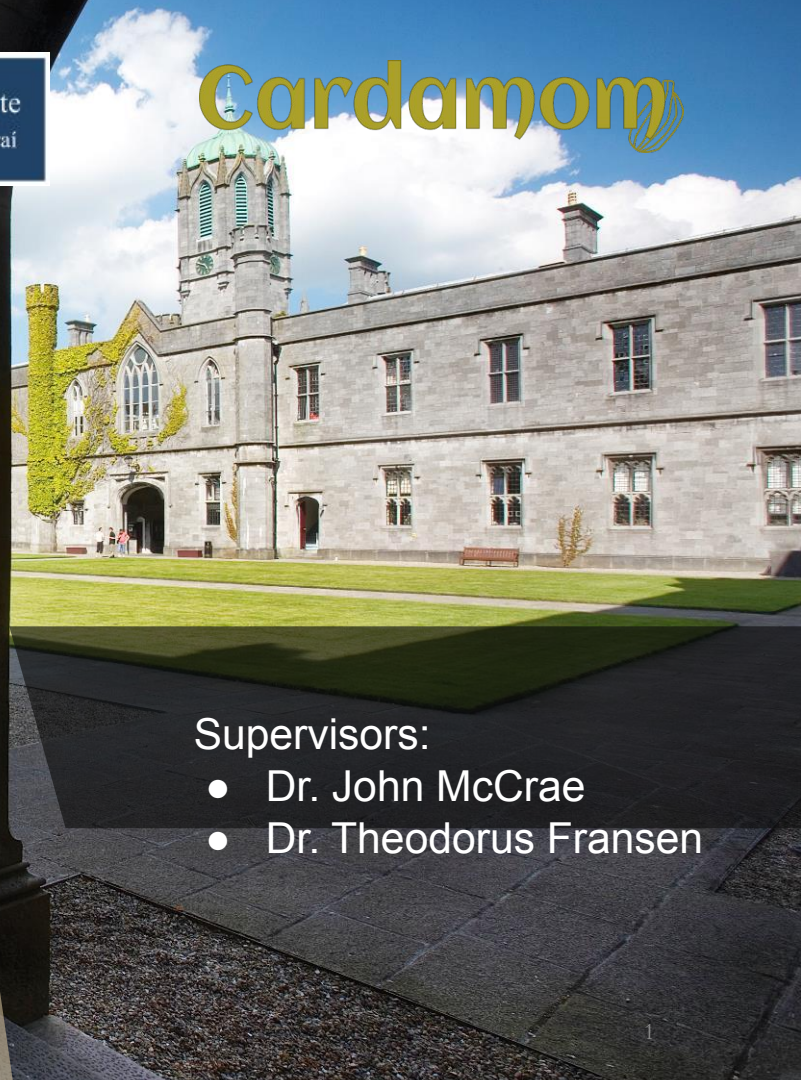
NUI Galway
OÉ Gaillimh

Data Science Institute
Institiúid Eolaíochta Sonraí

Cardamom

Unsupervised Representation Learning for Under-Resourced Languages

Name: Koustava Goswami
Unit for Linguistic Data
Data Science Institute
Date: 28/02/2022



Supervisors:

- Dr. John McCrae
- Dr. Theodorus Fransen

What are Under Resourced Languages

Non-standard Writing System

Lack of linguistic documentation

No large corpora

Limited presence on Web

Machine translation absent or poor

Some Under Resourced Languages

- Manx, Welsh, Breton, Irish, Scottish Gaelic (Indo-European - Celtic)
- Assamese, Bhojpuri, Gujarati (Indo-European - Indo-Aryan)
- Tamil, Telegu (Dravidian Family)
- Setswana, isiXhosa, isiZulu (African Languages)

