A (de)tour through bias mitigation and analogy based ML

Défi HyAIAI

G. Alves (ex-PhD) V. Bhargava (ex-intern) **M. Couceiro** L. Galarraga
E. Marquer (PhD) A. Napoli R. Querish (PhD) G. Zervakis (ex-PhD) ...

Overview

Part I: Bias mitigation...

Part II: Analogy based ML...

PART I: (Harmful) bias mitigation...

Harmful bias of ML Models

ML models: *designed* to have some bias that *guide* them in their tasks

Expected bias:

Credit card default prediction	(good) *credit payment history*	$\uparrow$
Hate speech prediction	(presence of) *offensive terms*	$\uparrow$

Harmful bias:

Credit card default prediction	*ethnicity* (minority)	$\downarrow$
Hate speech prediction	*language variant*	$\downarrow$

Harmful bias lead to **unfair** algorithmic decisions & **discrimination**

Discrimination: "unjust or prejudicial treatment of different categories of people, especially, on the grounds of race, age, or sex"

Motivation: unfair algorithmic decisions

COMPAS[1] (Tabular data)

Chatbot Tay[2] (Text)

Other Critical applications of algorithmic decisions: loan requests, job applications, Stop & Frisk, etc.

Need of fairness: Unfair outcomes not only affect human rights, but they undermine public trust in ML & AI.

[1] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[2] https://www.bbc.com/news/technology-35902104

Addressing "fairness" in ML...

Based on **decision outcomes**, fairness can be assessed through:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's reliance on "sensitive features" (e.g., salient features such as race, age, or sex,. . .)

Two main approaches to tackle ML unfairness:

- Enforce fairness constraints while learning, e.g.:

$$P(y_{\text{pred}} \neq y_{\text{true}} | race = Black) = P(y_{\text{pred}} \neq y_{\text{true}} | race = White)$$

  Drawback: Complexity, "fairness overfitting"

- Exclude sensitive/salient features

  Drawback: Decreased accuracy!

Addressing "fairness" in ML...

Based on **decision outcomes**, fairness can be assessed through:

- **Fairness metrics**: individual & group fairness, equal opportunity, demographic parity, equal accuracy, etc.
- **Process fairness**: model's reliance on "sensitive features" (e.g., salient features such as race, age, or sex,...)

Two main approaches to tackle ML unfairness:

- Enforce fairness constraints while learning, e.g.:

  $$P(y_{\text{pred}} \neq y_{\text{true}} | race = Black) = P(y_{\text{pred}} \neq y_{\text{true}} | race = White)$$

  Drawback: Complexity, "fairness overfitting"
- Exclude sensitive/salient features

  Drawback: Decreased accuracy!

Fairness through unawareness...

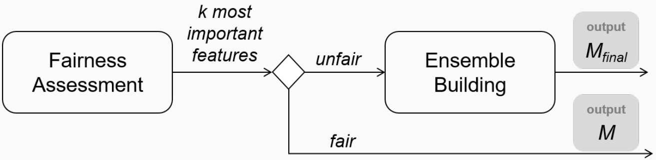
FixOut (FaIrness through eXplanations and feature dropOut)[3]

Goal: reduce model's dependence on sensitive/salient features while keeping (or improving) its performance

Fair Model: if its outcomes do not depend on sensitive features

FixOut: Human-centered approach to deal process fairness

Input: model M, dataset D, sensitive features F, explanation method E
Output: M if fair, otherwise a fairer and more accurate M_{final}



[3] https://fixout.loria.fr/

Example: FixOut with LIME (RF on German)[4]

Original

Feature	Contrib.
foreignworker	2.664899
otherinstallmentplans	-1.354191
housing	-1.144371
savings	0.984104
property	-0.648104
purpose	-0.415498
existingchecking	0.371415
telephone	0.311451
credithistory	0.263366
duration	-0.223288

