Machine Learning and Knowledge Engineering to enable sensemaking: Case Studies in Robotics and News Analytics

Prof Enrico Motta

Knowledge Media Institute, The Open University, United Kingdom

Media Futures Centre, University of Bergen, Norway

Preamble

Machine Learning (ML) is currently the dominant paradigm in AI. Nonetheless, explicit Knowledge Representation (KR) and, more in general, Knowledge Engineering (KE), still play a crucial role in AI research and AI system development.

In this presentation I will describe two scenarios in which we combine ML and KR.

In the first one, **robotics**, KR is used to improve the performance of a robot trying to make sense of the surrounding world.

In the second one, **news analytics**, KR is used to formalise the domain in hand prior to the implementation of computational solutions for news analytics.

Knowledge Media Institute

The Open University

Case study #1: The Intelligent Health and Safety Inspector

KMi's Agnese Chiatti wins prestigious L'Oréal-UNESCO Award for Women in Science

Enrico Motta, Monday 13 Jun 2022

At a ceremony held this morning in Milan, Agnese Chiatti has received the "L'Oréal-UNESCO For Women in Science" prize, in recognition of the excellence of her research in Robotics.

This award, which was established in October 2002 by L'Oréal Italia, in collaboration with the Italian National Commission for UNESCO, provides six scholarships of €20,000 each to female researchers under the age of 35, who have completed or are currently pursuing a PhD in the following scientific-disciplinary areas: Life Sciences; Environmental Sciences; Mathematics; Computer and Information Science; Physics; Chemistry; Engineering and Technologies. This year, over 240 applications were received, from which 6 winners were selected.

Hybrid AI: Combining Machine Learning with Knowledge Engineering

Informed Machine Learning – A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems

Laura von Rueden, Sebastian Mayer, Katharina Beckh, Bogdan Georgiev, Sven Giesselbach, Raoul Heese, Birgit Kirsch, Julius Pfrommer, Annika Pick, Rajkumar Ramamurthy, Michal Walczak, Jochen Garcke, Christian Bauckhage, *Member, IEEE*, and Jannis Schuecker

Combining Machine Learning and Semantic Web: A Systematic Mapping Study

ANNA BREIT, Semantic Web Company, Austria
LAURA WALTERSDORFER, TU Wien, Austria
FAJAR J. EKAPUTRA, Vienna University of Economics and Business (WU) and TU Wien, Austria
MARTA SABOU, Vienna University of Economics and Business (WU), Austria
ANDREAS EKELHART, University of Vienna and SBA Research, Austria
ANDREEA IANA, University of Mannheim, Germany
HEIKO PAULHEIM, University of Mannheim, Germany
JAN PORTISCH, University of Mannheim, Germany
ARTEM REVENKO, Semantic Web Company, Austria
ANNETTE TEN TEIJE, Vrije Universiteit (VU) Amsterdam, Netherlands
FRANK VAN HARMELEN, Vrije Universiteit (VU) Amsterdam, Netherlands

In line with the general trend in artificial intelligence research to create intelligent systems that combine learning and symbolic components, a new sub-area has emerged that focuses on combining machine learning (ML) components with techniques developed by the Semantic Web (SW) community – Semantic Web Machine Learning (SWeML for short). Due to its rapid growth and impact on several communities in the last two decades, there is a need to better understand the space of these SWeML Systems, their characteristics, and trends. Yet, surveys that adopt principled and unbiased approaches are missing. To fill this gap, we performed a systematic study and analyzed nearly 500 papers published in the last decade in this area, where we focused on evaluating architectural, and application-specific features. Our analysis identified a rapidly growing interest in SWeML Systems, with a high impact on several application domains and tasks. Catalysts for this rapid growth are the increased application of deep learning and knowledge graph technologies. By leveraging the in-depth understanding of this area acquired through this study, a further key contribution of this paper is a classification system for SWeML Systems which we publish as ontology.

CCS Concepts: • **Information systems** → **Semantic web description languages**; • **Computing methodologies** → **Knowledge representation and reasoning**; **Machine learning**.

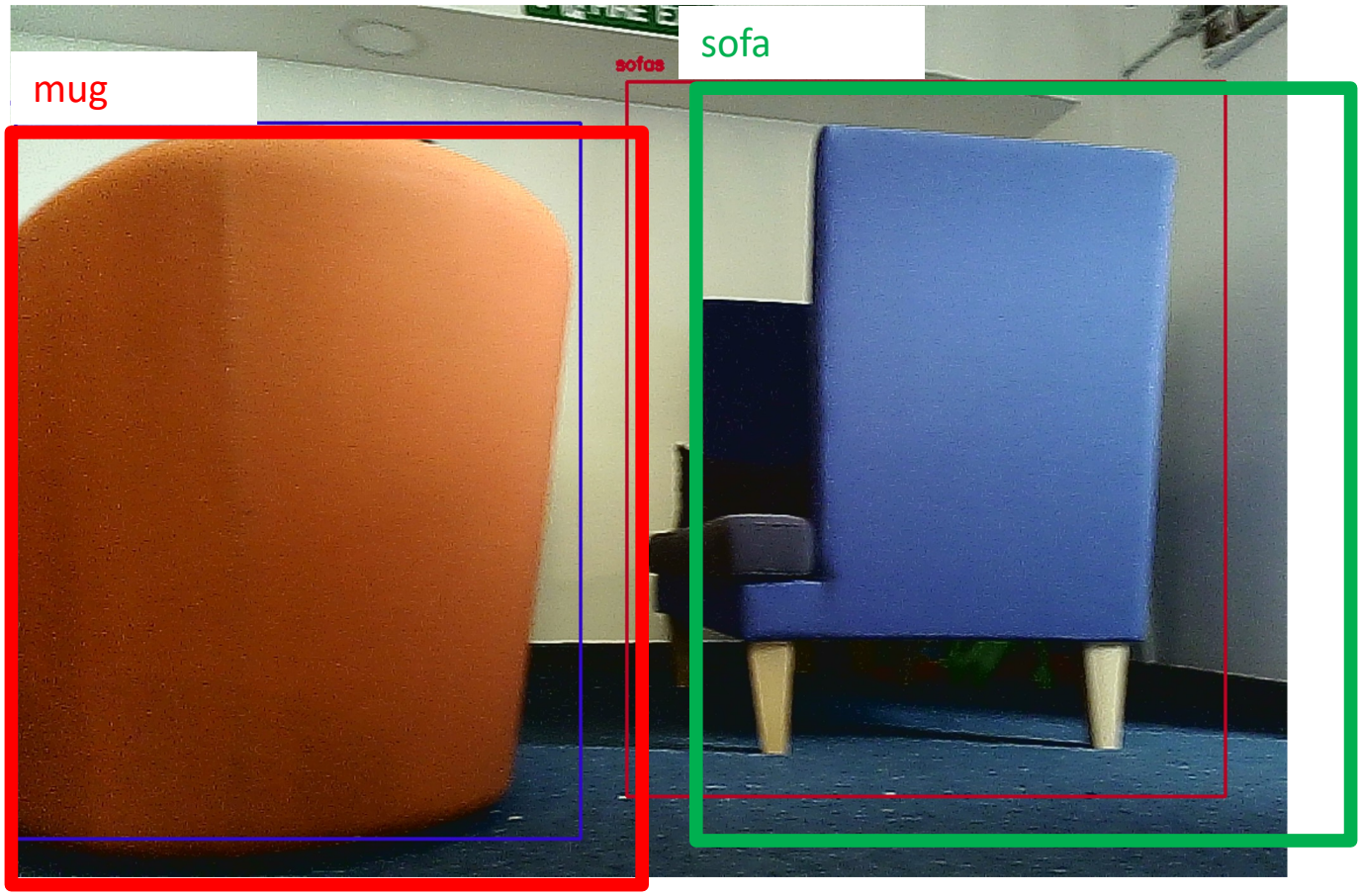
Additional Key Words and Phrases: semantic web, machine learning, artificial intelligence, knowledge graph, knowledge representation and reasoning, neuro-symbolic integration, systematic mapping study

Authors' addresses: Anna Breit, anna.breit@semantic-web.com, Semantic Web Company, Austria; Laura Waltersdorfer, laura.waltersdorfer@tuwien.ac.at, TU Wien, Austria; Fajar J. Ekaputra, fajar.ekaputra@wu.ac.at, Vienna University of Economics and Business (WU) and TU Wien, Austria; Marta Sabou, marta.sabou@wu.ac.at, Vienna University of Economics and Business (WU), Austria; Andreas Ekelhart, andreas.ekelhart@univie.ac.at, University of Vienna and SBA Research, Austria; Andreea Iana, andreea.iana@uni-mannheim.de, University of Mannheim, Germany; Heiko Paulheim, heiko@informatik.uni-mannheim.de, University of Mannheim, Germany; Jan Portisch, jan@informatik.uni-mannheim.de, University of Mannheim, Germany; Artem Revenko, artem.revenko@semantic-web.com, Semantic Web Company, Austria; Annette ten Teije, annette.ten.teije@vu.nl, Vrije Universiteit (VU) Amsterdam, Netherlands; Frank van Harmelen, frank.van.harmelen@vu.nl, Vrije Universiteit (VU) Amsterdam, Netherlands.

Key limitations of machine learning approaches (incl. LLMs)

- **Data hungry**
- **Brittle** – i.e., ML works best under the closed-world assumption and does not handle novel and dynamic scenarios well
  - e.g., object recognition often fails in real world scenarios where objects may change position, light may change, etc...
- Learning is **pattern-based** – inability to learn concepts
  - People learn **concepts** (e.g., the concept of a car), while ML programs simply learn **patterns from data (Lake et al., 2017)**. This aspect is a source of brittleness and limits the possibilities for intelligent behaviour.
  - For instance, object recognition focuses primarily on geometric shape rather than a holistic understanding of an object in its context

Armchair mistaken for mug



Deep Learning-based object recognition system lacks common sense knowledge about size of mugs

Service Robotics

Robots that are able to act (semi-) autonomously to perform service tasks in real-world scenarios

HanS: The Health and Safety Inspector

Situation: A book is next to a portable heater

- - - - - - - - - - - - - - - - - - -

Portable heater is an electric device

Electric devices can overheat when switched on

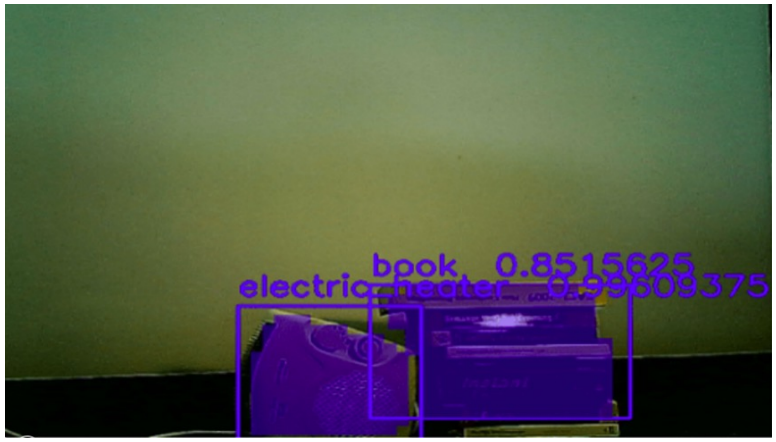
A book is made of paper

Paper is flammable

Health and Safety violation

Required epistemological capabilities

- Knowledge of H&S rules
  - E.g., Flammable material cannot be situated close to electrical appliances
- Ability to identify objects in the environment
  - E.g., recognising an electric heater under different lighting conditions
- Knowledge about relevant domain entities
  - E.g., a book is made of paper and paper is a flammable material
- Knowledge about spatial relations
  - E.g., ability to recognise that the heater is close to the book