Benefit of NLP for Under Resourced Languages

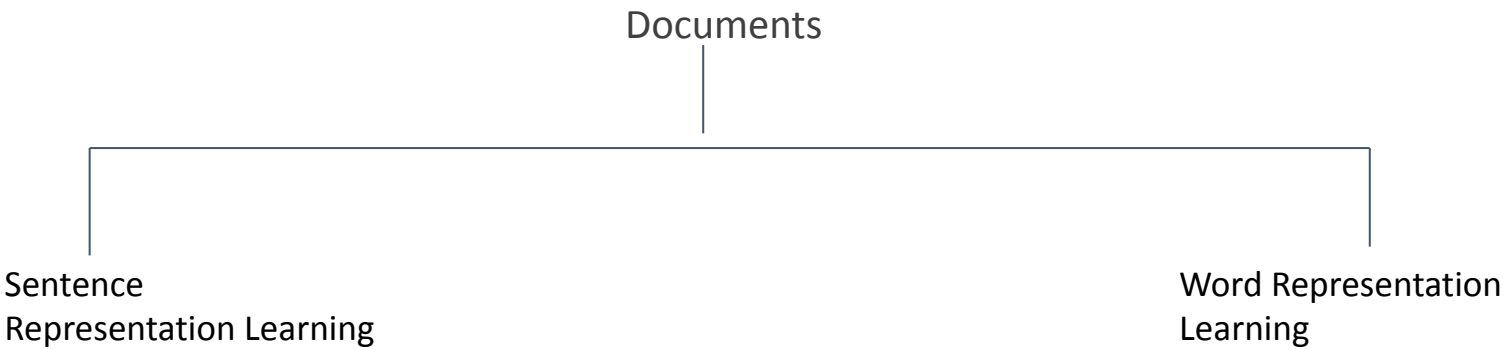
- Automatic Machine Translation helps native speakers to communicate with the outer world.
- Computer Aided Language Learning(CALL) model can be extremely beneficial where learning resources are not available.
- Languages where native speakers are not present can be revived.

Limitations of Deep Models in Under Resourced Languages

- Deep Learning algorithms are very data hungry as a result it is very hard to implement Deep Neural Networks for different Under Resourced Languages.
- It is very hard to find resources to label data in corpus.
- Texts which can be found on social network sites are full of code-mixed sentences which makes it very hard to identify languages of same family group where very few dictionary resources present.

Are you a Cross-Lingual Speaker of Under-Resourced Languages?

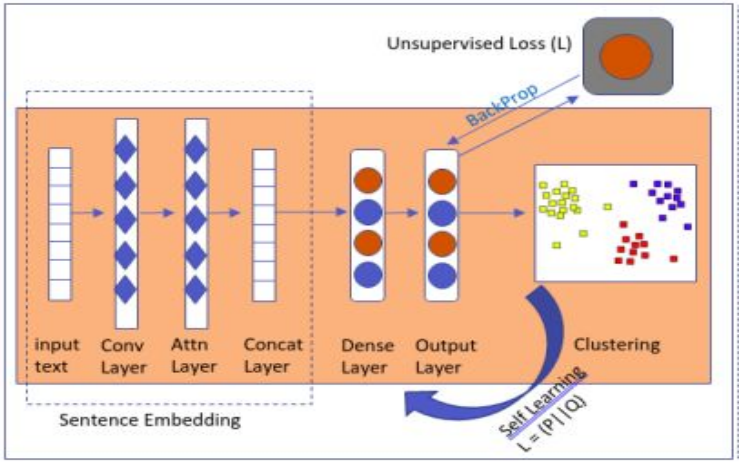
Representation Learning of Cross-lingual Documents



Question

- Is an unsupervised deep neural model capable of identifying languages as accurately as supervised language identification models for code-mixed under-resourced and closely-related languages?

Unsupervised Deep Language and Dialect Identification for Short Texts (UDLDI)

