Towards an omnilingual model for solving morphological analogies

Safa AlSaidi, Amandine Decker, Stephanie Monteiro

M2 – Software Project

What is an analogy?

Analogy

$$A : B :: C : D$$

"A is to B as C is to D"
Same relation between the pair A and B and the pair C and D

Order

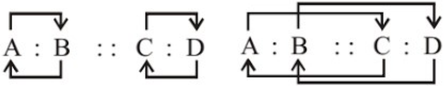


Figure: Symmetry and central permutation

Mathematical examples

- $1 : 2 :: 5 : 6 \leftrightarrow 1 - 2 = 5 - 6$ (Arithmetic)
- $1 : 2 :: 2 : 4 \leftrightarrow 1/2 = 2/4$ (Geometric)

What is morphology?

Morphemes

- Root (or base)
- Affixes: prefixes, suffixes

Prefixation

$\text{inactive}_A \rightarrow \text{in} + \text{active}_A$
$\text{disappear}_V \rightarrow \text{dis} + \text{appear}_V$

Suffixation

$\text{useful}_A \rightarrow \text{use}_N + \text{ful}$
$\text{reads}_V \rightarrow \text{read}_V + \text{s}$

Prefixation & suffixation

$\text{unconsciousness}_N \rightarrow \text{un} + \text{conscious}_A + \text{ness}$

What is a morphological analogy?

Example: "cat is to cat**s** as star is to star**s**"

Comparison in terms of the presence and absence of affixes to determine the validity and correctness of analogies

Typical tasks

1. Classification

Valid examples

“cat”:“cats”::“apple”:“apples”
“cat”:“apple”::“cats”:“apples”

Invalid examples

“cat”:“apples”::“cats”:“apple”
“cat”:“cat”::“apple”:“apples”

2. Solving Analogies
   - when one word is missing/unknown
   - “cat”:“cats”::“apple”: $X \rightarrow X =$ “apples”

Solving multilingual morphological analogies

- Analogy solving task based on transfer

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

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

Datasets

SIGMORPHON 2016 [Cotterell et al., 2016] and the Japanese Bigger Analogy Test Set [Karpinska et al., 2018].

- contain triples (lemma, target features, target word)
- analogies are generated based on triples sharing the same features ($F = F'$).

$$\langle A, F, B \rangle, \langle A', F', B' \rangle$$

Valid examples

cat	pos=N, num=PL	cats
apple	pos=N, num=PL	apples

cat:cats::apple:apples is valid

Invalid examples

cat	pos=N, num=PL	cats
sleep	pos=V, tense=PRS, per=3, num=SG	sleeps

cat:sleep::cats:sleeps is invalid

Language choice

We kept languages that shared features with other languages.

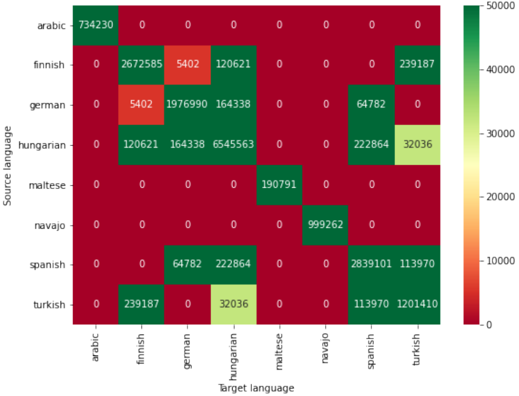
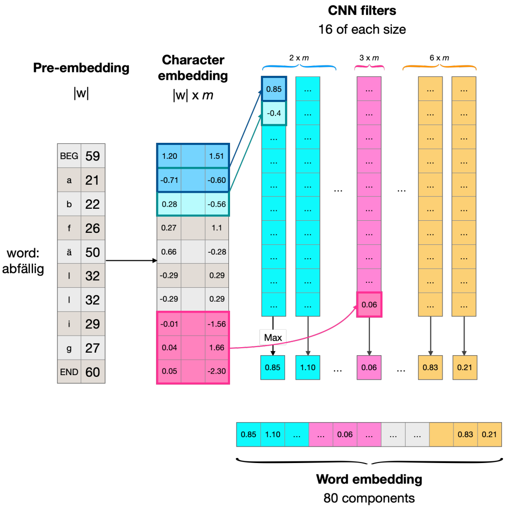


