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

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 – Software Project

Reminder of our aim

- Apply decoder to the regression task (solving analogies)

$$A : B :: C : X \xrightarrow{X=?} A : B :: C : D$$
$$e.g.\ \ star : stars :: cat : X \rightarrow cats$$

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

What we managed to do?

1. Move all codes to PyTorch Lightning
2. Research on morphology and variational auto-encoder
3. Build the decoder based on word embeddings
4. Train it on 11 Languages
5. Test with different parameters
6. Evaluate results with two metrics

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	• Cumulation • Fusion • Internal flection	• Morpheme ⇔1 meaning • Clear-cut boundary • Form not affected
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)

Results with different parameters

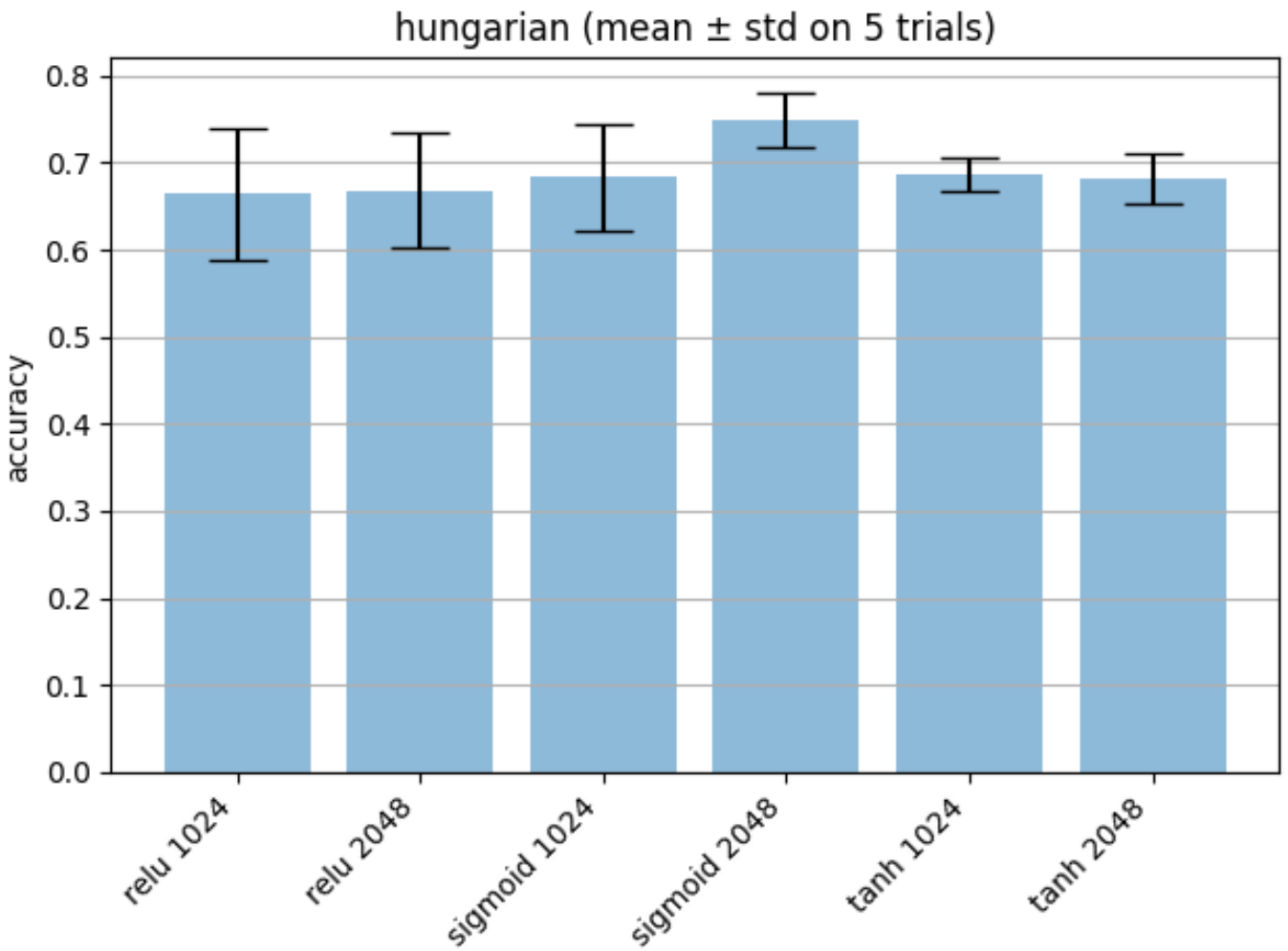

Figure: Mean accuracy (± standard deviation) on 5 trials for Hungarian

Results on all languages

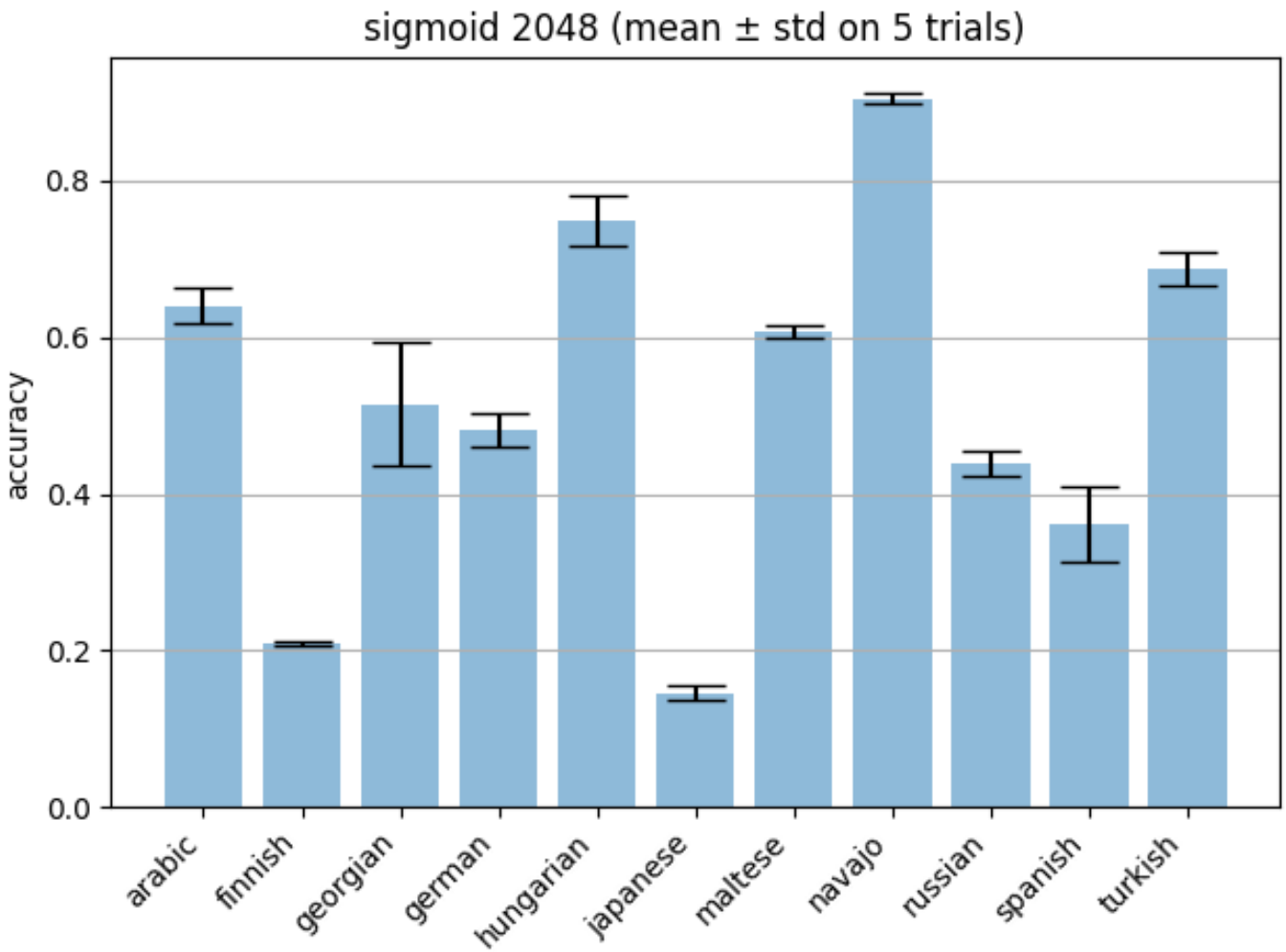

Figure: Mean accuracy (± standard deviation) on 5 trials with a sigmoid activation function and a hidden size of 2048