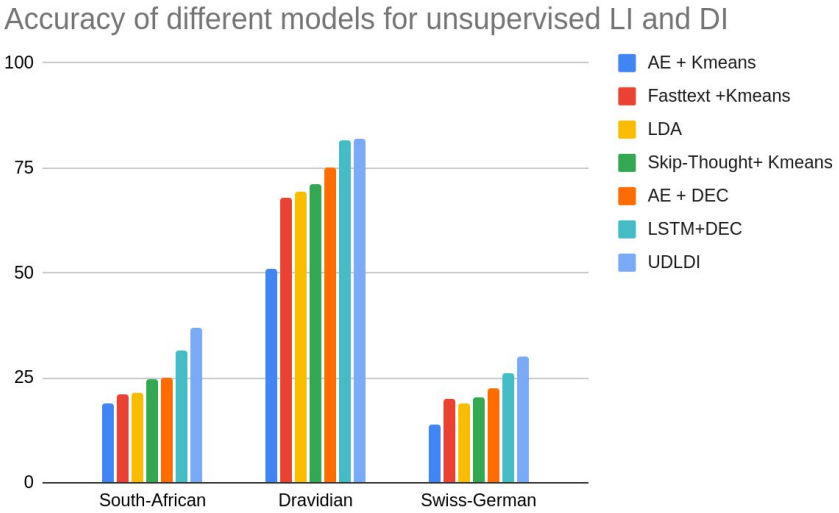


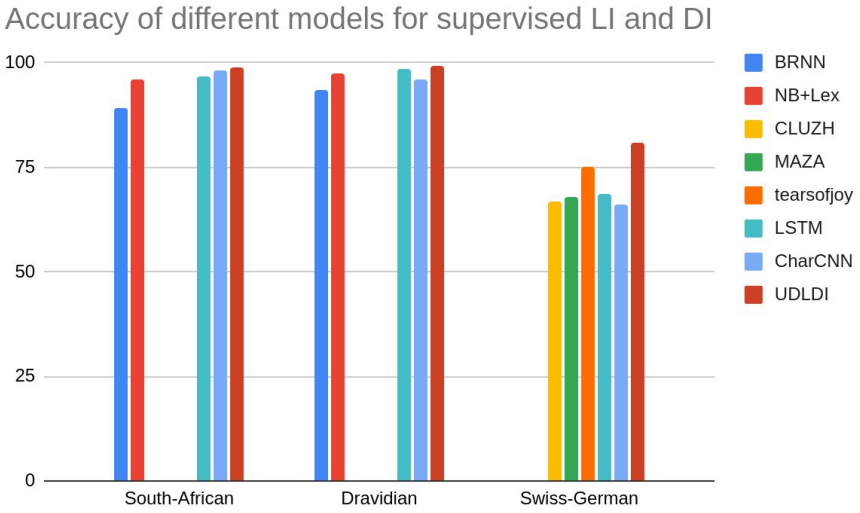
$$L_u = \sum_{i=1}^{N} \max_{j=1}^{i} p_{ij} - \max_{i=1}^{N} \sum_{j=1}^{i} p_{ij}^2$$

1. The model is designed based on two-way deep backpropagation method with joint learning.
2. Introduced new unsupervised loss function MLC (Maximum Likelihood Categorization) which maximizes the probability distribution of feature assignments on each class (or cluster).
3. The iterative clustering process fine-tunes the sentence embedding and ***enhances the cluster assignment*** in an unsupervised way.

Goswami, Koustava, et al. "Unsupervised deep language and dialect identification for short texts." ***28th International Conference on Computational Linguistics*****. 2020**.

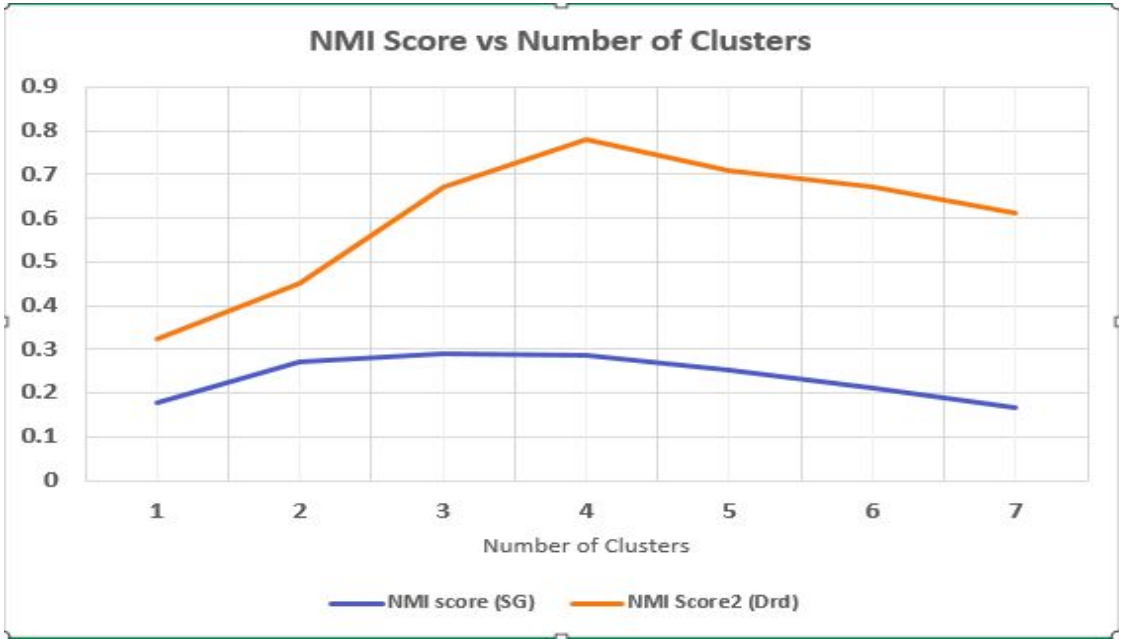
Unsupervised Deep Language and Dialect Identification for Short Texts (UDLDI)