Figure: Number of possible analogies for each pair of languages

Custom embedding model

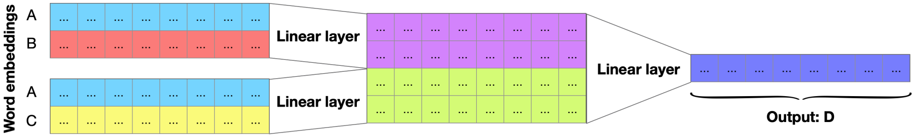


This model was inspired by [Kim et al., 2016]

Analogy solver model

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

$$X = g(f_1(A, B), f_2(A, C))$$



This model was inspired by [Lim et al., 2019]

Monolingual analogy solving models

Language	**ANNr (previous)** (mean ± std.)	**diff. parameters** (mean ± std.)
Arabic	77.97 ± 16.03	61.13 ± 0.83
Finnish	37.78 ± 9.28	**76.46 ± 1.58**
Georgian	**94.66 ± 1.13**	84.67 ± 2.78
German	86.38 ± 0.45	**88.70 ± 0.58**
Hungarian	53.83 ± 3.12	**78.72 ± 0.53**
Maltese	75.00 ± 5.08	78.04 ± 1.44
Navajo	31.74 ± 0.90	**45.74 ± 0.99**
Russian	**75.15 ± 0.44**	72.23 ± 0.44
Spanish	86.27 ± 0.71	**91.72 ± 0.43**
Turkish	61.95 ± 10.86	**80.37 ± 1.00**
Japanese	61.60 ± 1.33	**72.58 ± 2.47**

Table: Accuracy (in %) of the analogy solving models (10 runs for the new results, 3 for previous ones)

Omnilingual analogy solving model

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	61.85±1.79	3.22±1.73	28.58±2.69	/	55.23±3.85
German	3.15±1.47	64.38±1.27	66.91±4.49	62.07±3.58	/
Hungarian	36.26±4.52	55.25±1.47	73.33±1.31	78.36±1.33	32.00±3.03
Spanish	/	61.16±2.54	74.05±1.77	69.38±1.65	70.67±4.03
Turkish	54.12±1.48	/	25.38±3.94	65.72±6.09	52.23±1.09

