Cross-Lingual Information Retrieval

Information Retrieval and Web Search

Minh Duc Bui   Jakob Langenbahn   Niklas Sabel

June 14, 2021

Team 4, University of Mannheim

Outline

1. Task & Data Description
2. Unsupervised Ranking
3. Supervised Ranking
4. Evaluation and Zero-Shot Performance
5. Discussion

Task & Data Description

Introduction: Translation Retrieval

Query (EN)	Corpus (DE)
A sentence example	Guten Tag.
	Mannheim ist cool.
	Ein Satzbeispiel.
	⋮

- Task: Given a query in one language, recognize its translation from a large collection of sentences in another language

Solutions: Concepts

Unsupervised Solution:

- Represent query and document corpus in the same embedding space by inducing a cross-lingual word embedding
- Rank according to the similarity of query and documents (e.g. cosine similarity)

Supervised Solution:

- Train a classifier on a binary sentence pair translation task
- Use the trained model and calculate the confidence for each query and document pair
- Rank according to confidence

Europarl: Extract Train, Validation and Testset

Train Set (Binary):

- 220,000 sentence pairs from the EN-DE corpus
- 20,000 pairs are correct translations
- 200,000 pairs are wrong translations

Validation Set (Ranking):

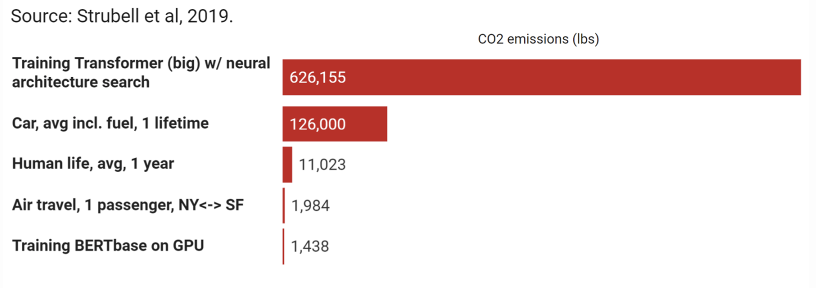
- EN-DE corpus
- Query collection size: 100
- Document collection size: 5,000

Test Set (Ranking):

- EN-DE, EN-PL, EN-IT corpus
- Same sizes as the validation set
- All sentences are unseen during training and validation

Motivation: Are Transformers necessary?

- Can we beat complex, computational expensive transformer models with cheap traditional machine learning approaches?



- Idea: Use text-based feature and extract features from a cross-lingual word embedding and feed them into a simple machine learning model

Goal: Beat the Transformer!

XLM-R Downsampling:

- Downsample training set to the minority class
- Randomly sample to create batches
- ... $\sim$ 220M parameters!

XLM-R Weighted:

- Use weighted loss to combat class imbalance:

$$\mathscr{L} = \frac{1}{2}\left(\mathscr{L}_{+} + \mathscr{L}_{-}\right),$$

- Each batch consists of the translation and the negative translations for each source sentence
- ... $\sim$ 220M parameters!

Unsupervised Ranking

Inducing Cross-Lingual Word Embeddings

Projection-based Methods (used in supervised models)

- `PROC(∼5k)`: Align monolingual word embeddings by solving the Procrustes problem
- `PROC-B(∼1k)`: Augment initial dictionary and solve procrustes problem
- `VecMap`: Build the seed dictionary in unsupervised fashion

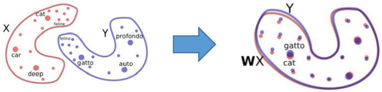


Image from [Conneau et al., '18]

XLM-R Multilingual Embedding (not used in supervised models)

- `XLM-R(Layer 1)`
- `XLM-R(Layer 12)`

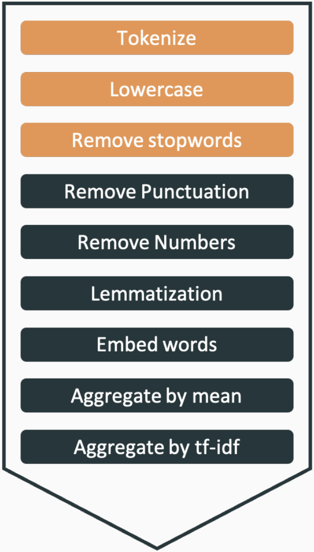
Model Performance Comparison

- Performance of unsupervised models on validation set (EN-DE)
- Jaccard Coefficient of direct translation significantly outperforms other methods!