Goswami, Koustava, et al. "Unsupervised deep language and dialect identification for short texts." ***28th International Conference on Computational Linguistics*****. 2020**.

Unsupervised Deep Language and Dialect Identification for Short Texts (UDLDI)



Unsupervised cluster assignment accuracy

Goswami, Koustava, et al. "Unsupervised deep language and dialect identification for short texts." ***28th International Conference on Computational Linguistics*****. 2020**.

Unsupervised Deep Language and Dialect Identification for Short Texts (UDLDI)

Text		Language
Text1	Text2	
und mir hei ja pro bateljoon nume zwei ikaa ghaa	wasmer do zeersch überchoo hei weissi nid sömer	BE
und mir hend glik ghaa isch immer es basler batelioon isch choo	oder die hend det unde	LU
und mir hä also s famillelääbe	oder woo wo de wo ebe de alarm gsii isch hä isch häi mir öisi	BS
und mir händ asoo seer mager müse dure	unfäll ghaa händ und züüg und sache soll me das fiire	ZH

Figure 2: Attention visualization of dialect-specific words pointed out by the model

- The dialects are very closely related from four different parts —Basel (BS), Bern (BE), Lucerne (LU), and Zurich (ZH).
- Consisting of the same characters even though they represents two different dialects.
- Model is also able to identify dialectal (pronunciation) variants for an inflected form of a verb ex:-
  a. in case of BE, it is written as “hei” whereas in LU it is written as “hend” for English word “have”.

Goswami, Koustava, et al. "Unsupervised deep language and dialect identification for short texts." ***28th International Conference on Computational Linguistics*****. 2020**.

Hmm!!! Interesting.. but wait, can we learn a better Cross-Lingual Sentence Representation?

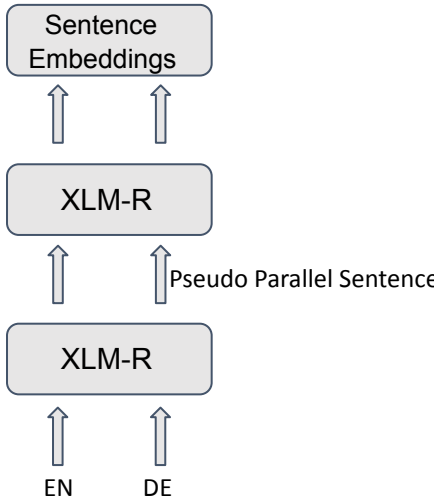
Question

- Does an unsupervised deep sentence embedding framework generate efficient sentence embeddings in cross-lingual domains for under-resourced languages without the use of parallel corpora for downstream natural language tasks?

Related Work

Unsupervised Sentence Embedding Model

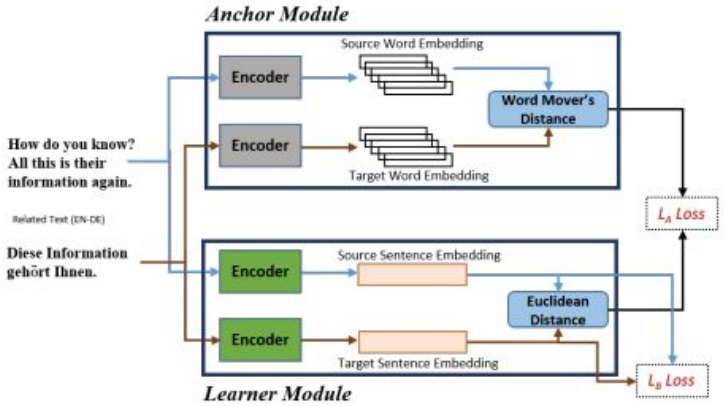
- The model relies on unsupervised machine translated alignment produced by XLM-R.
- The model performed less efficiently while tested for parallel sentence mining for low-resourced languages.
- Does not understand semantic similarity between sentences efficiently.