Table: Accuracy (in %) of 10 runs of the omnilingual analogy solver model

Valid analogy

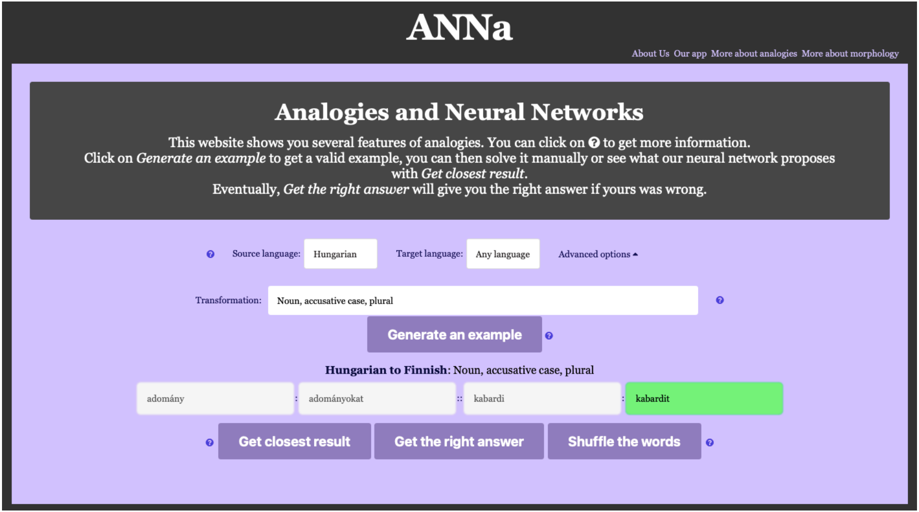


Figure: Preview of our software

Invalid analogy

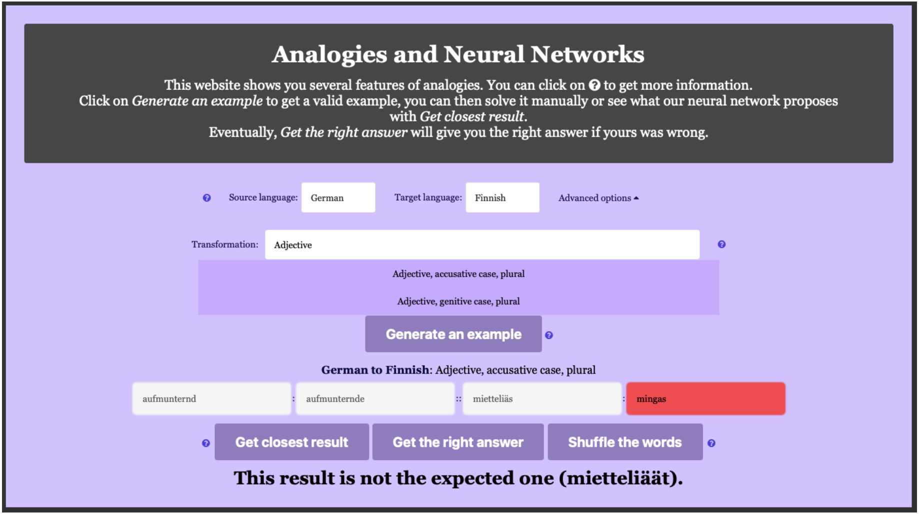


Figure: Preview of our software

Final Remarks

1. **What did we manage to do?**
   - Improve analogy solving results
   - Explore bilingual analogies
   - Train an omnilingual model
   - Develop a software
2. What challenges did we encounter?
   - Language proximity
3. What can be improved?
   - Qualitative analysis
   - Decoder
   - Extend to other languages

Final Remarks

1. What did we manage to do?
   - Improve analogy solving results
   - Explore bilingual analogies
   - Train an omnilingual model
   - Develop a software
2. **What challenges did we encounter?**
   - Language proximity
3. What can be improved?
   - Qualitative analysis
   - Decoder
   - Extend to other languages

Final Remarks

1. What did we manage to do?
   - Improve analogy solving results
   - Explore bilingual analogies
   - Train an omnilingual model
   - Develop a software
2. What challenges did we encounter?
   - Language proximity
3. **What can be improved?**
   - Qualitative analysis
   - Decoder
   - Extend to other languages

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.

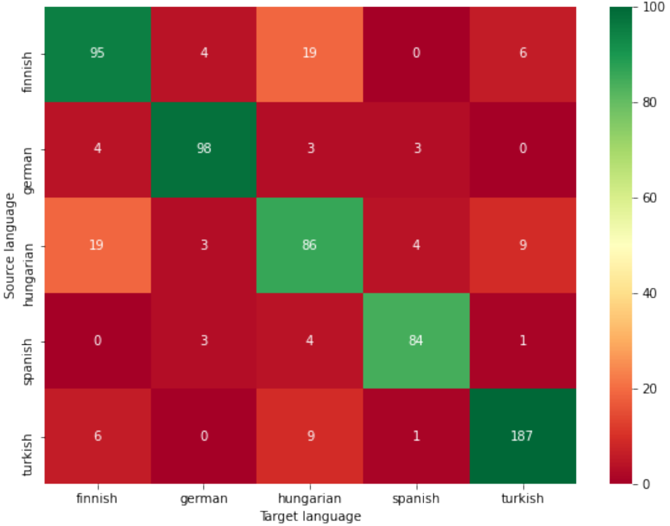
Karpinska, M., Li, B., Rogers, A., and Drozd, A. (2018).
Subcharacter Information in Japanese Embeddings: When Is It Worth It?
In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 28–37, Melbourne, Australia. Association for Computational Linguistics.

Kim, Y., Jernite, Y., Sontag, D., and Rush, A. (2016).
Character-aware neural language models.
In Proceedings of the AAAI Conference on Artificial Intelligence, 30(1).

References II

Lim, S., Prade, H., and Richard, G. (2019).
Solving word analogies: A machine learning perspective.
In Kern-Isberner, G. and Ognjanovic, Z., editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty, 15th European Conference, ECSQARU 2019, Belgrade, Serbia, September 18-20, 2019, Proceedings*, volume 11726 of *Lecture Notes in Computer Science*, pages 238–250. Springer.

