Proceedings of the 2nd Workshop on the Interactions between Analogical Reasoning and Machine Learning

Miguel Couceiro, Pierre-Alexandre Murena, Stergos Afantenos

Proceedings

IJCAI Workshop

"Interactions between Analogical Reasoning and Machine Learning"

IARML 2023

August 21, 2023

Macau, China

Editors

Miguel Couceiro (University of Lorraine, CNRS, Loria)

Stergos Afantenos (Université Paul Sabatier, IRIT)

Pierre-Alexandre Murena (Hamburg University of Technology)

http://iarml2023-ijcai.loria.fr/

Preface

This volume contains the proceedings of the 2nd edition of the workshop IARML@IJCAI. The first edition took place at IJCAI-ECAI 2022, Vienna, Austria[1] that counted with the participation of several colleagues from Europe, America and Asia. This year we expanded our audience to the 5 continents. As in the 1st edition, we will organize a Springer special volume in *Annals of Mathematics and Artificial Intelligence*.

Analogical reasoning is a remarkable human capability used to solve hard reasoning tasks. It consists in transferring knowledge from a source domain to a different, but somewhat similar, target domain by relying simultaneously on similarities and differences. Analogies have preoccupied humanity at least since antiquity (cf. the works of Atistotle, Theon of Smyrna, among others) and have been in more recent years characterized as being "at the core of cognition" (Hofstadter 2001) showing that they permeate almost every aspect of cognition (Hofstadter and Sanders, 2013). According to Hofstadter and the Fluid Analogies Research Group, analogy making is intimately related with abstraction and the search of a "common essence", which can lead to deep understanding of any concept or situation.

Analogies have been tackled from various angles. Traditionally, *analogical proportions*, i.e., statements of the form "A is to B as C is to D", are the basis of analogical inference. They contributed to *case-based reasoning* and to multiple *machine learning* tasks such as classification, decision making and machine translation with competitive results. Also, analogical extrapolation can support dataset augmentation (analogical extension) for model learning, especially in environments with few labeled examples. Other approaches include the *Structure Mapping* approach of Dedre Gentner that is based on logical descriptions (in the form of predicate-argument structures) of two domains: the more relational similarity one has between the two domains, the more analogous they can be considered.

Recent neural techniques, such as representation learning, enabled efficient approaches to detecting and solving analogies in domains where symbolic approaches had shown their limits. Transformer architectures trained using vast amounts of data have given us Large Language Models (LLMs) such as Chat-GPT, seem to exhibit human-like conversational and analogy making capacities (Webb et al. 2022). However, better evaluation metrics are needed in order to measure elusive concepts such as intelligence and understanding (Mitchel 2023). More than ever we need to understand the role that analogies, abstraction and similarities between concepts play in language and cognition.

The purpose of this series of workshops is to bring together AI researchers at the cross roads of machine learning, natural language processing, knowledge representation and reasoning, who are interested in the various applications of analogical reasoning in machine learning or, conversely, of machine learning techniques to improve analogical reasoning.

The contributions to this 2nd edition of IARML@IJCAI focused on the following:

- Machine learning for analogical reasoning: representation learning, Advanced similarity measures, analogical transfer, neuro-symbolic models for analogical inference.
- Analogical reasoning for machine learning: classification using analogical reasoning, case-Based Reasoning, creativity and data augmentation.
- Analogies in Large Language Models (LLMs): probing LLMs for analogies, evaluating capacities of LLMs for analogies, creativity in language through analogies.

[1] `https://iarml2022-ijcai-ecai.loria.fr/`

- Applications: to visual domains, to Natural Language Processing, etc.

The workshop welcomed submissions of research papers on all topics at the intersection of analogical reasoning and machine learning. The submissions were subjected to a strict double-blind reviewing process that resulted in the selection of five original contributions and one invited talk, in addition to the two plenary keynote talks.

Plenary talks:

Accelerating Innovation and Discovery through Analogy Mining (Dafna Shahaf)

Similarity measures at the core of analogical transfer and case-based prediction (Marie-Jeanne Lesot)

Invited talks:

Multimodal Analogical Reasoning over Knowledge Graphs (Ningyu Zhang, Lei Li, Xiang Chen, Xiaozhuan Liang, Shumin Deng)

IARML@IJCAI'23 takes place on August 21, 2022 in Macau (China), and we are truly thankful to the IJCAI workshop chairs for their help in the organization of this event. We are greatly indebt to the scientific committee for their reviews and suggestions for improving the accepted contributions.

Miguel Couceiro
Stergos Afantenos
Pierre-Alexandre Murena

Organising Committee

Miguel Couceiro (University of Lorraine, CNRS, Loria, FR)

Stergos Afantenos (Université Paul Sabatier, IRIT, FR)

Pierre-Alexandre Murena (Hamburg University of Technology, DE)

Scientific Committee

Fadi Badra (Université Sorbonne Paris Nord, LIMICS, FR)

Nelly Barbot (Université de Rennes 1, IRISA, FR)

Tarek R. Besold (DEKRA DIGITAL, Eindhoven University of Technology, NL)

Myriam Bounhas (LARODEC-ISGT, TU, UAE)

Adrien Coulet (Inria Paris, FR)

Sebastien Destercke (CNRS, Université de Technologie de Compiegne, Heudiasyc, FR)

Claire Gardent (University of Lorraine, CNRS, LORIA, FR)

Eyke Hullermeier (University of Munich, DE)

Mehdi Kaytoue (Infologic, FR)

David B. Leake (Indiana University, USA)

Yves Lepage (Waseda University, JA)

Jean Lieber (University of Lorraine, CNRS, LORIA, FR)

Esteban Marquer (University of Lorraine, CNRS, LORIA, FR)

Laurent Miclet (Université de Rennes, FR)

Pierre Monnin (Orange, FR)

Amedeo Napoli (University of Lorraine, CNRS, LORIA, FR)

Henri Prade (CNRS, Université Paul Sabatier, IRIT, FR)

Irina Rabkina (OXY Occidental College, USA)

Steven Schockaert (Cardiff University, IR)

Table of Contents

Plenary Talks

Accelerating innovation and discovery through analogy mining

Dafna Shahaf

Abstract: Large repositories of products, patents and scientific papers offer an opportunity for building systems that scour millions of ideas and help users discover inspirations. However, idea descriptions are typically in the form of unstructured text, lacking key structure that is required for supporting creative innovation interactions. In this talk, we will discuss several recent works that explore how to support creative innovation with analogies and idea representations. We propose novel representations that automatically extract different kinds of useful structure from idea descriptions, and demonstrate how these representations can be used to support creative tasks such as ideation, functional search for ideas, and exploration of the design space around a focal problem.

Short Biography: Dafna Shahaf is an Associate Professor in computer science at the Hebrew University of Jerusalem. Prof. Shahaf's research uses digital traces of human activity to better understand human capacities such as humor and creativity, and to develop computer systems that can support these capacities. She received her PhD from Carnegie Mellon University, and was a postdoctoral fellow at Stanford University and at Microsoft Research. Prof. Shahaf has won multiple awards, including best research paper awards at KDD 2010 and KDD 2017, an ERC starting grant, IJCAI Early Career Award. a Microsoft Research Fellowship, a Siebel Scholarship, Wolf's Foundation Krill Award, as well as MIT Tech Review's "Most thought-provoking paper of the week".

Similarity measures at the core of analogical transfer and case-based prediction

Marie-Jeanne Lesot

Abstract: Case-based prediction applies the plausible inference principle of analogical transfer, according to which if two cases are similar with respect to some criteria, in particular in the situation space, then it is plausible that they are also similar with respect to other criteria, in particular in the outcome space. In a first part, the presentation will review some existing approaches to case-based prediction, distinguishing them according to the type of knowledge used to measure the compatibility between the two sets of similarity relations. In a second part, the presentation will discuss the very notion of similarity measure, highlighting their variety both for numerical and categorical descriptive features. It will finally present some equivalence results that allow to define a reduced number of similarity families, providing some guidance for their selection.

Short Biography: Marie-Jeanne Lesot is an associate professor in the Computer Science Lab of Sorbonne Université, LIP6, and a member of the Learning and Fuzzy Intelligent systems (LFI) group. Her research interests focus on fuzzy machine learning with an objective of data interpretation and semantics integration, within the eXplainable Artificial Intelligence framework; they include similarity measures, fuzzy clustering, linguistic summaries and information scoring. She is also interested in approximate reasoning and the use of non classical logics, in particular weighted variants with increased expressiveness that are close to natural human reasoning processes.

Formulae for the solution of an analogical equation between Booleans using the Sheffer stroke (NAND) or the Pierce arrow (NOR)

Yves Lepage

Waseda University, Hibikino 2-7, Kitakyushu, 808-0135, Japan

Abstract

This paper gives a formula for the solution of an analogical equation between Booleans using the Sheffer stroke (NAND). Naturally, a counterpart using the Pierce arrow (NOR) is also given. Although not so intuitive, these formulae are somewhat elegant. The formulae are obtained in the following way: a rapid review on analogies between sets is given. The result on sets is transposed to Booleans. This result is rewritten using solely the operators mentioned above and simplified.

Keywords

Boolean analogies, Analogies on sets, Sheffer stroke (NAND), Pierce arrow (NOR)

1. Introduction

An axiomatic approach (Section 2) that postulates reflexivity ($A : B :: A : B$) and symmetry ($C : D :: A : B$) of conformity (::), in addition to the exchange of the means ($A : C :: B : D$), for any analogy $A : B :: C : D$, allows to define analogy on commutative magmas and commutative monoids (Section 3). The additional postulate of contiguity (the same analogy should hold on the inverse of objects) allows to define analogies on commutative groups (Section 4). Adding the postulate of similarity (all features in A should appear in B or C) is used to determine the solution of analogical equations between sets in [1] (Section 5). With all the above, the analogy induced by (a) the structure of the commutative groups $(\mathcal{P}(E), \triangle)$ or $(\mathcal{P}(E), \triangledown)$[1] is the same as the analogy induced by (b) the two monoids $(\mathcal{P}(E), \cup)$ and $\mathcal{P}(E), \cap)$ holding at the same time, under the condition

$$A \subset B \cup C \quad \wedge \quad B \cap C \subset A \tag{1.1}$$

(Section 5). This condition eliminates two cases of discrepancy between the analogies induced by (a) and (b). The solution D of an analogy between sets $A : B :: C : D$ is then:

$$D = ((B \cup C) \setminus A) \cup (B \cap C). \tag{1.2}$$

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China

yves.lepage@waseda.jp (Y. Lepage)

0000-0002-3059-4271 (Y. Lepage)

[1] $\triangle$ for symmetrical difference on sets (corresponding to XOR on Booleans), and $\triangledown$ for its counterpart corresponding to logical equivalence on Booleans.

		General	Magma	Group
	Definition		$A \star D = C \star B$	$A \star B^{-1} = C \star D^{-1}$
(0)	Reflexivity of conformity	$A : B :: A : B$	$A \star B = A \star B$	$A \star B^{-1} = A \star B^{-1}$
For any $A : B :: C : D$			$A \star D = C \star B$	$A \star B^{-1} = C \star D^{-1}$
(I)	Symmetry of conformity	$C : D :: A : B$	$C \star B = A \star D$	$C \star D^{-1} = A \star B^{-1}$
(II)	Inversion of ratios	$B : A :: D : C$	$B \star C = D \star A$	$B \star A^{-1} = D \star C^{-1}$
(III)	Inversion of objects (contiguity)	$A^{-1} : B^{-1} :: C^{-1} : D^{-1}$	undefined	$A^{-1} \star B = C^{-1} \star D$
(IV)	Distribution in objects (similarity)	any feature in A must appear in either B or C or both.	undefined	undefined
(V)	Exchange of the extremes	$D : B :: C : A$	$D \star A = C \star B$	$D \star B^{-1} = C \star A^{-1}$
(VI)	Exchange of the means	$A : C :: B : D$	$A \star D = B \star C$	$A \star C^{-1} = B \star D^{-1}$

Table 1
Postulates for analogy. The last two columns transcribe the definitions to the analogy naturally induced by the structures of a magma and a group.

The purpose of this paper is to transcribe (1.2) to analogy between Booleans (Section 6). As the Sheffer stroke (Section 7) is known to be functionally complete, the formulation uses only this operator (Section 10). The same is done with the Pierce arrow (Section 8).

2. Postulates for analogy

The classical way of writing down an analogy with $A : B :: C : D$ involves two basic articulations denoted by the signs : for ratio and :: that we choose to call conformity[2]. The four terms are traditionally divided into the means B and C, and the extremes A and D. Studies in the notion of analogy in its technical sense (not in its vernacular sense of mere similarity or comparison, as in analogical reasoning) extract two underlying notions, those of similarity and contiguity.

Conformity can be postulated to be reflexive and symmetric.[3]. The ratios can be thought to be inversible[4]. From the Greek antiquity, it is considered that analogy (in its strict technical meaning) cannot go without the exchange of the means[5]. All this leads to the postulates given in Table 1.

[2]The character : (U+2236) is named ratio in the ISO 10646 standard (Unicode) and :: (U+2237) is named proportion.
[3]I.e., a dependency relation. An equivalence relation requires transitivity in addition.
[4]*Invertendo* in the Latin tradition.
[5]*Permutando* or *alternando* in the Latin tradition.

Analogy		Corners of the square		
Transformation	Equivalent form	Transformation		D_8
identity	$A : B :: C : D$	identity	$\begin{matrix} A & B \\ C & D \end{matrix}$	e
counter-clockwise rotation	$B : D :: A : C$	rotation by $\pi/2$	$\begin{matrix} B & D \\ A & C \end{matrix}$	a
inverse of reading	$D : C :: B : A$	rotation by $2\pi/2 = \pi$	$\begin{matrix} D & C \\ B & A \end{matrix}$	a^2
clockwise rotation	$C : A :: D : B$	rotation by $3\pi/2 = -\pi/2$	$\begin{matrix} C & A \\ D & B \end{matrix}$	a^3
exchange of the means	$A : C :: B : D$	symmetry first diagonal	$\begin{matrix} A & C \\ B & D \end{matrix}$	x
inversion of ratios	$B : A :: D : C$	symmetry vertical axis	$\begin{matrix} B & A \\ D & C \end{matrix}$	ax
exchange of the extremes	$D : B :: C : A$	symmetry second diagonal	$\begin{matrix} D & B \\ C & A \end{matrix}$	a^2x
symmetry of conformity	$C : D :: A : B$	symmetry horizontal axis	$\begin{matrix} C & D \\ A & B \end{matrix}$	a^3x

Table 2
Bijection between the eight equivalent forms of an analogy and the eight elements of the dihedral group D_8, i.e., the transformations of the corners of the square.

Consecutive applications of (I), (II), (V) or (VI) in any number and in any order lead to only eight equivalent forms of the same analogy [2] which correspond to the eight possible transformations of the corners of a square, known as the dihedral group D_8 where the internal operation is composition. This bijection is given in Table 3. In the dihedral group, the choice of the two distinguished elements, a and x among the seven non-identity elements, is not totally free. The possible choices, expressed for analogy, are visualized in Figure 1. (II) Inversion of ratios and (VI) Exchange of the means is a possible choice. (I) Symmetry of conformity and (VI) Exchange of the means is another possible choice. For this last choice, it means that the postulates (II) and (V) are indeed dispensable.

3. Analogy induced on commutative magmas and monoids

Let $(\mathcal{E}, \star)$ be a magma, i.e., a set equipped with an internal law, without any specific property. To define analogy on such a structure, the only device offered is its internal operation. Drawing a parallel with numbers, where, for arithmetic and geometric analogies, one has $a + d = b + c$ and $a \times d = b \times c$, it is natural to posit the following equivalence to induce analogy from a

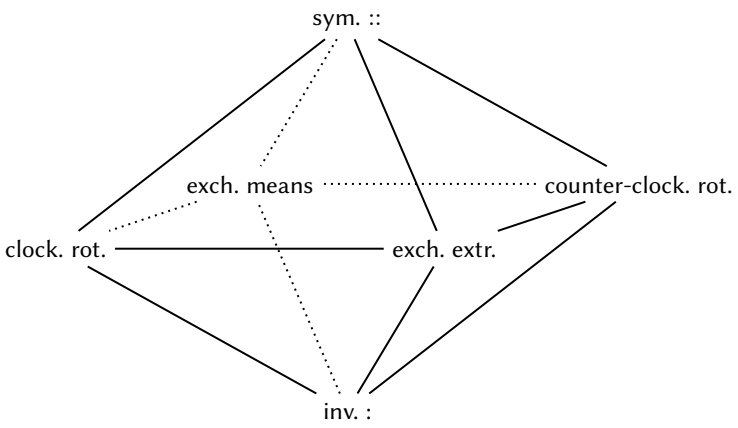


Figure 1: Any edge in the picture is a possible choice for a pair of transformations of analogy from which to get the eight equivalent forms from any analogy $A : B :: C : D$. This corresponds to selecting the elements usually denoted by a and x in the dihedral group D_8.

magma. However, observe that there are two possibilities, because the internal operation could be non-commutative.

$$\forall (A, B, C, D) \in \mathcal{E}^4, \quad A : B :: C : D \quad \overset{\text{def}}{\iff} \quad A \star D = B \star C \tag{3.1}$$

$$\text{or} \quad \forall (A, B, C, D) \in \mathcal{E}^4, \quad A : B :: C : D \quad \overset{\text{def}}{\iff} \quad A \star D = C \star B \tag{3.2}$$

With (3.1), the axiom of reflexivity of conformity would impose immediately that $\star$ be commutative because

$$\forall (A, B) \in \mathcal{E}^2, \qquad A : B :: A : B \quad \Leftrightarrow \quad \forall (A, B) \in \mathcal{E}^2,\ A \star B = B \star A.$$

For (3.2), the expression of each postulate is shown in Table 1. (0) and (1) hold because, equality being an equivalence relation, it is a dependency relation. The inverse of objects and the distribution in objects are left undefined. For all other axioms, a sufficient condition for them to be met is that $\star$ be commutative.

