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

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 – Software Project

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

Reminder of our aim

- Regression task based on transfer

$$A : B :: C : X \xrightarrow{X=?} A : B :: C : D$$
$$e.g.\ \ dog : dogs :: chat : X \rightarrow chats$$

- Input: A and B in language 1, C in language 2
- Output: D in language 2
- Same transformation for A, B and C, D

First trial results

	hungarian, german	turkish, finnish	hungarian, finnish
Cosine similarity	**58.9**	**39.5**	**18.9**
Euclidean distance	57.7	39.1	16.8

Table: Accuracy for the regression task on the three (source, target) language pairs.

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

3 approaches

1. Comparable data
2. Omnilingual model
3. Sigmorphon 2019

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

New results

Table: Accuracy (in %) of 3 runs of the regression model.

Language	**ANNr (previous)** (mean ± std.)	**actual**
Arabic	**77.97** ± 16.03	**61.13** ± 0.83
Finnish	**37.78** ± 9.28	**77.56** ± 1.78
Georgian	**94.66** ± 1.13	**86.40** ± 0.62
German	**86.38** ± 0.45	**86.93** ± 0.78
Hungarian	**53.83** ± 3.12	**78.98** ± 0.50
Maltese	**75.00** ± 5.08	**79.66** ± 1.11
Navajo	**31.74** ± 0.90	**45.88** ± 0.24
Russian	**75.15** ± 0.44	**70.53** ± 0.37
Spanish	**86.27** ± 0.71	**91.12** ± 1.06
Turkish	**61.95** ± 10.86	**80.34** ± 0.79
Japanese	**61.60** ± 1.33	**37.54** ± 37.33

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

Bilingual analogies

In our dataset: $\textsc{Word}_1$ $\textsc{Features}$ $\textsc{Word}_2$

An analogy: $\textsc{Word}_{1,A}$:$\textsc{Word}_{2,A}$::$\textsc{Word}_{1,B}$:$\textsc{Word}_{1,B}$
where $\textsc{Features}_A = \textsc{Features}_B$

Bilingual analogies: $\textsc{Language}_A \neq \textsc{Language}_B$

$\rightarrow$ keep only the subset of *shared features*

Shared features

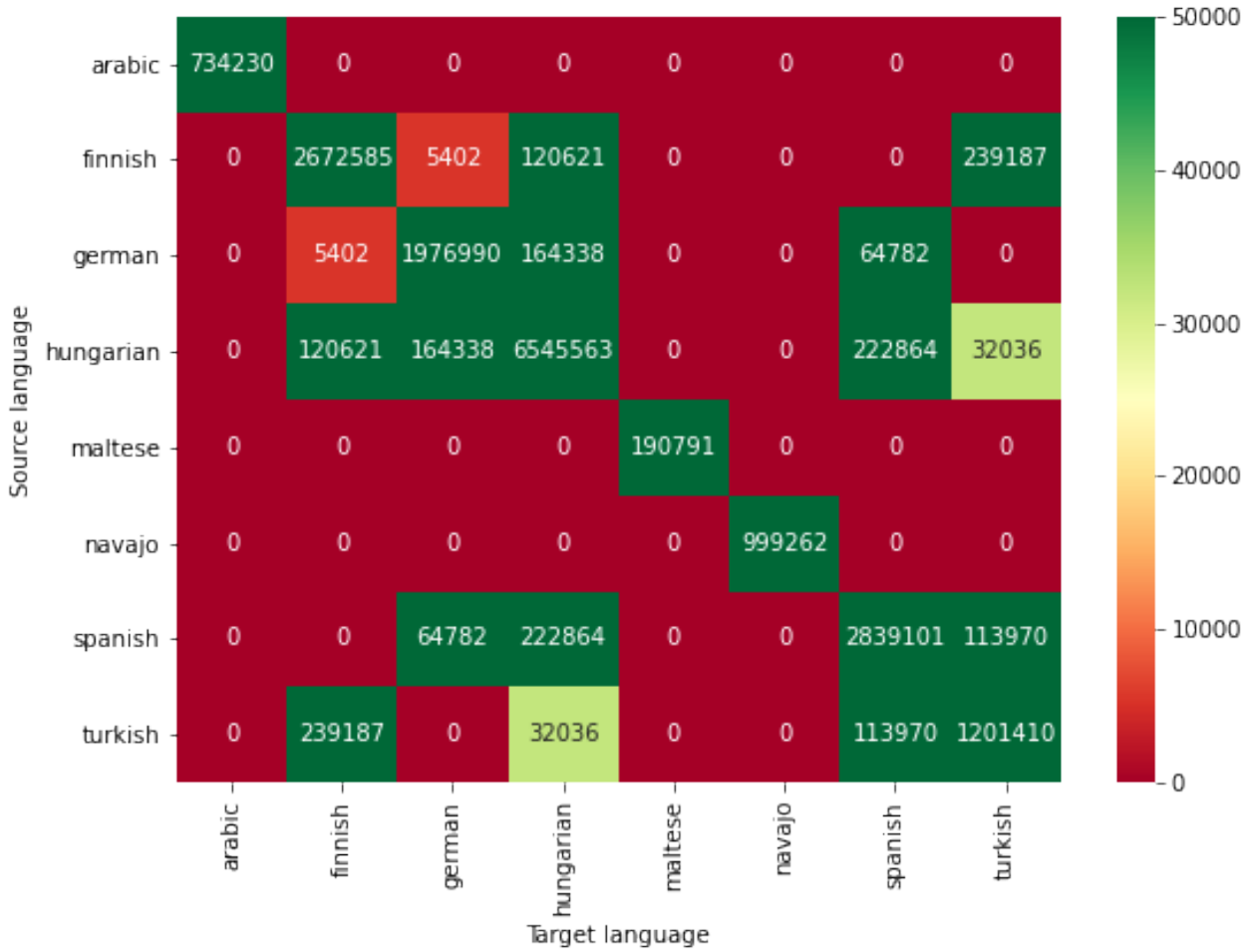

Figure: Number of possible analogies for each pair of languages

Comparison between monolingual and bilingual results

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	43.96±1.48	80.93±1.94	/	82.00±1.90
German	92.63±0.10	/	68.17±3.12	68.17±3.12	/
Hungarian	43.07±0.48	85.92±0.83	/	85.92±0.83	40.92±2.46
Spanish	/	93.97±0.25	93.97±0.25	/	94.05±0.31
Turkish	65.89±1.59	/	71.76±0.92	93.18±1.90	/

