Cross-lingual supervised text classification

Hauke Licht

Cologne Center for Comparative Politics
hauke.licht@wiso.uni-koeln.de

Tutorial prepared for COMPTEXT 2022

May 5, 2022

Background

Text-as-data

Goal

- study political behavior and communication using quantitative methods
- ultimately, answer substantively interesting research questions about politics

Premise

- text generation is political $\rightsquigarrow$ text is important artifact of political behavior
- texts indicates their authors' political preferences, attitudes and beliefs, and strategies

Cross-lingual quantitative text analysis

Study political behavior and communication across languages

- multilingual institutional contexts (e.g., EP or UN)
- overcome language barriers to text-as-data applications in *comparative* research

The central challenge

> *develop an alignment between the conceptual representations of the model across languages so that we know a particular scaling, topic, or class in one language is comparable with the representation in another language.* (Lucas el al. 2015, 259)

$\rightsquigarrow$ obtain **identical measurements** for documents that indicate the **same concept**

Cross-lingual quantitative text analysis

The central challenge

obtain **identical measurements** for documents that indicate the **same concept**
$\rightsquigarrow$ even if they are written **in different languages**

Why this is challenging

Tower of Babel problem (Chan el al. 2020, Maier el al. 2021)
different vocabularies, words do not co-occur, similar words have dissimilar contexts
$\rightsquigarrow$ makes language-independent inference hard

Cross-lingual measurement equivalence
the substance of political discourses varies across polito-linguistic contexts
$\rightsquigarrow$ might necessitate to account for context

Cross-lingual supervised text classification

Text-as-data approaches and tasks

Manual content analysis

- text classification (coding/categorization)
- text scaling (through pairwise comparison)

Quantitative text analysis

Text classification

- dictionary analysis
- **supervised text classification**
- topic modeling

Text scaling

- unsupervised (*Wordfish*, *Wordshoal*)
- semi-supervised (*Wordscores*, LSA)

Cross-lingual supervised text classification

Text classification

Assign each document in a corpus to a one of several pre-defined categories

Supervised text classification

"Learn" how to assign documents to categories based on a (small) subset of documents for which you know which categories they belong to

How? apply supervised machine learning techniques

Cross-lingual supervised text classification

"Learn" to assign documents to categories when they are written in different languages

Today's running example

Lehmann and Zobel (2018) data set

- corpus of human-coded election manifestos
- quasi-sentences coded into **issue categories**: immigration, integration, "others"
- manifestos of parties from 14 countries $\rightsquigarrow$ **8 languages**

$\rightsquigarrow$ view the data set **online**

Our goal

discriminate between the **immigration/integration** and the "others" category

Relevance

- study competition on the immigration/integration issue
- learn to extrapolate (expensive) human codings
  - to new countries (cross-lingual transfer)
  - to new domain (à la Osnabrügge, Ash and Morelli 2021)

Approaches

How can we apply supervised text classification to a multilingual text corpus?

Separate analysis

1. split multilingual corpus by language
2. train separate, language-specific classifiers
3. apply them to classify unlabeled documents

$\leadsto$ pool resulting measurements for cross-lingual comparison

Input alignment ← our focus today

1. transfer documents' representations to a "common denominator" (Lind et al. 2021)
2. train a single, cross-lingual classifier
3. apply it to classify unlabeled documents

$\leadsto$ use resulting measurements in cross-lingual comparison

Input alignment

A unifying idea

- transfers documents to a common denominator (i.e., "align" them)
- enable their joint analysis within a single model (Lucas et al. 2015)
  - direct cross-lingual comparison
  - information-sharing across languages
  - resource-efficient

How to achieve this

Machine translation

Translate all documents to a single **target language**

Multilingual embedding

Represent all documents in a joint, multilingual **embedding space**

Input alignment approaches

Machine translation

Idea

- overcome Babel problem by translating all documents to one target language
- the target language is the "common denominator"

Evidence

works quite very well for typical quantitative text analysis tasks

- topic modeling (Lucas el al. 2015; de Vries, Schoonvelde and Schumacher 2018; Reber 2018)
- dictionary analysis (Windsor, Cupit and Windsor 2019)
- supervised classification (Courtney el al. 2020)
- textual similarity (Düpont ant Rachuj 2021)

Implementation (I)

Machine translation

Human translators too expensive. Instead, rely on state-of-the-art **neural machine translation** (NMT) methods.