Results on all languages

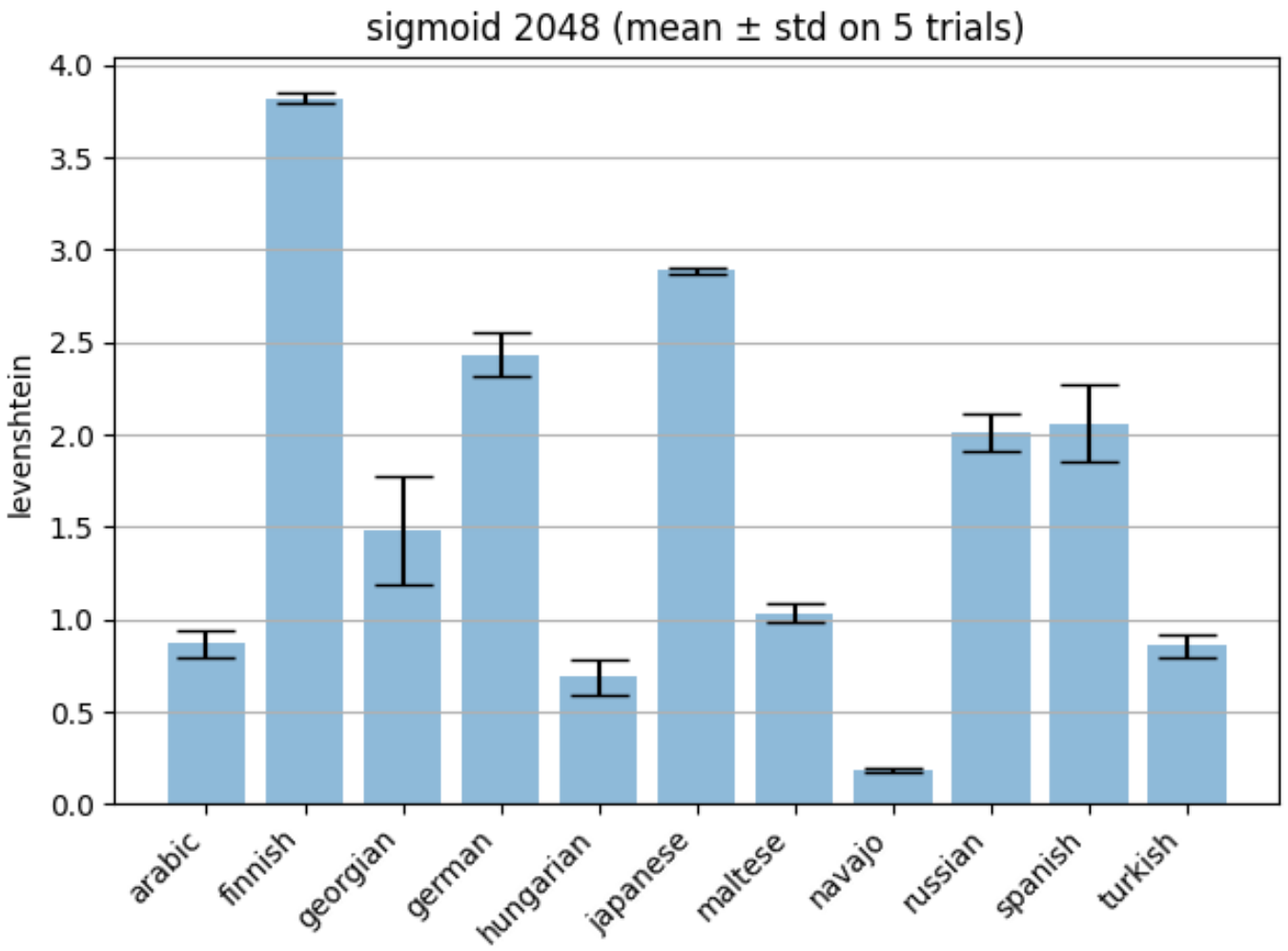

Figure: Mean levenshtein distance (± standard deviation) on 5 trials with a sigmoid activation function and a hidden size of 2048

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

What to improve?

