Cross Lingual Sentence Retrieval

Gundeti Ashwanth Kumar - CS17BTECH11017
Chintha Sai Sreenivas - CS17BTECH11012

Problem Statement

Problem Statement

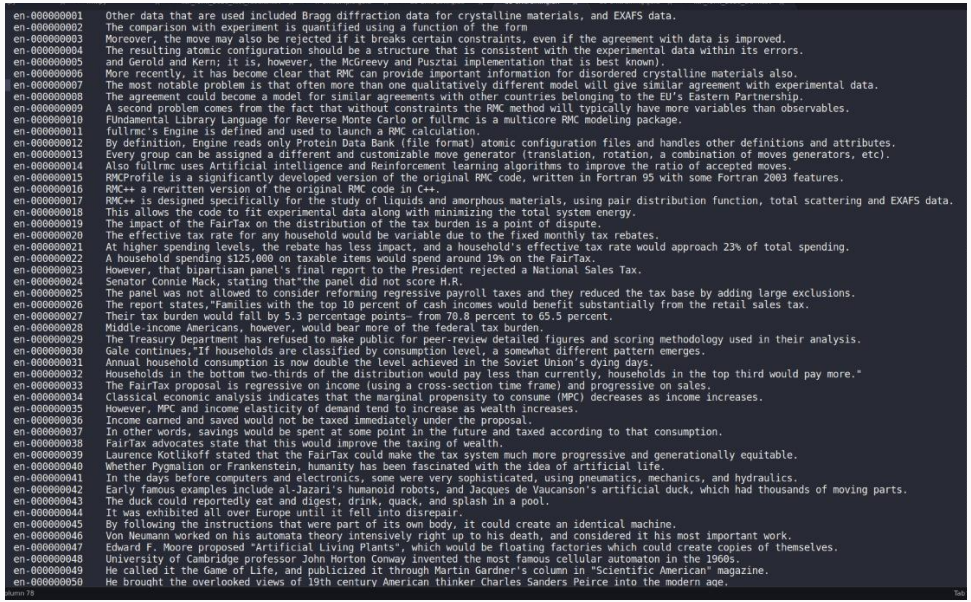
From a set of candidate sentences that are provided, our model has to rank the translations and pick the most relevant translation.

G. Ashwanth K

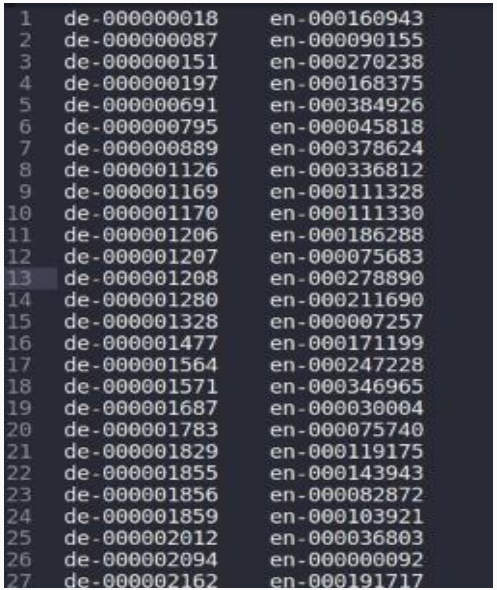
Description of Dataset

- Sentences from different languages - assigned an ID
- Contains the original mappings (ID to ID)
- For training process we use the parallel datasets.
- For testing we take a set of sentences and try to find the similarity score with other sentences
- Used 4 languages - en, fr, de, ru
- Training dataset consists of 50,000 pairs of sentences for each of en-fr, en-de, en-ru.
- We have found the cosine similarity on test dataset which consists of 2,000 pairs of sentences for each of en-fr, en-de, en-ru
- For calculating the F1-score, we used the BUCC dataset en-fr, en-de, en-ru. Each dataset consists of 10,000 sentences of each language and 1,000 gold standard sentences.