Ensemble

Feature	Contrib.
otherinstallmentplans	-1.487604
housing	-1.089726
savings	0.679195
duration	-0.483643
foreignworker	0.448643
property	-0.386355
credithistory	0.258375
job	-0.252046
existingchecking	-0.21358
residencesince	-0.138818

Result: M_{final} is “fairer” & at least as accurate (from 0.783 to 0.786)

[4] Bhargava, et al. LimeOut: An Ensemble Approach to Improve Process Fairness. PKDD/ECML Workshop XKDD 2020: 475-491

Q: Impact of FixOut on w.r.t. fairness metrics?

What about Fairness metrics?

Idea: Separate instances into two groups w.r.t. a sensitive feature

E.g.: Non-white people (unprivileged) **versus** white people (privileged)

Demographic Parity (DP)[5] :

$DP = P(\hat{y} = pos|D = unp) - P(\hat{y} = pos|D = priv)$

Equal Opportunity (EO)[6]**:** $EO = \frac{TP_{unp}}{TP_{unp}+FN_{unp}} - \frac{TP_{priv}}{TP_{priv}+FN_{priv}}$

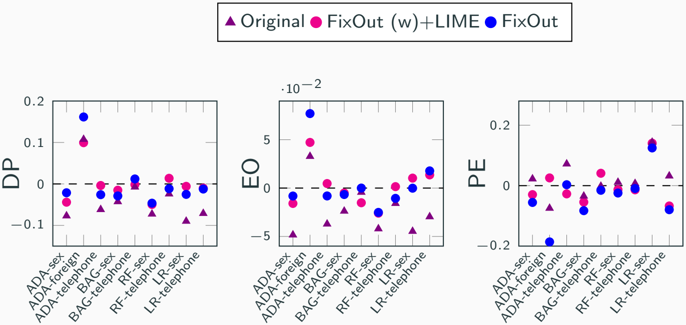
Predictive Equality (PE)[7]**:** $PE = \frac{FP_{unp}}{FP_{unp}+TP_{unp}} - \frac{FP_{priv}}{FP_{priv}+TP_{priv}}$

[5] Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data, 2017, 153–163.*

[6] Zafar, *et al.* Fairness beyond disparate treatment & impact: Learning classification without disparate mistreat. WWW 2017.

[7] Alves, *et al.* Making ML models fairer through explanations: the case of LimeOut. AIST 2020.

Fairness metrics



German dataset: Privileged groups

- “status sex”: “male single”
- “telephone”: “yes” (registered under the customers name)
- “foreign worker”: “no”

Conclusion

FixOut: A human-centered approach to mitigate harmful bias

1. On tabular data:

   Use of explanations and Control through aggregation

   Automated the choice of the most important features to be considered in the fairness assessment

2. On textual data:

   How to adapt *feature dropout* to *bag of words*.

   Reduced unintended bias of ML models on textual data[8]

3. Adaptation to neural classifiers:

   Feature dropout on the representations (embeddings)

[8] Alves, *et al.* Reducing Unintended Bias of ML Models on Tabular and Textual Data. DSAA 2021: 1-10.

Perspectives

FixOut: currently in **startup pre-maturation** (SATT & Incubateur Lorrain) to be followed by INRIA Startup Studio

Further results: Statistical approach (Hilbert-Schmidt IC) to detecting sensitive attributes (data-driven)[9]

Refine-LM: Reinforcement Learning for Harmful Bias Mitigation in LL Models

Further results: Portable bias filtering mechanism that is

- easy to train,
- adjustable to a multiple language models,
- adaptable to different bias contexts (gender, ethnicity, religion, etc.)

[9] Pelegrina, *et al.* A statistical approach to detect sensitive features in a group fairness setting. CoRR abs/2305.06994 (2023).