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

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

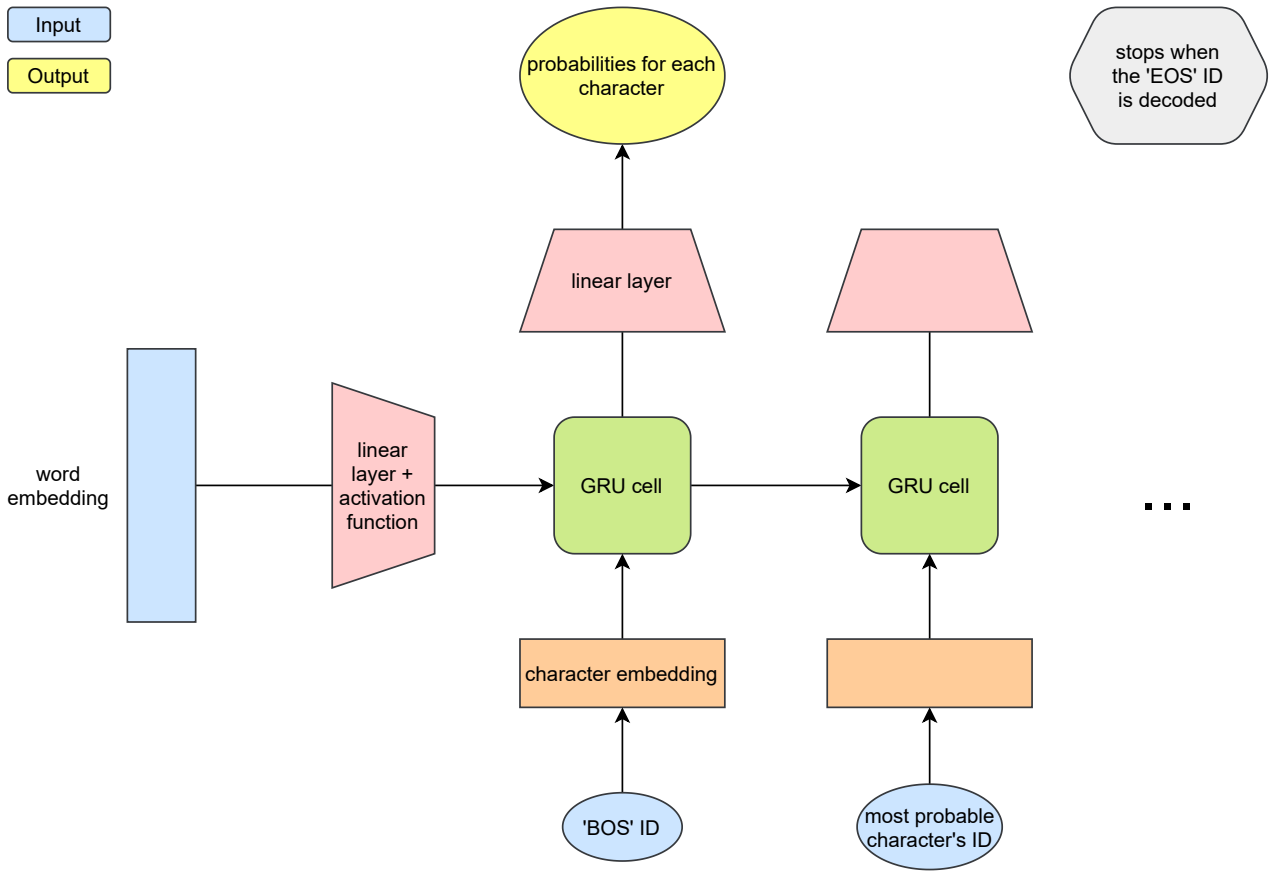

Figure: Our GRU based decoder

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

Results with different parameters (smaller hidden sizes)