To summarize, to naturally induce analogy from the structure of a magma, it suffices for the internal operation to be commutative. The two definitions (3.1) and (3.2) are then the same. The axioms of object inversion and distribution within objects can be left unspecified. Observe that neither conformity nor ratio are directly defined. Finally, nothing can be said in the general case for the problem of solving an analogical equation on a commutative magma: given a triplet $(A, B, C) \in \mathcal{E}^3$, find D such that $A : B :: C : D$, i.e., find D such that $A \star D = C \star B$,

On a commutative monoid, i.e., a magma with associativity of the internal operation and a neutral element, analogy can be naturally induced in the same way as for a commutative magma.

4. Analogy induced on commutative groups

Let $(\mathcal{E}, \star)$ be a group. Let a^{-1} denote the inverse element of a.

- Ratios can be defined directly:

$$\forall (A, B) \in \mathcal{E}^2, \quad A : B \stackrel{\text{def}}{=} A \star B^{-1}. \tag{4.1}$$

  Note that this definition of the ratio is very specific: the ratio between two elements of $\mathcal{E}$ is an element of $\mathcal{E}$. This is very different from the situation with magmas in which, generally speaking, we do not know what a ratio is.
- Conformity can be defined as equality.
- The definition of analogy can then be as follows:

$$\forall (A, B, C, D) \in \mathcal{E}^4, \qquad A : B :: C : D \quad \stackrel{\text{def}}{\Longleftrightarrow} \quad A \star B^{-1} = C \star D^{-1}. \tag{4.2}$$

The column marked Group in Table 1 gives the expression of each of the postulates using (4.2). Similarly as for magmas, conformity being equality, reflexivity and symmetry hold. Postulating the axiom of inversion of objects, i.e.,

$$\forall (A, B, C, D) \in \mathcal{E}^4, \qquad A : B :: C : D \quad \Leftrightarrow \quad A^{-1} : B^{-1} :: C^{-1} : D^{-1}, \tag{4.3}$$

has the consequence that A can be expressed in two ways in function of the other terms.

$$\begin{aligned} & A : B :: C : D \\ \Leftrightarrow\ & A \star B^{-1} = C \star D^{-1} \\ \Leftrightarrow\ & A = C \star D^{-1} \star B \end{aligned}$$

$$\begin{aligned} & A^{-1} : B^{-1} :: C^{-1} : D^{-1} \\ \Leftrightarrow\ & A^{-1} \star (B^{-1})^{-1} = C^{-1} \star (D^{-1})^{-1} \\ \Leftrightarrow\ & A^{-1} \star B = C^{-1} \star D \\ \Leftrightarrow\ & B^{-1} \star A = D^{-1} \star C \\ \Leftrightarrow\ & A = B \star D^{-1} \star C \end{aligned}$$

Commutativity on the entire group is sufficient to ensure the equality

$$A = B \star D^{-1} \star C = C \star D^{-1} \star B. \tag{4.4}$$

Hence, provided the group is commutative, the group structure entails all the axioms listed in Table 1 with the exception of the axiom of distribution in objects.

5. Analogy between sets

Let $\mathcal{E}$ be a set. The set of all subsets of $\mathcal{E}$ is noted $\mathcal{P}(\mathcal{E})$. Equipped with union, $(\mathcal{P}(\mathcal{E}), \cup)$ is a commutative monoid. Union is an internal operation in $\mathcal{P}(\mathcal{E})$ that is associative and commutative. The neutral element is $\emptyset$ ($\emptyset \cup A = A \cup \emptyset = A$). However there is no inverse element in general, i.e., for any set A in $\mathcal{P}(\mathcal{E})$, there is no set B such that $A \cup B = \emptyset$. Similarly, $(\mathcal{P}(\mathcal{E}), \cap)$ is a commutative monoid. The neutral element is $\mathcal{E}$.

The symmetrical difference on sets (noted $\triangle$ and corresponding to XOR on Booleans), and another operation noted $\triangledown$ (the counterpart of logical equivalence on Booleans) are defined as follows.

$$\forall (A,B) \in \mathcal{P}(\mathcal{E})^2, \quad A \triangle B = (A \cup B) \setminus (A \cap B) \tag{5.1}$$

$$A \triangledown B = \mathcal{E} \setminus (A \triangle B) \tag{5.2}$$

$(\mathcal{P}(\mathcal{E}), \triangle)$ is a commutative group. Symmetrical difference is an internal operation in $\mathcal{P}(\mathcal{E})$ that is associative and commutative. The neutral element is $\emptyset$ ($\emptyset \triangle A = A \triangle \emptyset = A$). The inverse element of any set A in $\mathcal{P}(\mathcal{E})$ is itself: $A \triangle A = \emptyset$. Similarly, $(\mathcal{P}(\mathcal{E}), \triangledown)$ is a commutative group, with $\mathcal{E}$ as the neutral element, and each element is its one inverse.

For any quadruple of sets in a power set $\mathcal{P}(\mathcal{E})$, if the two analogies induced by the two structures of commutative monoids $(\mathcal{P}(\mathcal{E}), \cup)$ and $(\mathcal{P}(\mathcal{E}), \cap)$ hold at the same time, then, the analogy induced by the structure of commutative group $(\mathcal{P}(\mathcal{E}), \triangle)$ holds too (and similarly for $(\mathcal{P}(\mathcal{E}), \triangledown)$).

$$\begin{aligned} A : B \overset{\cap}{::} C : D \ \wedge \ A : B \overset{\cup}{::} C : D \Leftrightarrow \ & (A \cap D) = (C \cap B) \ \wedge \ (A \cup D) = (C \cup B) \\ \Rightarrow \ & (A \setminus B) \cup (B \setminus A) = (C \setminus D) \cup (D \setminus C) \\ \Leftrightarrow \ & A \triangle B = C \triangle D \Leftrightarrow A : B \overset{\triangle}{::} C : D \\ \Leftrightarrow \ & A \triangledown B = C \triangledown D \Leftrightarrow A : B \overset{\triangledown}{::} C : D \end{aligned}$$

The second line above is only an implication. Now, the analogy induced by the structure of a commutative group of $(\mathcal{P}(\mathcal{E}), \triangle)$ (or, similarly, $(\mathcal{P}(\mathcal{E}), \triangledown)$) is the same as when the two analogies induced by the two commutative monoids $(\mathcal{P}(\mathcal{E}), \cup)$ and $(\mathcal{P}(\mathcal{E}), \cap)$ hold at the same time, under the condition $A \subset B \cup C \ \wedge \ B \cap C \subset A$. This is (1.1) given in the introduction. $A \subset B \cup C$ transcribes the postulate of distribution in objects (IV) for sets with the features being the elements. $B \cap C \subset A$ is obtained by taking the set complements, i.e., using the postulate of inversion of objects (III).

$$\begin{aligned} & A : B \overset{\triangle}{::} C : D \Leftrightarrow \\ & A : B \overset{\triangledown}{::} C : D \Leftrightarrow A \triangle B = C \triangle D \\ & \qquad \Leftrightarrow (A \setminus B) \cup (B \setminus A) = (C \setminus D) \cup (C \setminus D) \\ & \qquad \Leftrightarrow (A \cap D) = (C \cap B) \wedge (A \cup D) = (C \cup B) \\ & \qquad \Leftrightarrow A : B \overset{\cap}{::} C : D \ \wedge \ A : B \overset{\cup}{::} C : D \end{aligned}$$

Table 3 gives the explicit development of this correspondence.

In [1], it was shown that, under the condition (1.1), the solution of an analogical equation $A : B :: C : D$ of unknown D between sets is given by (1.2).

A	B	C	D	(a) $A \cap D = C \cap B$	(b) $A \cup D = C \cup B$	analogy induced by both monoids: (a) $\wedge$ (b)	(c) $A \triangle B = \mathcal{E} \setminus (A \triangledown B)$	(d) $C \triangle D = \mathcal{E} \setminus (C \triangledown D)$	analogy induced by group: (c) = (d)
F	F	F	F	T	T	T	F	F	T
F	F	F	T	T	F	F	F	T	F
F	F	T	F	T	F	F	F	T	F
F	F	T	T	T	T	T	F	F	T
F	T	F	F	T	F	F	T	F	F
F	T	F	T	T	T	T	T	T	T
F	T	T	F	F	F	F	T	T	T
F	T	T	T	F	T	F	T	F	F
T	F	F	F	T	F	F	T	F	F
T	F	F	T	F	F	F	T	T	T
T	F	T	F	T	T	T	T	T	T
T	F	T	T	F	T	F	T	F	F
T	T	F	F	T	T	T	F	F	T
T	T	F	T	F	T	F	F	T	F
T	T	T	F	F	T	F	F	T	F
T	T	T	T	T	T	T	F	F	T

Table 3
Correspondence, on sets, between the two analogies induced by the commutative monoids $(\mathcal{P}(\mathcal{E}), \cup)$ and $(\mathcal{P}(\mathcal{E}), \cap)$ holding at the same time and each of the analogies induced by the commutative groups $(\mathcal{P}(\mathcal{E}), \triangledown)$ or $(\mathcal{P}(\mathcal{E}), \triangledown)$.

6. Analogies between Booleans

There exists a correspondence between operations on sets and operations on Booleans. Here we use the correspondence between *union* and *or*, *intersection* and *and*, and the fact that the complement of a set in another one corresponds to taking the conjunction with the negation: $A \setminus B$ corresponds to $a \wedge \neg b$. With this, the solution of an analogy between Booleans, $a : b :: c : d$, transcribed from the solution of an analogy between sets, under the condition (transcribed from the condition on sets) that

$$a \Rightarrow b \vee c \quad \wedge \quad b \wedge c \Rightarrow a, \tag{6.1}$$

is:

$$d = ((b \vee c) \wedge \neg a) \vee (b \wedge c). \tag{6.2}$$

The condition corresponds to the cases in conflict in [3] and [4], and identified in [1], i.e., the problem of accepting or not T : F :: F : T and F : T :: T : F as valid analogies. Transposed on sets, this is tantamount to ask whether $\{e_1, e_2\} : \{e_2\} :: \{e_3\} : \{e_1, e_3\}$ is a valid analogy. The

condition given above for sets rejects this analogy by keeping the natural interpretation of sets as containers.

7. The Sheffer stroke

The Sheffer stroke (usually noted $|$, but noted $\uparrow$ here[6]) denotes the NAND operator. For two Boolean variables p and q,

$$p{\uparrow}q = \neg(p \wedge q). \tag{7.1}$$

It is known that the singleton containing the Sheffer stroke as sole Boolean operator is functionally complete. This means that any Boolean expression can be rewritten using solely the Sheffer stroke. For instance,

$$p \wedge q = (p{\uparrow}q){\uparrow}(p{\uparrow}q), \tag{7.2}$$
$$p \vee q = (p{\uparrow}p){\uparrow}(q{\uparrow}q), \tag{7.3}$$
$$\neg p = p{\uparrow}p. \tag{7.4}$$

Intuitive operators are associative and commutative as is the case for $+$ or $\times$ on numbers. However, remarkably, the Sheffer stroke is commutative

$$p{\uparrow}q = q{\uparrow}p, \tag{7.5}$$

but not associative, i.e., in general

$$(p{\uparrow}q){\uparrow}r \neq p{\uparrow}(q{\uparrow}r). \tag{7.6}$$

By virtue of $p{\uparrow}p = \neg p$, trivially,

$$(p{\uparrow}p){\uparrow}(p{\uparrow}p) = \neg(\neg p) = p. \tag{7.7}$$

The notation p^2 for $p{\uparrow}p$ can be introduced, and applying it twice, reduces (7.7) to:

$$(p^2)^2 = p. \tag{7.8}$$

8. The Pierce arrow

The Pierce arrow is the NOR operator, i.e.,

$$p{\downarrow}q = \neg(p \vee q). \tag{8.1}$$

It has similar properties as the Sheffer stroke: it is commutative, but not associative, negation is obtained by self-application

$$\neg p = p{\downarrow}p, \tag{8.2}$$

[6] As in [5] and other works, we prefer $\uparrow$ over $|$ for symmetry reasons due to the use of the Pierce arrow $\downarrow$.

and any Boolean formula can be rewritten using it solely, i.e., alone, it is functionally complete. There is a kind of symmetry with the Sheffer stroke for the expression of conjunction and disjunction, due to the fact that they are dual[7]:

$$p \wedge q = (p{\downarrow}p){\downarrow}(q{\downarrow}q) \tag{8.3}$$
$$p \vee q = (p{\downarrow}q){\downarrow}(p{\downarrow}q). \tag{8.4}$$

The notation p^2 can be used with the Pierce arrow with the same meaning and same value as with the Sheffer stroke:

$$p^2 = \neg p = p{\uparrow}p = p{\downarrow}p. \tag{8.5}$$

Consequently, (7.8) also holds for the Pierce arrow.

9. Relations between the Sheffer stroke and the Pierce arrow

The following properties can easily be established by using the expression of disjunction for the two operators:

$$b^2{\uparrow}c^2 = (b{\downarrow}c)^2, \tag{9.1}$$
$$a^2{\uparrow}(b^2{\uparrow}c^2) = (a{\downarrow}(b{\downarrow}c))^2. \tag{9.2}$$

Rather than using p, q and r for variable names, we used a, b and c on purpose, to ease the reading of Section 10. The same can be done for conjunction:

$$b^2{\downarrow}c^2 = (b{\uparrow}c)^2, \tag{9.3}$$
$$a^2{\downarrow}(b^2{\downarrow}c^2) = (a{\uparrow}(b{\uparrow}c))^2. \tag{9.4}$$

10. Formulae for the solution of a Boolean analogy

The rewriting of the solution of an analogy between Booleans into an expression that involves only the Sheffer stroke can be worked out by hand from (6.2). It is safer to rely on a program to automatically perform this rewriting. We give such a program in Figure 2. It starts from a tree representation of (6.2), i.e., (6.2) in Polish notation.

The result is as follows, with spaces for clarity.

$$\begin{aligned} d = (\ (\ (\ (((b{\uparrow}b){\uparrow}(c{\uparrow}c)){\uparrow}(a{\uparrow}a))\ {\uparrow}\ (((b{\uparrow}b){\uparrow}(c{\uparrow}c)){\uparrow}(a{\uparrow}a))\ )\ {\uparrow}\\ (\ (((b{\uparrow}b){\uparrow}(c{\uparrow}c)){\uparrow}(a{\uparrow}a))\ {\uparrow}\ (((b{\uparrow}b){\uparrow}(c{\uparrow}c)){\uparrow}(a{\uparrow}a))\ )\ )\ {\uparrow}\\ (((b{\uparrow}c){\uparrow}(b{\uparrow}c)){\uparrow}((b{\uparrow}c){\uparrow}(b{\uparrow}c)))\ ) \end{aligned}$$

This lengthy formula can be simplified by

- locating occurrences of (7.8), i.e., $(p{\uparrow}p){\uparrow}(p{\uparrow}p) = p$,
- introducing the p^2 notation, and

[7]The dual f^d of an operator f is defined as follows [5]: $f^d(a_1, a_2, \ldots a_n) = (f(a_1^2, a_2^2, \ldots a_n^2))^2$.

```python
def and_(p, q):
    return f'(({p}↑{q})↑({p}↑{q}))'
def or_(p, q):
    return f'(({p}↑{p})↑({q}↑{q}))'
def not_(p):
    return f'({p}↑{p})'
def a():
    return 'a'
def b():
    return 'b'
def c():
    return 'c'

# d = '((b or c) and non(a)) or (b and c)'

d = or_( and_(or_(b(), c()), not_(a())), and_(b(),
↪  c()) )

print(d)
```

Figure 2: Program for automatic generation of the solution of an analogy between Booleans using the Sheffer stroke only.