Perspectives

FixOut: currently in **startup pre-maturation** (SATT & Incubateur Lorrain) to be followed by INRIA Startup Studio

Further results: Statistical approach (Hilbert-Schmidt IC) to detecting sensitive attributes (data-driven)[9]

Refine-LM: Reinforcement Learning for Harmful Bias Mitigation in LL Models

Further results: Portable bias filtering mechanism that is

- easy to train,
- adjustable to a multiple language models,
- adaptable to different bias contexts (gender, ethnicity, religion, etc.)

[9] Pelegrina, *et al.* A statistical approach to detect sensitive features in a group fairness setting. CoRR abs/2305.06994 (2023).

Part II: Analogy based ML...

Example: Analogical proportion (a is to b as c is to d)

(a)  : (b)  :: (c)  : (d) 

Analogies simultaneously exploit similarities and dissimilarities

3 key cognitive processes: Abstraction, Inference and Creativity

Detecting/mining analogies: Given a, b, c, and d,

- is (a, b, c, d) a valid analogy?

Solving analogies: Given a, b, c,

- find x **s.t.** (a, b, c, x) a valid analogy

Reasoning and integrating analogical reasoning (AR):

- Depending on the concrete application and ML&AI task

Example: Analogical proportion (a is to b as c is to d)

(a)  : (b)  :: (c)  : (d) 

Analogies simultaneously exploit similarities and dissimilarities

3 key cognitive processes: Abstraction, Inference and Creativity

Detecting/mining analogies: Given a, b, c, and d,

- is (a, b, c, d) a valid analogy?

Solving analogies: Given a, b, c,

- find x **s.t.** (a, b, c, x) a valid analogy

Reasoning and integrating analogical reasoning (AR):

- Depending on the concrete application and ML&AI task

Different views on analogies

Axiomatic: As 4-ary relations satisfying certain postulates
Examples: reflexivity, (certain) permutations, etc.

Relational: $R(a, b, c, d) \equiv P(P_1(a, b), P_1(c, d))$, for P, P_1 predicates
Example: $R(\mathit{wine}, \mathit{France}, \mathit{beer}, \mathit{Germany})$

Functional: $R(a, b, c, d)$ if $b = T(a)$ and $d = T(c)$, for some T
Example: $R(\mathit{go}, \mathit{went}, \mathit{make}, \mathit{made})$

Model Theoretic: Relying on structural transformations and "rewriting"
Examples: *Structure mapping theory* and *Justifications*

NB: Different ways to define analogies depending on the data, the underlying structure and **the task** at hand...

Different views on analogies

Axiomatic: As 4-ary relations satisfying certain postulates
Examples: reflexivity, (certain) permutations, etc.

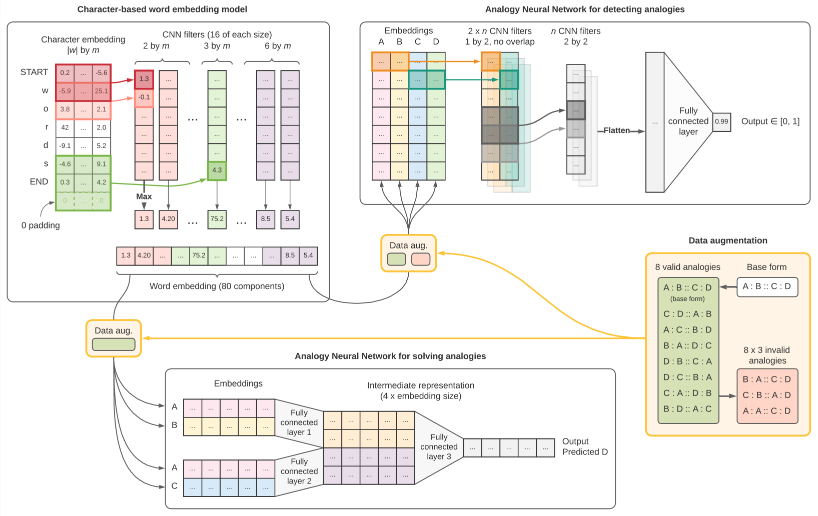
Relational: $R(a, b, c, d) \equiv P(P_1(a, b), P_1(c, d))$, for P, P_1 predicates
Example: $R(\mathit{wine}, \mathit{France}, \mathit{beer}, \mathit{Germany})$