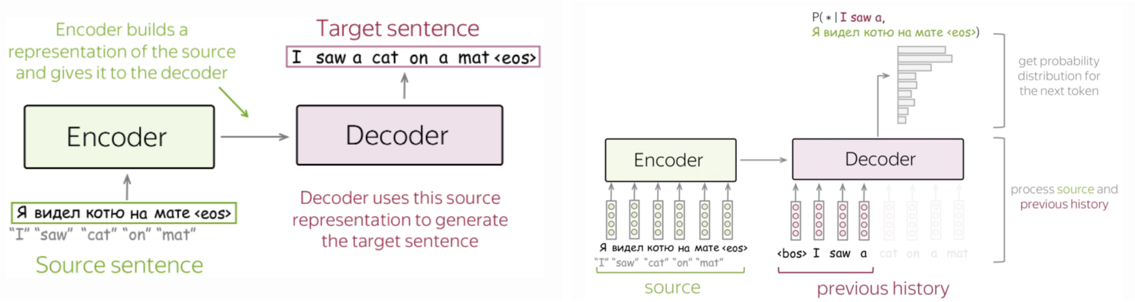


Figure 1: Translation as seq-to-seq problem. *Source:* Lena Voita's "NLP Course" (**online**)

Implementation (II)

Machine translation

Human translators too expensive. Instead, rely on state-of-the-art **neural machine translation** (NMT) methods.

Approaches

Using commercial service

- state-of-the-art NMT technology (e.g., *Google Translate* or *DeepL*)
- extensively evaluated in PolSci/CommSci literature
- "black box" (cf. Chan el al. 2020)
- quite expensive

Use open-source NMT model ← our approach today

- open-source ⇝ reproducible and free of charge
- massively pre-trained (e.g., Fan et al.'s M2M)

Free machine translation

pip install easynmt

The easyNMT python package provides a simple interface to download and use several large pre-trained NMT models:

```
from easynmt import easyNMT

# download and instantiate a pre-trained M2M model
model = easyNMT("m2m_100_418M")

# translate a single sentence
model.translate("Guten Tag liebe Freunde!", target_lang="en")
```

⇝ see **this** Colab Notebook for an illustration

Multilingual embedding

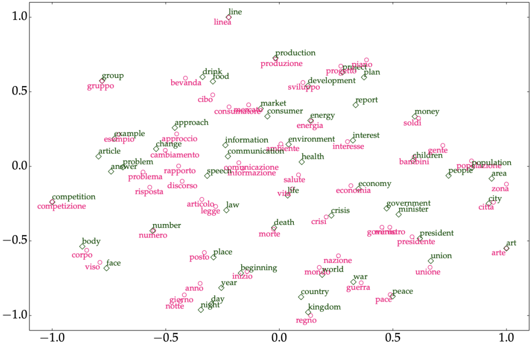


Figure 2: Low-dimensional representation of aligned English and Spanish word embeddings (Figure 1 in Ruder, Vulić and Søgaard 2019)

Idea

align documents by representing them in a multilingual embedding space

Existing applications

- topic modeling (Chan el al. 2020)
- text scaling (Glavaš, Nanni and Ponzetto 2017b; Goist 2021)
- textual semantic similarity (Radford, Dai and Golder 2021)
- supervised classification (Glavaš, Nanni and Ponzetto 2017a; Dai and Radford 2019; Licht 2022)