Kvapilíková, Ivana, et al. "Unsupervised Multilingual Sentence Embeddings for Parallel Corpus Mining." In ACL 2020

Can we use knowledge transfer to build an unsupervised sentence embedding model?

Introducing Anchor-Learner Machine Learning Framework and Unsupervised Sentence Embedding Model

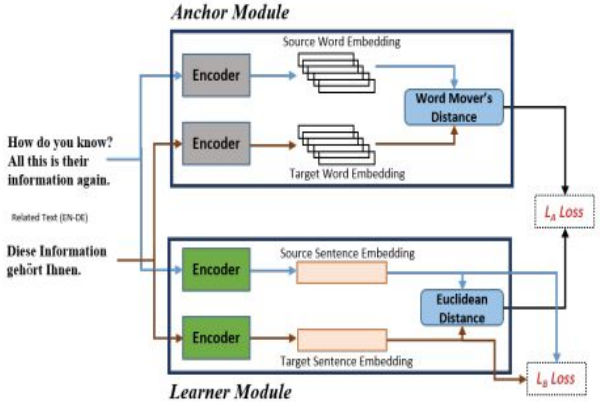


Sentence Embedding Framework

Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." ***2021 Conference on Empirical Methods in Natural Language Processing*****. 2021. (EMNLP)**

Unsupervised Sentence Embedding Model

- The Anchor works as a stabiliser in the system providing its prior knowledge on word level.
- The Learner learns the best alignment in the cross-lingual vector space.
- The semantic similarity and relatedness between sentences are being learned using multi-task learning.
- Automatic knowledge distillation process is introduced which does not need any manual supervision.



Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." *2021 Conference on Empirical Methods in Natural Language Processing*. 2021. (EMNLP)

Unsupervised Sentence Embedding Model

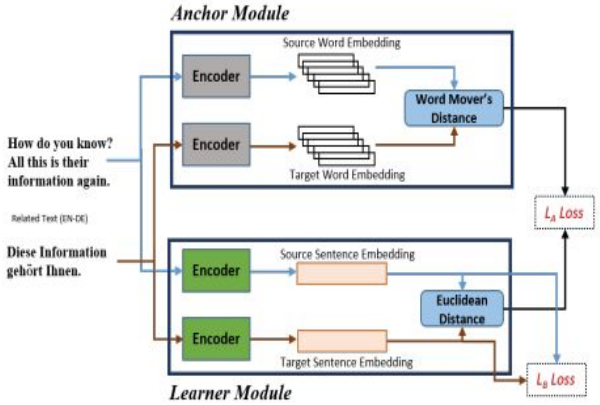
- *Learner* is trained to generate sentence embedding in *multi-task* setup
  - *Unsupervised* Loss function L_A captures *semantic* relationship between sentence pairs
  - Loss function L_B helps to map correct *translation pairs*

- We introduce *Word Mover's Distance* in loss function L_A
  - Preserve *relative semantic distances* between sentence pairs
  - Minimise *Euclidean* distance with the knowledge of *semantic relation* at *word level* from *anchor* model

$$\mathcal{L}_{\mathcal{A}} = \frac{1}{N}\sum_{i=1}^{N} \exp^{|\exp^{-d_{euc}(s'_i,t'_i)} - exp^{-d_{wmd}(s_i,t_i)}|}$$