Functional: $R(a, b, c, d)$ if $b = T(a)$ and $d = T(c)$, for some T
Example: $R(\mathit{go}, \mathit{went}, \mathit{make}, \mathit{made})$

Model Theoretic: Relying on structural transformations and "rewriting"
Examples: *Structure mapping theory* and *Justifications*

NB: Different ways to define analogies depending on the data, the underlying structure and **the task** at hand...

Example: detecting and solving morphological analogies



Recently: solving morphological analogies through generation [C22b]

ANNa: `https://anna.loria.fr/`

Applications in ML & AI

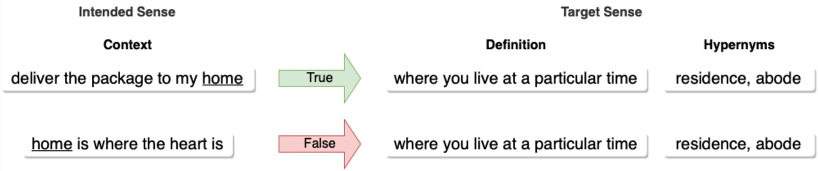
Some application domains

- NLP & translation [L03, S05,M20,A21a,M22,C22]
- Classification & recom. [B07,B17, C17,C18,C20b,C23b]
- CB & Machine reasoning [F89,G83,L21,L19a,L19b,L21,M21]
- Transfer learning [B19,C20a, A21b,F23,M23]
- VisualQA, ScholasticAP, TSV, Explainability [S15,P19,Z22,H20]
- ...and even humor: pun and meme generation (WIP)

Target Sense Verification (TSV)[10]

Here: Target Sense Verification (TSV)

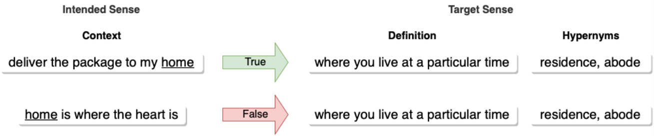
PhD: Georgios Zervakis (defended March, 2023)
Enriching large language models with semantic lexicons and analogies



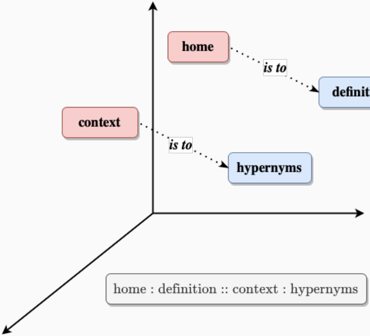
Question: Is the *intended sense* of home the same as in the *target sense*?

[10] Breit et al. (2021) WiC-TSV: An evaluation benchmark for target sense verification of words in context. EACL