Unsupervised	Similarity	Aggregating	Final
PROC(∼5k)	COS	AVG	0.4833
PROC(∼5k)	COS	TFIDF	0.5509
PROC(∼5k)	Jaccard	-	0.7515
PROC-B(∼0.5k)	COS	AVG	0.4417
PROC-B(∼0.5k)	COS	TFIDF	0.5309
PROC-B(∼0.5k)	Jaccard	-	0.7376
VecMap	COS	AVG	0.5721
VecMap	COS	TFIDF	0.6234
VecMap	Jaccard	-	0.7366
XLM-R (Layer 1)	COS	AVG	0.0355
XLM-R (Layer 12)	COS	AVG	0.0060

Supervised Ranking

Sentence Based Features



Sentence Based Features

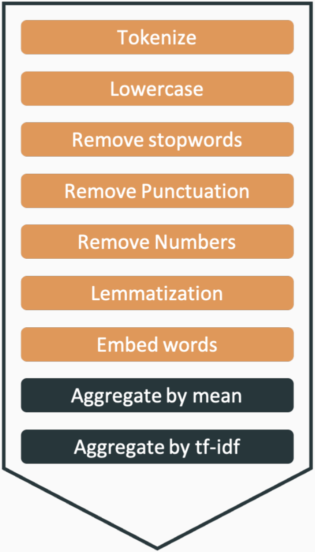
Absolute and relative comparison of

- Number of (unique) words, characters, punctuation marks
- Number of POS Tags and verb tenses
- Jaccard coefficient of named numbers in sentence

Embedding Based Features

- Jaccard coefficient of direct translations
- Euclidean distance of sentence embeddings
- Cosine Similarity of sentence embeddings

Word Embedding Based Features



Sentence Based Features

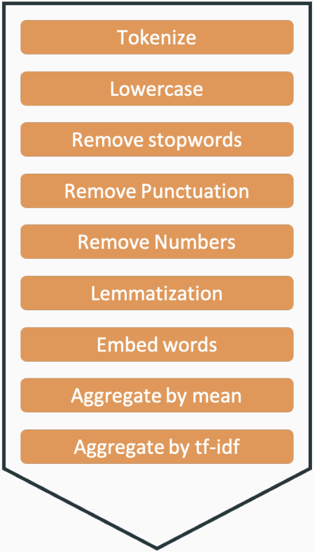
Absolute and relative comparison of

- Number of (unique) words, characters, punctuation marks
- Number of POS Tags and verb tenses
- Jaccard coefficient of named numbers in sentence

Embedding Based Features

- Jaccard coefficient of direct translations.
- Euclidean distance of sentence embeddings
- Cosine Similarity of sentence embeddings

Sentence Embedding Based Features



Sentence Based Features

Absolute and relative comparison of

- Number of (unique) words, characters, punctuation marks
- Number of POS Tags and verb tenses
- Jaccard coefficient of named numbers in sentence

Embedding Based Features

- Jaccard coefficient of direct translations.
- Euclidean distance of sentence embeddings
- Cosine Similarity of sentence embeddings

Model Performance Comparison

- Run forward feature selection and grid search with the validation set

	All features	Forward selection	Final
Naive Bayes	0.3244	0.8068	0.8068
Logistic Regression	0.6661	0.8321	0.8323
XGBoost	0.7100	0.8330	0.8357
MLPClassifier	0.6198	0.8477	0.8477

- Feature subset sizes

	All features	Final	Embedding-based	Text-based
Naive Bayes	97	12	4	8
Logistic Regression	97	14	7	7
XGBoost	97	8	3	5
MLPClassifier	97	9	5	4

Evaluation and Zero-Shot Performance

Evaluation and Zero-Shot Performance

- Performance on different languages in sentence-level tasks

	EN-DE	EN-IT	EN-PL	Average all
VecMap + Jaccard	0.7778	0.7945	0.7752	0.7825
Naive Bayes	0.8128	0.7947	0.8242	0.8106
Logistic Regression	0.8367	0.8311	0.8381	0.8353
XGBoost	0.8431	0.8686	0.8665	0.8594
MLPClassifier	0.8459	0.8725	0.8691	0.8625
XLM-R Downsampling	0.9287	0.8849	0.9235	0.9124
XLM-R Weighted	0.9351	0.9040	0.9155	0.9182

- Performance on document-level task

	Final models
VecMap + Jaccard	0.1181
Naive Bayes	0.0003
Logistic Regression	0.0079
XGBoost	0.0022
MLPClassifier	0.0004
XLM-R Weighted	0.0004

Discussion

Discussion

Observations

- Simple approaches perform reasonable well, however are still lacking behind Transformer models
- Zero-shot transfer into different languages possible
- Poor performance on document-level task.

Future Work

- Exhaustive search of hyperparameter, especially MLPClassifier.
- Zero-shot performances for more distant language pairs.
- Use of more sophisticated distance measures.