- L_B addresses translation ranking problem using *Cosine similarity*

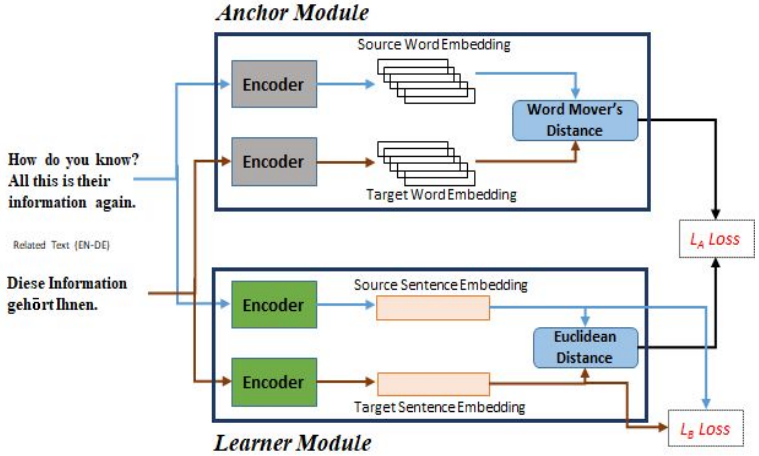
$$\mathcal{L}_{\mathcal{B}} = \frac{1}{N}\sum_{i=1}^{N} cossim(s'_i, t'_i)$$

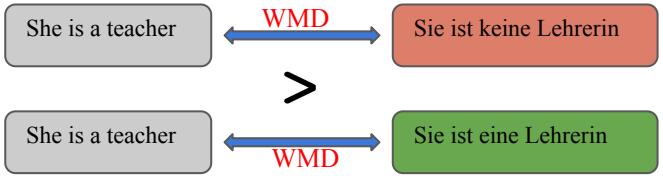


Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." *2021 Conference on Empirical Methods in Natural Language Processing*. 2021. (EMNLP)

Unsupervised Sentence Embedding Model

- Inclusion of ***Word Mover's Distance*** is ***advantageous*** for ***unsupervised*** learning
  - Closer representations for similar sentences
  - Dissimilar sentences have embeddings that are apart in the embedding space
- Efficiently captures ***negation*** in sentence pairs while understanding semantic relatedness





Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." *2021 Conference on Empirical Methods in Natural Language Processing*. 2021. (EMNLP)

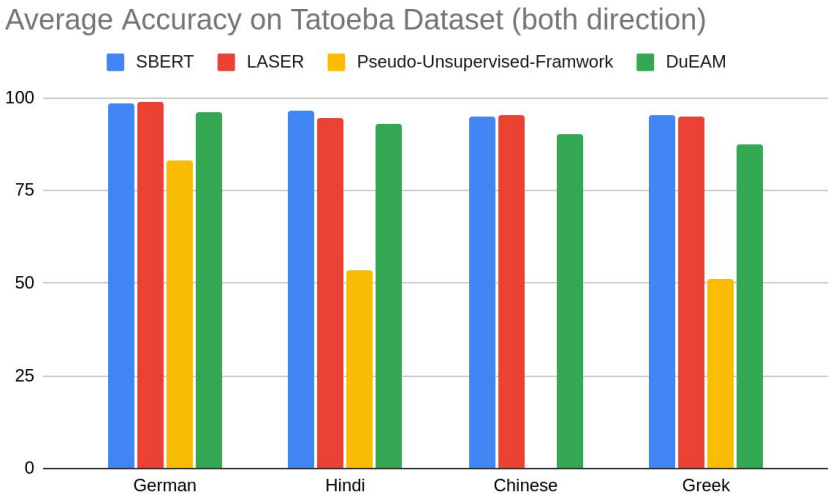
Datasets used to train the model

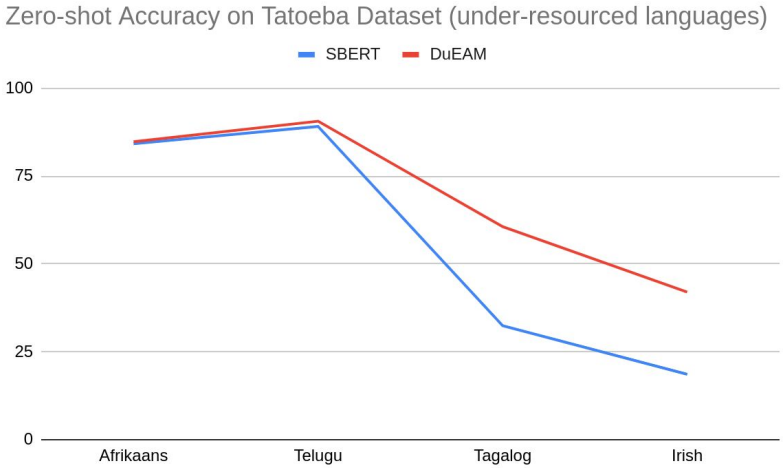
- We have trained the model based on Multilingual Natural Language Inference Dataset.
- Training does not involve any cross-lingual parallel datasets.
- The training dataset contains both monolingual and cross-lingual datasets.
  - In case of cross-lingual datasets building we keep premises from the source language and replace the hypothesis with random hypothesis sentences, and vice-versa.
- The training process does not take any annotated levels into account.
- We have trained our model on 13 languages.

