



HAL
open science

A Neural Approach for Detecting Morphological Analogies

Safa Alsaidi, Amandine Decker, Puthineath Lay, Esteban Marquer,
Pierre-Alexandre Murena, Miguel Couceiro

► **To cite this version:**

Safa Alsaidi, Amandine Decker, Puthineath Lay, Esteban Marquer, Pierre-Alexandre Murena, et al.. A Neural Approach for Detecting Morphological Analogies. IEEE DSAA 2021 - The 8th IEEE International Conference on Data Science and Advanced Analytics, Oct 2021, Porto / Online, Portugal. , IEEE DSAA 2021. hal-03328841

HAL Id: hal-03328841

<https://inria.hal.science/hal-03328841>

Submitted on 30 Aug 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

A Neural Approach for Detecting Morphological Analogies

Safa Alsaïdi*, Amandine Decker*, Puthineath Lay*, Esteban Marquer*, Pierre-Alexandre Murena**, Miguel Couceiro*

*Université de Lorraine, CNRS, LORIA, F-54000, Nancy, France; **HIIT, Aalto University, Helsinki, Finland



Proportional Analogy

Notation:

► $A : B :: C : D$ (read "A is to B as C is to D")

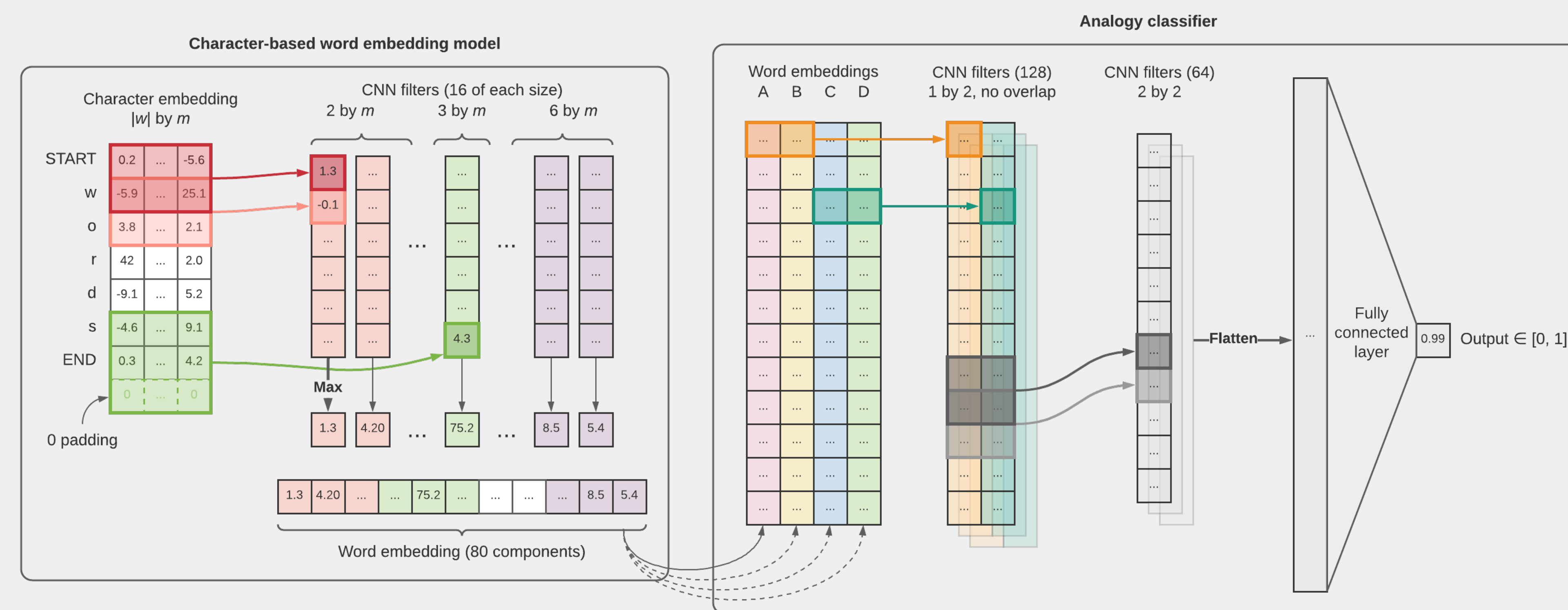
3 Core properties:

- 1 $A : B :: C : D \rightarrow C : D :: A : B$ (symmetry)
- 2 $A : B :: C : D \rightarrow A : C :: B : D$ (central permutation)
- 3 $A : B :: A : B$ (reflexively)

Morphology:

- "read is to readable as count is to countable"
- *aalto : aalloksi :: spirituaali : spirituaaliksi* (Finnish)

Proposed Approach: Morphological Embedding + CNN Classifier



Baselines

1 Classifier:

- Lepage Classifier [3]
 - matrix based approach to align characters between words and find common sub-words

2 Solvers:

- find missing x making $A : B :: C : x$ valid; if expected D in top 1 or 10 solutions, valid analogy
- Kolmogorov Complexity [6]:
 - minimize complexity of f such that $B = f(A)$ and $f(C)$ is computable; $x = f(C)$
- Alea [5]:
 - random permutations of the characters of B not in A and those of C (Monte-Carlo method)

Training and Evaluation Procedure

Analogies obtained by aligning morphological transformations:

- *dog*, to NUM=PLURAL, *dogs*
- *cat*, to NUM=PLURAL, *cats*
- resulting analogy: *dog : dogs :: cat : cats*

For each analogy $A : B :: C : D$ extracted from the data (use max 50000 of the available analogies):