The Open University

KMi Knowledge Media Institute

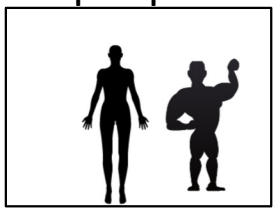
Knowledge Media Institute

The Open University

Starting point: NN-based Object Recognition

Object recognition in KMi

people

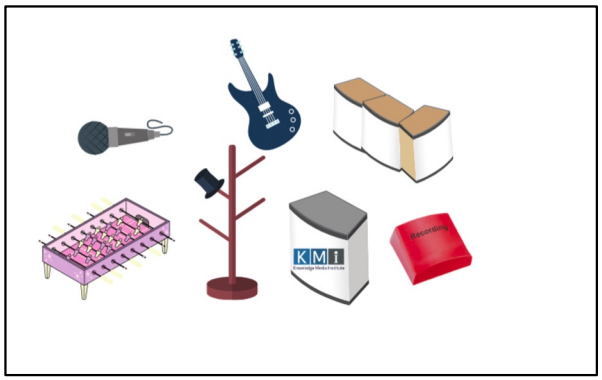


robots

Common objects in office environments

Health & Safety specific

Other miscellaneous stuff



60 distinct classes

1642 instances (i.e., observations)

Performance of best NN reasoners on KMi dataset

Training Set = 240 object regions; Test Set = 1342 object regions

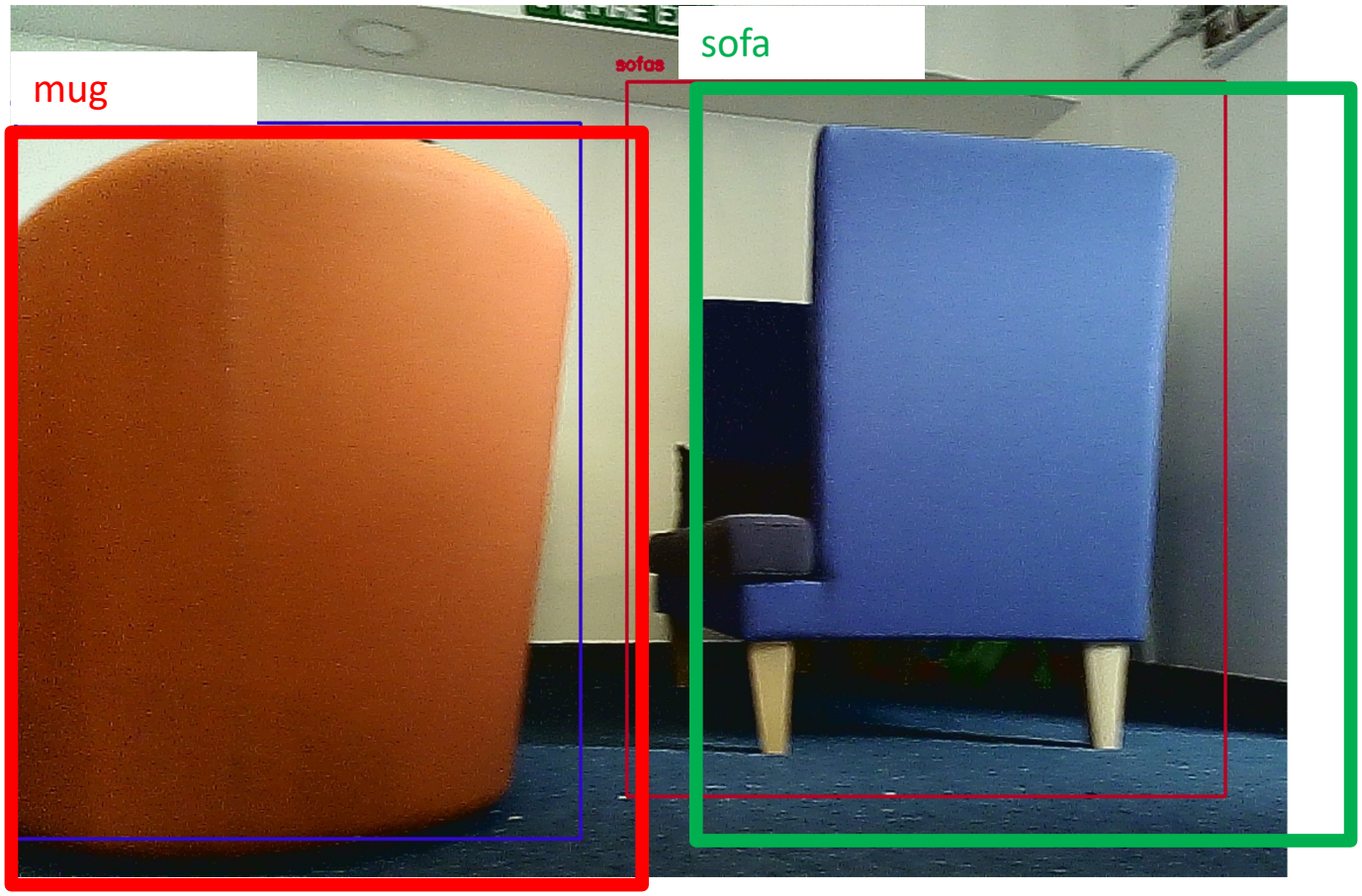
Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
N-net (Zeng et al., 2022)	.45	.34	.40	.31	.62	.45	.47	.33	.36	.63
K-net (Zeng et al., 2022)	**.48**	**.39**	**.40**	**.34**	**.68**	**.48**	**.50**	**.38**	**.41**	**.65**

nDCG@5 -> Normalised Discounted Cumulative Gain of top-5 ranking

Correct predictions appearing lower in the ranking are penalised

"Normalised", i.e., DCG is divided by the score of a perfect/ideal ranking, to obtain a score between 0. and 1.

Armchair mistaken for mug



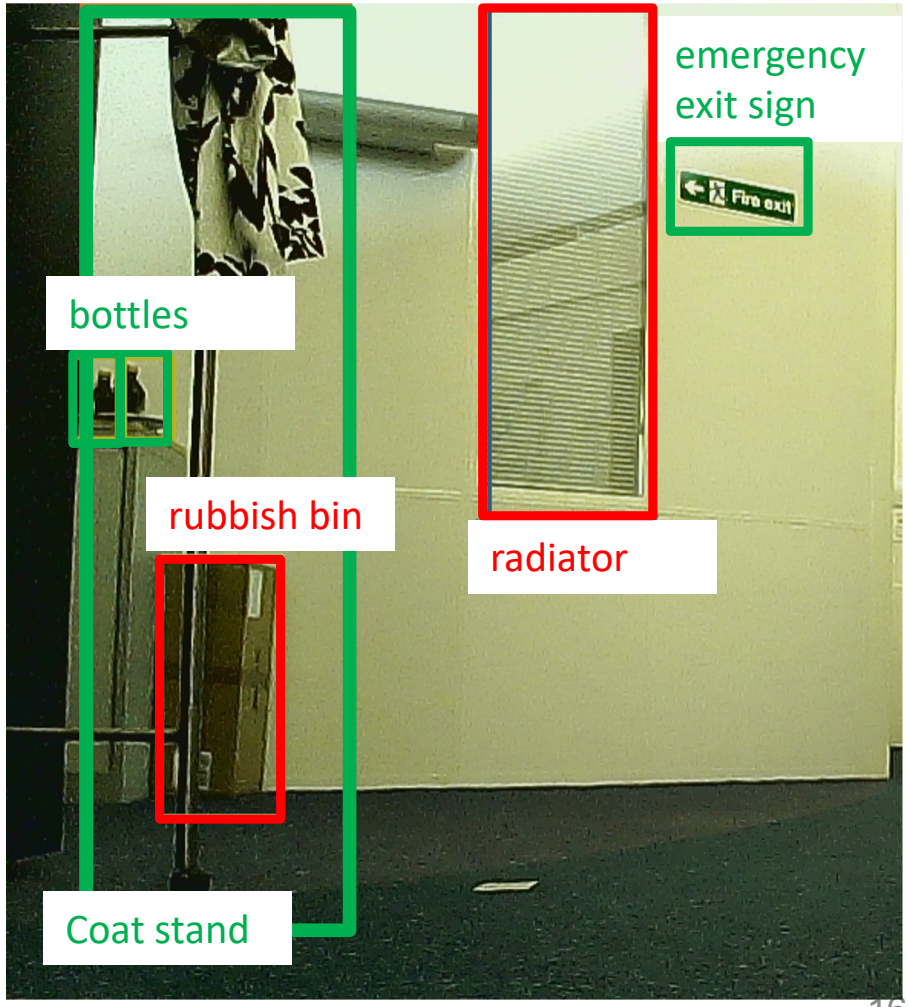
Deep Learning-based object recognition system lacks common sense knowledge about size of mugs

Window mistaken for radiator

However, different observations over time may correctly identify object as a window – in particular, when there is no reflection from the blinds.

Common sense knowledge tells us that a window is unlikely to be replaced by a radiator!

Need for **common sense knowledge about motion/temporal properties of objects** (i.e., a window does not move and does not change over time!)



Research Questions

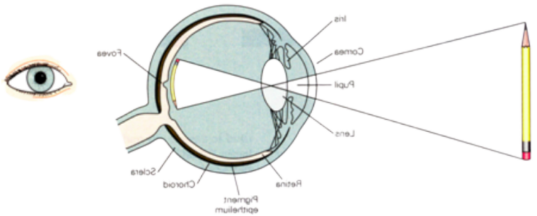
- Hypothesis: Equipping HanS with common sense knowledge ought to improve its object recognition performance
- Research Questions:
  - What types of common sense knowledge are required for visual intelligence?
  - How can we represent these different types of common sense knowledge?
  - How can we integrate these different types of common sense knowledge with a deep learning module for object recognition?

The Open University

KMi Knowledge Media Institute

Cognitive Foundations of Visual Intelligence

Chiatti, A., Motta, E., & Daga, E. (2020). Towards a Framework for Visual Intelligence in Service Robotics: Epistemic Requirements and Gap Analysis. KR 2020.