- reestablishing the order of appearance of a, b and c by commutativity of the Sheffer stroke.

$$\begin{aligned} d &= (((b{\uparrow}b){\uparrow}(c{\uparrow}c)){\uparrow}(a{\uparrow}a)){\uparrow}(b{\uparrow}c) \\ &= (\ (b^2{\uparrow}c^2) \uparrow a^2\ ) \ \uparrow\ (b{\uparrow}c) \\ &= (\ a^2 \uparrow (b^2{\uparrow}c^2)\ ) \ \uparrow\ (b{\uparrow}c) \end{aligned} \tag{10.1}$$

For the Pierce arrow, the formula output by a similar program is as follows. Similarly, it can be simplified.

$$\begin{aligned} d = {} & ((((((b{\downarrow}c){\downarrow}(b{\downarrow}c)){\downarrow}((b{\downarrow}c){\downarrow}(b{\downarrow}c))) \downarrow ((a{\downarrow}a){\downarrow}(a{\downarrow}a))){\downarrow}((b{\downarrow}b){\downarrow}(c{\downarrow}c)))\ \downarrow \\ & ((((((b{\downarrow}c){\downarrow}(b{\downarrow}c)){\downarrow}((b{\downarrow}c){\downarrow}(b{\downarrow}c))) \downarrow ((a{\downarrow}a){\downarrow}(a{\downarrow}a))){\downarrow}((b{\downarrow}b){\downarrow}(c{\downarrow}c)))) \\ = {} & \big(\ ((((((b{\downarrow}c){\downarrow}(b{\downarrow}c)){\downarrow}((b{\downarrow}c){\downarrow}(b{\downarrow}c))) \downarrow ((a{\downarrow}a){\downarrow}(a{\downarrow}a))){\downarrow}((b{\downarrow}b){\downarrow}(c{\downarrow}c))\ \big)^2 \\ = {} & \big(\ ((((b{\downarrow}c){\downarrow}a){\downarrow}((b{\downarrow}b){\downarrow}(c{\downarrow}c)))\ \big)^2 \\ = {} & \big(\ ((b{\downarrow}c){\downarrow}a) \downarrow (b^2{\downarrow}c^2)\ \big)^2 \\ = {} & \big(\ (a{\downarrow}(b{\downarrow}c)) \downarrow (b^2{\downarrow}c^2)\ \big)^2 \end{aligned} \tag{10.2}$$

This second formula could have been obtained directly from (10.1) by exploiting the relations seen in Section 9, i.e., the duality between the two operators.

a	b	c	$b\uparrow c$	$b^2\uparrow c^2$	$a\uparrow(b\uparrow c)$	$a^2\uparrow(b^2\uparrow c^2)$	d in (10.1)	d in (10.3)
F	F	F	T	F	T	T	F	F
F	F	T	T	T	T	F	T	T
F	T	F	T	T	T	F	T	T
F	T	T	F	T	T	F	T	T
T	F	F	T	F	F	T	F	F
T	F	T	T	T	F	T	F	F
T	T	F	T	T	F	T	F	F
T	T	T	F	T	T	T	T	T

Table 4
True value tables of (10.1) and (10.3).

$$
\begin{aligned}
&(a^2\uparrow(b^2\uparrow c^2))\uparrow(b\uparrow c) && && (10.1)\\
&\quad = \quad (a\downarrow(b\downarrow c))^2\uparrow(b\uparrow c) && \text{by (9.2)} &\\
&\quad = \quad (a\downarrow(b\downarrow c))^2\uparrow(b^2\downarrow c^2)^2) && \text{by (9.3)} &\\
&\quad = \quad \big(\ (a\downarrow(b\downarrow c))\downarrow(b^2\downarrow c^2)\ \big)^2 && \text{by (9.1)} & (10.2)
\end{aligned}
$$

Remarkably, (10.1) is equivalent to the following formula, where the whole is squared and variables are squared.[8]

$$d = \big(\ (a\uparrow(b\uparrow c))\uparrow(b^2\uparrow c^2)\ \big)^2 \tag{10.3}$$

The equivalence between (10.1) and (10.3) is shown by the table of truth values for the two formulae in Table 4. The grayed-out lines are the two lines corresponding to the cases where condition (6.1) is not verified. In this table, the symmetry around the central line says that the value of d is negated by taking the negation of each of the variables a, b and c. This just states that, considered as an operator on three variables, the solution of an analogy is self-dual:

$$d(a^2, b^2, c^2) = d(a, b, c)^2.$$

This follows intuition as, d being the solution of an analogy, the postulate of inversion of objects (III) should hold. For the same reason, an equivalent form to (10.2) is:

$$d = (\ a^2\downarrow(b^2\downarrow c^2)\ )\ \downarrow\ (b\downarrow c) \tag{10.4}$$

Thus, remarkably, the formulae using the Pierce arrow (NOR) are the same as the ones using the Sheffer stroke (NAND). That is, (10.4) is the same as (10.1) and (10.2) is the same as (10.3), except for the operator.

[8]The submitted version of this paper contained a regrettable error in the justification of this equivalence. We fortunately became aware of it before the feedback of the reviewers, who, of course, spotted it. We thank one of them for suggesting a proof of this equivalence.

11. Conclusion

This paper gave formulae for the solution of an analogical equation between Booleans using solely the Sheffer stroke (NAND) or the Pierce arrow (NOR).

These formulae were obtained by transposing a formula on sets to Booleans. To justify this first formula, we reminded postulates for analogy and briefly showed how analogy can be induced from some algebraic structures (see also [6]). We then gave a rapid review on analogies between sets and stressed the fact that there is a discrepancy between analogy induced by union or intersection and analogy induced by symmetrical difference. Transposed to Booleans, this discrepancy tantamounts to ask whether T : F :: F : T and F : T :: T : F (by inversion of ratios (II)) should be considered valid analogies.

Although not so intuitive, the formulae for Booleans using the Sheffer stroke or the Pierce arrow are somewhat elegant. They reflect the self-duality of the solution of a Boolean analogical equation. Any of the two operators, Sheffer stroke or Pierce arrow, can indifferently be used for them. It is an open question whether these formulae are the most economical ones in terms of number of occurrences of operators or variables, i.e., whether their efficiency is the best possible [5].

12. Acknowledgments

This work has been supported in part by a research grant from JSPS Kakenhi Kiban C n° 21K12038 entitled "Theoretically founded algorithms for the automatic production of analogy test sets in NLP."

References

[1] Y. Lepage, De l'analogie rendant compte de la commutation en linguistique, Mémoire d'habilitation à diriger les recherches, Université de Grenoble, 2003.

[2] J. J. Rallier des Ourmes, Proportion, in: Encyclopédie ou Dictionnaire raisonné des sciences, des arts et des métiers, par une société de gens de lettres, volume XIII, Chez Briasson, David, Le Breton ou Durand, Paris, 1765, pp. 466a–467b.

[3] S. Klein, Culture, mysticism and social structure and the calculation of behavior, in: Proceedings of the European Conference on Artificial Intelligence (ECAI 1982), 1982, pp. 141–146.

[4] L. Miclet, H. Prade, Handling analogical proportions in classical logic and fuzzy logics settings, in: C. Sossai, G. Chemello (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty, Springer, Berlin, Heidelberg, 2009, pp. 638–650.

[5] M. Couceiro, E. Lehtonen, P. Mercuriali, R. Péchoux, On the efficiency of normal form systems for representing Boolean functions, Theoretical Computer Science 813 (2020) 341–361. URL: https://inria.hal.science/hal-02153506. doi:10.1016/j.tcs.2020.01.009.

[6] N. Stroppa, Définitions et caractérisation de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles, Thèse de doctorat, École nationale supérieure des télécommunications, 2005. URL: https://hal.archives-ouvertes.fr/tel-00145147/.

Some Perspectives on Similarity Learning for Case-Based Reasoning and Analogical Transfer

Fadi Badra[1,*,†], Marie-Jeanne Lesot[2,†], Esteban Marquer[3,†] and Miguel Couceiro[3,†]

[1]Université Sorbonne Paris Nord, Laboratoire d'Informatique Médicale et d'Ingénierie des Connaissances en e-Santé, LIMICS, Sorbonne Université, INSERM, F-93000, Bobigny, France

[2]Sorbonne Université, CNRS, LIP6, F-75005 Paris, France

[3] University of Lorraine, CNRS, Loria, Nancy, France

Abstract

In this paper we investigate interactions between recent advances in the modeling of analogical transfer and similarity learning. Indeed, a unifying principle of case-based prediction methods was recently established, according to which the plausible inference principle of analogical transfer can be interpreted as a transfer of similarity knowledge from a situation space to an outcome space. Following this principle, the task of analogical transfer can be addressed using a global indicator of the compatibility between two similarity measures. Such an indicator can also be used to assess the quality of the situation space similarity measure with respect to the case-based prediction task. We discuss several perspectives opened by such an interpretation of the task of analogical transfer as the optimisation of the compatibility criterion: we explore interactions with similarity learning, as well as with energy function optimisation.

Keywords

Case-Based Reasoning, Analogical transfer, Similarity learning, Quality measure

1. Introduction

Analogical transfer is a cognitive process that allows to derive some new information about a target situation by applying a plausible inference principle, according to which if two situations are similar with respect to some criteria, then it is plausible that they are also similar with respect to other criteria [1]. Case-based reasoning (CBR) systems implement analogical transfer in order to infer some information about a new situation directly by comparing it to a set of past experiences (called cases) stored in memory [2]. In that process, similarity knowledge is a critical component and is dependent on the task and data considered. For instance, several approaches have been proposed to measure similarities between data represented as Boolean vectors and between sequences in the context of analogical reasoning, as described in [3].

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China

badra@sorbonne-paris-nord.fr (F. Badra); Marie-Jeanne.Lesot@lip6.fr (M. Lesot); esteban.marquer@loria.fr (E. Marquer); miguel.couceiro@loria.fr (M. Couceiro)

https://limics.fr (F. Badra); https://lip6.fr (M. Lesot); https://emarquer.github.io/ (E. Marquer); https://members.loria.fr/mcouceiro/ (M. Couceiro)

0000-0002-2437-8230 (F. Badra); 0000-0002-3604-6647 (M. Lesot); 0000-0003-2315-7732 (E. Marquer); 0000-0003-2316-7623 (M. Couceiro)

Recent work [4] showed that a common principle underlying case-based prediction methods is that they interpret the plausible inference principle of analogical transfer as a transfer of similarity knowledge from a situation space to an outcome space. This idea of modeling analogical transfer as a transfer of similarity knowledge is a powerful idea, that can have many implications. One of them is that learning a similarity measure can be framed as the problem of optimizing the compatibility between two similarity measures on a data set.

In this paper, we discuss some perspectives and directions that could be given to this line of research. A global indicator of the compatibility between two similarity measures has already been proposed in the CoAT method [5], and preliminary experiments showed that such an indicator can be used as an intrinsic indicator of the quality of the similarity measure with respect to the case-based prediction task [6]. A natural perspective to this research is to apply these results to similarity learning, and to design a similarity learning method that would optimise such an indicator on the data set. To this aim, we explore in this paper the connections between the CoAT method and existing work in the domain of similarity learning. We then show that interpreting CoAT in an energy-based model is quite straightforward, so that the similarity learning task can be stated as the task of learning an energy function.

The paper is organized as follows. In Section 2 we recall the previous work on the CoAT method. We then briefly survey in Section 3 some approaches to learning (dis)similarities that seem relevant to CoAT, and discuss to how to leverage CoAT to obtain suitable similarity measures. We also explore techniques based on the optimisation of energy function that we propose in Section 4 and discuss further perspectives in Section 5.

2. The CoAT Method

In the CoAT method [5, 6, 7], the analogical transfer inference is made by minimizing a global indicator of compatibility between two similarity measures. Such an indicator can also be used as an intrinsic indicator of the quality of the similarity measure w.r.t. the transfer task.

2.1. Definition of the Indicator

Let $\mathcal{S}$ denote an input space, and $\mathcal{R}$ an output space. An element of $\mathcal{S}$ is called a *situation*, and an element of $\mathcal{R}$ is called an *outcome*, or a result. A set $CB = \{(s_1, r_1), \ldots, (s_n, r_n)\}$ of elements in $\mathcal{S} \times \mathcal{R}$ is called a *case base*. An element $c = (s, r) \in CB$ is called a *source case*. In addition, the spaces $\mathcal{S}$ and $\mathcal{R}$ are respectively equipped with two similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$, that respectively denote the similarity measure on situations and on outcomes.

The compatibility of $\sigma_{\mathcal{R}}$ with $\sigma_{\mathcal{S}}$ is measured globally on the case base CB, by introducing a global indicator $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$. This indicator measures the compatibility of $\sigma_{\mathcal{R}}$ with $\sigma_{\mathcal{S}}$ from an ordinal point of view on the whole case base CB, by checking if the order induced by $\sigma_{\mathcal{R}}$is the same as the one induced by $\sigma_{\mathcal{S}}$. The following continuity constraint is tested on each triple of cases (c_0, c_i, c_j), with $c_0 = (s_0, r_0)$, $c_i = (s_i, r_i)$, and $c_j = (s_j, r_j)$:

$$\text{if } \sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j), \text{ then } \sigma_{\mathcal{R}}(r_0, r_i) \geq \sigma_{\mathcal{R}}(r_0, r_j). \qquad (C)$$

Constraint (C) expresses that anytime a situation s_i is more similar to a situation s_0 than situation s_j, this order should be preserved on outcomes. A triple (c_0, c_i, c_j) does *not* satisfy (C) if the

case c_i is more similar to the case c_0 (that we will refer to as *anchor*) than the case c_j for situations, but less similar for outcomes, *i.e.*, when $\sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j)$ and $\sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)$. Such a violation of the constraint is called an *inversion of similarity*. The indicator $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$ counts the total number of inversions of similarity observed on a case base CB:

$$\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB) = |\{((s_0, r_0), (s_i, r_i), (s_j, r_j)) \in CB \times CB \times CB \text{ such that } \sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j) \text{ and } \sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)\}|.$$

2.2. Inference

When the case base is fully known, except for the outcome r_t of one case $c_t = (s_t, r_t)$, the transfer inference consists in finding the outcome r_t that minimizes the value of the indicator:

$$r_t = \arg\min_{r \in \mathcal{R}} \Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB \cup \{(s_t, r)\}).$$

2.3. An Intrinsic Indicator of the Quality of a Similarity Measure

The indicator $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$ can be used to assess the quality of the situation space similarity measure $\sigma_{\mathcal{S}}$ with respect to the transfer task, independently of the algorithm used for the inference. We report here some first experiments made in [6] that show a strong correlation between the value of the $\Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$ indicator obtained for a chosen similarity measure $\sigma_{\mathcal{S}}$ and the corresponding performance of the CoAT prediction algorithm.

Experimental Protocol. The experiment is conducted on 200 instances extracted from the Balance Scale data set[1]. As the instances of these data sets are described only by d numeric features, each situation can be represented by a vector of $\mathbb{R}^d$. Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ be two such vectors. These data induce a classification task: the outcomes are categorical classes and the outcome similarity measure $\sigma_{\mathcal{R}}$ is the class membership, i.e. $\sigma_{\mathcal{R}}(u, v) = 1$ if $u = v$, and 0 otherwise. The performance of the CoAT algorithm is measured by generating 100 different classification tasks $\{(\sigma_i, \sigma_{\mathcal{R}}, CB)\}_{1 \leq i \leq 100}$, each of which is obtained by choosing for $\sigma_{\mathcal{S}}$ a decreasing function of a randomly weighted Euclidean distance. More precisely, a set of random linear maps $\{L_i : \mathbb{R}^d \longrightarrow \mathbb{R}^d\}_{1 \leq i \leq 100}$ are generated, and for each map L_i, σ_i is defined as a decreasing function of the Euclidean distance computed in the L_i's embedding space:

$$\sigma_i(\mathbf{x}, \mathbf{y}) = e^{-d_i(\mathbf{x}, \mathbf{y})} \text{ with } d_i(\mathbf{x}, \mathbf{y}) = \|L_i\mathbf{x} - L_i\mathbf{y}\|_2 = \sqrt{(\mathbf{x} - \mathbf{y})^T L_i^T L_i (\mathbf{x} - \mathbf{y})}.$$

The performance is also measured on the task $(\sigma_E, \sigma_{\mathcal{R}}, CB)$, in which $\sigma_E(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x} - \mathbf{y}\|_2}$ is a decreasing function of the Euclidean distance, which amounts to taking as linear map the identity matrix. For each task, the performance is measured by the prediction accuracy, with 10-fold cross validation.

[1] https://archive.ics.uci.edu/ml/datasets/balance+scale

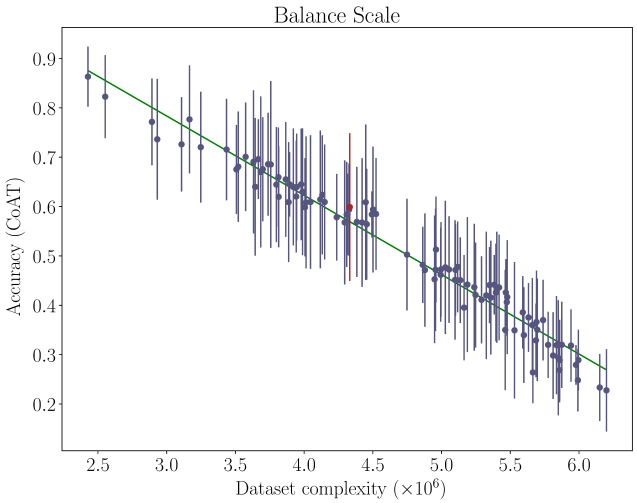


Figure 1: Relation between CoAT performance (accuracy) and value of the Γ indicator on the Balance Scale data set (as of [6]).

Results. Fig. 1 shows for each classification task the average accuracy and standard deviation of the CoAT algorithm according to the value of the Γ indicator ("Dataset complexity" axis on the figure). The blue points correspond to the randomly generated σ_i similarity measures. The red point gives the results for the σ_E similarity measure based on the standard Euclidean distance. The green line shows the result of a linear regression on the data. The Pearson's coefficient is -0.97. The results clearly show a correlation between the value of the indicator and the performance of the CoAT algorithm.