Approaches

Multilingual *word* embedding

- many different approaches (see Ruder, Vulić and Søgaard 2019)
- currently most common among ME-based contributions (e.g. Chan el al. 2020; Goist 2021)

Multilingual *sentence* embedding ← our focus today

- again different approaches (next slide)
- well-suited for applications when documents have sentence-like lengths
  ⇝ **supervised text classification** (Licht 2022)

Multilingual sentence embedding (I)

LASER

encoder–decoder model learns fixed-size embedding layer (Artexte and Schwenk 2019)
⇝ essentially like NMT architecture

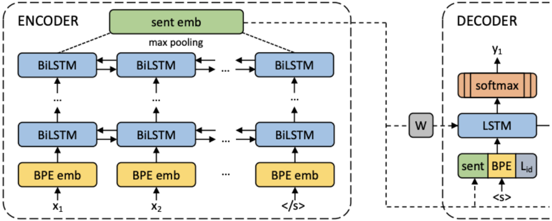


Figure 3: In Artexte and Schwenk (2019)

Knowledge distillation

extend pre-trained monolingual sentence embedding model to new languages (Reimers and Gurevych 2020)

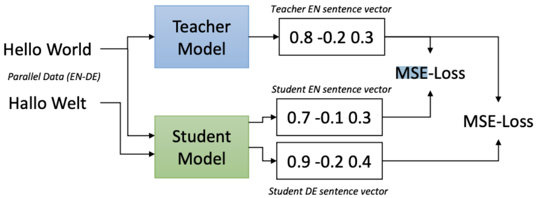


Figure 4: In Reimers and Gurevych (2020)

Multilingual sentence embedding (II)

pip install sentence-transformers

The sentence-transformers python package provides a simple interface to download and use several large pre-trained MSE models:

```
from sentence_transformers import SentenceTransformer

# download and instantiate a knowledge-distilled XLM-R model
model = SentenceTransformer("paraphrase-xlm-r-multilingual-v1")

# translate a single sentence
model.encode("Guten Tag liebe Freunde!")
```

⇝ see **this** Colab Notebook for an illustration

Application in cross-lingual supervised text classification

Supervised classification: recap (I)

Terminology

- **classes**: the set of outcome categories
- **label**: the class a document belongs to
- **sample**: a single document and its label
- **data set**: a collection of samples

```
   | label | text                                 |
   |-------|--------------------------------------|
-> |   +   | "Our government is doing a great job!"     | <-- a sample
   |   -   | "Those crooks in power are super corrupt!" |
   |  ...  |                                      |
```

Supervised classification: recap (II)

Training & Evaluation

- **training**: fit a model to a data set to optimize its classification accuracy
- **held-out samples**: samples not used to train a classifier
- **evaluation**: see how a classifier performs in held-out samples ("out of sample")
- ***k*-fold cross validation** (CV): train and evaluate a classifier on k partitions of a training data set
  $\rightsquigarrow$ estimate out-of-samples performance

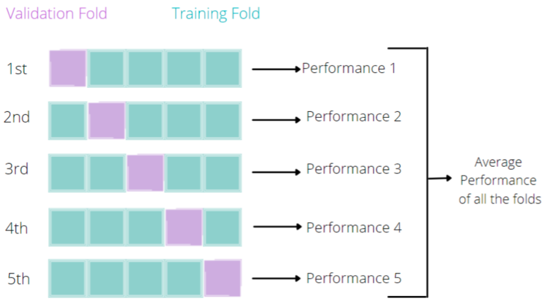


Figure 5: Cross validation procedure (source)

Cross-lingual supervised text classification

MT approach

1. **machine-translate** all documents into the target language
2. train classifier using **documents' target-language representations**
3. evaluate classifier on held-out documents