GENERIC 2D VIEWS

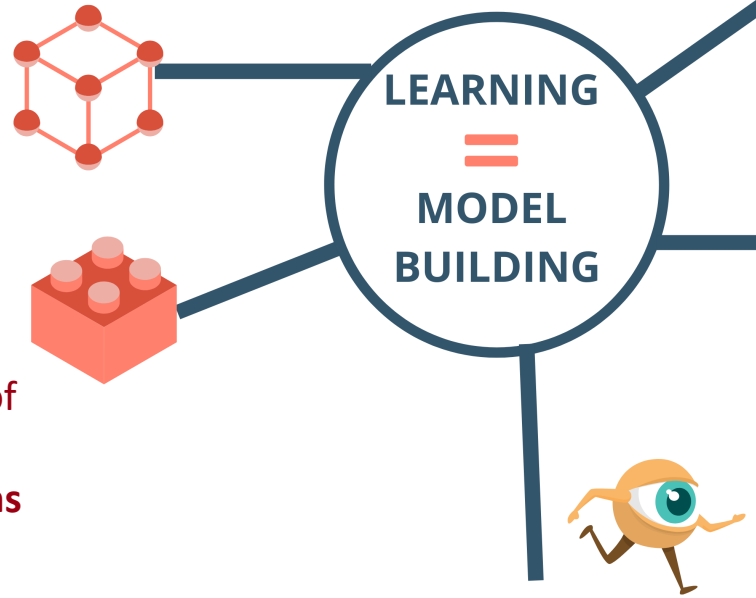
The images cast at the back of our eye are 2-dimensional. We construct the 3D mentally from **prototypical 2D shapes** (Rosch, 1999)

NAIVE PHYSICS

Infants can grasp basic physics principles (e.g., inertia), before 6 months

COMPOSITIONALITY

Humans are very good at processing objects in terms of their structural **subparts** and at identifying **spatial relations**



MOTION VISION

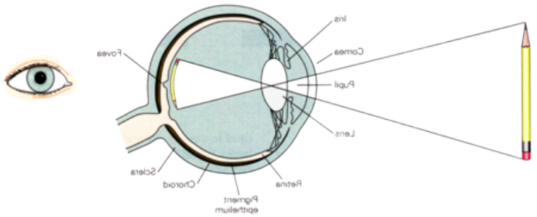
The human brain appears to maintain distinct representations for static and moving objects

FAST PERCEPTION

Our visual perception is extremely fast, and we can learn to recognize new objects from the very first exposure

Framework based on
- Lake et al., 2017
- Hoffman, 2000

Operationalizing the framework for visual intelligence

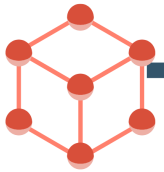


GENERIC 2D VIEWS

Use of synthetic 2D shapes provided with existing KBs – e.g., ShapeNet

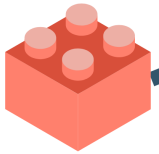
NAIVE PHYSICS

Physical properties of objects (e.g., size, natural orientation, etc.)



COMPOSITIONALITY

Spatial reasoning; part-whole relations; fine-grained segmentation;



LEARNING = MODEL BUILDING

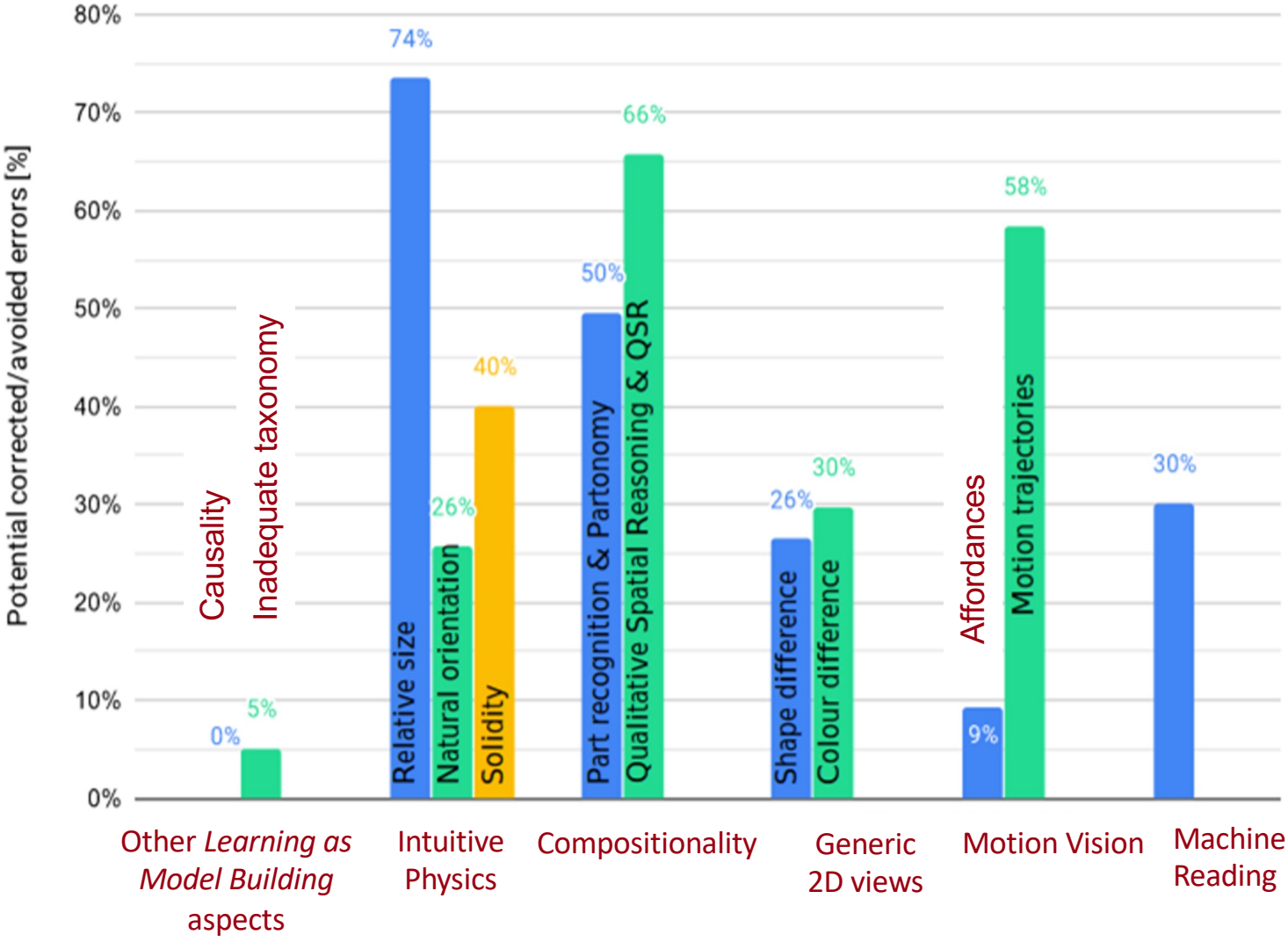
MOTION VISION

Object tracking and action recognition across temporally ordered frames

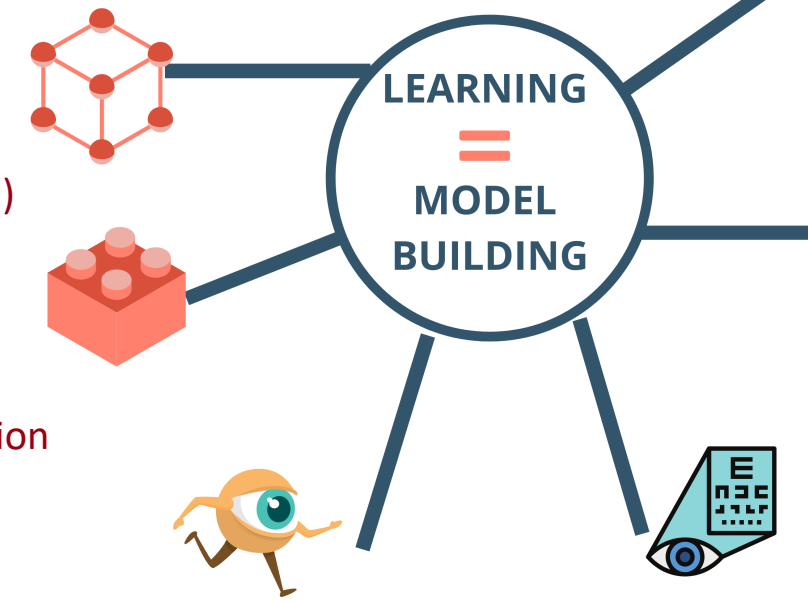
FAST PERCEPTION

Use of DL methods to provide very fast object recognition performance on known categories

Error analysis



Revised Framework for Visual Intelligence



NAIVE PHYSICS

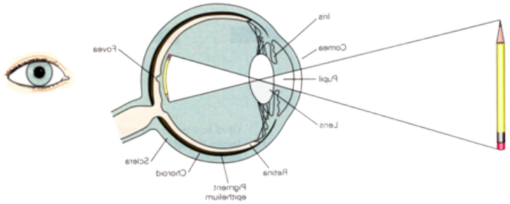
Physical properties of objects (e.g., size, natural orientation, etc.)

COMPOSITIONALITY

Spatial reasoning
part-whole relations
fine-grained segmentation

FAST PERCEPTION

Use of DL methods to provide very fast object recognition performance on known categories



GENERIC 2D VIEWS

Use of synthetic 2D shapes provided with existing KBs – e.g., ShapeNet

MOTION VISION

Object tracking and action recognition across temporally ordered frames

MACHINE READING √

Ability to read labelled objects and signs

Knowledge Media Institute

The Open University

Introducing a size reasoner

Representing size of objects in KMi: Approach

- Collect object dimensions from external sources
  - ShapeNet, Amazon
- Removing outliers and erroneous data
  - e.g., a 142 cm x 90 cm x 43 cm hat
- Synoptic Representation
  - From hundreds of individual chairs to a synoptic representation of chair sizes
  - 3d representation (h, w, d) -> 2d (area and depth)