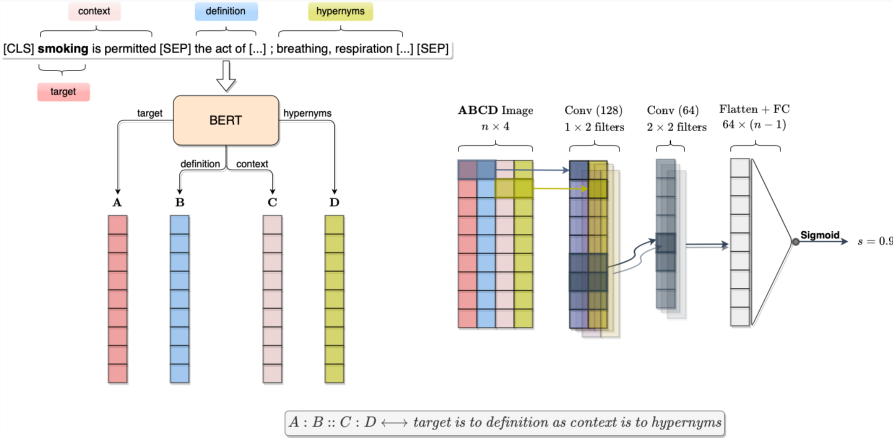
TSV as analogy detection



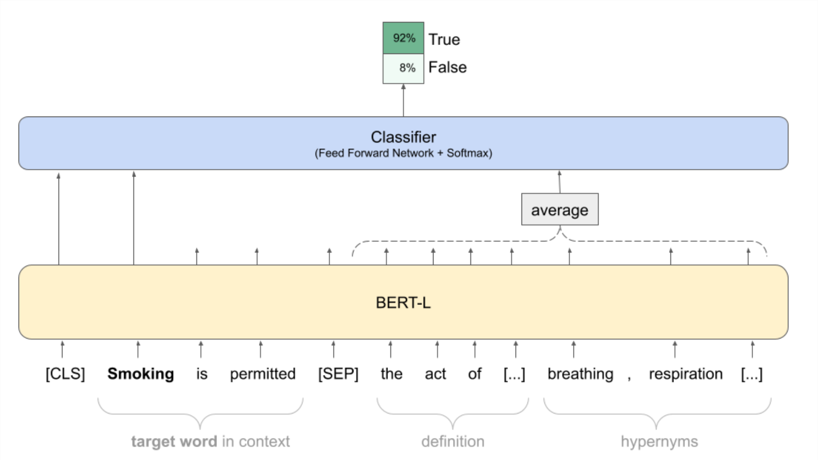
Idea: Formulate the question as an analogy and check whether it is valid.



AB4TSV architecture



Baselines (Breit et al.: WiC-TSV 2021)



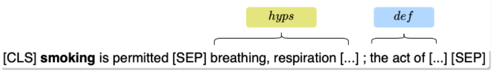
BERT + [CLS] + target word + average(defnition, hypernyms) $\implies$ Classifier

Choice of input encoding and analogy relation

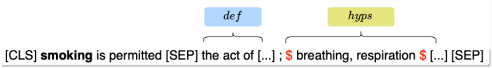
default:

Alternative input encoding operations

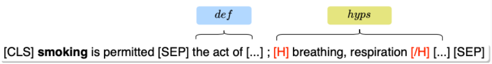
swap:



fc[11]**:**



em[12]**:**

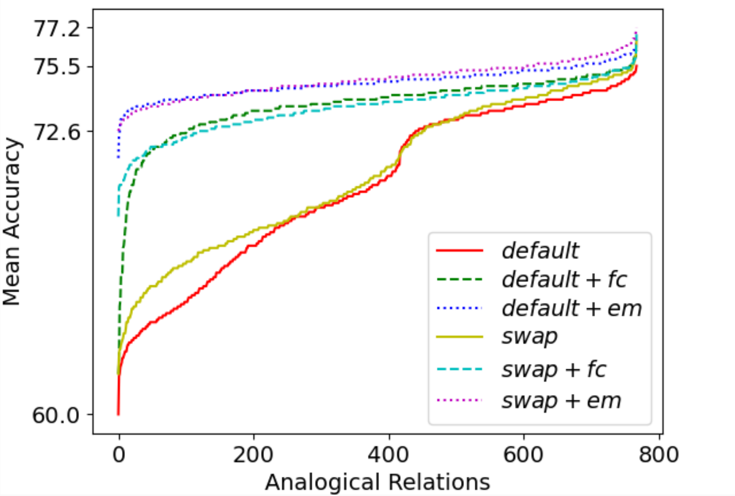


[11] Huang et al. (2019) GlossBERT: BERT for word sense disambiguation with gloss knowledge. EMNLP-IJCNLP.