Premise	Hypothesis	Type
How do you know? All this is their information again.	This information belongs to them.	*Monolingual (EN-EN)*
- woher weißt du das ? All das sind ihre Informationen.	Diese Information gehört Ihnen.	*Monolingual (DE-DE)*
How do you know? All this is their information again.	Diese Information gehört Ihnen.	*Cross-lingual (EN-DE)*
- woher weißt du das ? All das sind ihre Informationen.	This information belongs to them.	*Cross-lingual (DE-EN)*

**Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning."
2021 Conference on Empirical Methods in Natural Language Processing. 2021. (EMNLP)**

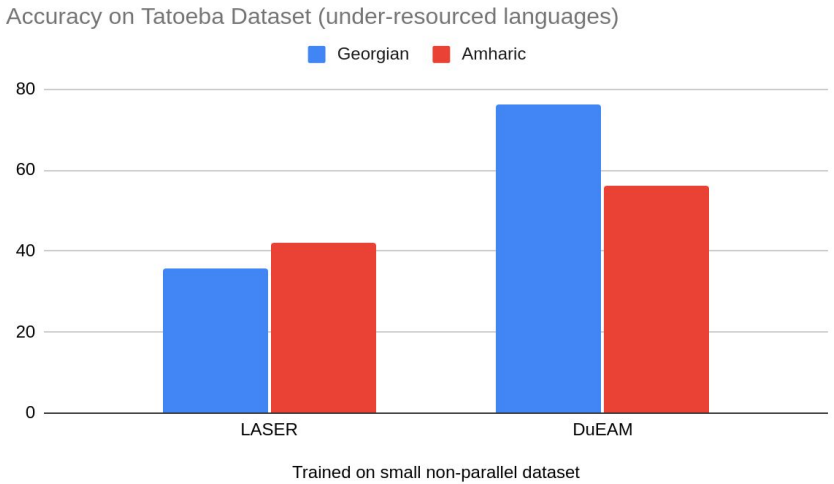
Unsupervised Sentence Embedding Model





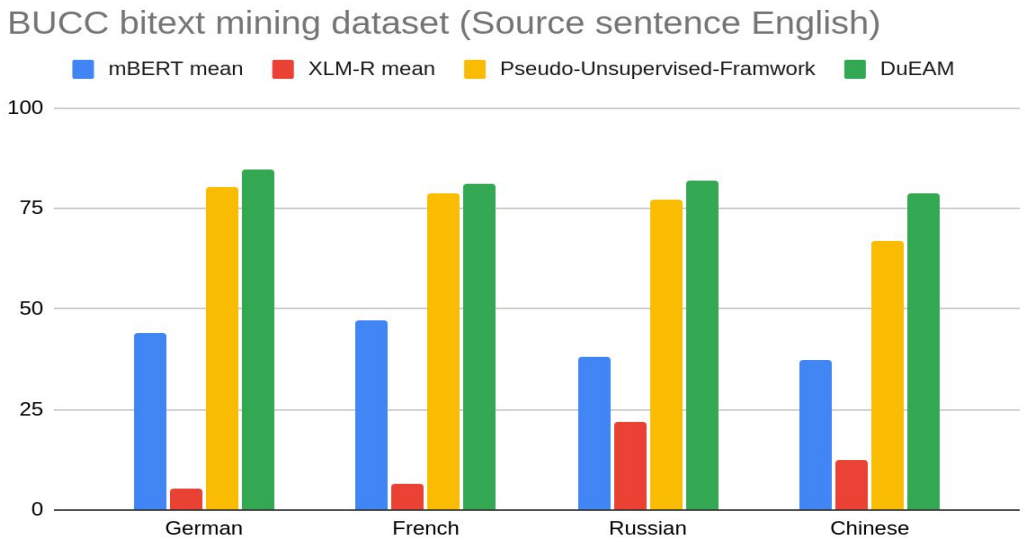
Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." ***2021 Conference on Empirical Methods in Natural Language Processing*****. 2021. (EMNLP)**