3. Perspectives on Learning (Dis)similarity Measures

While it is possible to use CoAT to quantify the suitability of a similarity measure for a CBR task, we argue that it should be possible to adapt CoAT to learn suitable similarity measures. Below we describe some existing methods to learning similarity (or dissimilarity) that appear relevant to adapt CoAT, before discussing how optimizing the indicator of CoAT relates to these (dis)similarity measure learning methodologies.

In what follows, we will not make distinction between similarity and dissimilarity measures since they are the counterpart of one another. It is possible to define one from the other, for instance, given a dissimilarity $d(u, v)$ defined on $\mathbb{R}^+$ we define the similarity $\sigma(u, v)$ on $[0, 1]$ with the inverse $\sigma(u, v) = \frac{1}{1+d(u,v)}$ or the exponential $\sigma(u, v) = e^{-d(u,v)}$.

3.1. Related Works on (Dis)similarity Measure Learning

Constructing a similarity measure for a given task is difficult and time-consuming, especially if domain knowledge is to be taken into account into the process. It is possible to use data to

support and facilitate this process, either to guide the design of the measure [8] or to learn suitable parameters for a similarity measure.

Designing or learning (dis)similarity measures from data has long been studied [9]. Here, we briefly discuss three approaches, namely, by combining local similarities, by unsupervised approaches based on clustering techniques, and by supervised or semi-supervised metric learning approaches. Note that there is a particular focus in CBR on the explainability of the similarity measures as well as on using complex data (*i.e.*, heterogeneous or structured), which constrains the learning of (dis)similarity measures.

Combining local (dis)similarities. Computing (dis)similarities in heterogeneous data can be performed by transforming the input dataset into a homogeneous one. An interesting approach is to consider the overall similarity measure as a weighted sum of ad-hoc measures. For instance, the *k-Prototypes* algorithm [10] computes a dissimilarity $d(x, y)$ between two instances x and y as

$$d(x, y) = d_E(x, y) + \lambda d_C(x, y), \tag{1}$$

where $d_E(x, y)$ is the Euclidean distance for a subset of continuous attributes, $d_C(x, y)$ the number of mismatched categorical attributes, and where λ a weighting parameter. Gower's similarity [11] is a popular measure that works in a similar fashion.

More generally, it is possible to rely on existing similarity measures for each aspect of the data, and combine them to obtain a global similarity. For instance, [8, 12, 13] learn the weights of linear combinations of local similarity functions for CBR tasks. Another example is [14], in which a set of local similarities estimated by artificial neural networks are aggregated. Note that the above mentioned weights can be thought as the importance that each local measure has, and thus used for explanation and fairness purposes [14, 15, 16, 17, 18, 19].

The main drawback of combining local dissimilarities is that it requires additional preprocessing and learning as well as supervision.

Unsupervised learning of (dis)similarities. Shi and Horvath [20] proposed a method to compute dissimilarities between instances in unsupervised settings using Random Forest (RF). RF [21] is a popular algorithm for supervised learning tasks, and is widely used in many applied fields, *e.g.,* in biology [22] and in image recognition [23]. Essentially, it is an ensemble method that combines decision trees in order to obtain better classification results in supervised learning on high-dimensional data.

The algorithm begins by creating several new training sets, each one being a bootstrap sample of elements from the initial data set X. A decision tree is built on each training set, using a random sample of m_{try} features at each split. The prediction task is then performed by a majority vote or by averaging the results of the decision trees, according to the problem at hand (classification or regression). This approach leads to better accuracy and generalization capacity of the model compared to single decision trees, while reducing the variance [24]. However, this ensemble approach requires labelled data.

The adaptation of RF to unsupervised settings was made possible by the generation of synthetic instances, that enable a binary classification between the latter and the observed (unlabelled) instances. The use of Unsupervised Random Forest (URF) for measuring (dis)similarity

presents several advantages. For instance, instances described by mixed types of variables as well as missing values can be handled. In fact, this method has been successfully used in many applications [25, 26, 27, 28].

Albeit its appealing character, the method suffers from two main drawbacks. Firstly, the generation step is not computationally efficient: since the obtained trees highly depend on the generated instances, it is necessary to construct many forests with different synthetic instances and average their results, leading to a computational burden. Secondly, the synthetic instances may bias the model being constructed to discriminate instances on specific features.

More recently, Ting *et al.* [29] proposed a similar approach to compute a mass-based dissimilarity between instances, based on isolation forests [30]. While their approach is similar, it differs on some key points, such as the fact that self-similarities are not constant in mass-based dissimilarity, since it they depend on the distribution of the data. This property is interesting and may lead to good results in cases where clusters are of varying density. However, this method does not apply to heterogeneous data.

Following the tracks of [20] and [31], [32] proposed a method, called Unsupervised Extremely Randomised Trees (UET), to compute similarities on unlabelled data. The main idea is to randomly split the data in an iterative fashion until a stopping criterion is met, and to compute a similarity based on the co-occurrence of instances in the leaves of each generated tree. It was shown to provide tailor made multidimensional similarity measures for complex and heterogeneous data [33] and to be easily adaptable to structured data such as labelled graphs [34]. The empirical study of UET showed that it outperforms existing methods (such as URF) in terms of computational time, while giving better cluster results and, consequently, more relevant similarities. Moreover, it has interesting invariance properties such as such as invariance under monotonic transformations of variables and robustness to correlated variables and noise, that drastically reduces preprocessing.

Despite of producing tailor made measures for data at hand, the main drawback of UET is that it computes similarities on each space (the situation and outcome) without establishing links between the two.

Metric learning. Learning dissimilarity measures from data has been tackled in the field of metric learning (for an extended introduction, see [35, 36]) by learning the parameters of parametric distance functions d_θ, following either relative (ordinal) constraints or link/cannot link (similarity/dissimilarity) constraints. Metric learning techniques have been used for representation learning: combining a parametric representation model with a simple non-parametric distance function (typically the Euclidean distance) allows to learn a representation model suitable to preserve the relative or link/cannot link constraints. These constraints are usually implemented by minimizing the triplet loss or the contrastive loss as follows.

On the one hand, contrastive loss [37] is used to enforce link/cannot link constraints on training pairs s_i, s_j associated with labels r_i, r_j. If s_i, s_j, associated with labels r_i, r_j, is a pair of similar elements ($r_i \approx r_j$), then we want to minimize $d_\theta(s_i, s_j)$, and we want to maximize the latter if the pair is not similar ($r_i \neq r_j$). The contrastive loss is defined as

$$L(s_i, s_j) = \sigma_{\mathcal{R}}(r_i, r_j) d_\theta(s_i, s_j) - (1 - \sigma_{\mathcal{R}}(r_i, r_j)) d_\theta(s_i, s_j),$$

where $\sigma_{\mathcal{R}}$ is the already mentioned class membership similarity measure, such that $\sigma_{\mathcal{R}}(u, v) = 1$ if $u = v$, and 0 otherwise.

On the other hand, triplet loss [38, 39] methods use training triplets s_0, s_i, s_j associated with labels r_0, r_i, r_j. that are selected such that r_0 (called the *anchor* as in CoAT) is closer to r_i than r_j. For such triplets, it is desired that $d_\theta(s_0, s_i) < d_\theta(s_0, s_j)$ which translates into the triplet loss

$$L(s_0, s_i, s_j) = \max(d_\theta(s_0, s_i) - d_\theta(s_0, s_j) + \alpha, 0)$$

where the *margin* α is used to enforce a gap between the clusters of situations.

To implement relative constraints with triplet loss, it is enough to have r_0, r_i, r_j verify an ordinal relation of the form $r_0 \leq r_i < r_j$ or $r_0 \geq r_i > r_j$. In a classification setting, the labels are classes that do not necessarily have an order defined, so the link/cannot link constraint $r_0 = r_i \neq r_j$ is used instead. This latter constraint corresponds to $\sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)$, where $\sigma_{\mathcal{R}}$ is the class membership similarity measure mentioned above.

Note that while metric learning was initially designed to use class labels, making it a supervised methodology, semi-supervised and unsupervised variants have been also proposed [40, 41].

3.2. Links Between CoAT and Metric Learning Approaches

The Γ indicator defined in Section 2.1 measures how suitable a similarity measure is for a particular CBR task. As such, it could be used to identify or, following metric learning methodology, to learn a similarity measure or a suitable representation space. To help make such a parallel, we propose to leverage striking similarities between CoAT and triplet loss.

Indeed, as in triplet loss methods, the CoAT method considers similarity judgements that are data triplets of the form $\{(s_0, s_i, s_j) \mid \sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)\}$, but then counts the number of triplets violating the constraint (C), *i.e.*, such that $\sigma_{\mathcal{S}}(s_0, s_i) \geq \sigma_{\mathcal{S}}(s_0, s_j)$ and $\sigma_{\mathcal{R}}(r_0, r_i) < \sigma_{\mathcal{R}}(r_0, r_j)$. In triplet loss terminology, this corresponds to counting the number of hard negatives among all possible triplets formed with instances of the data set. Semi-hard negatives (*i.e.*, triplets such that $\sigma_{\mathcal{S}}(s_0, s_i) + \alpha \geq \sigma_{\mathcal{S}}(s_0, s_j)$ for some margin α) are excluded from this procedure. Therefore, when applied to classification settings, the contribution of a triplet to the CoAT indicator Γ can be seen as a simplified version of the loss $L(s_0, s_i, s_j)$ used in triplet loss methods, that would take value 1 if the triplet is a hard negative, and 0 otherwise.

However, the idea of the CoAT method is to sum up these contributions on all possible triplets of a case base. Although in our first experiments, the case base consisted in the whole data set, a more case-based approach would require crafting a (preferably small but informative) case base for the task before attempting to learn a similarity measure. Moreover, one contribution of the work done on the CoAT method has been to show that the prediction for a new case depends only on the *new* similarity relations that result from the addition of the new case to the case base [6]. This suggests that learning should be done by carefully selecting a case base from whole data set, and training for a test case (t, r) by minimizing

$$\Delta\Gamma(t, r, \sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB) = \Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB \cup \{(t, r)\}) - \Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB).$$

This could lead to giving additional theoretical justification of triplet loss methods and give new insights on how to solve the sampling issue (*i.e.*, which training triplets to select).

4. Perspectives on Learning an Energy Function

This section discusses another perspective opened by the analogical inference interpretation as the optimisation of the proposed Γ indicator, as established in Section 2.2. Indeed, this view allows to exploit the formalism of energy-based models proposed for machine learning tasks by [42] reminded below. As detailed in the following, the interpretation of CoAT in an energy-based model is quite straightforward: the global indicator Γ of the CoAT approach can be seen as an energy function, that measures the compatibility between two similarity measures $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ on the case base CB. In this perspective, CoAT's transfer strategy is an energy-based inference, that consists in completing the description of the case base in order to minimize its energy, and learning the energy function (and hence, the similarity measure) could be achieved by optimizing a contrastive loss function.

Energy-Based Models. Inspired from statistical physics, energy-based models specify a probability distribution

$$p(x;\theta) = \frac{e^{-E_\theta(x)/T}}{\int e^{-E_\theta(x)/T} dx}$$

directly via a parameterized scalar-valued function $E_\theta(x)$ called an *energy function*. In machine learning, energy-based models are trained to be optimized on the data manifold: the energy function is learned to give low values to training data, and higher values to data points that are far from the data manifold [42]. In its conditional version, the definition of an energy function $E_\theta : \mathcal{X} \times \mathcal{Y} \longrightarrow \mathbb{R}$ assumes the existence of an input space $\mathcal{X}$, an output space $\mathcal{Y}$, and a set of parameters θ. The energy function E_θ associates to each pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ a scalar value $E_\theta(x, y)$ that represents the compatibility between the input x and the output y under the set of parameters θ. The energy function E_θ takes low values when y is compatible with x, and higher values when y and x are less compatible. The goal of the energy-based *inference* is to find, among a set of outputs $\mathcal{Y}$, the output $y^* \in \mathcal{Y}$ that minimizes the value of the energy function:

$$y^* = \arg\min_{y \in \mathcal{Y}} E_\theta(x, y).$$

Given a family of energy functions $E_\theta(x, y)$ indexed by a set of parameters θ, the goal of *learning* is to optimize the θ parameters in order to "push down" (*i.e.*, assign lower energy values to) the points on the energy surface that are around the training samples, and to "pull up" all other points. Contrastive divergence [43] is a common learning strategy that consists in optimizing a contrastive loss function such as the hinge loss, which is defined, for a training sample (x_k, y_k) and a generated out of distribution sample $(x_k, \hat{y})$ by:

$$\ell(\theta, x_k, y_k) = \max(0, \beta + E_\theta(x_k, y_k) - E_\theta(x_k, \hat{y})).$$

The hinge loss associates a loss value to a training sample (x_k, y_k) whenever its energy is not lower by at least a margin β than the energy of the incorrect sample $(x_k, \hat{y})$.

An Energy-Based Model of Analogical Transfer. The input space $\mathcal{X}$ (from which similarity knowledge is transferred) is the situation space $\mathcal{S}$. The output space $\mathcal{Y}$ (to which similarity knowledge is transferred) is the outcome space $\mathcal{R}$. The situation space $\mathcal{S}$ is equipped with a similarity measure $\sigma_{\mathcal{S}}$, and the outcome space is equipped with a similarity measure $\sigma_{\mathcal{R}}$. The energy function $E_\theta : \mathcal{S} \times \mathcal{R} \longrightarrow \mathbb{R}$ measures the compatibility of the outcome similarities with the added situation similarities when a potential new case $\hat{c}_t = (t, r)$ is added to the case base. The energy function E_θ is parameterized by a hyperparameter $\theta = (\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$, which includes the case base CB. Indeed, assuming that $\sigma_{\mathcal{S}}$ and $\sigma_{\mathcal{R}}$ are defined on different sets of attributes, the compatibility between two similarity measures can not be evaluated *per se*, but only relatively to a given set of case pairs. For a new situation t, the goal of the energy-based *inference* is to find, among a set of potential outcomes $r \in \mathcal{R}$, the outcome r_t that minimizes the value of the energy function:

$$r_t = \arg\min_{r \in \mathcal{R}} E_\theta(t, r).$$

Among the three parameters of $\theta = (\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB)$, the case base CB and the outcome similarity measures $\sigma_{\mathcal{R}}$ are usually fixed, so that learning θ amounts to learning the situation similarity measure $\sigma_{\mathcal{S}}$ for the task at hand. This can be done by contrastive divergence using the hinge loss defined as follows: for a training sample $(s_k, r_k) \in \mathcal{S} \times \mathcal{R}$ and a chosen outcome $\hat{r} \in \mathcal{R}$,

$$\ell(\theta, s_k, r_k) = \max(0, m + E_\theta(s_k, r_k) - E_\theta(s_k, \hat{r})).$$

The CoAT case-based prediction method directly implements this energy-based model by taking as energy function the global indicator Γ:

$$E_\theta^{\text{CoAT}}(t, r) = \Gamma(\sigma_{\mathcal{S}}, \sigma_{\mathcal{R}}, CB \cup \{(t, r)\}).$$

Illustration on Some Synthetic Data Sets. Fig. 2 gives some examples of energy maps that are obtained for different synthetic data sets on a binary classification task. On each figure, the dataset size is the same ($|CB| = 100$), but the instances span differently on the 2D description space. The instances are equally split into two classes (orange and blue). The similarity measure on situations $\sigma_{\mathcal{S}}$ is a decreasing function of the Euclidean distance as in Sec. 3 (*i.e.*, $\sigma_{\mathcal{S}} = \sigma_E$), except for Fig. 2 d, where $\sigma_{\mathcal{S}}$ is constructed from a linear transformation of the Euclidean distance, by choosing from a set of 100 randomly generated transformations, the one that minimizes the energy of the case base. Let us denote by σ^* the resulting similarity measure. The similarity measure on outcomes $\sigma_{\mathcal{R}}$ represents class membership as previously.On the figures, the colors indicate for each point of space the class that would be predicted by the CoAT algorithm : green for the blue class, and orange for the orange class. The color saturation is proportional to the difference between the energy of the predicted class and the energy of the other class.