Number of different features per languages



Omnilingual vs Bilingual

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	3.22±1.73	28.58±2.69	/	**55.23±3.85**
German	3.15±1.47	/	**66.91±4.49**	**62.07±3.58**	/
Hungarian	36.26±4.52	55.25±1.47	/	78.36±1.33	32.00±3.03
Spanish	/	**61.16±2.54**	74.05±1.77	/	**70.67±4.03**
Turkish	**54.12±1.48**	/	**25.38±3.94**	65.72±6.09	/

Table: Accuracy (in %) of 10 runs of the omnilingual analogy solving model (diff. parameters)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	**66.01±33.01**	**36.28±0.76**	/	28.34±0.48
German	**64.43±32.23**	/	30.53±0.57	11.92±3.30	/
Hungarian	**50.61±2.26**	**77.98±1.24**	/	**94.02±0.29**	33.89±0.77
Spanish	/	32.42±17.01	**78.99±0.13**	/	40.45±1.52
Turkish	46.43±1.24	/	16.07±0.90	70.86±0.04	/

Table: Accuracy (in %) of 5 runs of the bilingual analogy solving models (diff. parameters)

Omnilingual vs Monolingual

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	61.85±1.79	3.22±1.73	28.58±2.69	/	55.23±3.85
German	3.15±1.47	64.38±1.27	66.91±4.49	62.07±3.58	/
Hungarian	36.26±4.52	55.25±1.47	73.33±1.31	78.36±1.33	32.00±3.03
Spanish	/	61.16±2.54	74.05±1.77	69.38±1.65	70.67±4.03
Turkish	54.12±1.48	/	25.38±3.94	65.72±6.09	52.23±1.09

Table: Accuracy (in %) of 10 runs of the omnilingual analogy solving model (diff. parameters)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	**76.46±1.58**	**39.75±2.67**	**80.33±1.95**	/	**79.89±2.43**
German	**94.26±0.63**	**88.70±0.58**	70.55±2.61	**70.55±2.61**	/
Hungarian	43.93±3.44	**85.39±1.76**	**78.72±0.53**	**85.39±1.76**	**40.98±3.42**
Spanish	/	**94.26±0.53**	**94.26±0.53**	**91.72±0.43**	**95.83±0.24**
Turkish	**64.98±2.76**	/	**70.74±2.23**	**94.03±3.70**	**80.37±1.00**

Table: Accuracy (in %) of 10 runs of the monolingual analogy solver models (evaluated on shared features only)

Omnilingual model (diff. parameters): different training datasets

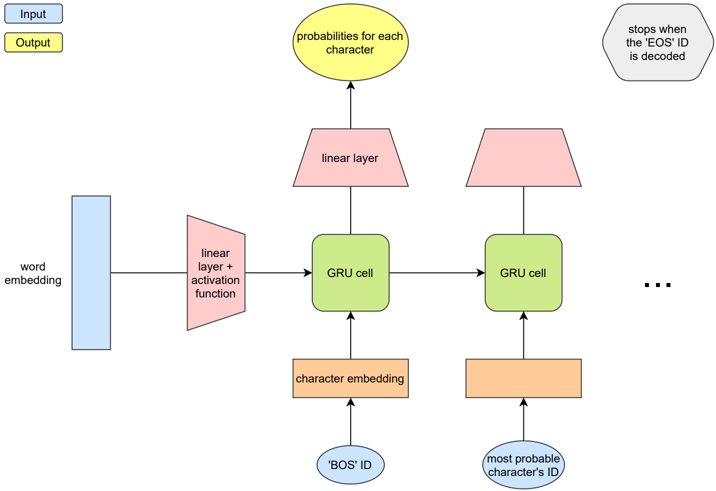
	Finnish	German	Hungarian	Spanish	Turkish
Finnish	**61.85±1.79**	3.22±1.73	28.58±2.69	/	**55.23±3.85**
German	3.15±1.47	**64.38±1.27**	66.91±4.49	**62.07±3.58**	/
Hungarian	36.26±4.52	55.25±1.47	**73.33±1.31**	78.36±1.33	32.00±3.03
Spanish	/	**61.16±2.54**	74.05±1.77	**69.38±1.65**	70.67±4.03
Turkish	54.12±1.48	/	25.38±3.94	65.72±6.09	**52.23±1.09**