Unsupervised Sentence Embedding Model



Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." ***2021 Conference on Empirical Methods in Natural Language Processing*****. 2021. (EMNLP)**

Unsupervised Sentence Embedding Model



Goswami, Koustava, et al. "Cross-lingual Sentence Embedding using Multi-Task Learning." *2021 Conference on Empirical Methods in Natural Language Processing*. 2021. (EMNLP)

That is interesting.. What about word representation learning?

Current Research Work

Question

- Does an unsupervised deep neural model learn morphological paradigm relatedness without any prior linguistic information for closely related and low-resourced languages?

Unsupervised Paradigm Discovery Problem

- This work treats the paradigm discovery problem (PDP)—the task of learning an inflectional morphological system from unannotated sentences.
- The system makes use of word embeddings and string similarity to cluster forms by cell and by paradigm.
- They have released gold standard dataset for 8 languages.

Gold Grid	cell 1	cell 2	cell 3	cell 4	cell 5
paradigm 1	«watch»	«watches»	watching	watched	watched
paradigm 2	«follow»	follows	«following»	followed	followed
paradigm 3	see	«sees»	«seeing»	«saw»	seen

Alexander, et al. "The Paradigm Discovery Problem" In ACL 2020

Unsupervised paradigm clustering task

For example, if the tokenized Bible text is: " **peace be with you ! as the father has sent me , I am sending you .** ", then the output format is:

peace	be,am	with	you	as	the
father	has	sent,sending	me,I		

Unsupervised Morphological Typology Learning (currently in progress)

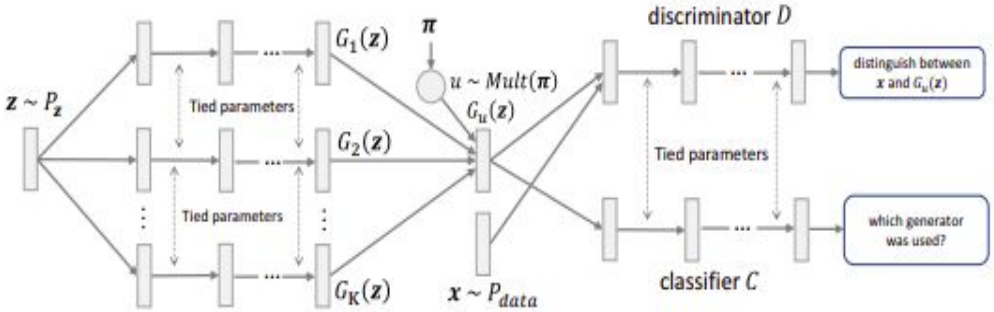


Figure 1: MGAN's architecture with K generators, a binary discriminator, a multi-class classifier.

- No rule extraction is needed.
- Language independent.
- Providing only list of words as corpora will be sufficient.
- Can be extended to n number of languages.

Reference Papers for the talk

- Goswami, K., Rani, P., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. (2020, December). ULD@ NUIG at SemEval-2020 Task 9: Generative Morphemes with an Attention Model for Sentiment Analysis in Code-Mixed Text. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation@LREC2020* (pp. 968-974).

- Goswami, K., Sarkar, R., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. (2020, December). Unsupervised Deep Language and Dialect Identification for Short Texts. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1606-1617).

- Goswami, K., Dutta, S., Assen, H., Fransen, T., & McCrae, J. P. (2021). Unsupervised Cross Lingual Sentence Embedding using Multitask Learning (IJCAI 2021) (Currently Under Review).

- Rani, P., Suryawanshi, S., Goswami, K., Chakravarthi, B. R., Fransen, T., & McCrae, J. P. (2020, May). A comparative study of different state-of-the-art hate speech detection methods in Hindi-English code-mixed data. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying@LREC2020* (pp. 42-48).

- Sarkar, R., Goswami, K., Arcan, M., & McCrae, J. P. (2020, December). Suggest me a movie for tonight: Leveraging Knowledge Graphs for Conversational Recommendation. In *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 4179-4189).

Thanks

Questions?