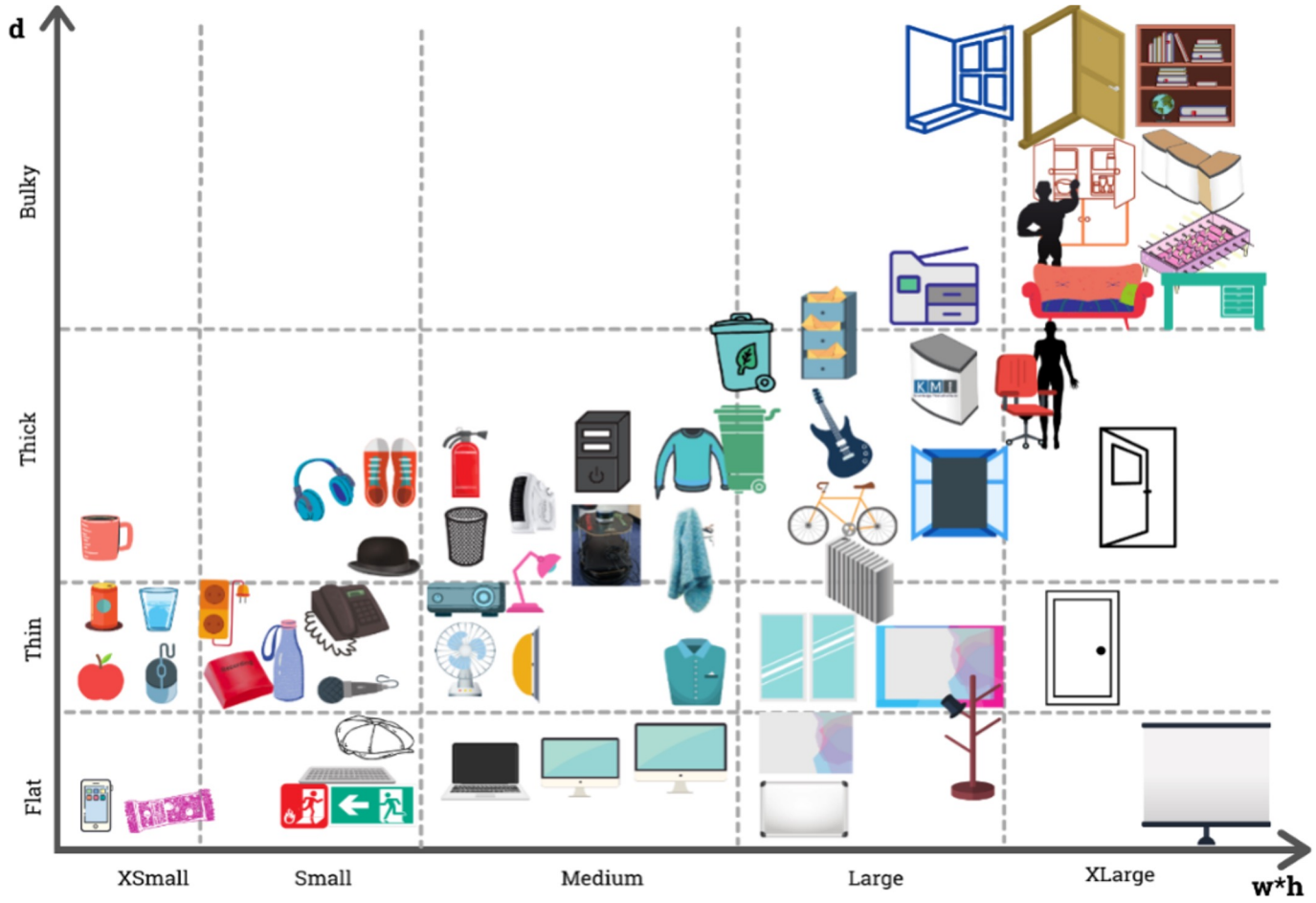
The Open University

KMi Knowledge Media Institute

Our proposed representation

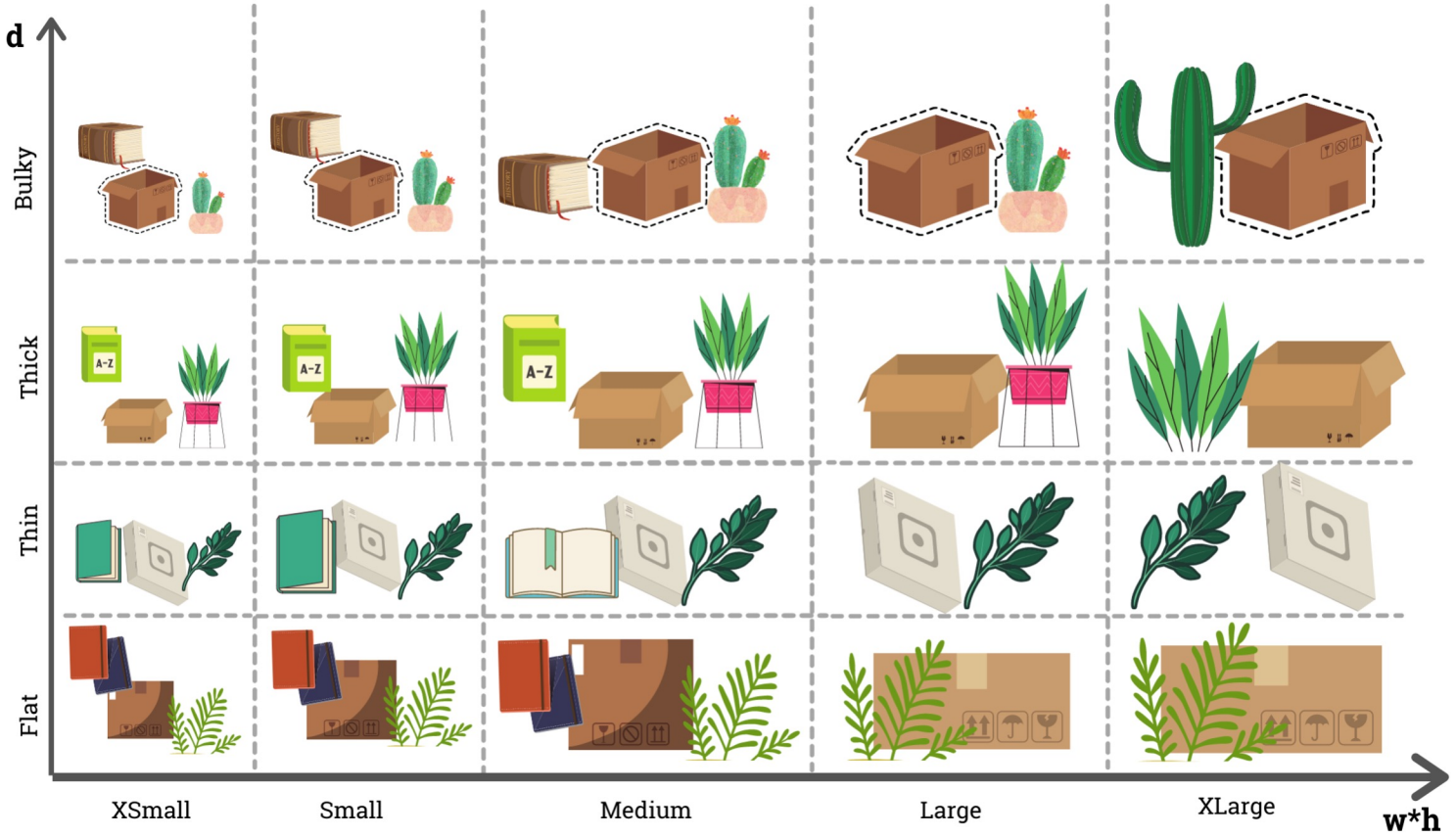
- Allocation of **60 objects** commonly found in KMi to each bin

- **Relative sorting** of objects within the same bin



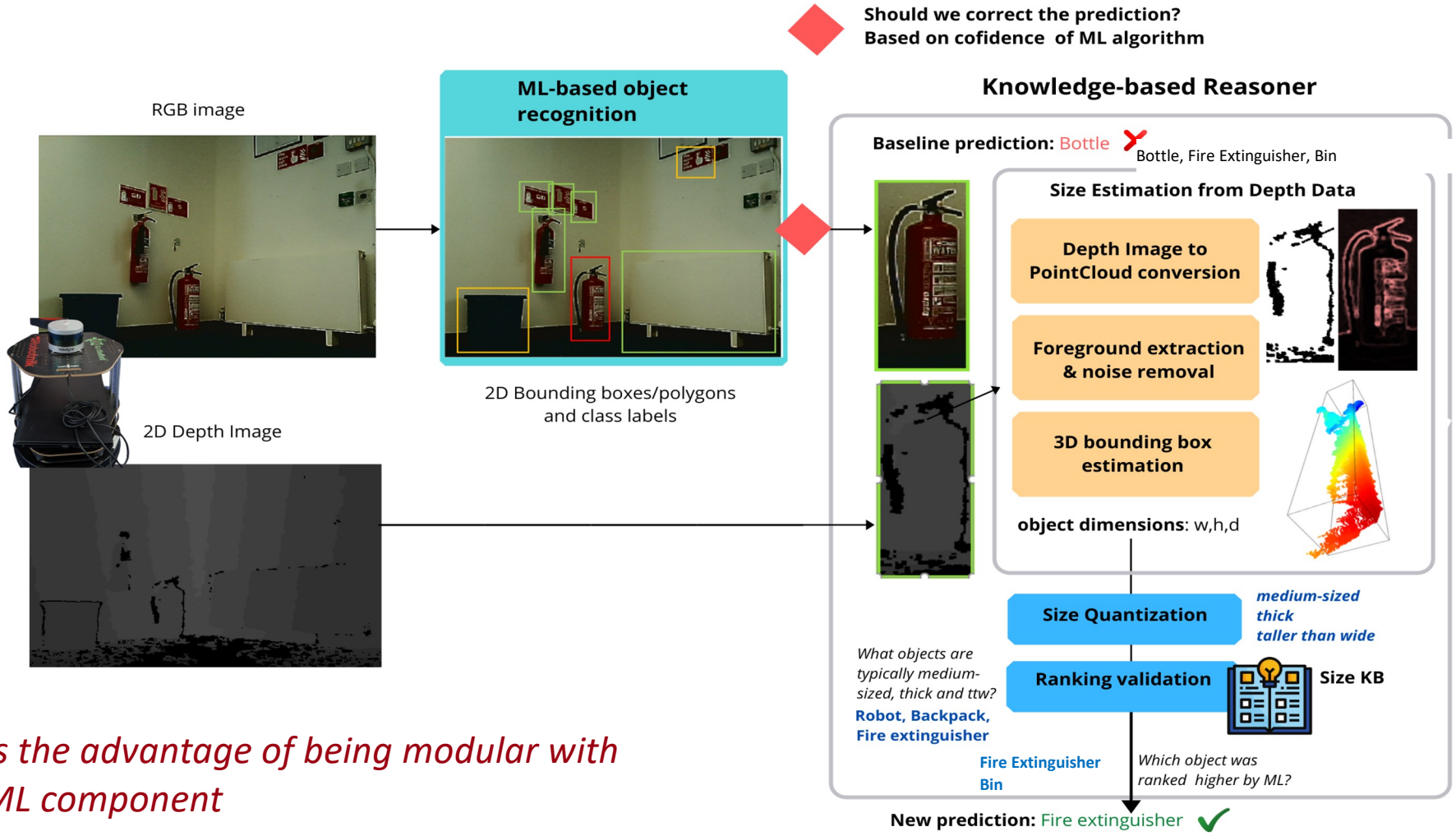
Our proposed representation

- Bin membership is **non-exclusive**
- Representation allows us to handle categories which are **extremely variable in size**



Advantages: performance and explainability

Integration of size reasoner with ML algorithm



Solution has the advantage of being modular with respect to ML component

Performance of best NN reasoners on KMi dataset

Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
N-net (Zeng et al., 2018)	.45	.34	.40	.31	.62	.45	.47	.33	.36	.63
K-net (Zeng et al., 2018)	**.48**	**.39**	**.40**	**.34**	**.68**	**.48**	**.50**	**.38**	**.41**	**.65**

nDCG@5 -> Normalised Discounted Cumulative Gain of top-5 ranking

Correct predictions appearing lower in the ranking are penalised

"Normalised", i.e., DCG is divided by the score of a perfect/ideal ranking, to obtain a score between 0. and 1.