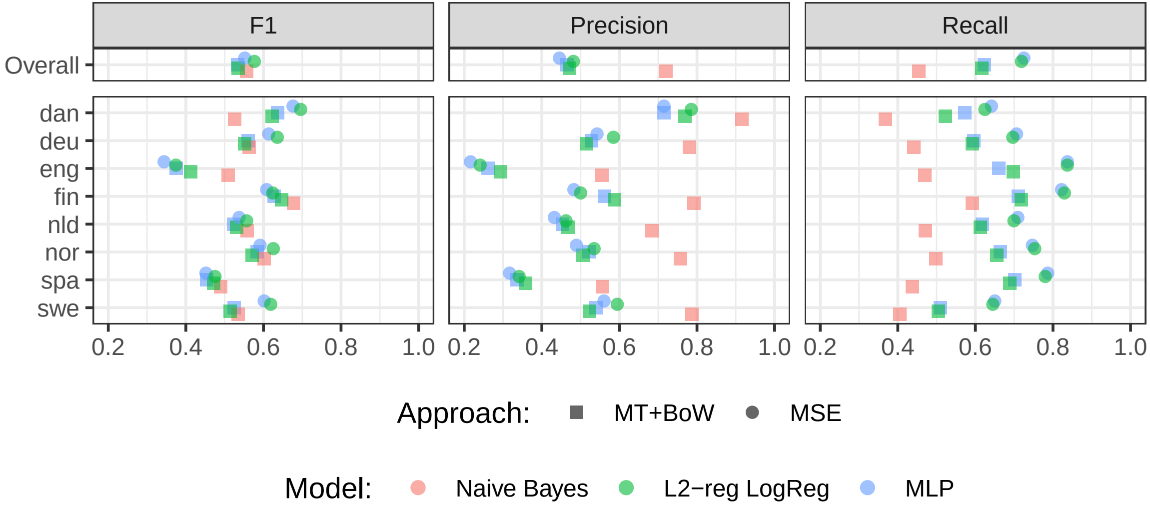
⇝ see **this notebook** for an illustration

MSE approach

1. **embed** all documents using a pre-trained MSE model
2. train classifier on **documents' embeddings**
3. evaluate classifier on held-out documents

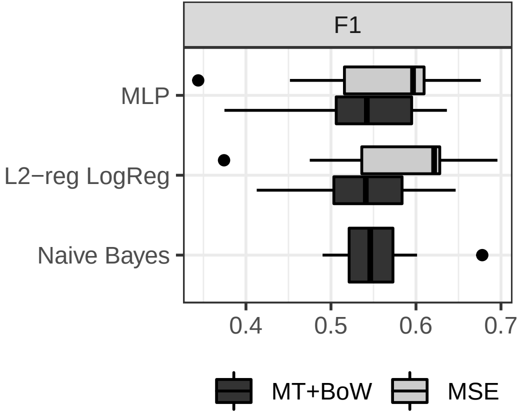
⇝ see **this notebook** for an illustration

Results



F1
Precision
Recall
Overall
dan
deu
eng
fin
nld
nor
spa
swe
0.2
0.4
0.6
0.8
1.0
Approach:
MT+BoW
MSE
Model:
Naive Bayes
L2−reg LogReg
MLP

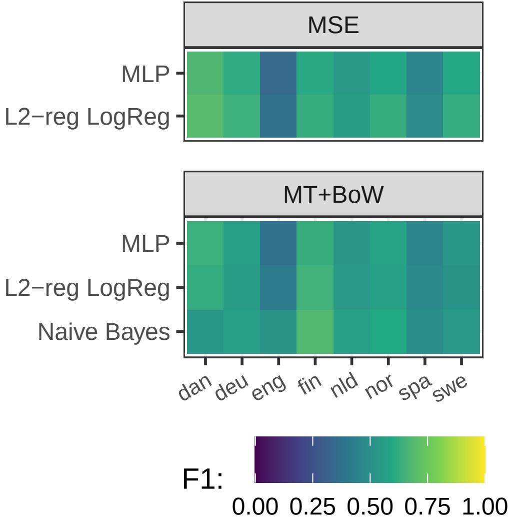
Results



Language-specific F1 scores

- MSE approach "better" overall
- much variation in F1 scores for all classifiers but the Naive Bayes
- but the Naive Bayes classifier massively overshoots (see prev. slide)