Results In Fig. 2 a, the two classes are well separated, and no instance is more similar to an instance of a different class than it is to an instance of the same class, hence, $E(\sigma_E, \sigma_{\mathcal{R}}, CB) = 0$. In Fig. 2 b, the two classes are closer, and even overlap, and some inter-class similarities

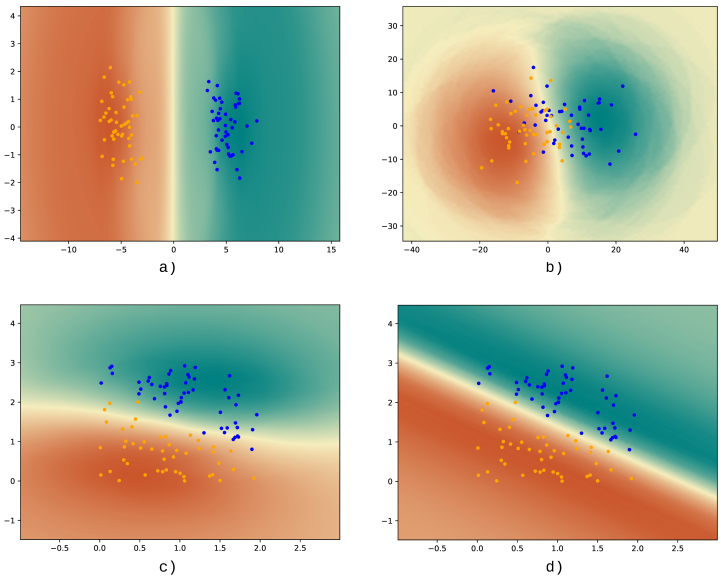


Figure 2: Energy maps illustrating the confidence values associated to each class by the CoAT algorithm for different synthetic datasets in a binary classification scenario. Green areas correspond to areas where new instances would be predicted as belonging to the blue class, and red areas correspond to areas where new instances would be predicted as belonging to the orange class. The color saturation is proportional to the difference between the energy of the predicted class and the energy of the other class. On figures (a), (b), and (c), $\sigma_{\mathcal{S}}$ is constructed from the Euclidean distance. On the lower right figure (d), $\sigma_{\mathcal{S}}$ is optimized to minimize the energy of the case base.

happen to be lower than some intra-class similarities, leading to the non-zero data set energy $E(\sigma_E, \sigma_{\mathcal{R}}, CB) = 86,786$. Fig. 2 c and d show a data set with two linearly separable classes. In Fig. 2 c, $\sigma_{\mathcal{S}}$ is set to the (inverse of) the Euclidean distance, which leads to sub-optimal prediction performance: the prediction frontier does not correspond to the real class frontier, and some instances are misclassified. The energy of the case base is $E(\sigma_E, \sigma_{\mathcal{R}}, CB) = 43,264$. In Fig. 2 d, the similarity measure $\sigma_{\mathcal{S}}$ is optimized by choosing a similarity measure σ^* that minimizes the energy of the case base. The resulting prediction performance is improved: the prediction frontier corresponds to the real class frontier, and no instance of the case base are misclassified. The energy of the case base is $E(\sigma^*, \sigma_{\mathcal{R}}, CB) = 17,146$.

5. Conclusion

In this paper we investigated interactions between analogical transfer and similarity learning, in the framework of CoAT. In particular, we identified similarities between the Γ indicator and the triplet loss of metric learning, that may be used to obtain suitable similarities for analogical transfer. We also proposed an interpretation of the CoAT method in the formalism of energy-based models, so that the similarity learning task can be expressed as the task of learning an

energy function.

The established connections allow to envision other applications. For instance, it could be used for case base construction and maintenance. Indeed, if we consider the indicator as an energy function, the competence of a case should relate to its ability, when it is added to the case base, to lower the energy of other cases. Reasoning with a small but competent case base would solve one of the actual limitations of the CoAT method, which is the quadratic computational complexity of the inference procedure.

An additional direction for future works concerns the integration of expert knowledge, to promote interaction with domain experts when processing a case base. We envision this integration at two levels: the design of the similarity measure and the choice of suitable cases. We envision a semi-automatic approach to reach a suitable compromise between available data, expert input, and selection of competent cases.

References

[1] T. R. Davies, S. J. Russell, A logical approach to reasoning by analogy, in: IJCAI, 1987. arXiv:1011.1669v3.

[2] A. Aamodt, E. Plaza, Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches, AI Communications 7 (1994) 39–59.

[3] L. Miclet, S. Bayoudh, A. Delhay, Analogical Dissimilarity: Definition, Algorithms and Two Experiments in Machine Learning, JAIR 32 (2008) 793–824.

[4] F. Badra, M.-J. Lesot, Case-Based Prediction – A Survey, IJAR (2023).

[5] F. Badra, A Dataset Complexity Measure for Analogical Transfer, in: IJCAI, 2020, pp. 1601–1607.

[6] F. Badra, M.-J. Lesot, Theoretical and Experimental Study of a Complexity Measure for Analogical Transfer, in: ICCBR, 2022, pp. 175–189.

[7] F. Badra, M.-J. Lesot, CoAT-APC: When Analogical Proportion-based Classification Meets Case-Based Prediction, in: ATA@ICCBR, CEUR-WS, 2022.

[8] D. Verma, K. Bach, P. J. Mork, Similarity Measure Development for Case-Based Reasoning–A Data-Driven Approach, in: NAIS, volume 1056, Springer, Cham, 2019, pp. 143–148.

[9] M. M. Deza, E. Deza, Encyclopedia of distances, in: Encyclopedia of Distances, Springer, 2009, pp. 1–583.

[10] Z. Huang, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery 2 (1998) 283–304.

[11] J. Gower, A general coefficient of similarity and some of its properties, Biometrics (1971) 857–871.

[12] W. Cheng, E. Hüllermeier, Learning Similarity Functions from Qualitative Feedback, in: ECCBR, volume 5239, Springer, 2008, pp. 120–134.

[13] A. Jaiswal, K. Bach, A Data-Driven Approach for Determining Weights in Global Similarity Functions, in: ICCBR, volume 11680, Springer International Publishing, 2019, pp. 125–139.

[14] T. Gabel, E. Godehardt, Top-Down Induction of Similarity Measures Using Similarity Clouds, in: ICCBR, volume 9343, Springer, Cham, 2015, pp. 149–164.

[15] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: 22nd SIGKDD, ACM, 2016, pp. 1135–1144.
[16] S. M. Lundberg, S. Lee, A unified approach to interpreting model predictions, in: NIPS 2017, 2017, pp. 4765–4774.
[17] G. Alves, M. Amblard, F. Bernier, M. Couceiro, A. Napoli, Reducing unintended bias of ML models on tabular and textual data, in: 8th DSAA, IEEE, 2021, pp. 1–10.
[18] J.-B. Lamy, B. Sekar, G. Guezennec, J. Bouaud, B. Séroussi, Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach, Artificial Intelligence in Medicine 94 (2019) 42–53.
[19] K. Bach, P. J. Mork, On the Explanation of Similarity for Developing and Deploying CBR Systems, in: AAAI, 2020.
[20] T. Shi, S. Horvath, Unsupervised learning with random forest predictors, Journal of Computational and Graphical Statistics 15 (2006) 118–138.
[21] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.
[22] B. Percha, Y. Garten, R. B. Altman, Discovery and explanation of drug-drug interactions via text mining, in: PSB, 2012, pp. 410–421.
[23] M. Pal, Random forest classifier for remote sensing classification, Int. J. Remote Sensing 26 (2005) 217–222.
[24] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, Springer series in statistics New York, 2001.
[25] H. L. Kim, D. Seligson, X. Liu, N. Janzen, M. Bui, H. Yu, T. Shi, A. S. Belldegrun, S. Horvath, R. Figlin, Using tumor markers to predict the survival of patients with metastatic renal cell carcinoma, The Journal of urology 173 (2005) 1496–1501.
[26] M. Abba, H. Sun, K. Hawkins, J. Drake, Y. Hu, M. Nunez, S. Gaddis, T. Shi, S. Horvath, A. Sahin, *et al.*, Breast cancer molecular signatures as determined by sage: correlation with lymph node status, Molecular Cancer Research 5 (2007) 881–890.
[27] S. Rennard, N. Locantore, B. Delafont, R. Tal-Singer, E. Silverman, J. Vestbo, B. Miller, P. Bakke, B. Celli, P. Calverley, *et al.*, Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the eclipse cohort using cluster analysis, Annals of the American Thoracic Society 12 (2015) 303–312.
[28] K. Peerbhay, O. Mutanga, R. Ismail, Random forests unsupervised classification: The detection and mapping of solanum mauritianum infestations in plantation forestry using hyperspectral data, IEEE J. Sel. Top. Appl. Earth Obs Remote Sens 8 (2015) 3107–3122.
[29] K. Ting, Y. Zhu, M. Carman, Y. Zhu, T. Washio, Z. Zhou, Lowest probability mass neighbour algorithms: Relaxing the metric constraint in distance-based neighbourhood algorithms, Machine Learning (2018).
[30] F. Liu, K. Ting, Z. Zhou, Isolation forest, in: 8th ICDM, IEEE, 2008, pp. 413–422.
[31] P. Geurts, D. Ernst, L. Wehenkel, Extremely randomized trees, Machine learning 63 (2006) 3–42.
[32] K. Dalleau, M. Couceiro, M. Smaïl-Tabbone, Unsupervised extremely randomized trees, in: 22nd PAKDD 2018, volume 10939 of *LNCS*, Springer, 2018, pp. 478–489. URL: https://doi.org/10.1007/978-3-319-93040-4_38. doi:10.1007/978-3-319-93040-4_38.
[33] K. Dalleau, M. Couceiro, M. Smaïl-Tabbone, Unsupervised extra trees: a stochastic approach to compute similarities in heterogeneous data, Int. J. Data Sci. Anal. 9 (2020) 447–459.

[34] K. Dalleau, M. Couceiro, M. Smaïl-Tabbone, Computing vertex-vertex dissimilarities using random trees: Application to clustering in graphs, in: 18th IDA, volume 12080 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 132–144.

[35] A. Bellet, A. Habrard, M. Sebban, A Survey on Metric Learning for Feature Vectors and Structured Data (2014). arXiv:1306.6709.

[36] A. Bellet, A. Habrard, M. Sebban, Metric Learning, AIM, Springer, 2015. URL: https://hal.science/hal-01121733. doi:10.2200/S00626ED1V01Y201501AIM030.

[37] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: CVPR, volume 1, 2005, pp. 539–546. doi:10.1109/CVPR.2005.202.

[38] F. Schroff, D. Kalenichenko, J. Philbin, FaceNet: A Unified Embedding for Face Recognition and Clustering, in: CVPR, 2015, pp. 815–823. arXiv:1503.03832.

[39] D. P. Vassileios Balntas, Edgar Riba, K. Mikolajczyk, Learning local feature descriptors with triplets and shallow convolutional neural networks, in: Proceedings of the British Machine Vision Conference (BMVC), BMVA Press, 2016, pp. 119.1–119.11. URL: https://dx.doi.org/10.5244/C.30.119. doi:10.5244/C.30.119.

[40] W. Liu, S. Ma, D. Tao, J. Liu, P. Liu, Semi-supervised sparse metric learning using alternating linearization optimization, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10, ACM, 2010, p. 1139–1148. URL: https://doi.org/10.1145/1835804.1835947. doi:10.1145/1835804.1835947.

[41] S. Kim, D. Kim, M. Cho, S. Kwak, Self-taught metric learning without labels, in: CVPR, IEEE, 2022, pp. 7421–7431. URL: https://doi.org/10.1109/CVPR52688.2022.00728. doi:10.1109/CVPR52688.2022.00728.

[42] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. J. Huang, A Tutorial on Energy-Based Learning, in: Predicting Structured Data, 2006, p. 59.

[43] G. Hinton, S. Osindero, M. Welling, Y.-W. Teh, Unsupervised Discovery of Nonlinear Structure Using Contrastive Backpropagation, Cognitive Science 30 (2006) 725–731.

Can LLMs solve generative visual analogies?

Shrey Pandit[1], Gautam Shroff[2], Ashwin Srinivasan[1] and Lovekesh Vig[2]

[1]BITS Pilani, K.K. Birla Goa Campus, India
[2]TCS Research,New Delhi, India

Abstract

Recent experiments with large language models (LLMs) have provided some evidence that these models can perform abstract analogical reasoning [1], including textual puzzles similar to Raven's progressive matrices. We consider a visual analogical reasoning task that was solved using neuro-symbolic techniques in [2], and investigate how LLMs fare on this task. The task involves learning a sequence of transformations by which a sample input/output pair of images are related so as to analogously transform a test input. Note that unlike the analogical reasoning tasks in [1], this task involves *generating* an output as opposed to selecting from a set of choices. We evaluated various LLMs including GPT-4, GPT 3.5-turbo (ChatGPT), and GPT3 on this task for differing lengths of the sequence of transformations relating the input and output. Our results suggest that GPT-4 performs the best overall, while GPT 3.5-turbo and GPT3 perform strongly on shorter program lengths. At the same time, the performance of LLMs for this task falls far short of the neuro-symbolic approach used earlier, and we speculate as to why this may be the case, at least as of now.

Keywords

Large language models, GPT-4, Visual analogy, Neural analogical reasoning

1. Introduction

As in [2] we consider the class of visual reasoning problems as demonstrated in Figure 1c in which each task involves a *functional analogy* (i.e., $x : f(x) :: y : f(y)$) wherein each shape in the input image (here just one) is transformed to one or more positions in the output image via a sequence of elementary transformations, e.g., shifts in the 3x3 grid. Given a solved example, constructing the analogous output for a test input can be cast as a program synthesis problem where we seek to discover a program consisting of one or more sequences of shifts that need to be applied to each input shape in order to generate the output image. This program can then be applied to a text image to generate an *analogous* output.

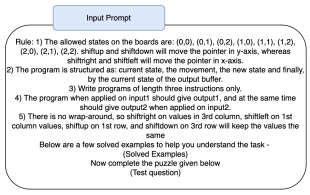
The neuro-symbolic approach in [2] achieved 100% success on a large collection of such problems when presented in symbolic form and 94.7% success when given images. In contrast, pure deep-learning approaches, including meta-learning could achieve 74% success at best as also documented in [2].

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China
*Corresponding author.
pandit.shrey.01@gmail.com (S. Pandit)

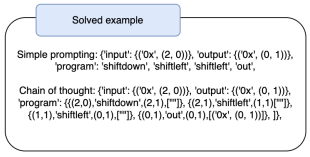
2. Experiments

We applied LLMs to solve such a visual analogy task, with the image translated into symbolic form. We use the trained models provided by OpenAI API and give a few solved examples in the prompt to help the LLMs learn.

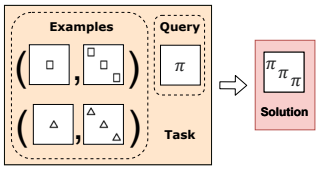
Prompting We prompt the LLMs with a set of rules for the task, which includes information on the allowed positional shifts, the set of permissible states, and the non-wrapping state of the grid. As a hint, we also specify the expected program length in the prompt. We also provide a set of solved examples to guide the LM's learning process. Figure 1a illustrates a representative example of the prompt.



(a) Prompt with rules, solved examples and test input.



(b) Chain-of-thought vs. simple prompting



(c) Given example input-output image pairs, generate the analogous output for the given query image.

Figure 1: Figure 1a showing the prompt rules, Figure 1b showing the comparison of COT vs. simple prompting, and the Figure 1c showing overview of the entire process.

		Program length 3		Program length 5	
	Best of n outputs	Simple prompting	Chain of thoughts	Simple prompting	Chain of thoughts
GPT-3	1	14%	20%	5.4%	4.16%
	3	26%	30%	11.1%	12.8%
	5	30%	42%	23%	18.9%
GPT-3.5-turbo	1	10%	6%	6.12%	8%
	3	16.3%	14%	22%	18%
	5	33.3%	22%	12%	14%
GPT-4	1	18.36%	26%	22%	12%
	3	38.77%	52%	26.5%	22%
	5	40%	46%	39.58%	34%

Table 1
Table comparing the performance of GPT3, GPT 3.5-turbo, and GPT4 over program lengths 3 and 5, with various different numbers of outputs and ways of prompting over 50 trials. In the above case, we are providing ten solved examples.

Sampling The top prediction generated by a LM may not always be the optimal choice, so we also evaluate results using the top-n predictions for consideration in determining the correctness of the task. The task is deemed successful if *any* of the n predictions are accurate.

GPT3			
Simple Prompting	Tokens	2847	3690
	Accuracy	48%	32%
Chain-of-thought prompting	Tokens	2345	3813
	Accuracy	42%	50%

Table 2
Chain-of-thought vs more examples for same token length.

Chain-of-thought prompting: Previous works such as [3] have shown that language-model performance increases drastically in reasoning tasks when given chain-of-thought prompts. In our context we provide a chain-of-thought by providing, with each step of the solved example, the current state on the grid, positional shift, next position on grid, and the state of the output buffer.
Changing the program length: We experimented with different program lengths (3 & 5); empirically, increasing the program length makes the task more difficult.

3. Results and Conclusions

Referring to Table 1 we observe the following: (i) Chain-of-thought improves performance over simple prompting, which was expected. (ii) Further, chain-of-thought prompting is also better (albeit slightly) than providing more examples for similar token lengths, see Table 2. (ii) Sampling more examples improves performance, also expected. (iii) Analogies involving longer sequences (programs) are more difficult as expected; however we observe a drastic drop in performance for GPT3 and GPT 3.5-turbo but only a marginal drop is observed for GPT4. While the input prompts do affect the LLMs' performance, it's important to mention that a uniform prompt template was used for all the analyzed LLMs in this study.

Overall the performance of LLMs for our simple visual analogy task fall far short of the neuro-symbolic techniques used in [2]. We note that [2] relied *search* over possible sequences that could successfully transform a text input to its output. LLMs do not explicitly search over potential outputs. We speculate that incorporating elements of explicit search may enable LLMs to perform better at generative analogies.

References

[1] W. et. al, Emergent analogical reasoning in large language models, arXiv:2212.09196 (2022).
[2] S. et. al, Solving visual analogies using neural algorithmic reasoning, AAAI (Student Abstract) (2022).
[3] J. W. et. al, Chain of thought prompting elicits reasoning in large language models, NeurIPS 2022 (2022).