Snapshots of Dataset



```
en-000000001  Other data that are used included Bragg diffraction data for crystalline materials, and EXAFS data.
en-000000002  The comparison with experiment is quantified using a function of the form
en-000000003  Moreover, the move may also be rejected if it breaks certain constraints, even if the agreement with data is improved.
en-000000004  The resulting atomic configuration should be a structure that is consistent with the experimental data within its errors.
en-000000005  and Gerold and Kern; it is, however, the McGreevy and Pusztai implementation that is best known).
en-000000006  More recently, it has become clear that RMC can provide important information for disordered crystalline materials also.
en-000000007  The most notable problem is that often more than one qualitatively different model will give similar agreement with experimental data.
en-000000008  The agreement could become a model for similar agreements with other countries belonging to the EU's Eastern Partnership.
en-000000009  A second problem comes from the fact that without constraints the RMC method will typically have more variables than observables.
en-000000010  FUndamental Library Language for Reverse Monte Carlo or fullrmc is a multicore RMC modeling package.
en-000000011  fullrmc's Engine is defined and used to launch a RMC calculation.
en-000000012  By definition, Engine reads only Protein Data Bank (file format) atomic configuration files and handles other definitions and attributes.
en-000000013  Every group can be assigned a different and customizable move generator (translation, rotation, a combination of moves generators, etc).
en-000000014  Also fullrmc uses Artificial intelligence and Reinforcement learning algorithms to improve the ratio of accepted moves.
en-000000015  RMCProfile is a significantly developed version of the original RMC code, written in Fortran 95 with some Fortran 2003 features.
en-000000016  RMC++ a rewritten version of the original RMC code in C++.
en-000000017  RMC++ is designed specifically for the study of liquids and amorphous materials, using pair distribution function, total scattering and EXAFS data.
en-000000018  This allows the code to fit experimental data along with minimizing the total system energy.
en-000000019  The impact of the FairTax on the distribution of the tax burden is a point of dispute.
en-000000020  The effective tax rate for any household would be variable due to the fixed monthly tax rebates.
en-000000021  At higher spending levels, the rebate has less impact, and a household's effective tax rate would approach 23% of total spending.
en-000000022  A household spending $125,000 on taxable items would spend around 19% on the FairTax.
en-000000023  However, that bipartisan panel's final report to the President rejected a National Sales Tax.
en-000000024  Senator Connie Mack, stating that"the panel did not score H.R.
en-000000025  The panel was not allowed to consider reforming regressive payroll taxes and they reduced the tax base by adding large exclusions.
en-000000026  The report states,"Families with the top 10 percent of cash incomes would benefit substantially from the retail sales tax.
en-000000027  Their tax burden would fall by 5.3 percentage points— from 70.8 percent to 65.5 percent.
en-000000028  Middle-income Americans, however, would bear more of the federal tax burden.
en-000000029  The Treasury Department has refused to make public for peer-review detailed figures and scoring methodology used in their analysis.
en-000000030  Gale continues,"If households are classified by consumption level, a somewhat different pattern emerges.
en-000000031  Annual household consumption is now double the level achieved in the Soviet Union's dying days.
en-000000032  Households in the bottom two-thirds of the distribution would pay less than currently, households in the top third would pay more."
en-000000033  The FairTax proposal is regressive on income (using a cross-section time frame) and progressive on sales.
en-000000034  Classical economic analysis indicates that the marginal propensity to consume (MPC) decreases as income increases.
en-000000035  However, MPC and income elasticity of demand tend to increase as wealth increases.
en-000000036  Income earned and saved would not be taxed immediately under the proposal.
en-000000037  In other words, savings would be spent at some point in the future and taxed according to that consumption.
en-000000038  FairTax advocates state that this would improve the taxing of wealth.
en-000000039  Laurence Kotlikoff stated that the FairTax could make the tax system much more progressive and generationally equitable.
en-000000040  Whether Pygmalion or Frankenstein, humanity has been fascinated with the idea of artificial life.
en-000000041  In the days before computers and electronics, some were very sophisticated, using pneumatics, mechanics, and hydraulics.
en-000000042  Early famous examples include al-Jazari's humanoid robots, and Jacques de Vaucanson's artificial duck, which had thousands of moving parts.
en-000000043  The duck could reportedly eat and digest, drink, quack, and splash in a pool.
en-000000044  It was exhibited all over Europe until it fell into disrepair.
en-000000045  By following the instructions that were part of its own body, it could create an identical machine.
en-000000046  Von Neumann worked on his automata theory intensively right up to his death, and considered it his most important work.
en-000000047  Edward F. Moore proposed "Artificial Living Plants", which would be floating factories which could create copies of themselves.
en-000000048  University of Cambridge professor John Horton Conway invented the most famous cellular automaton in the 1960s.
en-000000049  He called it the Game of Life, and publicized it through Martin Gardner's column in "Scientific American" magazine.
en-000000050  He brought the overlooked views of 19th century American thinker Charles Sanders Peirce into the modern age.
```



