

# Filling the Gap: Decoding of Word Embeddings for Generation of Coherent New Words

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 — Software Project



# 1 State of the Project

2 Languages study

3 Results

4 Discussion

5 What to improve?

6 Future work

# Reminder of our aim

- Apply decoder to the regression task (solving analogies)

$$A : B :: C : X \xrightarrow{X=?} A : B :: C : D$$

e.g. *star : stars :: cat : X → cats*

- Current output: vectors ( $\neq$  word)
- Aim: transform these vectors into words

# What we managed to do?

- ➊ Move all codes to PyTorch Lightning
- ➋ Research on morphology and variational auto-encoder
- ➌ Build the decoder based on word embeddings
- ➍ Train it on 11 Languages
- ➎ Test with different parameters
- ➏ Evaluate results with two metrics

- 1 State of the Project
- 2 Languages study**
- 3 Results
- 4 Discussion
- 5 What to improve?
- 6 Future work

# Language family

Language family	Languages
Indo-European	German, Russian, Spanish
Afro-Asiatic	Arabic, Maltese
Uralic	Finnish, Hungarian
Altaic	Turkish, Japanese
Caucasian	Georgian
Na-Dene	Navajo

Figure: Classification according to language families

# Morphological typology (1)



Figure: Classification according to the degree of internal complexity

# Morphological typology (2)

Morphological type	Flectional	Agglutinating
Characteristics	<ul style="list-style-type: none"> <li>● Cumulation</li> <li>● Fusion</li> <li>● Internal flection</li> </ul>	<ul style="list-style-type: none"> <li>● Morpheme <math>\Leftrightarrow</math> 1 meaning</li> <li>● Clear-cut boundary</li> <li>● Form not affected</li> </ul>
Languages	German, Russian, Spanish, Arabic, Maltese	Finnish, Hungarian, Turkish, Japanese, Georgian, Navajo

Figure: Classification according to the technique



# Inflectional morphology

Affixes	Suffixes++	Suffixes+	=	Prefixes+	Prefixes++
Languages	German Russian Spanish Arabic Maltese Finnish Hungarian Turkish Japanese	Georgian			Navajo

Figure: Affixes used in inflectional morphology [Dryer, 2013]



Arabic, Maltese: templatic morphology (root-and-pattern strategy)

- 1 State of the Project
- 2 Languages study
- 3 Results**
- 4 Discussion
- 5 What to improve?
- 6 Future work

# Results with different parameters

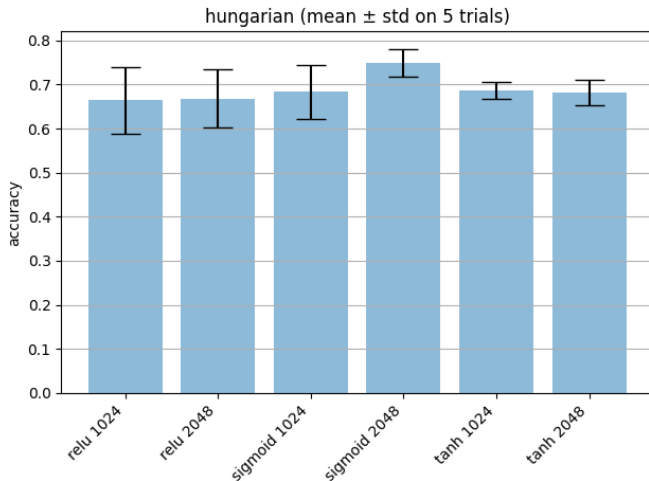


Figure: Mean accuracy ( $\pm$  standard deviation) on 5 trials for Hungarian

# Results on all languages

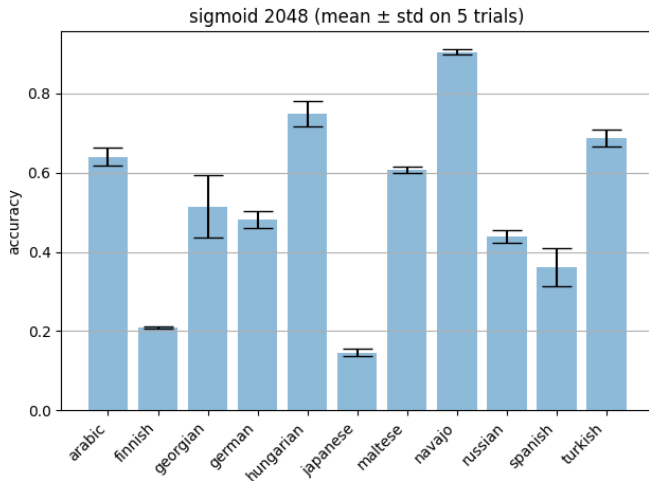


Figure: Mean accuracy ( $\pm$  standard deviation) on 5 trials with a sigmoid activation function and a hidden size of 2048