- 1 embed A, B, C , and D
- 2 permute embeddings using properties of proportional analogy:
 - Train. & eval.: 8 valid analogies
 - Training: 3 invalid analogies
 - Evaluation: 3×8 invalid analogies and their equivalent forms
- 3 aggregate over all permutations:
 - Training: loss (Binary Cross-Entropy)
 - Evaluation: accuracy

$A : B :: C : D$ (base form) $C : D :: A : B$
 $A : C :: B : D$ $B : A :: D : C$ $D : B :: C : A$
 $D : C :: B : A$ $C : A :: D : B$ $B : D :: A : C$

Table: 8 equivalent forms per analogy

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

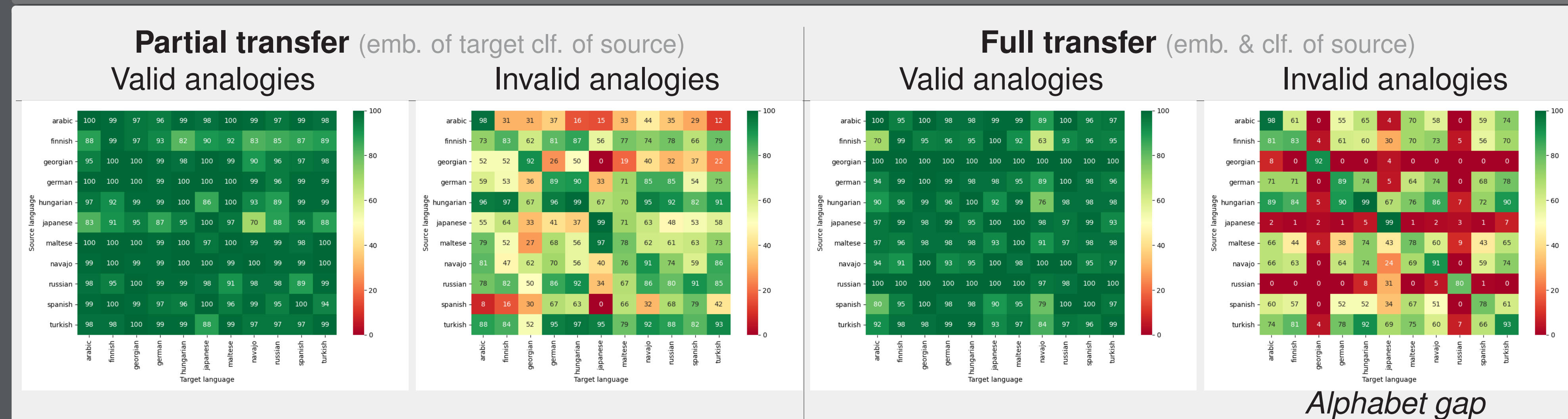
Table: 3 invalid analogies per analogy

Classification Results

Language	CNN (ours)		Best Baseline		Number of analogies in the data	
	Valid	Invalid	Valid	Invalid	Training	Test
<i>Sigmorphon2016 task 1 (inflection) [2]</i>						
Arabic	99.89	97.52	34.21 (Alea@10)	97.79 (Kolmo@1)	373240	555312
Finnish	99.44	82.62	25.60 (Lepage)	98.78* (Alea@1)	1342639	4691453
Georgian	99.83	91.71	93.20 (Kolmo@10)	95.21 (Alea@1)	3553763	8368323
German	99.48	89.01	86.90 (Alea@10)	97.19 (Alea@1)	994740	1480256
Hungarian	99.99	98.81	36.80 (Kolmo@10)	98.40 (Kolmo@1)	3280891	66195
Maltese	99.96	77.83	78.05 (Alea@10)	69.29 (Kolmo@1)	104883	3707
Navajo	99.53	90.82	21.45 (Kolmo@10)	94.93 (Kolmo@1)	502637	4843
Russian	97.95	79.85	42.37 (Alea@10)	93.88 (Lepage)	1965533	6421514
Spanish	99.94	78.33	85.90 (Alea@10)	86.62 (Lepage)	1425838	4794504
Turkish	99.48	92.63	44.76 (Alea@10)	91.40 (Kolmo@1)	606873	11360
<i>Japanese Bigger Analogy Test Set [4]</i>						
Japanese	99.99	98.65	19.20 (Kolmo@10)	98.13 (Lepage)	18487	7923

123 no significant difference
123 best result, significant difference between baselines and ours
 * obtained on 4000 analogies (too slow on 50000)

Transferability



Perspectives

- ✓ explore transferability [1]
- ✓ balancing data
- ✓ regression (analogy solving)
- qualitative analysis of embedding model
- other domains (images, text, explanations, etc.)

Bibliography

- [1] Safa Alsaïdi et al. "On the Transferability of Neural Models of Morphological Analogies". In: *AIMLAI, ECML PKDD 2021*. 2021.
- [2] Ryan Cotterell et al. "The SIGMORPHON 2016 Shared Task—Morphological Reinflection". In: *SIGMORPHON, 2016*. ACL, Aug. 2016.
- [3] Rashel Fam and Yves Lepage. "Tools for The Production of Analogical Grids and a Resource of N-gram Analogical Grids in 11 Languages". In: *11th LREC*. ELRA, 2018.
- [4] Marzena Karpinska et al. "Subcharacter Information in Japanese embeddings: when is it worth it?". In: *Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*. ACL, 2018.
- [5] Philippe Langlais et al. "Improvements in Analogical Learning: Application to Translating Multi-Terms of the Medical Domain". In: *12th EACL*. ACL, 2009.
- [6] Pierre-Alexandre Murena et al. "Solving Analogies on Words based on Minimal Complexity Transformation". In: *29th IJCAI*. 2020.