1	de-000000018	en-000160943
2	de-000000087	en-000090155
3	de-000000151	en-000270238
4	de-000000197	en-000168375
5	de-000000691	en-000384926
6	de-000000795	en-000045818
7	de-000000889	en-000378624
8	de-000001126	en-000336812
9	de-000001169	en-000111328
10	de-000001170	en-000111330
11	de-000001206	en-000186288
12	de-000001207	en-000075683
13	de-000001208	en-000278890
14	de-000001280	en-000211690
15	de-000001328	en-000007257
16	de-000001477	en-000171199
17	de-000001564	en-000247228
18	de-000001571	en-000346965
19	de-000001687	en-000030004
20	de-000001783	en-000075740
21	de-000001829	en-000119175
22	de-000001855	en-000143943
23	de-000001856	en-000082872
24	de-000001859	en-000103921
25	de-000002012	en-000036803
26	de-000002094	en-000000092
27	de-000002162	en-000191717

Algorithm

Baseline Algorithm - Naive Approach

- We have used the **mBert** pre-trained model for finding the embeddings of the sentences.
- For the similarity value between two sentences we use the cosine-similarity metric.

  sQ, sC : query sentence and candidate sentence respectively

  eQ, eC : query embedding and candidate embedding respectively

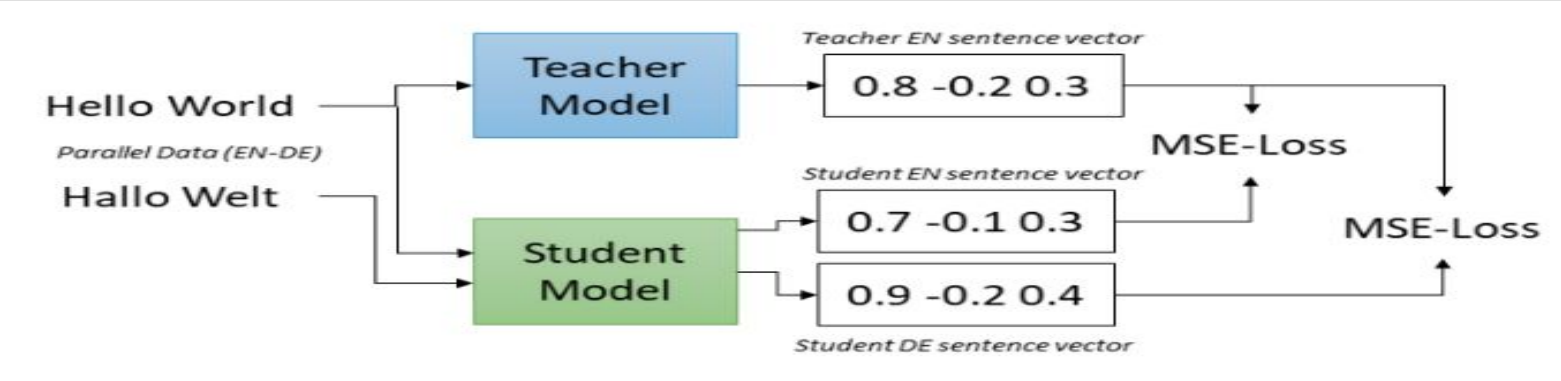
  similarity(sQ, sC) = cosine-similarity(eQ, eC)