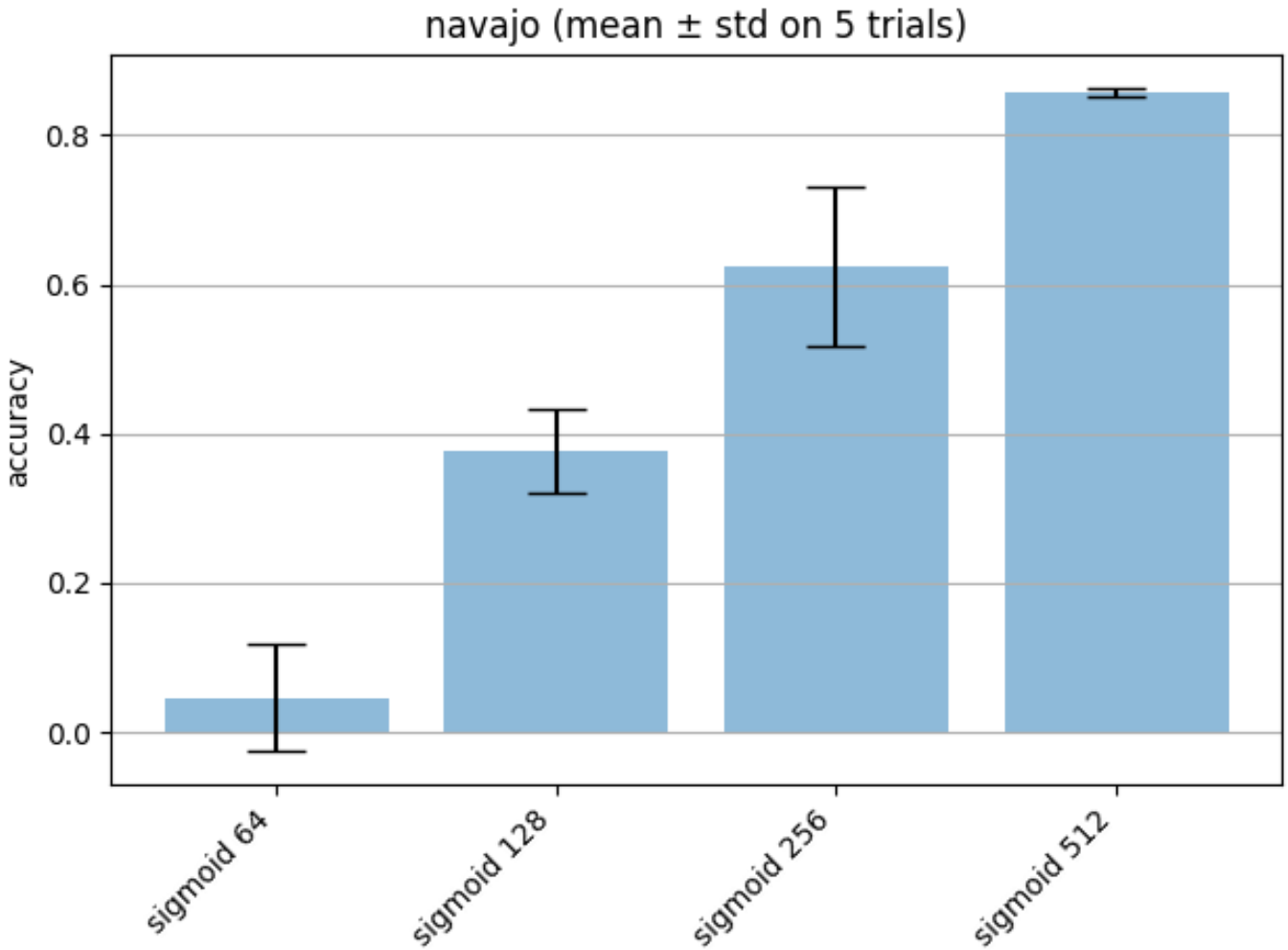

Figure: Mean accuracy (± standard deviation) on 5 trials for Navajo