[12] Baldini Soares et al. (2019) Matching the blanks: Distributional similarity for relation learning. ACL.

Impact of input encoding

Mean accuracy: 6 input encodings × 768 analog. relations × 4 random seeds



Analogical relation optimization results

Encoding	Analogy	Dev *Acc*	Dev *F*1
default	*cls* : *descr* :: *cls* : *ctx*	74.5 ± 0.015	77.0 ± 0.016
default+fc	*cls* : *def* :: *ctx* : *cls*	74.9 ± 0.010	77.3 ± 0.006
default+em	*tgt* : *descr* :: *cls* : *def*	75.4 ± 0.027	**77.8** ± 0.023
swap	*def* : *cls* :: *cls* : *ctx*	75.4 ± 0.016	77.7 ± 0.016
swap+fc	*def* : *ctx* :: *cls* : *hyps*	**75.8** ± 0.013	77.7 ± 0.013
swap+em	*hyps* : *def* :: *cls* : *ctx*	**75.8** ± 0.017	77.7 ± 0.012
	Baselines		
default		74.0 ± 0.014	76.9 ± 0.007
default+fc		73.9 ± 0.018	76.3 ± 0.018
default+em	HyperBert3	73.1 ± 0.031	75.2 ± 0.032
swap		73.8 ± 0.015	76.3 ± 0.015
swap+fc		73.5 ± 0.011	75.6 ± 0.013
swap+em		74.4 ± 0.011	75.7 ± 0.024

NB1: AB4TSV >> Baselines **NB2:** [CLS] token matters

Comparative results

Approach	Test *Acc*	Test *F*1
CTLR[13]	78.3	78.5
V[14]	71.9	76.2
BERT_{Base}[15]	76.6	78.2
BERT_{Large}[15]	76.3	77.8
FastText[15]	53.4	63.4
AB4TSV+swap+em	75.7	77.5
AB4TSV+swap+fc	**78.6**	**79.8**
AB4TSV_{pi}+default+em	**78.6**	79.4
U-dBERT[15]	61.2	51.3
U-BERT[15]	60.5	51.9
MIRRORWIC[16]	73.7	—

[13] Moreno et al. (2021) CTLR@WiC-TSV
[14] Vandenbussche et al. (2021) SemDeep-6
[15] Breit et al. (2021) WiC-TSV
[16] Liu et al. (2021) MirrorWiC

Conclusions

Analogy and BERT for target sense verification (AB4TSV)

1. Combining LLMs with analogy classifiers improves results on TSV
2. Marking the input text with special characters can boost the performance.
3. Integrating the properties of analogies offers gains in interpretability.
4. AB4TSV shows OOD generalization and transfer learning capabilities.

ANR AT2TA 2023-2026 (ANR-22-CE23-0023)

AT2TA

Analogies: from Theory to Tools and Applications

https://at2ta.loria.fr/

PRCE Axis E.2, CES 23: Intelligence artificielle et science des données

Partners:

Loria — Laboratoire lorrain de recherche en informatique et ses applications

IRIT

Inria

institut imagine — GUÉRIR LES MALADIES GÉNÉTIQUES

INFOLOGIC ENGINEERING

orange™

General objective and Challenges

AT2TA General Objective: propose an ML framework that integrates analogical reasoning (AR), easily adaptable to different real use cases.

(C1) Bridging the gap between ML and KRR

(C2) Analogy modeling and representation learning for AR

(C3) AR adaptation across domains

(C4) Platform for multimodal/multi-domain AR

Some actions and events

Events & dissemination:

- Annual workshops with proceedings:
  IARML@IJCAI-ECAI 2022 & ATA@ICCBR 2022 (both published)
  IARML@IJCAI 2023 & ATA@ICCBR 2023 (upcoming!)
- Springer special issues (yearly):
  Analogies: from Mathematical Foundations to Applications and Interactions with ML and AI (S722: Analogical reasoning)
  in Annals of Mathematics and Artificial Intelligence (AMAI)
- Shared Tasks (upcoming)
- Other actions (to discuss)

Merci de votre attention!

