

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

- 1 Reminder
- 2 Our approach
- 3 Regression task
- 4 Results & discussion**
- 5 Software
- 6 Future work

Bilingual analogies

In our dataset: $\text{WORD}_1 \text{ FEATURES } \text{WORD}_2$

An analogy: $\text{WORD}_{1,A}:\text{WORD}_{2,A}::\text{WORD}_{1,B}:\text{WORD}_{2,B}$
where $\text{FEATURES}_A = \text{FEATURES}_B$

Bilingual analogies: $\text{LANGUAGE}_A \neq \text{LANGUAGE}_B$

→ keep only the subset of *shared features*

Shared features

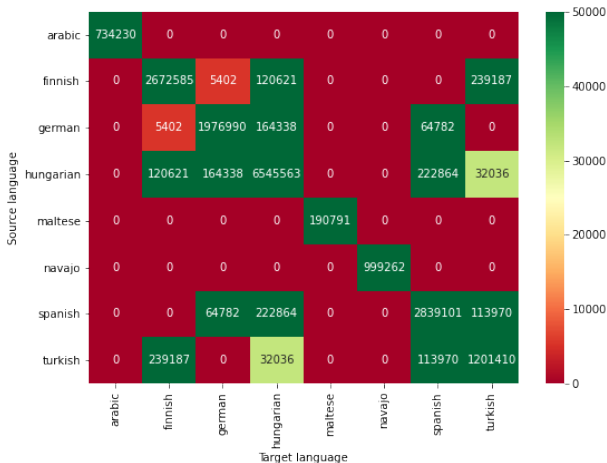


Figure: Number of possible analogies for each pair of languages

Comparison between monolingual and bilingual results

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

Table: Monolingual analogies: Accuracy (\pm 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 \pm 3.19	78.41 \pm 1.59	/	91.62	33.93
Spanish	/	17.63	83.26	/	40.63
Turkish	45.81 \pm 0.17	/	16.17	70.27	/

Table: Bilingual analogies: Accuracy (\pm 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 (\pm std) on 5 runs

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

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

→ 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

The interface is a web application for generating bilingual analogies. It features a light purple background. At the top, there are two input fields for 'Source language:' and 'Target language:', both containing the word 'Hungarian'. To the right of these fields is the text 'Bilingual analogies?'. Below these fields is a large purple button labeled 'Generate an example'. To the right of this button is the text 'What is a valid example?'. Below the button, there are four input fields arranged horizontally. The first three are white, and the fourth is grey. They are separated by colons. Below the first three fields is the text 'How is an analogy solved?'. Below the fourth field is the text 'Does the order matter?'. Between the first and second fields is a purple button labeled 'Get closest result'. Between the third and fourth fields is a purple button labeled 'Shuffle the words'.

Source language: Hungarian Target language: Hungarian Bilingual analogies?

Generate an example What is a valid example?

How is an analogy solved? Get closest result Shuffle the words Does the order matter?

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.