Table: Accuracy (in %) of 10 runs of the omnilingual analogy solving model trained on the full dataset

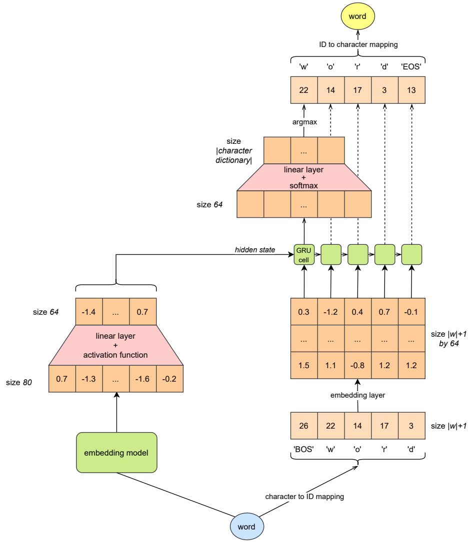
	Finnish	German	Hungarian	Spanish	Turkish
Finnish	52.41±1.35	**24.79±1.89**	**38.22±2.26**	/	48.81±1.33
German	**18.40±4.44**	31.24±1.25	**79.08±1.85**	55.48±1.09	/
Hungarian	**64.11±2.06**	**68.52±1.10**	10.56±0.37	79.29±1.81	**41.26±2.96**
Spanish	/	51.27±1.50	**80.04±1.64**	28.91±0.65	74.57±5.75
Turkish	**59.89±1.54**	/	26.52±2.69	**83.49±2.41**	21.86±1.04

Table: Accuracy (in %) of 10 runs of the omnilingual analogy solver model trained on shared features only

Decoder model: use



Decoder model: training



Bilingual results

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	66.01±33.01	36.28±0.76	/	28.34±0.48
German	64.43±32.23	/	30.53±0.57	11.92±3.30	/
Hungarian	50.61±2.26	77.98±1.24	/	94.02±0.29	33.89±0.77
Spanish	/	32.42±17.01	78.99±0.13	/	40.45±1.52
Turkish	46.43±1.24	/	16.07±0.90	70.86±0.04	/

Table: Accuracy (in %) of 5 runs of the bilingual analogy solver models (diff. parameters)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	83.24±0.35	35.58±0.74	/	28.14±1.14
German	80.13±0.57	/	30.14±0.29	12.92±6.03	/
Hungarian	51.21±3.09	78.09±0.74	/	94.05±0.12	34.55±0.60
Spanish	/	36.79±11.74	79.11±0.63	/	41.83±0.85
Turkish	47.23±0.91	/	15.29±0.85	70.79±0.06	/

Table: Accuracy (in %) of 5 runs of the bilingual analogy solver models (shared parameters)

Monolingual VS Bilingual (different parameters)

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	39.75±2.67	**80.33±1.95**	/	**79.89±2.43**
German	94.26±0.63	/	**70.55±2.61**	**70.55±2.61**	/
Hungarian	43.93±3.44	**85.39±1.76**	/	85.39±1.76	**40.98±3.42**
Spanish	/	**94.26±0.53**	**94.26±0.53**	/	**95.83±0.24**
Turkish	**64.98±2.76**	/	**70.74±2.23**	**94.03±3.70**	/

Table: Accuracy (in %) of 10 runs of the monolingual analogy solver models

	Finnish	German	Hungarian	Spanish	Turkish
Finnish	/	66.01±33.01	36.28±0.76	/	28.34±0.48
German	64.43±32.23	/	30.53±0.57	11.92±3.30	/
Hungarian	**50.61±2.26**	77.98±1.24	/	**94.02±0.29**	33.89±0.77
Spanish	/	32.42±17.01	78.99±0.13	/	40.45±1.52
Turkish	46.43±1.24	/	16.07±0.90	70.86±0.04	/

Table: Accuracy (in %) of 5 runs of the bilingual analogy solver models