Results from hybrid reasoner

***Realistic scenario**: ML predictions are selected based on automatically determined confidence threshold – nfold cross-validation*

Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
Hybrid (area)	.50	.40	**.40**	.36	.66	.50	.52	.41	.43	.68
Hybrid (area+flat/not-flat)	.50	.41	.39	.36	.66	.50	.52	.40	.43	.66
Hybrid (area+thickness)	**.51**	.45	.39	.39	.65	**.51**	.54	**.42**	**.44**	**.69**
Hybrid (area+flat +aspect ratio)	.49	.43	.39	.37	.69	.49	.53	.40	.42	.66
Hybrid (area+thickness +aspect ratio)	**.51**	**.47**	.39	**.40**	**.69**	**.51**	**.55**	**.42**	**.44**	.68

Results from hybrid reasoner

***Best-case scenario**:*
*the predictions to be corrected are known – i.e., use **ground truth** rather than confidence of ML reasoner*

Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
Hybrid (area)	.55	.47	.46	.43	.69	.55	.56	.38	.41	.65
Hybrid (area+flat)	.58	.55	.49	.47	.72	.58	.61	.42	.45	.71
Hybrid (area+thickness)	.60	.60	.51	.52	.72	.60	.63	.44	.47	.74
Hybrid (area+flat +aspect ratio)	.59	.60	.51	.50	**.76**	.59	.63	.43	.46	.73
Hybrid (area+thickness +aspect ratio)	**.62**	**.64**	**.52**	**.54**	**.76**	**.62**	**.66**	**.45**	**.48**	**.76**

Result on ARC set

Amazon Robotic Challenge (ARC): 562 test images, representing
41 **known** and 20 **novel object categories – size reasoner integrated with mixed n-net/k-net architecture – thickness not important on Amazon DB**

Method	Top-1 Accuracy			Top-5 results unweighted (class-based)		
	Known	Novel	Mixed	Mean P@5	Mean nDCG@5	Hit ratio
N-net (Zeng et al.,2018)	.57	**.82**	.65	.62	.63	.73
K-net (Zeng et al.,2018)	**1.**	.30	.78	.74	.75	.82
Hybrid (area)	.95	.72	**.88**	**.83**	**.84**	**.90**
Hybrid (area+flat)	.95	.72	**.88**	**.83**	**.84**	**.90**
Hybrid (area+thickness)	.82	.39	.69	.65	.66	.70

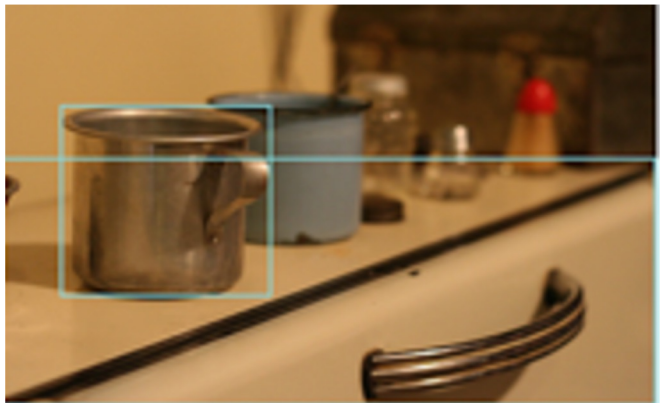
Knowledge Media Institute

The Open University

Introducing a Spatial Reasoner

Objectives of our work

- To link a formal QSR representation to the concrete operational frameworks used in the robotic community (Deeken et al., 2018)
- To account for the informal language people use to describe spatial relations (e.g., A is on B)
  - This is needed to use spatial relations in existing KBs - e.g., Visual Genome, ConceptNet, SpatialSense, and others.
  - It also provides a foundation for HRI when exchanging info about spatial relations ("bring me the book on the coffee table")

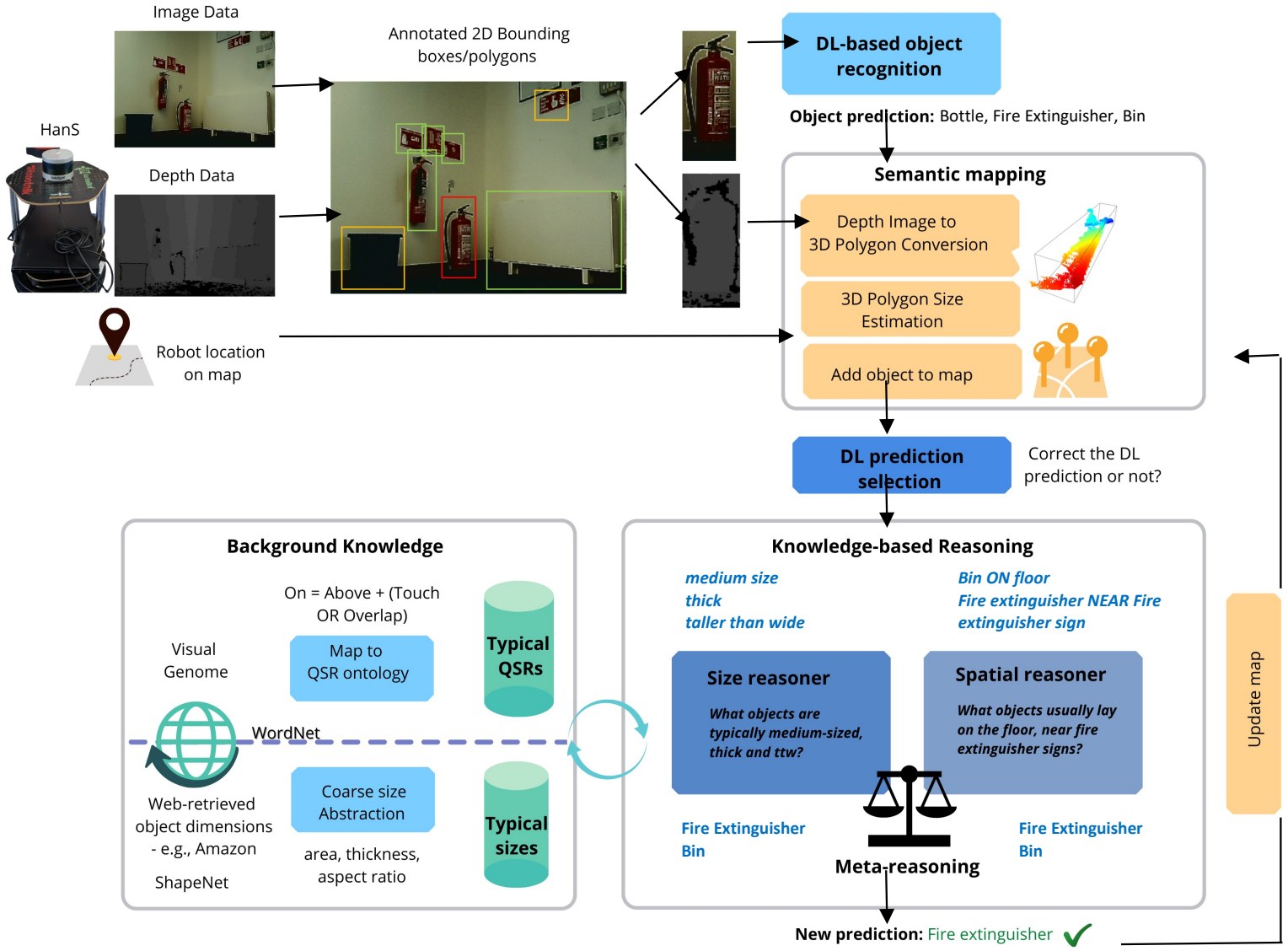


Examples from Visual Genome https://visualgenome.org/VGViz/explore

Formalising spatial relations

- Typical spatial relations involve two objects, where one is the reference (or landmark)
  - E.g., bike near house
- We specified a logical theory (defined in terms of 40 axioms), which formalizes all spatial relations needed to reason about the relative locations of objects. The model builds on the foundational notions of **geometric point** and **proper spatial region**, the latter defined as a connected set of geometric points.
- A key issue in the formalizations of QSRs is the **frame of reference**.
  - Relations can be expressed with respect to different frames of reference. E.g., the mouse is on the right of the keyboard → "right of" depends on the frame of reference

Modified architecture integrating **two** common sense reasoners



Experiments with size and spatial reasoners

The Open University

All experiments are on the same data collected in KMi earlier.

Experiment	**Do we know which ML predictions are wrong, i.e., need to be corrected through reasoning?**	**In the QSRs linked to the object to be classified, are the nearby objects represented by ground truth labels?**
A	✓	✓
B	✓	✗
C	✗	✓
D	✗	✗

Results from hybrid (NN+ Size reasoner) architecture

***Realistic scenario**: ML predictions are selected based on automatically determined confidence threshold – nfold cross-validation*

Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
Hybrid (area)	.50	.40	**.40**	.36	.66	.50	.52	.41	.43	.68
Hybrid (area+flat/not flat)	.50	.41	.39	.36	.66	.50	.52	.40	.43	.66
Hybrid (area+thickness)	**.51**	.45	.39	.39	.65	**.51**	.54	**.42**	**.44**	**.69**
Hybrid (area+flat/not flat +aspect ratio)	.49	.43	.39	.37	.69	.49	.53	.40	.42	.66
Hybrid (area+thickness +aspect ratio)	**.51**	**.47**	.39	**.40**	**.69**	**.51**	**.55**	**.42**	**.44**	.68

Results: Experiment D (realistic case)

Method	Top-1 Acc.	Top-1 unweighted (class-based)			Top-1 weighted (instance-based)			Top-5 results unweighted (class-based)		
		P	R	F1	P	R	F1	Mean P@5	Mean nDCG@5	Hit ratio
Hybrid size-only (area+thickness +aspect ratio)	.51	.47	.39	.40	.69	.51	.55	**.42**	**.44**	**.68**
Hybrid spatial-only	.48	.40	.41	.35	.70	.48	.51	.39	.41	.65
Hybrid size+ spatial (cascade)	**.50**	**.50**	.39	.40	**.71**	.50	.54	**.42**	**.44**	**.68**
Hybrid size+spatial (parallel)	**.54**	.48	**.40**	**.41**	**.71**	**.54**	**.58**	**.42**	**.44**	**.68**

Knowledge Media Institute

The Open University

HanS in action: Representing H&S rules

Fire Warden Monthly Inspection Form

Section 1 Personal Details					
Site		Building		Department	
Detail of area inspected	Floors:		Rooms:	Stairwells:	
Fire Warden (s)			Date		