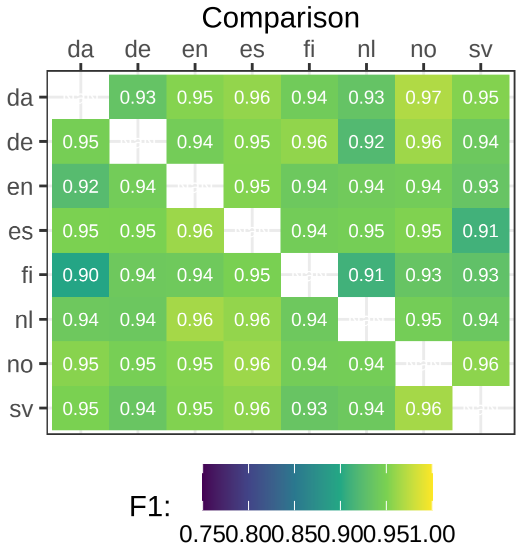
Results



Language-specific F1 scores

- both approaches systematically under-perform in English and perform rel. well in Finnish
  - Naive Bayes classifier is exception
- compare to prevalence of positive samples in corpus (overall 0.039):
  - English: 0.014 (lowest)
  - Spanish: 0.024 (2nd lowest)
  - Finnish: 0.034
  - Danish: 0.094 (highest)

Results



Language-independent measurement?

One potential way to assess language-independence:

1. take set of *parallel* texts
2. translate/embed them
3. predict their labels
4. compare consistency

Example: Figure shows results for MSE-based classifier (**code**)

Room for improvement

Class imbalance

- we have artificially down-sampled negative samples
  - label distribution in training and test data differs
  - discards relevant info
- maybe better: make misclassifying positive samples more costly $\rightsquigarrow$ use **class weights**

Text representations

- BoW representations are very sparse $\rightsquigarrow$ use pre-trained (English) word embeddings
- MSEs are "frozen" (not specialized for the classification task)
  $\rightsquigarrow$ fine-tune multilingual Transformer (e.g., mBERT, XLM-R)

Open issues

- How to account for fact that prevalence varies across countries?
- How to asses cross-lingual measurement equivalence?

Resources

Links to Google Colab notebooks

- inspect the Lehmann and Zobel (2019) data set (**link**)
- *input alignment*
  - how to machine-translate with the `easyNMT` package (**link**)
  - how to sentence-embed with the `sentence-transformers` package (**link**)
- *supervised classification*
  - sample the train, CV, and test indices (**link**)
  - train MT+BoW classifiers (**link**)
  - train MSE-based classifiers (**link**)

Data

- the cleaned Lehmann and Zobel data set (incl. machine-translated texts, **link**)
  - XLM-R sentence embeddings (**link**)
- the train, CV, and test indeces configuration JSON (**link**)

Back-up slides

References I

Artetxe, Mikel and Holger Schwenk (2019). "Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond". In: *Transactions of the Association for Computational Linguistics* 7, pp. 597–610. DOI: 10.1162/tacl_a_00288.

Chan, Chung-Hong et al. (2020). "Reproducible Extraction of Cross-Lingual Topics (Rectr)". In: *Communication Methods and Measures* 14.4, pp. 285–305. DOI: 10.1080/19312458.2020.1812555.

Courtney, Michael et al. (2020). "Automatic Translation, Context, and Supervised Learning in Comparative Politics". In: *Journal of Information Technology & Politics* 17.3, pp. 208–217. DOI: 10.1080/19331681.2020.1731245.

Dai, Yaoyao and Benjamin J. Radford (2019). "Multilingual Word Embedding for Zero-Shot Text Classification". URL: https://yaoyaodai.github.io/files/Dai_0BlinC.pdf.

