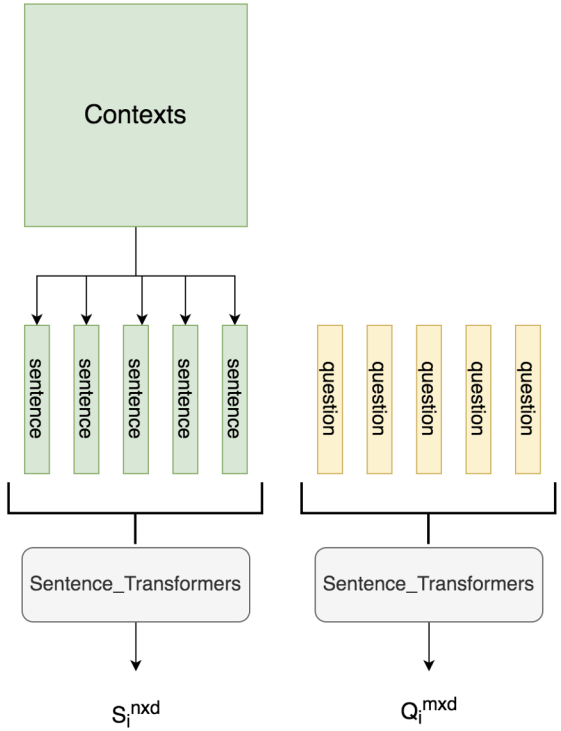


Andreas Poyiatzis, *"Extract contextualized word embeddings from BERT"*, 2019

Approach 2 : Sentence embeddings

- Split each dataset into contexts and question
- Split each context into sentences
- Feed questions and sentences into sentence_transformers
- Compute the Wasserstein distance between each dataset:

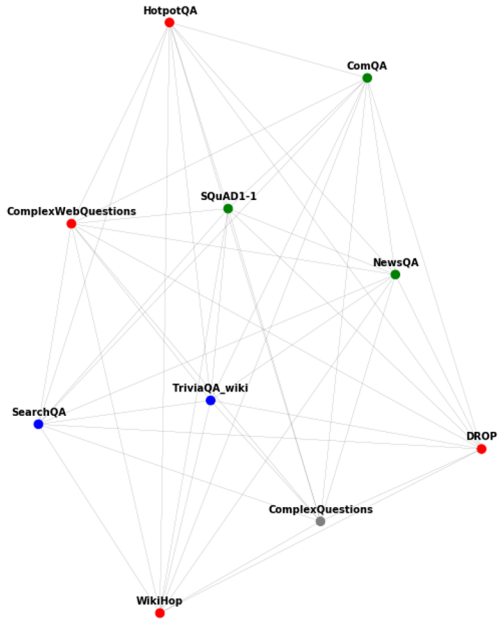
$$W_{i,j} = \frac{sinkhorn(S_i, S_j) + sinkhorn(Q_i, Q_j)}{2}$$



Experiments & Results

- Computing distances between 10 QA datasets
- Datasets sharing similar features tend to cluster
- For both approaches :

  Efficient computation time for datasets embedding + distance calculation



Wasserstein distance for generalization of Q&A tasks

- Comparing results with Talmor and Berant 2019
  - Lower Wasserstein distance = good generalization performance
  - Clear correlation for 50% of the datasets
  - Many possible research direction to improve results

	CQ	CWQ	ComQA	Wikihop	Drop	SQuAD	NewsQA	SearchQA	TriviaQA	HotpotQA
SQuAD	23.6	12	20	4.6	5.5	-	31.8	8.4	33.4	11.8
NewsQA	24.1	12.4	18.9	7.1	4.4	60.4	-	10.1	28.4	8.0
SearchQA	30.3	18.5	25.8	12.4	2.8	23.3	12.7	-	35.4	5.2
TriviaQA	30.3	16.5	23.6	12.6	5.1	35.5	19.4	27.8	-	8.7
HotpotQA	27.7	15.5	22.1	10.2	9.1	54.5	25.6	19.6	34.9	-

Talmor and Berant. *"An Empirical Investigation of Generalization and Transfer in Reading Comprehension"*, 2019

Minimal distance

Most successful generalization

Conclusion

- Promising approach for identifying Q&A datasets similarities efficiently
- Method can be related to generalization (zero-shot transfer learning)
- Several improvements possible:
    - Embedding approach
    - Considering labels (answers)
    - Modifying cost function of Wasserstein