Some Preliminary Results on Analogies Between Sentences Using Contextual and Non-Contextual Embeddings

Thomas Barbero[1], Stergos Afantenos[2]

[1]IRIT, University of Toulouse, France

Abstract

Analogies have been characterized as fundamental to abstraction, concept formation, and perception, and are traditionally expressed as quadruplets in the form of proportional analogies $a : b :: c : d$ read "a is to b as c is to d". While Natural Language Processing (NLP) has primarily focused on word analogies and SAT problems, recent research has started exploring analogies between sentences and even documents. In this paper we explore the potential of identifying analogies between pairs of sentences via the identification of common latent relations between them. We exploit three different datasets generating pairs of sentences which can either share the same latent relation—forming thus an analogy—or not. We encode phrases into a higher dimensional vector space using embeddings from GloVe, BERT, and RoBERTa which we then feed to both a Multi Layer Perceptron (MLP) and a Convolutional Neural Network (CNN). Results show that architectures using contextual embeddings as inputs outperform those based on static embeddings.

1. Introduction

Analogies have preoccupied humanity at least since antiquity [1]. In recent years they have been characterized as being at "the core of cognition" [2] and have even been considered as being the fundamental mechanism via which abstraction, concept formation and perception are achieved [3, 4].

Traditionally analogies have been expressed as a quadruplet $a : b :: c : d$ read "a is to b as c is to d". Such quadruplets then form valid analogies if pairs (a, b) and (c, d) share the same underlying relation, forming thus a *proportional analogy*. The underlying relation has been viewed as the symbolic counterpart of arithmetic or geometric proportions: $a - b = c - d$ and $\frac{a}{b} = \frac{c}{d}$ respectively[1] [5].

In Natural Language Processing (NLP) various approaches adopt the framework of quadruplets focusing mostly on word analogies, such as *man is to woman as king is to queen* [6, 7, 8, 9], morphology [10] or on SAT problems [11]. More recently several researchers have focused on the problem of identifying analogies between sentences [12, 13, 14] or even documents [15].

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China

*This work was performed while the first author was working at IRIT, University of Toulouse, France. The first author is the corresponding author.

[1]Note though that any kind of latent relation can be used in order to form a valid analogy. For example, $2 : 4 :: 3 : 9$ is a valid analogy because $2 : 2^2 :: 3 : 3^2$

In this paper we are interested in further exploring the potential of analogies between sentences via the identification of common latent relations between them. We exploit three different datasets, namely the Microsoft Research Paraphrases Corpus (MSRP) [16], the Penn Discourse TreeBank (PDTB) [17] as well as the Stanford Natural Language Inference (SNLI) corpus [18] and we use GloVe [19], or transformer-based architectures such as BERT [20] and RoBERTa [21] for the encoding of phrases into a higher dimensional vector space. We show that architectures that are based on contextual embeddings outperform ones that are based on static embeddings.

The rest of the paper is structured as follows. In Section 2 we present the related work. In Section 3 we present the datasets that we have used in order to perform our experiments. The methodology for these experiments is described in Section 4 while the results are presented in Section 5. We conclude in Section 6.

2. Related Work

Initial work on analogies in NLP was performed by [11] who introduced *Latent Relational Analysis (LRA)* in order to identify analogies in the context of the Scholastic Aptitude Test (SAT), testing this approach in 20 scientific and metaphorical examples.

More recently Mikolov et al. [22, 23] have used analogies as a means to test the quality of static vectors representing word embeddings produced with *word2vec* for use in neural architectures. The authors showed that such embeddings could preserve the parallelogram rule that is found in analogies, evaluating thus the intrinsic qualities of such embeddings. Later work though has shown that this is not sufficient since most models appear to take shortcuts; no evidence exists of abstraction and analogical mapping, as one would expect from such claims. More precisely, [24] show the Google analogy test set that we used by [22, 23] is not well balanced and thus does not allow us to draw any safe conclusions concerning the underlying embeddings. They show that the vector offset approach is not enough to claim that the proposed method captures analogies. The authors thus introduce the Bigger Analogy Test Set (BATS). In this more sophisticate dataset the authors show that derivational and lexicographic relations remain a challenge. Similar conclusions are drawn by [25] both for the vector offset approach as well as the 3CosAdd [26]. They argue that such datasets cannot be used to evaluate the intrinsic qualities of such datasets.

In terms of word analogy classification [27] used the Google dataset [6] which they extended using permutation properties of analogies, presented in the same article. They then apply a Convolutional Neural Network using as input Glove embeddings representing each word. A similar approach was also adopted by [10] in the context of detecting morphological analogies. We also adopt this approach in this paper.

Recently, several researchers have explored sentential analogies. [12] explore analogies between sentences in order to identify D from a predefined set of possible candidates, given (A, B) and C such that $A : B :: C : D$ is a valid analogy. They use syntactic and semantic datasets and test various embedding methods. In a similar vein [28] perform a similar task but generate D instead. Both approaches show that analogies based on syntactic analogies obtain better results than semantic ones. [13, 14] explore sentential analogies based purely on semantic

information.

In another approach, [15] view analogies via the prism of the *Structure Mapping Theory* [29]. Their goal is to identify analogies in procedural texts focusing on the structural similarities between the texts. Underlying texts describe procedures in two different domains. The authors extract entities and their relationships. The latter are sets of ordered verbs. They extract those based on question answer pairs. The similarity measures that they propose reflect the fact that the two sets share more relations. Bert vectors representing the questions via which entities were extracted, are used in order to measure cosine similarity and thus identify potential mappings.

3. Data used

In order to perform our experiments we used three well known datasets. In what follows we provide a detailed description of the corpora used as well as the procedure which lead us to the creation of analogical quadruplets that were later used in our experiments. We should mention that we used the input datasets as they were released, no further additions or modifications were performed from us.

3.1. Paraphrases

The first corpus that we used was the Microsoft Research Paraphrases Corpus (MSRP) [16] which is composed of 5801 pairs of sentences labeled as paraphrase or not. The pairs are distilled from a database containing more than 13 million sentences pairs, itself extracted from a more than 9 million sentences corpus [16]. The 9M sentences corpus is composed of sentences extracted from +32k news clusters from internet. This corpus then has been largely reduced to contain sentences with a credited author only, leaving 49375 individual sentence pairs. So this corpus is composed of naturally occurring, non handcrafted sentences pairs. Sentences pairs with minimal variations such as typography error have been removed as they could have constituted "low quality" paraphrases.

A Support Vector Machine-Classifier (SVM-Classifier) is then used to identify a set of possible paraphrases from the 49375 sentences pairs. This set is validated by human annotators later. The SVM-Classifier is trained on a 10000 sentences pairs training set annotated by 2 human judges, and a 3rd who served the function of judge in case of disagreement. The distribution of this training set is 2968 positives examples and 7032 negatives. The classifier considered multiples features: string similarity, morphological variants, synonyms mapping with WordNet Lexical Mapping and Encarta Thesaurus, and finally composite features. The SVM-Classifier allowed to extract 20574 sentences pairs as possible paraphrases from the 4959375 previously considered. The number is high because the classifier's role was to separate possible sentences pairs to be evaluated by human judgment and not discriminate all non-paraphrases pairs, so the classifiers tend to classify inputs as positive rather than negative, at the assumed cost of having more false-positives.

Human judgment was applied to a 5801 subset of the 20574 previously extracted sentence pairs. Two judges annotated each sentence and a third one was used in case of disagreement. Each judge was asked if the pairs' sentences were semantically equivalent. About 3900 (67%) of the sentences pairs were labeled as semantically equivalent.

3.2. PDTB

The second corpus that we used was the Penn Discourse TreeBank (PDTB) [17] corpus which contains discourse annotations between sentences clauses extracted from the Wall Street Journal Corpus containing over 1 million words. The corpus describes a total of 36592 relations [17]. Discourse annotations can be triggered by an explicit or implicit discourse connective. The former are extracted from syntactically defined classes and are separated in 3 grammatical classes subordinating conjunctions, coordinating conjunctions and discourse adverbs. Explicit connectives can be connected to more than 1 clause or sentence, but the minimality principle is applied which requires minimum information to complete the interpretation. In the case of an implicit connection between the two clauses the annotators have been instructed to insert an explicit connective. Three other labels were available in order to correctly annotate three cases that prevented the annotators from inserting a coherent explicit connective. The AltLex indicating that the relation was already explicited by a non-connective expression, the insertion of a connective would then lead to a redundancy. The entRel indicating the existence of an entity based coherence relation between the two clauses, but no other relation. And finally noRel in case of no relation between the two clauses. PDTB relations are ordered hierarchically into class, type and subtype. For our experiments we used the first level of the hierarchy.

The inter-annotator agreement was high: 90,2% for explicit relations and 85,1% for implicit when exact match metric was considered; and respectively 94,5% and 92,6% when partial match metric was considered. Class level disagreement was resolved by a team of 3 experts, disagreement at lower levels were resolved by providing a tag for the direct higher level. Agreement for the class level reached 94%, 84% for type level and 80% for subtype level.

3.3. SNLI

The Stanford Natural Language Inference (SNLI) corpus [18] labels pairs of sentences as Contradiction, Entailment or semantic neutrality [18]. It contains 570k pairs of sentences by humans. Construction of the corpus was done using Mechanical Turks who were presented with a premise in the form of a sentence and were asked to provide three hypotheses, in a sentential form, for contradiction, entailment and semantic similarity. 10% of the corpus was validated by trusted Mechanical Turks. Overall a Fleiss κ of 0.70 was achieved.

The indeterminacies of event and entity co-reference are two well known issues during labeling of NLI data degrading the quality of the annotated corpus. They represent respectively a possible confusion between an Entailment and a Neutral relation, and between a Contradiction and a Neutral relation. This confusion comes from the fact that an assumption may or not have been made.

In order to solve this problem the annotation process was made in a grounded scenario aiming to reduce assumptions. Annotators were then able to generate sentences in the same scenario in order to illustrate the relations instead of relying on automatic data augmentation techniques. The work of 2500 employees permitted the data collection phase. When presented with an image caption without the matching image, the annotators had to write three sentences, one for each relation (the exact instructions are described in the SNLI paper). The image captions came from the Flickr30k corpus containing 160k captions from 30k individuals images. The

validation phase is completed on 10% of the 570k pairs of sentences by a set of 30 trusted workers. They were presented pairs of sentences and had to label them, each pair being presented to 4 annotators so there is 5 judgments considering the label from the data collection phase. The gold-label has been assigned to the pairs with at least a 3-annotators consensus, representing 98% of the data. The corpus is then separated in three individual files : test and dev (10k pairs each), train (the rest of the pairs).

3.4. Generation of analogical quadruplets

In order to create our analogical quadruplets we proceeded as follows. For each of the aforementioned datasets we randomly selected two pairs of sentences each one linked with a relation. Since our input datasets do not contain relations that have as arguments the same sentences, we never have analogies of the form $a : a :: b : b$. In case the relation linking the two pairs is the same we have a positive instance of an analogy otherwise a negative instance. For the SNLI corpus we considered neutral as not being a relation. For each input dataset we create a balanced training, test and development datasets containing the same number of positive and negative instances. Training consists of 400K instances while testing and development 40K instances each.

4. Methodology

Our problem can be formalized as follows. Given a set of quadruplets of sentences $a : b :: c : d$ which can either form an analogy (pairs $a : b$ and $c : d$ share the same latent relation) or not we need to estimate a function that predicts whether a new instance of four sentences is an analogy or not. Each quadruplet is represented by the input tokens of its sentences $s = \{w_1^s, \ldots, w_{|s|}^s\}$ with $s \in \{a, b, c, d\}$ and $|s|$ representing the length of the sentence. With each quadruplet we associate a $y \in \{0, 1\}$ which represents whether the quadruplet is an analogy or not. For each quadruplet we obtain embeddings using GloVe [19], BERT [20] and RoBERTa [21] which we then pass to two different architectures, a Multi-layer perceptron (MLP) and a Convolutional Neural Network (CNN).[2]

4.1. Embeddings

In order to perform classification we need to provide embeddings for each sentence. In the case of GloVe[3] static embeddings are provided for each word, while in the case of BERT[4] and RoBERTa[5] embeddings are dynamic. In order to obtain embeddings that represent sentences from the ones representing words a common approach [20, for example] is to take the mean of the embeddings representing each word. This is the approach that we have used as well. For

[2]Our code is available at https://github.com/ThomasBARBERO/EXPLO_ANALOGIE
[3]The Glove embeddings that we used are the following: https://huggingface.co/sentence-transformers/average_word_embeddings_glove.6B.300d
[4]https://huggingface.co/bert-base-uncased
[5]https://huggingface.co/roberta-base

GloVe	300
BERT	768
RoBERTa	1024

Table 1
Embedding dimensions of different encoders.

each sentence $s \in \{a, b, c, d\}$ we obtain an embedding

$$\mathbf{s} = \frac{1}{|s|} \sum_{w \in s} \texttt{emb}(w)$$

with $\texttt{emb} \in \{\texttt{glove}, \texttt{bert}, \texttt{roberta}\}$. Thus four different embeddings $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ are obtained for each of the sentences a, b, c and d. In the case of BERT we have also examined the use of the representation obtained for the final hidden state of the special symbol [CLS]. Embedding dimensions for each method are shown in Table 1. No further fine-tuning was performed on BERT or RoBERTa.

4.2. Classifiers

Multi-layer perceptron (MLP) The first classifier that we use is a multi-layer perceptron. The MLP takes as input the concatenation of the representations for the four sentences $\mathbf{a}, \mathbf{b}, \mathbf{c}$ and $\mathbf{d}$ as a vector $[\mathbf{a}; \mathbf{b}; \mathbf{c}; \mathbf{d}]$ and has two hidden layers, the first has a dimension of 100 and the second of 50.

$$\mathbf{z_1} = \mathbf{W_1}^T[\mathbf{a}; \mathbf{b}; \mathbf{c}; \mathbf{d}] + \mathbf{b_1}$$
$$\mathbf{z_2} = \mathbf{W_2}^T\mathbf{z_1} + \mathbf{b_2}$$

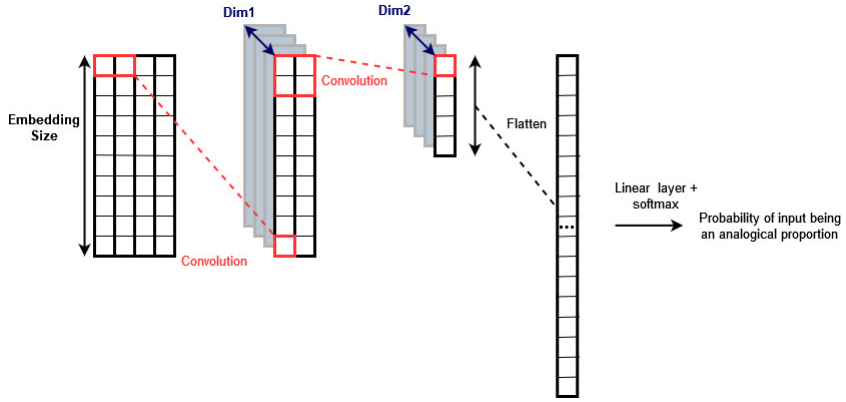
with $\mathbf{W_1}, \mathbf{b_1}, \mathbf{W_2}, \mathbf{b_2}$ learnable matrices. The output layer is of dimension 1 and we use a sigmoïd function providing a score for the final prediction

$$\hat{y} = \frac{1}{1 + e^{-\mathbf{z_2}}}$$

Convolutional Neural Network (CNN) The second classifier architecture that we have used are Convolutional Neural Networks which are widely used for image and audio processing, but are useful for Natural Language Processing tasks too as well, including analogies [27, 10, *inter alia*]. CNNs aim to recognize patterns, extracting features from the initial tensor given as input. The core of CNNs is their convolutional layers applying filters called kernels on the whole input. Kernels' weights and bias are learnt parameters, convolution between a kernel and a tensor allows to extract a learnt feature from the tensor. Parameters define the method for the application of kernels: the kernels' size, the stride which indicates the spatial distance between two kernel applications and padding which indicates the number of pixels to add on the considered tensor's borders. Our CNN's implementation is as follows, illustrated in Fig. 1:

1. The input goes through a first convolutional layer with 2×1 kernels and 2×1 stride allowing to firstly get feature maps for the pairs (a, b) and (c, d). At the end of this process (a, b) and (c, d) are reduced to one dimension regarding the width while the other dimension represents the embedding size. The size of the output is $2 \times EMBEDDING\_SIZE$.
2. We fed the output to a second convolutionnal layer with 2×2 kernels and 2×2 stride so the features maps of (a, b) and (c, d) are now reunited in one single dimension across width, the embedding size is divided by 2 too.
3. We then apply dropout and feed the output to a singular linear layer, we use sigmoid activation function to compute a confidence score for the 2 sentence pairs being in an analogical proportion relation.

Figure 1: Diagram representing the CNN architecture, each Convolutional layer is followed by a ReLU layer. Dropout is applied before linear layer. $Dim_1 = Embedding\_size * 2$ is the number of feature maps after passing through the first convolutional layer, $Dim_2 = Embedding\_size/2$ is the number of feature maps after passing through the second convolutional layer. The flatten operation outputs a tensor of size Dim_2^2.



5. Experiments and Results

For both architectures we ranged learning rate between 10^{-4} and 10^{-5}, and dropout from 0.1 to 0.3. We used Adam optimizer with default PyTorch settings and Binary Cross Entropy Loss. Results for both architectures and combinations of embeddings are shown in Table 2.