# Results on all languages

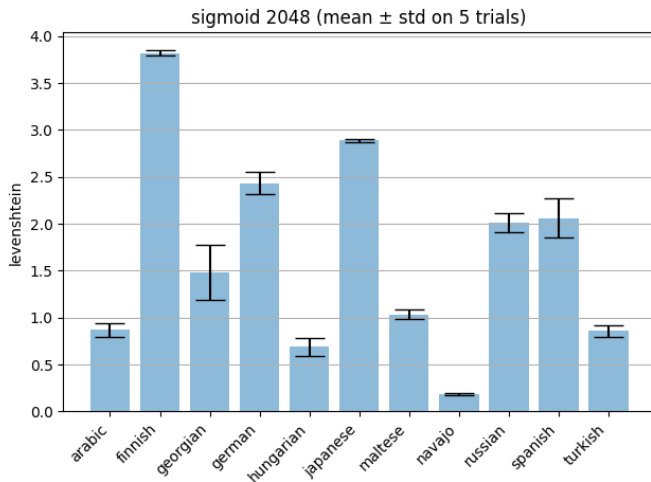


Figure: Mean levenshtein distance ( $\pm$  standard deviation) on 5 trials with a sigmoid activation function and a hidden size of 2048

- 1 State of the Project
- 2 Languages study
- 3 Results
- 4 Discussion**
- 5 What to improve?
- 6 Future work

# Leads to explain the results

- ① content of the embeddings
  - ▶ subwords ? (root-and-pattern strategy)
  - ▶ amount of different subwords
  - ▶ proximity of the subwords
- ② morphological features of the languages

- 1 State of the Project
- 2 Languages study
- 3 Results
- 4 Discussion
- 5 What to improve?**
- 6 Future work



# What to improve?

- ❶ Find a better evaluation metrics e.g.:
  - ❶ search for a new metrics that deals with word lengths
- ❷ Have a better understanding of the content of the embeddings:
  - ❶ subwords = morphemes ?
  - ❷ decoded words: real for some languages

- 1 State of the Project
- 2 Languages study
- 3 Results
- 4 Discussion
- 5 What to improve?
- 6 Future work**

# Future work

- 22 Nov - Regression model + decoder / Variational auto-encoder
- 10 Dec - Qualitative analysis / Multilingual model
- 14 Jan - Application docker & webpage
- 3 Feb - Report

شكرا جزيلا

Thank you

Merci

អរគុណ

Obrigado

# References I

-  Booij, G. E., Lehmann, C., Mugdan, J., and Skopeteas, S. (2008).  
*Morphologie: Ein internationales Handbuch zur Flexion und Wortbildung*.  
De Gruyter Mouton.
-  Dryer, M. S. (2013).  
Prefixing vs. suffixing in inflectional morphology.  
In Dryer, M. S. and Haspelmath, M., editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, Leipzig.
-  Eifring, H. and Theil, R. (2005).  
Linguistic typology.  
*Linguistics for students of Asian and African languages*.

# Our decoder structure

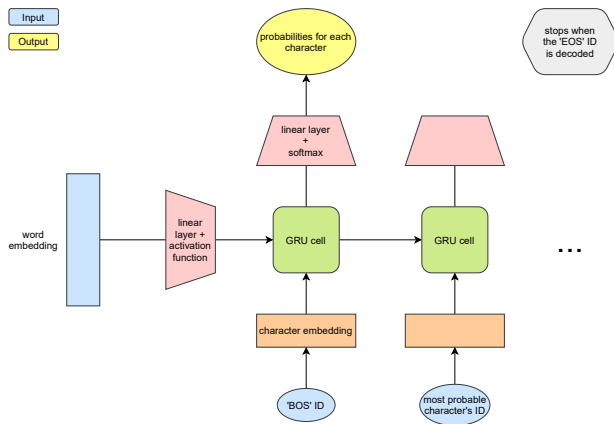


Figure: Our GRU based decoder

Inspired by this blogpost <https://rajatvd.github.io/Generating-Words-From-Embeddings/>

## Results with different parameters (smaller hidden sizes)

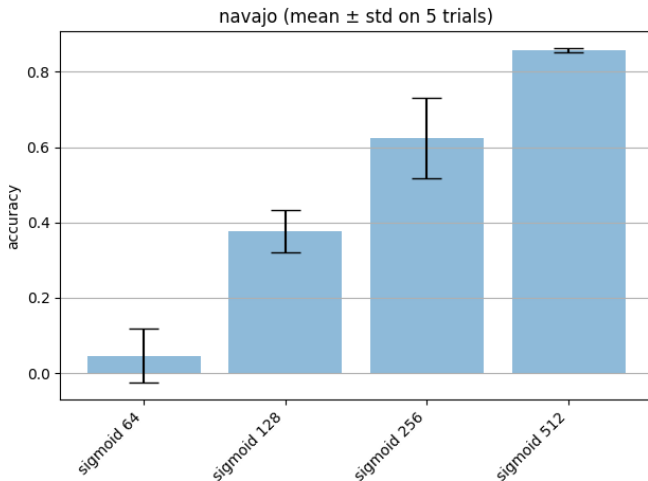


Figure: Mean accuracy ( $\pm$  standard deviation) on 5 trials for Navajo