Table: Monolingual analogies: Accuracy (±std) on 3 runs

	Finnish	German	Hungarian	Spanish	Turkish
finnish	/	**81.88**	35.88	/	30.19
german	80.31	/	30.41	35.10	/
hungarian	48.83±3.19	78.41±1.59	/	**91.62**	33.93
spanish	/	17.63	83.26	/	40.63
turkish	45.81±0.17	/	16.17	70.27	/

Table: Bilingual analogies: Accuracy (±std) on 3 runs

Omnilingual model

Languages which share features with at least one other language: Finnish, German, Hungarian, Turkish, Spanish

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	60.30±1.26	3.08±0.86	31.78±1.79	/	52.62±1.84
German	3.08±0.86	63.27±0.68	57.71±0.48	62.47±2.41	/
Hungarian	31.78±1.79	57.71±0.48	71.12±1.04	62.89±1.86	24.73±1.26
Spanish	/	62.47±2.41	62.89±1.86	66.82±1.34	62.20±6.57
Turkish	52.62±1.84	/	24.73±1.26	62.20±6.57	49.73±0.82

Table: Accuracy (±std) on 5 runs

Next time: Sigmorphon 2019 [McCarthy et al., 2019]

88 languages: 8/10 from Sigmorphon 2016 [Cotterell et al., 2016]

$\rightarrow$ Arabic, Finnish, German, Hungarian, Russian, Spanish, Turkish, Maltese (Georgian and Navajo missing)

Aim: apply trained models to the new dataset

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

What we want

- Solving analogies: monolingual and bilingual
- Use the omnilingual model

What it looks like

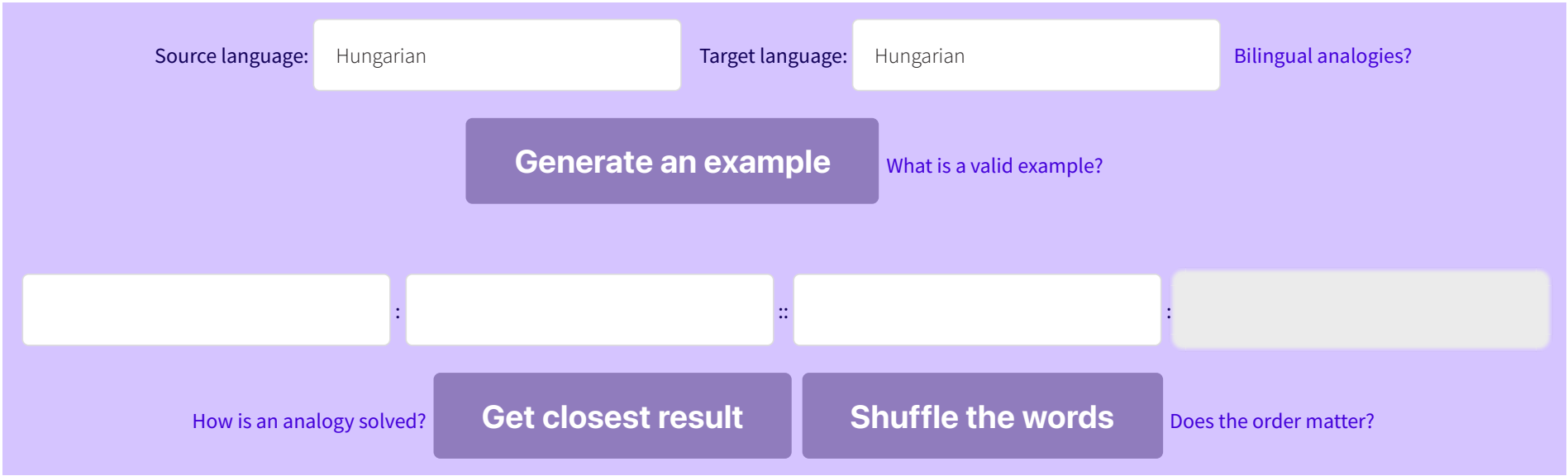

Figure: Preview of our software

1 Reminder

2 Our approach

3 Regression task

4 Results & discussion

5 Software

6 Future work

Future work

- Run final experiments
- Improve and adapt our software
- Continue writing the report

شكرا جزيلا

Thank you

Merci

អរគុណ

Obrigado

References I

Cotterell, R., Kirov, C., Sylak-Glassman, J., Yarowsky, D., Eisner, J., and Hulden, M. (2016).
The sigmorphon 2016 shared task—morphological reinflection.
In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

McCarthy, A. D., Vylomova, E., Wu, S., Malaviya, C., Wolf-Sonkin, L., Nicolai, G., Kirov, C., Silfverberg, M., Mielke, S. J., Heinz, J., Cotterell, R., and Hulden, M. (2019).
The SIGMORPHON 2019 shared task: Morphological analysis in context and cross-lingual transfer for inflection.
In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.