Section 2 Fire Prevention				
		Yes	No	n/a
1	Does electrical equipment appear free from damage and defect?	☐	☐	☐
2	Are electrical extension cables kept to a minimum and sockets not overloaded by the use of multiplugs?	☐	☐	☐
3	Is all personal electrical equipment appropriate and PAT tested (*NB this does not include computers)?	☐	☐	☐
4	Are small electrical kitchen appliances e.g. toasters, kettles being used only in a kitchen/designated area?	☐	☐	☐
5	Have electric heaters been issued by Estates and are they being used in a safe manner, away from confined areas?	☐	☐	☐
6	Is waste and rubbish kept in a designated area and collected regularly?	☐	☐	☐
7	Have combustible materials been reduced to a minimum and stored in a suitable designated space away from sources of ignition?	☐	☐	☐

Section 3 Safety Documentation				
		Yes	No	n/a
1	Are there fire action notices displayed at all exit points on all floors?	☐	☐	☐
2	Are the names, locations and contact details of the fire wardens clearly displayed and up to date on each floor and generic notices displayed in public areas?	☐	☐	☐

Section 4 Fire Alarms & Call Points				
		Yes	No	n/a
1	Are the sounders clearly audible in work areas? (ask staff if they are clearly heard during weekly alarm testing)	☐	☐	☐
2	Are all fire alarm call points clearly signed and easily accessible?	☐	☐	☐

Section 5 Fire Escape Routes

H&S Rules

Rule nickname	Natural language description
waste_check	Is waste and rubbish kept in a designated area?
ignition_check	Have combustible materials been stored away from sources of ignition?
fire_call_check	Are all fire alarm call points clearly signed and easily accessible?
fire_escape_check1	Are all fire exit signs in place and unobstructed?
fire_escape_check2	Are fire escape routes kept clear?
fall_check	Is the condition of all flooring free from trip hazards?
door_check	Are fire doors kept closed, i.e., not wedged open?
extinguisher_check1	Are portable fire extinguishers clearly labelled?
extinguisher_check2	Are all portable fire extinguishers readily accessible and not restricted by stored items?
extinguisher_check3	Are portable fire extinguishers either securely wall mounted or on a supplied stand?

H&S Rules

WASTE_CHECK. *Is waste and rubbish kept in a designated area?* For a given spatial object o and area of interest a:

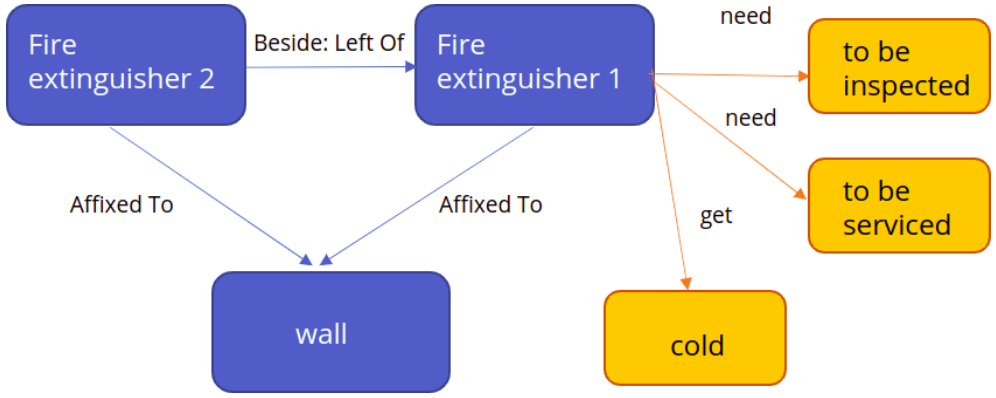
$$\exists o, a\ \mathit{HasCategory}(o, \mathit{rubbish\ bin}) \wedge \mathit{ComplCont}(o, a) \wedge \mathit{HasCategory}(a, \mathit{waste\ collection\ area}).$$

IGNITION_CHECK. *Have combustible materials been stored away from sources of ignition?* This statement implies that the rule is violated if a flammable object is in contact with an ignition source. Hence, given two spatial objects, o_1 and o_2:

$$\exists o_1, \neg\exists o_2\ \mathit{HasProperty}(o_1, \mathit{flammable}) \wedge \mathit{Touches}(o_1, o_2) \wedge \mathit{HasProperty}(o_2, \mathit{ignition}).$$

(a)

(b)



(c)

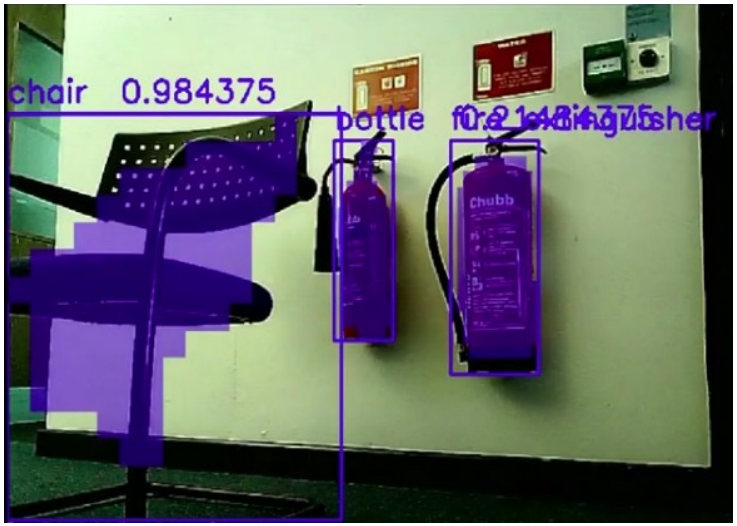
```
Are portable fire extinguishers either securely wall mounted or on a supplied stand?
Rule check passed for anchor 15 of class fire_extinguisher
Rule check passed for anchor 14 of class fire_extinguisher
Are all portable fire extinguishers readily accessible and not restricted by stored items?
Rule check passed for anchor 15 of class fire_extinguisher
Rule check passed for anchor 14 of class fire_extinguisher
Are portable fire extinguishers clearly labelled?
[Warning!] rule check not passed for anchor 15 of class fire_extinguisher
[Warning!] rule check not passed for anchor 14 of class fire_extinguisher
```

✓ Are portable fire extinguishers either securely wall mounted or on a supplied stand?

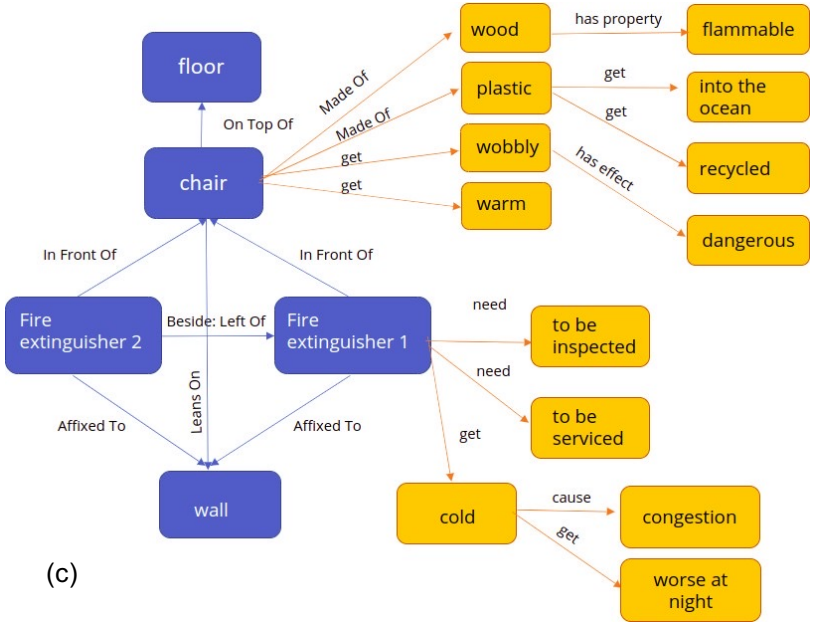
✓ Are portable fire extinguishers readily accessible and not restricted by stored items?

⚠ ✗ Are portable fire extinguishers clearly labelled?

(a)



(b)



(c)

```
Are portable fire extinguishers either securely wall mounted or on a supplied stand?
Rule check passed for anchor 59 of class fire_extinguisher
Rule check passed for anchor 58 of class fire_extinguisher
Are all portable fire extinguishers readily accessible and not restricted by stored items?
[Warning!] rule check not passed for anchor 59 of class fire_extinguisher
[Warning!] rule check not passed for anchor 58 of class fire_extinguisher
Are portable fire extinguishers clearly labelled?
[Warning!] rule check not passed for anchor 59 of class fire_extinguisher
[Warning!] rule check not passed for anchor 58 of class fire_extinguisher
Is the condition of all flooring free from trip hazards?
Rule check passed for anchor floor of class floor
```

Are portable fire extinguishers either securely wall mounted or on a supplied stand?

Are portable fire extinguishers readily accessible and not restricted by stored items?

Are portable fire extinguishers clearly labelled?

Is the condition of all flooring free from trip hazards?

https://www.youtube.com/watch?v=h8sZgLt_KQw

Conclusions

- Adding common sense reasoning components to our architecture significantly improves on ML baseline
  - In realistic scenarios: +5% using size reasoner, +8% using size and spatial reasoners
  - Promising results on a challenging, real-world dataset of robot-collected images
- Approach evidences value of combining AI paradigms, in particular deep learning with large-scale knowledge bases and common sense reasoning
- Separating DL from knowledge-based components has the advantage of making the architecture modular and explainable (compared to trying and embedding domain knowledge directly in ML component)

The Open University

KMi
Knowledge Media Institute

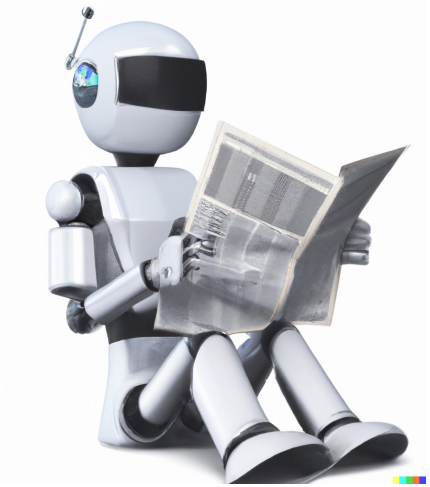
Knowledge Media Institute

The Open University

Case study #2: Using AI to Capture the Dynamics of Mainstream News

Research Context

- Digitalised news content is now available on a very large scale with commercial providers providing online access to thousands of news sources
  - e.g., see Aylien service at https:aylien.com
- This unprecedented availability of large-scale news content opens up a variety of opportunities for
  - large-scale news monitoring
    - e.g., to alert journalists to new relevant stories
  - large-scale analyses of the media landscape
    - e.g., to assess **fairness**, **balance**, **bias**, etc. in the context of a particular news provider or media landscape
  - new intelligent services for readers
    - e.g., sophisticated search and recommender solutions
  - etc.