5.1. Transformer-based Language Models vs GloVe

Transformer-based Language Models outperform GloVe almost constantly in terms of accuracy and F1-score (ability to recognize valid analogical proportions). While the scores are not significantly higher, we can still conclude that contextual embeddings provide better handling of latent relations and analogies between sentences in comparison to static embeddings. Let us note also that representing a sentence by the mean of its contextual word vectors outperforms the CLS sentence representation.

		Precision	Recall	F1	Accuracy
	PDTB				
GloVe-mean	class 1	59.35	53.473	56.259	53.855
	class 0	48.36	54.331	51.172	
BERT-base-mean	class 1	59.39	55.847	**57.564**	56.218
	class 0	53.045	56.639	54.783	
BERT-base-CLS	class 1	56.02	56.132	56.076	56.12
	class 0	56.22	56.108	56.164	
roBERTa-base-mean	class 1	49.835	57.364	53.335	**56.398**
	class 0	62.96	55.655	59.083	
roBERTa-base-CLS	class 1	42.995	55.256	48.361	54.09
	class 0	65.185	53.347	58.675	
	SNLI				
GloVe-mean	class 1	64.17	62.346	63.245	62.708
	class 0	61.245	63.09	62.154	
BERT-base-mean	class 1	70.215	64.32	67.138	65.633
	class 0	61.05	67.21	63.982	
BERT-base-CLS	class 1	71.065	62.654	66.595	64.353
	class 0	57.64	66.578	61.787	
roBERTA-base-mean	class 1	70.315	64.474	**67.268**	**65.785**
	class 0	61.255	67.358	64.162	
roBERTa-base-CLS	class 1	70.27	61.713	65.714	63.338
	class 0	56.405	65.484	60.607	
	MRPC				
GloVe-mean	class 1	60.43	65.066	**62.662**	63.992
	class 0	67.555	63.062	65.231	
BERT-base-mean	class 1	51.03	65.473	57.356	62.06
	class 0	73.09	59.88	65.829	
BERT-base-CLS	class 1	42.94	62.2	50.806	58.422
	class 0	73.905	56.431	63.997	
roBERTa-base-mean	class 1	56.35	66.372	60.952	**63.9**
	class 0	71.45	62.076	66.434	
roBERTa-base-CLS	class 1	50.995	61.307	55.677	59.405
	class 0	67.815	58.051	62.554	

(a) Results for CNN

		Precision	Recall	F1	Accuracy
	PDTB				
GloVe-mean	class 1	39.39	55.296	46.007	53.773
	class 0	68.155	52.93	59.585	
BERT-base-mean	class 1	48.705	57.66	52.805	**56.47**
	class 0	64.235	55.6	59.607	
BERT-base-CLS	class 1	50.285	56.662	**53.284**	55.913
	class 0	61.54	55.314	58.261	
roBERTa-base-mean	class 1	45.41	56.873	50.499	55.487
	class 0	65.565	54.567	59.563	
roBERTa-base-CLS	class 1	48.66	56.297	52.2	55.442
	class 0	62.225	54.792	58.273	
	SNLI				
GloVe-mean	class 1	68.96	62.224	65.419	63.547
	class 0	58.135	65.192	61.462	
BERT-base-mean	class 1	69.635	66.681	68.126	67.42
	class 0	65.205	68.227	66.682	
BERT-base-CLS	class 1	65.72	64.787	65.25	65.0
	class 0	64.28	65.219	64.746	
roBERTa-base-mean	class 1	68.515	67.887	**68.2**	**68.053**
	class 0	67.59	68.221	67.904	
roBERTa-base-CLS	class 1	73.18	61.496	66.831	63.68
	class 0	54.18	66.889	59.867	
	MRPC				
GloVe-mean	class 1	44.605	60.167	51.23	57.537
	class 0	70.47	55.989	62.4	
BERT-base-mean	class 1	57.285	58.757	58.012	58.537
	class 0	59.79	58.329	59.05	
BERT-base-CLS	class 1	58.99	57.771	58.374	57.935
	class 0	56.88	58.106	57.486	
roBERTa-base-mean	class 1	59.375	58.222	58.793	58.385
	class 0	57.395	58.554	57.969	
roBERTa-base-CLS	class 1	60.565	59.117	**59.832**	**59.34**
	class 0	58.115	59.575	58.836	

(b) Results for MLP

Table 2
Results

5.2. Performance across corpora

Overall scores for the SNLI dataset are the highest with accuracy ranging from 62.708 to 68.01 and F1-score peaking at 68.2 across MLP and CNN. Scores for MRPC are a bit lower with accuracy ranging from 57.537 to 63.992 and F1-score peaking at 62.662 considering CNN only as it constantly outperforms MLP. The classifiers had a harder time grasping analogies on the PDTB corpus with accuracy ranging from 53.773 to 56.47 and F1-score peaking at 57.564, F1-score being below 50 for roBERTa-base-CLS/CNN and GloVe-mean/MLP. The classifiers had a harder time grasping analogies on the PDTB corpus with accuracy ranging from 53.773 to 56.47 and F1-score peaking at 57.564, F1-score being below 50 for roBERTa-base-CLS/CNN and GloVe-mean/MLP. This can be explained by the fact that the number of latent relations that we had to handle in PDTB is much higher (5 latent relations) than the latent relations that we have in the MRPC or the SNLI corpora. We assume that providing more data will yield better overall results.

5.3. BERT vs roBERTa

One main difference between BERT and roBERTa is respectively the presence and absence of the Next sentence prediction training task. While BERT considered this task to be beneficial for the learning of long range dependencies roBERTa considered this task counter-productive.

Although roBERTa performs slightly better than BERT (considering the mean-pooling sentence representation method) we cannot draw a definitive conclusion about the utility of the Next Sentence Prediction training task for analogical properties learning. A bigger training set may have enforced the tendency. roBERTa outperformed BERT for SNLI and MRPC, the two corpora for which the sentences from the sentence pairs do not follow each other in a natural context. The next sentence prediction may be detrimental in this case.

5.4. CNN vs MLP

Both MLP and CNN are relevant for the classification task we performed as they have almost similar results with the CNNs usually outperforming the MLPs, but not always. However the MRPC's results show a large difference in performance between the 2 classifiers, the CNNs performing significantly better than the MLPs. Meaningful features were extracted from the sentences representations. As we described Section 4 features are first extracted from the (a, b) pair and the (c, d) pair in tandem. This could probably be attributed due to teh fact that paraphrases use semantically similar words which probably are closer to the vector space which is better captured by CNNs than MLPs, although further analysis is needed in order for this claim to be verified.

6. Conclusions and Future Work

In this paper we have focused on the problem of identifying analogies between pairs of sentences based on common latent relations that exist or not between the pairs. We have used both contextual embeddings (BERT en roBERTa) as well as static embeddings (GloVe). Both BERT and roBERTa outperformed GloVe at the binary classification task we performed. We believe an error analysis or a different classification task might shed more light on those results. In conclusion this work scratches the surface of Transformer-based Language Models' ability to encode analogical properties. Our experiments show that embeddings issued from Transformer-based architectures can better capture analogies via the identification of common latent relations, in comparison to static embedding approaches. Nonetheless it is premature to conclude that such architectures can indeed capture more broadly the mechanism of analogy making.

Acknowledgments

The authors would like to thank the anonymous reviewers for their thoughtful and constructive comments. This work has been partially funded by the ANR AT2TA project, grant number ANR-22-CE23-0023.

References

[1] Aristotle, Poetics, 384—322 BCE.

[2] D. R. Hofstadter, Analogy as the Core of Cognition, in: D. Gentner, K. J. Holyoak, B. N. Kokinov (Eds.), The Analogical Mind: Perspectives from Cognitive Science, The MIT Press, Cambridge, Massachusetts, 2001, pp. 499–538.

[3] D. Hofstadter, E. Sander, Surfaces and Essences: Analogy as the Fuel and Fire of Thinking, Basic Books, 2013.

[4] F. Chollet, On the measure of intelligence, 2019. arXiv:1911.01547.

[5] N. Barbot, L. Miclet, H. Prade, Analogy between concepts, Artificial Intelligence 275 (2019) 487–539.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: C. J. C. B. et al. (Ed.), Advances in Neural Information Processing Systems 26, Curran Associates Inc., 2013, pp. 3111–3119.

[7] T. Mikolov, W.-t. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, 2013, pp. 746–751. URL: https://aclanthology.org/N13-1090.

[8] S. Lim, H. Prade, G. Richard, Classifying and completing word analogies by machine learning, International Journal of Approximate Reasoning 132 (2021) 1–25. URL: https://www.sciencedirect.com/science/article/pii/S0888613X21000141. doi:https://doi.org/10.1016/j.ijar.2021.02.002.

[9] S. Lim, H. Prade, G. Richard, Solving word analogies: A machine learning perspective, in: European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty, volume 11726, Springer, 2019, pp. 238–250. URL: https://doi.org/10.1007/978-3-030-29765-7_20. doi:10.1007/978-3-030-29765-7_20.

[10] S. Alsaidi, A. Decker, P. Lay, E. Marquer, P.-A. Murena, M. Couceiro, A neural approach for detecting morphological analogies, in: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, pp. 1–10. doi:10.1109/DSAA53316.2021.9564186.

[11] P. D. Turney, The Latent Relation Mapping Engine: Algorithm and Experiments, Journal of Artificial Intelligence Research 33 (2008) 615–655.

[12] X. Zhu, G. de Melo, Sentence analogies: Linguistic regularities in sentence embeddings, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 3389–3400. URL: https://aclanthology.org/2020.coling-main.300. doi:10.18653/v1/2020.coling-main.300.

[13] S. D. Afantenos, T. Kunze, S. Lim, H. Prade, G. Richard, Analogies between sentences: Theoretical aspects - preliminary experiments, in: J. Vejnarová, N. Wilson (Eds.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 16th European Conference, ECSQARU 2021, Prague, Czech Republic, September 21-24, 2021, Proceedings, volume 12897 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 3–18. URL: https://doi.org/10.1007/978-3-030-86772-0_1. doi:10.1007/978-3-030-86772-0_1.

[14] S. D. Afantenos, S. Lim, H. Prade, G. Richard, Theoretical study and empirical investigation of sentence analogies, in: M. Couceiro, P. Murena (Eds.), Proceedings of the Workshop on the Interactions between Analogical Reasoning and Machine Learning (International Joint

Conference on Artificial Intelligence - European Conference on Artificial Intelligence (IJAI-ECAI 2022)), Vienna, Austria, July 23, 2022, volume 3174 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022, pp. 15–28. URL: http://ceur-ws.org/Vol-3174/paper2.pdf.

[15] O. Sultan, D. Shahaf, Life is a circus and we are the clowns: Automatically finding analogies between situations and processes, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 3547–3562. URL: https://aclanthology.org/2022.emnlp-main.232.

[16] W. B. Dolan, C. Brockett, Automatically constructing a corpus of sentential paraphrases, in: Proceedings of the Third International Workshop on Paraphrasing (IWP2005), 2005. URL: https://aclanthology.org/I05-5002.

[17] R. Prasad, N. Dinesh, A. Lee, E. Miltsakaki, L. Robaldo, A. Joshi, B. Webber, The Penn Discourse TreeBank 2.0., in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/754_paper.pdf.

[18] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2015.

[19] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543. URL: http://www.aclweb.org/anthology/D14-1162.

[20] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[21] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. URL: http://arxiv.org/abs/1907.11692, cite arxiv:1907.11692.

[22] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in Neural Information Processing Systems, volume 26, Curran Associates Inc., 2013, pp. 3111–3119. URL: https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[23] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, in: International Conference on Learning Representations, Workshop, 2013. URL: https://arxiv.org/abs/1301.3781. doi:10.48550/ARXIV.1301.3781.

[24] A. Gladkova, A. Drozd, S. Matsuoka, Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't., in: North American Chapter of the Association for Computational Linguistics, Student Research Workshop, Association for Computational Linguistics, San Diego, California, 2016, pp.

8–15. URL: https://aclanthology.org/N16-2002. doi:10.18653/v1/N16-2002.

[25] A. Rogers, A. Drozd, B. Li, The (too many) problems of analogical reasoning with word vectors, in: Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Vancouver, Canada, 2017, pp. 135–148. URL: https://aclanthology.org/S17-1017. doi:10.18653/v1/S17-1017.

[26] O. Levy, Y. Goldberg, Linguistic regularities in sparse and explicit word representations, in: Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Association for Computational Linguistics, Ann Arbor, Michigan, 2014, pp. 171–180. URL: https://aclanthology.org/W14-1618. doi:10.3115/v1/W14-1618.

[27] S. Lim, H. Prade, G. Richard, Solving word analogies: A machine learning perspective, in: Proc.15th Europ. Conf. Symb. & Quantit. Appr. to Reas. with Uncert. (ECSQARU), LNCS 11726, 238-250, Springer, 2019.

[28] L. Wang, Y. Lepage, Vector-to-sequence models for sentence analogies, in: International Conference on Advanced Computer Science and Information Systems, 2020, pp. 441–446. doi:10.1109/ICACSIS51025.2020.9263191.

[29] D. Gentner, Structure Mapping: A Theoretical Framework for Analogy, Cognitive Science 7 (1983) 155–170. URL: https://doi.org/10.1207/s15516709cog0702_3. doi:10.1207/s15516709cog0702_3.

A Framework for Neural Machine Translation by Fuzzy Analogies

Liyan Wang[1,*], Bartholomäus Wloka[2] and Yves Lepage[1]

[1]Waseda University, Kitakyushu, 808-0135, Japan
[2]University of Vienna, Vienna, 1190, Austria

Abstract
This paper introduces a novel translation technique, driven by modeling fuzzy analogies that capture approximate conformity to parallel transformations between fragments in sentences. We conduct preliminary experiments on English-Japanese translations with a data set of limited size. The results show the potential of using fuzzy analogies for translation, achieving an increase of about 6 BLEU points compared to NMT.

Keywords
Machine translation, Fuzzy analogy, Limited data

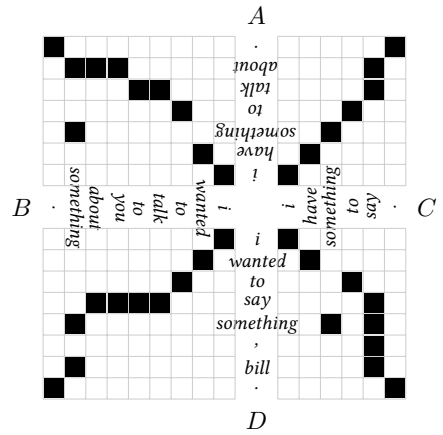
1. Introduction

Low resource settings pose significant challenges to modern Machine Translation (MT) systems [1, 2]. Neural MT (NMT) with large-scale models require large amounts of parallel data to fine-tune learnt weights of two language spaces [3]. MT by analogy (i.e, example-based MT) [4, 5], enables tracing translations by structuring knowledge from examples. It relies on strict analogies that involve ratios with the exact same transformation rule [6]. However, finding sentence analogies with strictness on form can be difficult, particularly in cases where there are less correlated sentences in relatively small sized corpora. In this paper, we propose to explore partial analogies between sentences, which capture approximate conformity between ratios relying on fuzzy matches, i.e., ratios which are partial transformations are matched. For example, *I feel ridiculous.* : *That is untrue.* :: *I feel funny.* : *That is funny.* is a quadruple that captures parallel transformation on sentence fragments. We call this **fuzzy analogy**.

2. Methodology

The proposed method is built on the indirect paradigm of example-based MT in [5]. Similar to this, given translation queries D, we first construct sentence analogies as $A : B :: C : D$,

IARML@IJCAI'2023: Workshop on the Interactions between Analogical Reasoning and Machine Learning, at IJCAI'2023, August, 2023, Macao, China
*Corresponding author.
wangliyan0905@toki.waseda.jp (L. Wang); bartholomaeus.wloka@univie.ac.at (B. Wloka); yves.lepage@waseda.jp (Y. Lepage)
0000-0002-9561-5037 (L. Wang); 0000-0002-7484-878X (B. Wloka); 0000-0002-3059-4271 (Y. Lepage)



i	:	*i*	::	*i*	:	*i*	1
have	:	*wanted*	::	*have*	:	*wanted*	1
to	:	*to*	::	*to*	:	*to*	1
talk	:	*talk*	::	*say*	:	*say*	1
something	:	*something*	::	*something*	:	*something*	1
				say	:	*,*	0
something	:	*something*	≠	*say*	:	*bill*	0
.	:	*.*	::	*.*	:	*.*	1

$$\frac{1+1+1+1+1+0+0+1}{|D|} = \frac{6}{8} = 0.75$$

Figure 1: Computation of analogical score. Black cells in matrices indicate alignment points between tokens. Each token quadruple linked by alignments, one for each (sub-)word in D, is checked for trivial analogy. We divide by the length of D to get the analogical score. If $A : B :: C : D$ is a strict analogy, the score is 1.

where A, B, and C are source examples retrieved from translation memory, that will maximize analogical score with D. By looking up the annotated translations of (A, B, C), we can obtain corresponding analogical equations in the target language. Following this, we exploit a previously learnt model to generate solutions of target analogies as translation results, i.e., $A' : B' :: C' : x \quad \Rightarrow \quad x = D'$.