- For every query sentence we will rank the candidate sentences based on the similarity values.

Step Forward

- Inspired from the paper - **Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation** , we have used this method as a next step.
- Main idea is that a translated sentence should be mapped to the same location in the vector space as the original sentence.
- We use the original (monolingual) model to generate sentence embeddings for the source language and then train a new model on translated sentences to mimic the original model.

Step Forward



- We use a teacher model M (monolingual) for a source language s.
- We need the set of translated sentences for s, let's say {t1,t2, ..ti,..}
- We train a student model M* such that

$$M^*(s) == M(s) \text{ and } M^*(ti) == M(s)$$

Step Forward

- We initialized
  - Teacher model with - **bert-base-nli-stsb-mean-tokens** (huggingface library) and english as a source language.
  - Student model with - **XLM-RoBERTa**
- We then tried :
  - Teacher Model : **bert-base-nli-stsb-mean-tokens**
  - Student Model : **xlm-r-100langs-bert-base-nli-mean-tokens**
- Used TED and BUCC dataset.

Evaluation and Results

Evaluation

- We compute this score for every pair and if this score is greater than a certain threshold we will mark that pair as matching pair.
- For the threshold we used the training set and tried 10 different values and picked out the value which results in maximum F1 score.

$$\text{score}(x, y) = \text{margin}(\cos(x, y), \sum_{z \in \text{NN}_k(x)} \frac{\cos(x, z)}{2k} + \sum_{z \in \text{NN}_k(y)} \frac{\cos(y, z)}{2k})$$

Results

- Mean Similarity : 0.91 over 1000 test sentences

```
French Sentence :  Et maintenant, je dois enlever mes chaussures pour monter à bord d'un avion !
English Sentence :  (Laughter) Now I have to take off my shoes or boots to get on an airplane!
Similarity :  0.88603795


======================


French Sentence :  --Rires Applaudissements-- Je vais vous raconter une petite histoire pour vous montrer ce que çà a été pour moi.
English Sentence :  (Laughter) (Applause) I'll tell you one quick story to illustrate what that's been like for me.
Similarity :  0.93396103


======================


French Sentence :  C'est une histoire vraie, dans tous ses détails.
English Sentence :  (Laughter) It's a true story -- every bit of this is true.
Similarity :  0.9343486


======================


French Sentence :  Après que Tipper et moi avons quitté la --Faux sanglot-- Maison Blanche --Rires-- nous étions en route pour une petite ferme que nous avons à 80 km à l'est de Nashville --
English Sentence :  Soon after Tipper and I left the -- (Mock sob) White House -- (Laughter) we were driving from our home in Nashville to a little farm we have 50 miles east of Nashville.
Similarity :  0.9595846


======================


French Sentence :  conduisant nous-mêmes.
English Sentence :  Driving ourselves.
Similarity :  0.7773765
```

Results : Cosine Similarity

Translation type	Average Similarity
EN-DE	0.914
EN-FR	0.919
EN-RU	0.876

Results : F1 Score

Model	Similarity Threshold	EN-DE	EN-FR	EN-RU
bert ← xlm-r-b	0.8	0.649	0.643	0.579
bert←xlm-r-100l-b	0.75	0.853	0.862	0.748
bert←xlm-r-100l-b	0.8	0.939	0.933	0.811