References II

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher (2018). "No Longer Lost in Translation: Evidence That Google Translate Works for Comparative Bag-of-Words Text Applications". In: *Political Analysis* 26.4, pp. 417–430. DOI: 10.1017/pan.2018.26.

Düpont, Nils and Martin Rachuj (2022). "The Ties That Bind: Text Similarities and Conditional Diffusion among Parties". In: *British Journal of Political Science* 52.2, pp. 613–630. DOI: 10.1017/S0007123420000617.

Fan, Angela et al. (2020). "Beyond English-Centric Multilingual Machine Translation". URL: http://arxiv.org/abs/2010.11125.

Glavaš, Goran, Federico Nanni, and Simone Paolo Ponzetto (2017a). "Cross-Lingual Classification of Topics in Political Texts". In: *Proceedings of the Second Workshop on NLP and Computational Social Science*, pp. 42–46. DOI: 10.18653/v1/W17-2906.

References III

Glavaš, Goran, Federico Nanni, and Simone Paolo Ponzetto (2017b). "Unsupervised Cross-Lingual Scaling of Political Texts". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 688–693. URL: https://www.aclweb.org/anthology/E17-2109.

Goist, Mitchell (2020). "The Radical Right in Parliament: A New Method and Application for Studying Political Text in Multiple Languages". PhD thesis. URL: https://etda.libraries.psu.edu/catalog/17825mlg307.

Hillard, Dustin, Stephen Purpura, and John Wilkerson (2008). "Computer-Assisted Topic Classification for Mixed-Methods Social Science Research". In: *Journal of Information Technology & Politics* 4.4, pp. 31–46. DOI: 10.1080/19331680801975367.

Lehmann, Pola and Malisa Zobel (2018). "Positions and Saliency of Immigration in Party Manifestos: A Novel Dataset Using Crowd Coding". In: *European Journal of Political Research* 57.4, pp. 1056–1083.

References IV

Licht, Hauke (2022). "Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings". URL: https://osf.io/384wr/?view_only=abcfb31cada64dbca8f7b43a59b1e696.

Lind, Fabienne et al. (2021). "Building the Bridge: Topic Modeling for Comparative Research". In: *Communication Methods and Measures* First View, pp. 1–19. DOI: 10.1080/19312458.2021.1965973.

Lucas, Christopher et al. (2015). "Computer-Assisted Text Analysis for Comparative Politics". In: 23.2, pp. 254–277. DOI: 10.1093/pan/mpu019.

Osnabrügge, Moritz, Elliott Ash, and Massimo Morelli (2021). "Cross-Domain Topic Classification for Political Texts". In: *Political Analysis* First view, pp. 1–22. DOI: 10.1017/pan.2021.37.

Radford, Benjamin J., Yaoyao Dai, and Matt Golder (2021). "Attributing Policy Influence in Multilingual Setting Using Semantic Textual Similarity". In: 2021 Annual Meeting of the American Political Science Asociation (APSA).

References V

Reber, Ueli (2018). “Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora”. In: *Communication Methods and Measures* 13.2, pp. 102–125. DOI: 10.1080/19312458.2018.1555798.

Reimers, Nils and Iryna Gurevych (2020). “Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Empirical Methods in Natural Language Processing (EMNLP), pp. 4512–4525. DOI: 10.18653/v1/2020.emnlp-main.365.

Ruder, Sebastian, Ivan Vulić, and Anders Søgaard (2019). “A Survey Of Cross-lingual Word Embedding Models”. In: *Journal of Artificial Intelligence Research* 65, pp. 569–631. DOI: 10.1613/jair.1.11640. arXiv: 1706.04902.

References VI

Windsor, Leah Cathryn, James Grayson Cupit, and Alistair James Windsor (2019). "Automated Content Analysis across Six Languages". In: *PloS One* 14.11, e0224425. DOI: 10.1371/journal.pone.0224425.