A research agenda for computational news analytics

The Open University

- Ultimate Goal:
  - To develop novel computational solutions **to better model what the news talk about**, both in terms of covered **topics** and the **viewpoints** presented on each topic
- Research Gap
  - Gap between computational solutions on one side and the needs, methods and concepts used by media scientists and practitioners to analyse news content
  - Current analyses and solutions limited with respect to **both scale and range of concepts**
- Approach
  - **Phase 1**. To use Knowledge Engineering techniques to characterise the task of **fine-grained news classification** - **COMPLETED**
  - **Phase 2.** To develop computational methods that leverage the concepts analysed in Phase 1 and effectively support analyses of the news dynamics - **STARTED**

Paper submitted to the Semantic Web Journal

The Open University

The Epistemology of Fine-Grained News Classification

Enrico Motta[a,b], Enrico Daga[a], Aldo Gangemi[c,d], Maia Lunde Gjelsvik[b], Francesco Osborne[a] and Angelo Salatino[a]

[a]*Knowledge Media Institute, The Open University, Walton Hall, Milton Keynes, United Kingdom*
[b]*MediaFutures Centre, University of Bergen, Lars Hilles gate 30, Bergen, Norway*
[c]*Department of Philosophy and Communication Studies, University of Bologna, Via Azzo Gardino, Bologna, Italy*
[d]*STLab, Institute for Cognitive Sciences and Technologies, National Research Council, 40126 Bologna, Italy*

Abstract. The process of news digitalization over the past decades has released massive amounts of news content, revolutionizing consumer access to news and disrupting traditional business models. These radical changes have also introduced new opportunities for media content analysis, potentially opening up new scenarios for ambitious large-scale media analytics initiatives, which can go well beyond the relatively small-scale studies currently carried out by media scholars and practitioners. However, take-up of computational methods to support media content analysis activities has been rather modest, reflecting a degree of disconnect between the needs of scholars and practitioners for task-specific and usable software solutions and the state of the art in computational techniques for news media analysis. In this paper we perform an initial step towards bridging this gap, by looking in detail at the task of *fine-grained news classification*. In particular, we propose a typology of *news topics,* which is formally specified and realised into a family of reusable ontologies. The proposed model has been validated empirically, through an analysis of a multilingual news corpus, as well as formally, in terms of the functional and logical properties of the ontologies. Our analysis brings together the media and computer science literature, connecting the formal definitions provided in this paper to the concepts used by media scholars.

Keywords: news classification, ontologies, knowledge engineering, formal specifications, semantic technologies

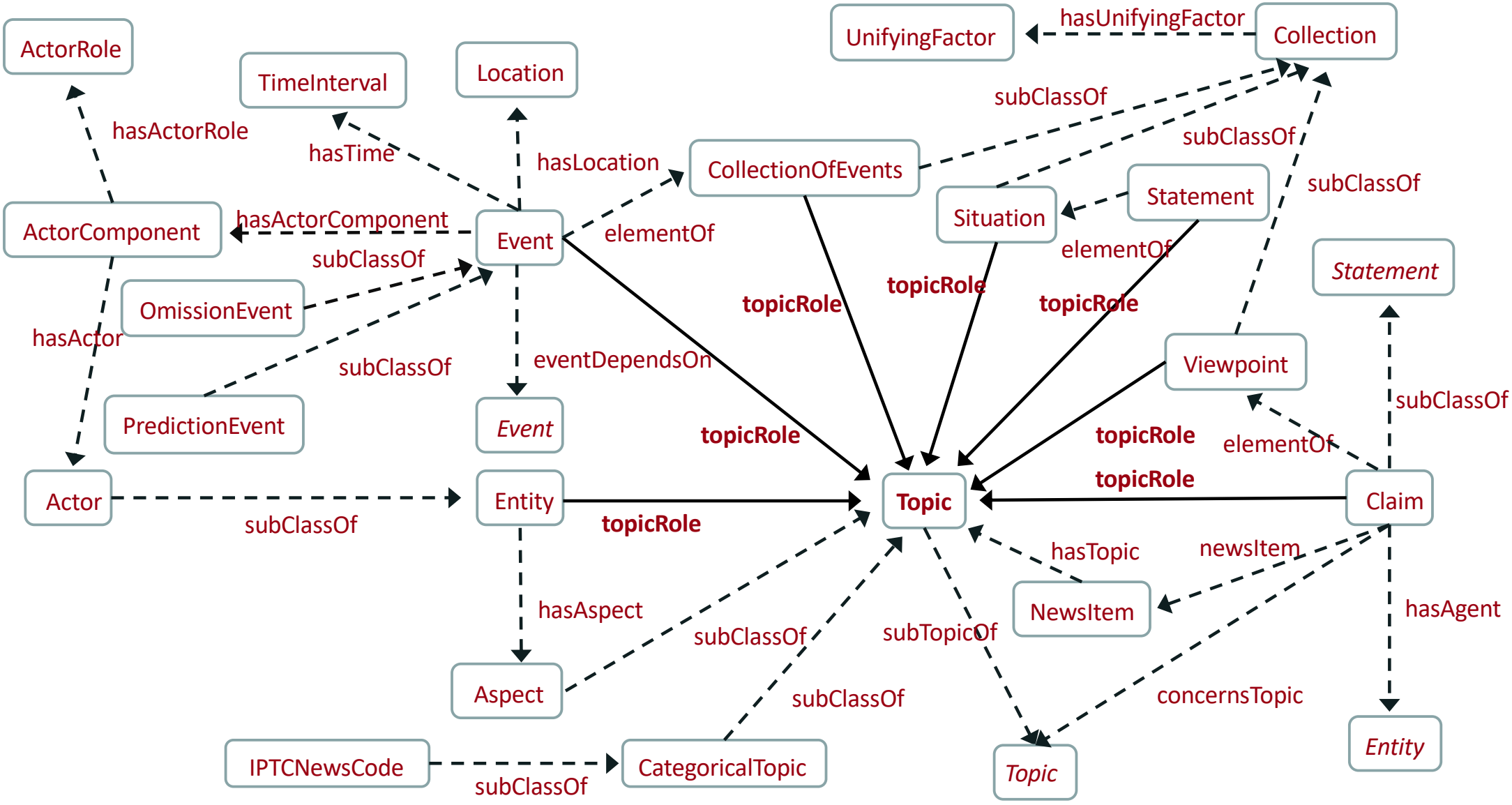
https://www.semantic-web-journal.net/content/epistemology-fine-grained-news-classification

Contributions presented in the SWJ paper

- **Framework** for news classification that includes five classes of relevant concepts
  - Entities, Events, Situations, Categorical Topics and the Commentary
  - Framework uses knowledge engineering techniques to characterise and clarify the concepts used in the media science literature
- An **initial empirical validation** of the framework carried out by classifying a corpus of news articles drawn from both English and Norwegian sources
- **Formalization** of the framework in first order logic
- **A set of OWL ontologies** that support the development of reusable knowledge bases informed by the framework
  - E.g., see http://data.open.ac.uk/ontology/newsclassification

Framework for characterizing topics in the news

Entities	
Individual Entity	Persons, organizations, fictional characters, etc..
Entity Aspect	A particular aspect of an entity (e.g., somebody's religious beliefs).
Relation between Entities	E.g., a relationship between a business person and a politician.
Events	
Individual Event	An individual event that is the focus a news story.
Collection of Events	Events that are grouped together (this is linked to the notion of **impact** in journalistic guidelines).
Negative Event	An event where an entity expresses agency by omitting to carry out an (expected) action.
Prediction	*A prediction (of an event or a situation).*
Dependency between Events	*Two main types of dependencies: super/sub-event (e.g., trial/verdict); preconditions or causality (a trial only takes place if a referral to trial is issued). The notion of dependency here is linked to the notion of context or background in journalistic guidelines.*
Situations	A state of affairs (e.g., no power in a city after an earthquake).
Viewpoints/Debate	Viewpoints define a macro concept that abstracts from a number of claims about an issue.
Categorical Issues	The broader categories that are useful to cluster together news focusing on different events in the same space (e.g., politics, poverty, immigration, etc…). These categories tend to be persistent and are covered by existing taxonomies, such as IPTC NewsCodes.


ActorRole
TimeInterval
Location
UnifyingFactor
hasUnifyingFactor
Collection
hasActorRole
hasTime
hasLocation
subClassOf
subClassOf
CollectionOfEvents
ActorComponent
hasActorComponent
Event
elementOf
Situation
Statement
subClassOf
subClassOf
elementOf
topicRole
topicRole
topicRole
Statement
OmissionEvent
hasActor
subClassOf
eventDependsOn
Viewpoint
PredictionEvent
Event
topicRole
topicRole
topicRole
elementOf
subClassOf
Actor
Entity
Topic
Claim
subClassOf
topicRole
topicRole
hasTopic
newsItem
NewsItem
hasAspect
subClassOf
subTopicOf
hasAgent
Aspect
subClassOf
concernsTopic
IPTCNewsCode
CategoricalTopic
Topic
Entity
subClassOf

Approach to formalising the model

- We model definitions as First Order Logic (FOL) statements, using a notation which mirrors standard representations for knowledge graphs, such as RDF.
- Hence, we limit ourselves to binary relations and we use *typeOf (?instance, class)* and *subclassOf (?class1 ?class2)* to represent taxonomies.
- We introduce an operator **T** to distinguish between an entity, say John Kennedy (former US president) and the topic "John Kennedy", which is discussed in the news. Hence, if we have a news item, *ni1234*, which is about John Kennedy, we would state:

  hasTopic (ni1234 T(JF_Kennedy))

  <JF_Kennedy is an individual in our KB denoting John Kennedy>

Reification

We also use reification to allow us to make statements about statements. Here we use the notation ***"id: statement"*** to indicate both that a statement is asserted in our knowledge base and also that ***id*** is the identifier reifying the statement. For example:

s1: hasTopic (ni1234 T(JF_Kennedy))

provides an abbreviated way to assert both the domain statement

hasTopic (ni1234 T(JF_Kennedy))

and also the following ones:

typeOf (s1 Statement)

hasSubject (s1 ni1234)

hasObject (s1 JF_Kennedy)

hasPredicate (s1 hasTopic)

Examples showing the need for reification

- Modelling relations between entities as topics

s2: hasBusinessConnection (politician1 businessperson1)

hasTopic (ni5534 T(s2))

- Modelling situations as collection of statements

subclassOf (Situation Collection)

type (?x Situation) ∧ elementOf (?s ?x) → type (?s Statement)

s4: quitsJob (executive1 company1)

s5: quitsJob (executive2 company1)

type (situation1 Situation)

elementOf (s4 situation1)

elementOf (s5 situation1)

Viewpoints in Media Scholarship

- Not all opinions necessarily define different viewpoints
- Viewpoints must “open up different perspectives” and “construct different meaning” (Baden and Springer, 2017),
- Example: analysis by Masini et al. identifies the following viewpoints on immigration, abstracting from various claims:
  - **Negative**: Immigrants carry diseases, commit crimes, etc.
  - **Administrative burden**: e.g., concerns about the management of the arrivals, food supply, etc.
  - **Victimisation**: e.g., immigrants are victims of unjust government policies, traffickers, etc.
  - **Positive**: e.g., immigration empowers work force, enhances “positive multiculturalism”, immigrants work hard, etc.).