Thank you for your attention!

Selected References

[A22] Ch. Antić. Analogical proportions. AMAI 90:6 (2022) 595–644

[A21a] S. Alsaidi, *et al.* A neural approach for detecting morphological analogies. DSAA21, 1-10

[A21b] S. Alsaidi, *et al.* On the transferability of neural models of morphological analogies. *AIMLAI@ECML–PKDD21*, 76–89

[B23] F.Badra, *et al.* Some Perspectives on Similarity Learning for Case-Based Reasoning and Analogical Transfer. To appear.

[B19]b Z. Bouraoui, *et al.* From Shallow to Deep Interactions Between Knowledge Representation, Reasoning and Machine Learning (Kay R. Amel group), 2019. `http://arxiv.org/abs/1912.06612`

[B17] M. Bounhas, *et al.* Analogy-based classifiers for nominal or numerical data. IJAR 91 (2017) 36–55

[B07] S. Bayoudh, *et al.* Learning by analogy: A classification rule for binary and nominal data. *IJCAI07*, 678–683

[C23] M. Couceiro, E. Lehtonen. Galois theory for analogical classifiers. In press AMAI, 2023. 21 p.

[C22] K. Chan, *et al.* Solving Morphological Analogies Through Generation. IARML@IJCAI-ECAI 2022, 29–39

[C20a] A. Cornuéjols,*et al.* Transfer learning by learning projections from target to source. *IDA20*, 119–131

[C20b] M. Couceiro, *et al.* When nominal analogical proportions do not fail. *SUM20*, 68–83

[C18] M. Couceiro, *et al.* Behavior of analogical inference w.r.t. Boolean functions. *IJCAI18*, 2057–2063

[C17] M. Couceiro, *et al.* Analogy-preserving functions: A way to extend Boolean samples. *IJCAI17*, 1575–1581

[G83] D. Gentner. Structure-Mapping: A Theoretical Framework for Analogy. Cognitive Science 7(2) (1983) 155–170

[H16] N. Hug, *et al.* Analogical classifiers: A theoretical perspective. *ECAI16*, 689–697

[H20] E. Hüllermeier. Towards Analogy-Based Explanations in ML. CoRR abs/2005.12800, 2020

Selected References (cont.)

[L21] J. Lieber, *et al.* When revision-based case adaptation meets analogical extrapolation. *ICCBR21*, 156–170

[L19a] J. Lieber, *et al.* Improving analogical extrapolation using case pair competence. *ICCBR19*, 251–265

[L19b] S. Lim, *et al.* Solving word analogies: A machine learning perspective. *ECSQARU19*, 238–250

[L03] Y. Lepage. De l'analogie rendant compte de la commutation en linguistique. HdR Université Joseph Fourier - Grenoble I, 2003

[M23] E. Marquer, *et al.* Less is Better: An Energy-Based Approach to Case Base Competence. To appear.

[M22] E. Marquer, *et al.* A Deep Learning approach to solving morphological analogies. *ICCBR22*, 159–174

[M21] M. Mitchell. Abstraction and analogy-making in artificial intelligence, 2021

[M20] P.-A. Murena, *et al.*. Solving analogies on words based on minimal complexity transformation. *IJCAI20*, 1848–1854

[P19] J. Peyre, *et al.* Detecting unseen visual relations using analogies. *ICCV19*

[S05] N. Stroppa, *et al.* Analogical learning and formal proportions: Definitions and methodological issues. Technical Report D004, ENST-Paris, 2005

[Z22] G. Zervakis, *et al.* An Analogy based Approach for Solving Target Sense Verification. NLPIR 2022