To **retrieve** sentence analogies, we first pre-compute candidate pools for terms A, B, and C by collecting the k nearest neighbors of D using cosine similarity between sentence embeddings. Theoretically, there will be a cubic number of possible combinations of sentence quadruples (A, B, C, D). To reduce the computational cost, we prune candidate quadruples. We leave out the quadruples with no lexical overlap between A and C, and between B and D. Finally, for each D, we rank the quadruples by analogical score, and select the first n ones. As in [7], we use alignments between (A, B, C, D) considered as sequences of (sub-)words. We count the number of trivial analogies of the form $a : a :: b : b$ or $a : b :: a : b$ for every aligned (sub-)word quadruple. Figure 1 illustrates the computation of analogical scores.

Next, we **train** a sequence-to-sequence model to solve analogies, so as to derive translation answers. Suppose $A : B :: C : D$ and $A' : B' :: C' : D'$ are a retrieved source analogy and its corresponding translation. We concatenate 7 sentences (excluding D') in two monolingual analogies as input X, to train the model to generate the solution D' by optimizing cross-entropy (CE) between probability distributions conditional on the context of input and preceding target tokens:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|D'|} \log P(D'_i | D'_{<i}, X) \tag{1}$$

To encourage the model to be more confident in reconstructing target fragments that are in analogical relationships, while being flexible to non-analogical relationships, we introduce a

weighting scalar in (1). Formally, the aim is to minimize weighted CE (WCE):

$$\mathcal{L}_{\text{WCE}} = -\sum_{i=1}^{|D'|} w_i \log P(D'_i | D'_{<i}, X) \tag{2}$$

where w_i takes the value of 1 for trivial analogies, and 0.5 else. For each target token, a weighted value is determined by its corresponding aligned token in D.

3. Preliminary Experiments

3.1. Datasets

We experiment with parallel sentences from the Japanese-English Subtitle Corpus[1], with 50,000 pairs for training, 2,000 for validation, and 2,000 for test. In this work, we primarily investigate the translation quality from English to Japanese. The source sentences contain approximately nine words on average. For each data set, we take source sentences as queries and look for fuzzy analogies from the source part of translation memory (i.e., the training set). The strictness in analogies depends on how closely the queries match the examples in memory. We assess the closeness between the data sets and the memory by computing the similarity using the length of longest common subsequence between sentences at the word level. Specifically, we compare the query sentence to the twenty most similar examples in the memory, excluding itself in the case of the training set. Table 1 shows the statistics of three data sets. On average, the three data sets exhibit similar characteristics, where source sentences are found to have an overlap of four words with their corresponding similar sentences in the memory.

Table 1
Data statistics for the English-to-Japanese translation task, specifically pertaining to the source side of the data sets. Closesness to memory indicates the average number of words that overlap with each of the twenty most similar examples retrieved from the memory.

	Training	Validation	Test
Number of parallel sentences	50,000	2,000	2,000
Sentence length	9 ± 3	9 ± 3	9 ± 3
Number of word types	24,689	3,425	3,348
Closeness to memory	3.89 ± 0.96	3.89 ± 0.95	3.91 ± 0.94

3.2. Implementation Details

In order to retrieve analogies from the corpus, we first use a Sentence-BERT [8] model[2] to represent sentences as vectors. Subsequently, for each query D, we collect twenty examples as the candidates of B and C, which are the nearest neighbors to D in the embedding space.

[1] https://nlp.stanford.edu/projects/jesc/
[2] https://www.sbert.net/docs/pretrained_models.html

Sentences A are selected from the twenty closest neighbors to each candidate for B. We pre-tokenize sentences into sub-words using a SentencePiece [9] model with the vocabulary size of 250,000[3]. We then enumerate (A, B, C, D) from collected candidates and filter possible quadruples by the overlap constraint between A and C, and between B and D at the sub-word level. Next, We use mGIZA [10] and Moses[4] to estimate sub-sentential alignments. Base on that, we compute analogical score for each possible quadruple. For each D, we select one fuzzy analogy for translation.

To learn from analogy, we fine-tune a pre-trained mBART [3] model[5] on fuzzy analogies that are retrieved from the training set. We utilize the large-scale mBART model consisting of a 12-layer encoder and a 12-layer decoder. The target sentences are generated using a beam size of 5 during decoding. To fine-tune the model, we freeze the encoder part and update the parameters of the last 6 layers of the decoder. The frozen model is trained using a batch size of 8 for a maximum of 20 epochs. In the case there are no improvements for three consecutive epochs, we halt the training process before completing all the epochs (early stopping). Finally, we save the model that demonstrates the best performance on the validation set.

3.3. Results and Analysis

We compare to an NMT system by fine-tuning the same pre-trained mBART model on the data sets of parallel sentences. The baseline NMT model is trained using the consistent settings as described above. On 50,000 parallel sentences, NMT obtains a BLEU score of only 2.9. Our system using (1) achieved an improvement of 5.6 and the use of (2) leads to a further gain of about 0.4 BLEU points. Even though fuzzy analogies relax the strictness, the inclusion of partial evidence in parallel transformations still helps in deducing possible translation.

In the retrieved analogies, query sentences are covered by examples under the analogy constraint to different extents with analogical scores ranging from 0 to 1. Figure 2 shows the number of fuzzy analogies constructed for the sentences in the three data sets, categorized by their respective scores. In general, three sets of analogy data demonstrate a comparable distribution in the extent of fuzzy matches between sentence transformations. The majority of analogies fall within the score range of 0.3 to 0.7. This indicates that approximately 30%-70% of tokens in query sentences are associated with examples in the analogy relationship.

Next, we examinate the model performance in inferring translation answers by solving fuzzy analogies with different scores. Figure 3 shows that our model is capable of reasoning analogies with lower scores, where less than half of a query sentence is linked to translation examples through analogical associations. This suggests that fuzzy analogies can capture relative knowledge of two languages, which can even assist in translating queries that are distant from memory. We also compare to an NMT baseline on translating each test sentence. In Figure 3, blue points (415 out of 2,000) indicate the cases where our model performs worse than NMT in BLEU. Relatively, there are fewer underperforming cases when analogies have higher scores ($>$0.7). In Table 2, we list examples of two methods in translating sentences that are either close

[3]To enable the learning model (e.g., mBART) to identify analogical transformations in quadruples, we use the SentencePiece model with the same tokenization as in mBART.

[4]http://www2.statmt.org/moses/

[5]https://huggingface.co/facebook/mbart-large-50

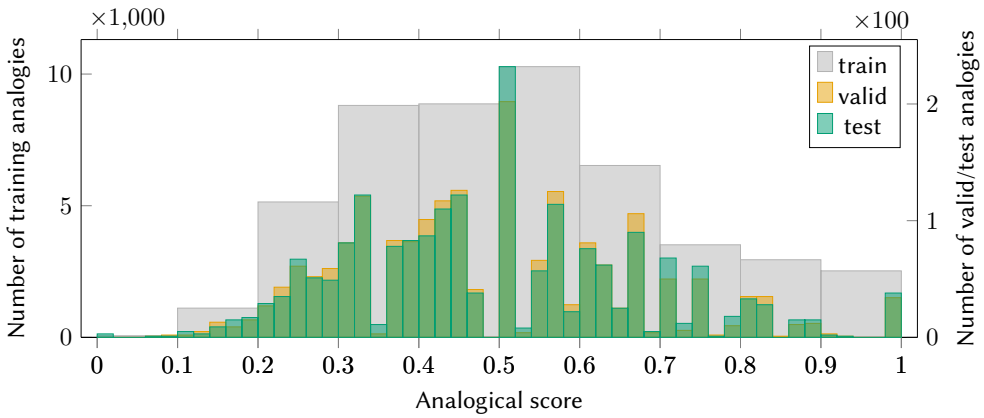


Figure 2: Distribution of analogical scores for fuzzy analogies retrieved from three data sets. Note that there are two different vertical scales: one for training, one for validation and test. The scales for training is ten times more than the second one.

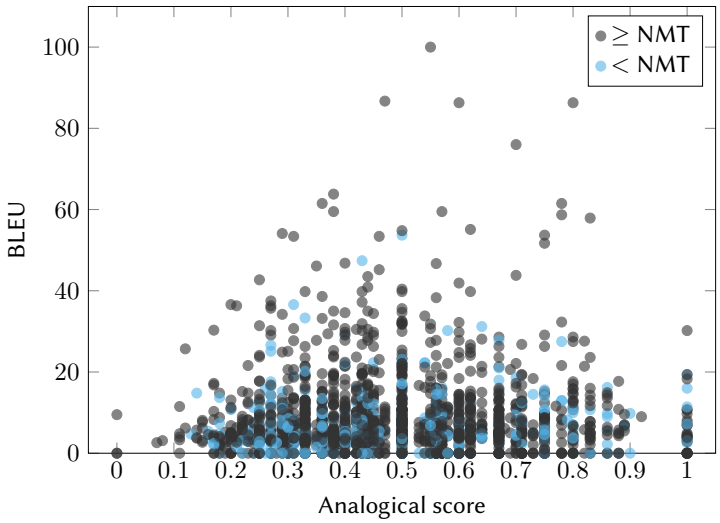


Figure 3: BLEU scores against analogical scores for test analogies. The test cases where our model outperforms or achieves equal performance to NMT are represented by gray points, while the remaining cases are denoted in blue.

to or distant from translation memory.

Do different source analogies constructed for the same query result in diverse translation outputs? We conduct additional experiments to address this question. For each test sentence, we retrieve five fuzzy analogies with the maximum scores and then employ the model trained specifically to handle one analogy per query to solve each of these analogies. Table 3 presents five distinct translations to a query sentence. As shown by the example, it is possible for the model to generate more idiomatic translations that closely convey the intended meaning, using

Table 2
Examples of translating sentences that are either close to or distant from the memory. For each generated Japanese answer, we also provide the corresponding English translation below, which has been translated using Google Translate. The underlines denote formal matchs to the references. For the upper instance, NMT achieves a higher BLEU score but it fails to specifically mention the content regarding worry, whereas our model effectivly captures the meaning of the original sentence. Regarding the translation of the distant query that involves specialized terms not present in the memory, NMT seems to draw upon knowledge from its pre-trained data. However, the translation is not accurate. Our model provides a translation for the word *"called"* by leveraging hints from the analogy, but does not convery an accurate translation for the term itself.

Test query close to the memory		BLEU
Query	*ah... i'm sorry i made you worry.*	
Ref.	あ... 心配かけてすみません。	
NMT	あ... ごめんなさいごめんなさい。 (*ah... i'm sorry, i'm sorry.*)	15.8
Ours	analogical score: 0.82 *i'm sorry for bothering you. thank you.* : *i... i'm sorry. i wasn't being careful enough.* :: *i'm fine. i'm sorry for making you worry.* : *ah... i'm sorry i made you worry.* すいませんでした。失礼します! : すす・すみませんでした私が十分な注意を払っていませんでした :: 大丈夫です。すいませんご心配おかけして : x ⇒ x = あ... ごめんね心配かけちゃって。 (*ah... i'm sorry i made you worry.*)	14.9
Test query distant from the memory		**BLEU**
Query	*called alpha lipoxanthine glucoside*	
Ref.	[スピーカ]「アルファー・リボキサンチン・グルコシド」n略して	
NMT	アルファリン酸オキシトリン酸グリシトリン酸 (*alpha phosphate oxyphosphate glycitrate*)	0.0
Ours	analogical score: 0.0 *that's the name.* : *it is called alexon biotech company.* :: *the name is abraham lincoln.* : *called alpha lipoxanthine glucoside* そう名付けることにしたよ。 : それはアレクソン・バイオテックという会社なんですけど :: 名はエイブラハム・リンカーン : x ⇒ x = アルパルス・グローブという名で (*under the name alpuls grove*)	0.0

analogies with less evidence. We speculate that enlarging the number of fuzzy analogies will facilitate models in acquiring more potential associations in two languages.

4. Conclusion and Future Work

In this paper, we introduced a novel translation approach based on the mechanism of using indirect analogies for translation. Unlike the work in [5], we proposed to handle partial analogies that capture approximate conformity between sentence transformations. We call that fuzzy analogies. To solve fuzzy analogies between sentences, we trained an mBART model to generate translations given source quadruples and three known translations in the target analogies. We conducted a comparison between our approach and an NMT baseline under low resource constraints. Additionally, we investigated the impact of analogical quality on translation.

In future work, we will conduct ablation studies to search for optimal configurations for modeling analogies. In addition, we will expand this work to different language pairs and directions, as well as investigate the influence of corpus size on performance.

Acknowledgments

The research reported in this paper was supported in part by a grant for Kakenhi (kiban C) from the Japanese Society for the Promotion of Science (JSPS), n° 21K12038 "Theoretically founded algorithms for the automatic production of analogy tests in NLP".

References

[1] R. Aharoni, M. Johnson, O. Firat, Massively multilingual neural machine translation, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3874–3884. URL: https://aclanthology.org/N19-1388. doi:10.18653/v1/N19-1388.

[2] J. Gu, Y. Wang, Y. Chen, V. O. K. Li, K. Cho, Meta-learning for low-resource neural machine translation, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 3622–3631. URL: https://aclanthology.org/D18-1398. doi:10.18653/v1/D18-1398.

[3] Y. Liu, J. Gu, N. Goyal, X. Li, S. Edunov, M. Ghazvininejad, M. Lewis, L. Zettlemoyer, Multilingual denoising pre-training for neural machine translation, Transactions of the Association for Computational Linguistics 8 (2020) 726–742. URL: https://aclanthology.org/2020.tacl-1.47. doi:10.1162/tacl_a_00343.

[4] M. Nagao, A framework of a mechanical translation between Japanese and English by analogy principle, Artificial and human intelligence (1984) 351–354.

[5] Y. Lepage, E. Denoual, Purest ever example-based machine translation: Detailed presentation and assessment, Machine Translation 19 (2005) 251–282. doi:10.1007/s10590-006-9010-x.

[6] P.-A. Murena, M. Al-Ghossein, J.-L. Dessalles, A. Cornuéjols, Solving analogies on words based on minimal complexity transformation, in: C. Bessiere (Ed.), Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20, Interna-

tional Joint Conferences on Artificial Intelligence Organization, 2020, pp. 1848–1854. URL: https://doi.org/10.24963/ijcai.2020/256. doi:10.24963/ijcai.2020/256, main track.

[7] L. Miclet, S. Bayoudh, A. Delhay, Analogical dissimilarity, J. Artif. Int. Res. 32 (2008) 793–824.

[8] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[9] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71. URL: https://aclanthology.org/D18-2012. doi:10.18653/v1/D18-2012.

[10] Q. Gao, S. Vogel, Parallel implementations of word alignment tool, in: Software Engineering, Testing, and Quality Assurance for Natural Language Processing, Association for Computational Linguistics, Columbus, Ohio, 2008, pp. 49–57. URL: https://aclanthology.org/W08-0509.

Table 3
Discrepency between translation answers for the query *"i think it's better if you frankly say your intentions."* by solving the first five fuzzy analogies with different analogical scores. The reference translation is "あの... ズバッと訴えたい事を伝えたほうがいいと思うんです。". We highlight the answer that closely matches the reference according to the human evaluation.

score	Analogy and solution							BLEU
0.47	*i think that's a good idea.*	:	*i think that's best.*	::	*i think it's great to do what you like.*	:	*i think it's better if you frankly say your intentions.*	
	はい。それがいいかと思います。	:	確かに、それが良いかもな	::	好きなことを頑張るっていいことだと思いますけど	:	x	
	⇒ x = じゃあはっきり言うなら... 。 (*well, to put it bluntly...*)							1.3
0.40	*i guess it's better on my side.*	:	*yep, i guess it's probably for the best.*	::	*i think that's best.*	:	*i think it's better if you frankly say your intentions.*	
	まだ俺の方がいいや。へへッ。何だよ? これ。痛っ。	:	あぁ、最良だったと思うよ	::	確かに、それが良いかもな	:	x	
	⇒ x = はっきり言うなら... 意地悪いのは悪いけど。 (*to put it bluntly... sorry for being mean.*)							2.4
0.40	*i think that's best.*	:	*yep, i guess it's probably for the best.*	::	*i think it's great to do what you like.*	:	*i think it's better if you frankly say your intentions.*	
	確かに、それが良いかもな	:	あぁ、最良だったと思うよ	::	好きなことを頑張るっていいことだと思いますけど	:	x	
	⇒ x = はっきり言うなら... 意図は正直に言うべきよ (*to put it bluntly... you should be honest about your intentions*)							2.0
0.40	*i guess it's better on my side.*	:	*i think that's best.*	::	*yep, i guess it's probably for the best.*	:	*i think it's better if you frankly say your intentions.*	
	まだ俺の方がいいや。へへッ。何だよ? これ。痛っ。	:	確かに、それが良いかもな	::	あぁ、最良だったと思うよ	:	x	
	⇒ x = はっきり言うなら君の意地悪は悪いけどな。 (*to put it bluntly, your meanness is bad.*)							2.0
0.40	*yep, i guess it's probably for the best.*	:	*i think that's best.*	::	*i guess it's better on my side.*	:	*i think it's better if you frankly say your intentions.*	
	あぁ、最良だったと思うよ	:	確かに、それが良いかもな	::	まだ俺の方がいいや。へへッ。何だよ? これ。痛っ。	:	x	
	⇒ x = 正直なところ... 意地悪いって言ったら (*honestly... if you say mean*)							1.8