Conceptualizing viewpoint diversity in news discourse

Christian Baden
The Hebrew University of Jerusalem, Israel

Nina Springer
Ludwig Maximilians University Munich, Germany

Measuring and explaining the diversity of voices and viewpoints in the news

Masini, A, Van Aelst, P, Mancini, P, Zerback, T, Teineman, C, Mazzoni, M, Damiani, M and Coen, S

http://dx.doi.org/10.1080/1461670X.2017.1343650

Example: A claim by the (former) home secretary

Priti Patel blasts government's 'secretive' five-year plan to house asylum seekers on RAF base

Suella Braverman used an 'emergency' planning bypass for a 12-month development at RAF Wethersfield, but leaked memo says it will be used for five years

Lizzie Dearden Home Affairs Editor • Monday 14 August 2023 20:03 BST • 169 Comments

Claim: <Utterance, actor, news item, news source, date, topic>

Claims

subclassOf (Claim Statement)

typeOf (?c Claim) → ∃ ?a hasAgent (?c ?a)

typeOf (?c Claim) → ∃ ?t concernsTopic (?c ?t)

typeOf (?c Claim) → ∃ ?n claimInNewsItem (?c ?n)

type (?c Claim) ∧ hasJustification (?C ?j) → type (?j Justification)

Viewpoints

subclassOf (Viewpoint Collection)

typeOf (?v Viewpoint) $\wedge$ *elementOf (?s ?v)* $\rightarrow$ *typeOf (?s Claim)*

typeOf (?x Viewpoint) $\rightarrow \exists$ *?uf hasUnifyingFactor (?x ?uf)*

hasClaim (?ni ?c) $\wedge$ *elementOf (?c ?v)* $\wedge$ *type (?v Viewpoint)* $\rightarrow$ *hasViewpoint (?ni ?v)*

hasTopic (?ni T(?c)) $\wedge$ *type (?c Claim)* $\wedge$ *elementOf (?c ?v)* $\wedge$ *type (?v Viewpoint)*
$\rightarrow$ *hasTopic (?ni T(?v))*

A viewpoint comprises the set of all claims that satisfy the criterion associated with the viewpoint in question

Experiments on claim and viewpoint detection

- Dataset extracted from the Aylien service (now Quantexa)
  - Topic: immigration; Dates: 01/06 to 31/08; 11 news sources; 603 articles
- Extracted 4123 statements made by 1473 actors, leading to about ~4000 claims
- Claims are extracted from both direct or indirect quotes
  - However, we only consider claims associated with explicitly named actors in a news item
- We use primarily GPT-4 (but also carried out some tests with Zephyr, Gemini and Llambda)
- If we focus on the UK immigration debate, we have 778 statements, which are aggregated around 591 distinct claims

Capturing the viewpoint dynamics in the news: Approach

The Open University

1. Generate a news corpus about a particular topic
2. Automatically extract **claims** from news corpus
3. Use an LLM to suggest **dimensions** that can be used to classify the perspectives associated with the extracted claims
   e.g., "immigrants as a threat" or "economic benefits of immigration"
4. Use a human expert to finalise set of viewpoint dimensions
5. Situate each individual claim in a n-dimensional viewpoint space
6. *Use geometric distance to identify clusters of claims that may form a coherent viewpoint (TBD)*

Viewpoint dimensions extracted from analysis of statements

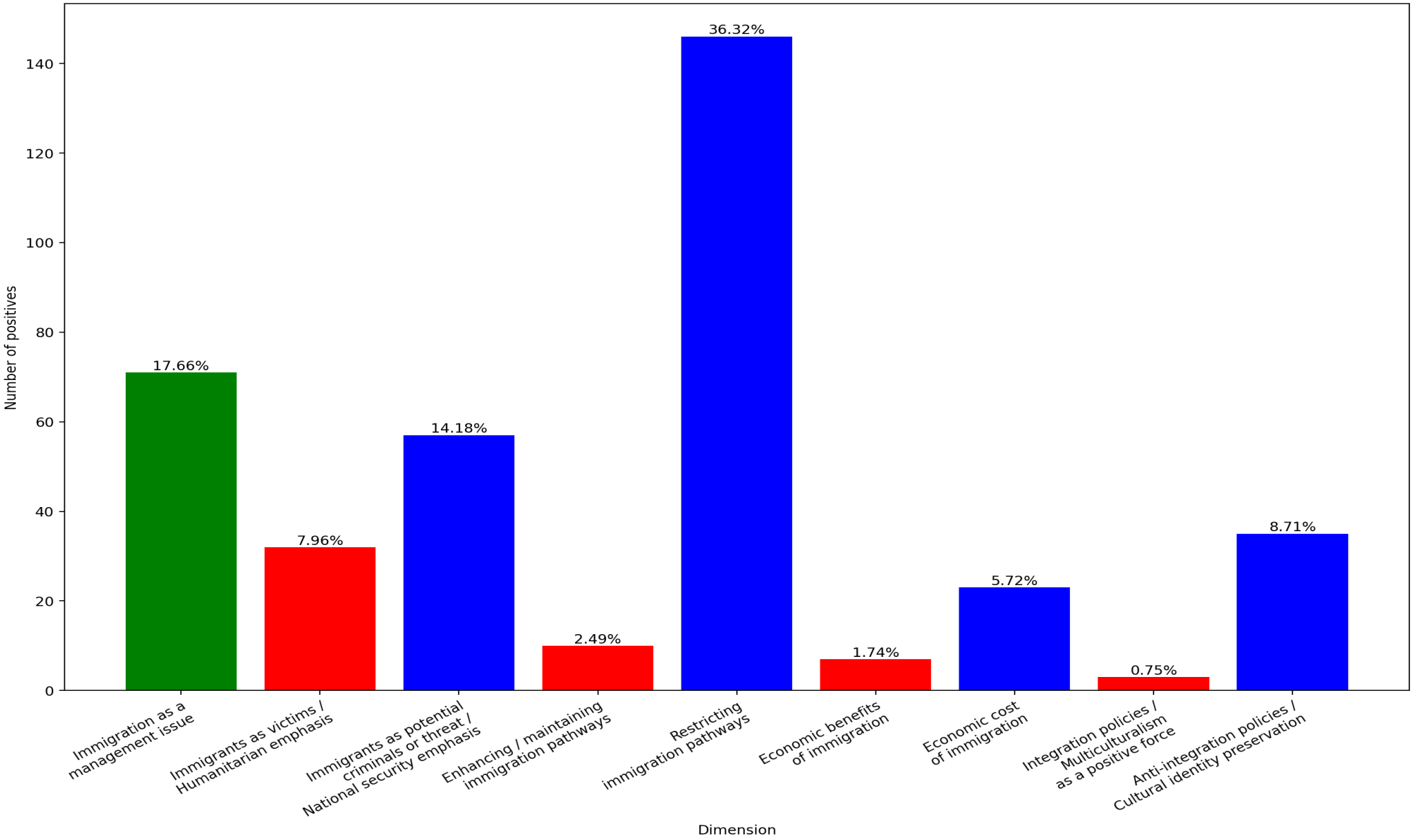
Immigration as a Management Issue

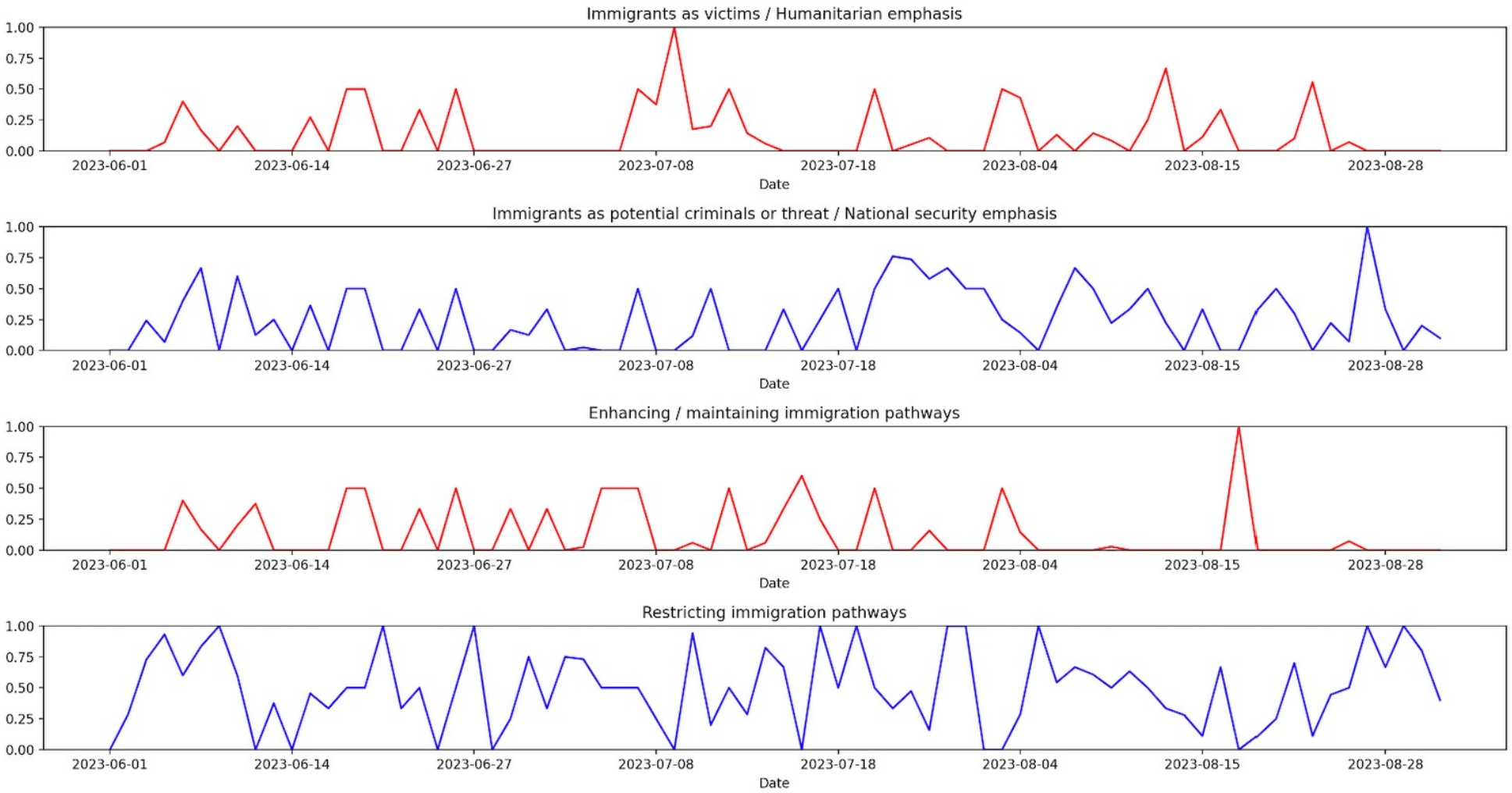
Immigrants as victims / Humanitarian Emphasis vs
Immigrants as potential criminals or otherwise a threat / National Security Emphasis

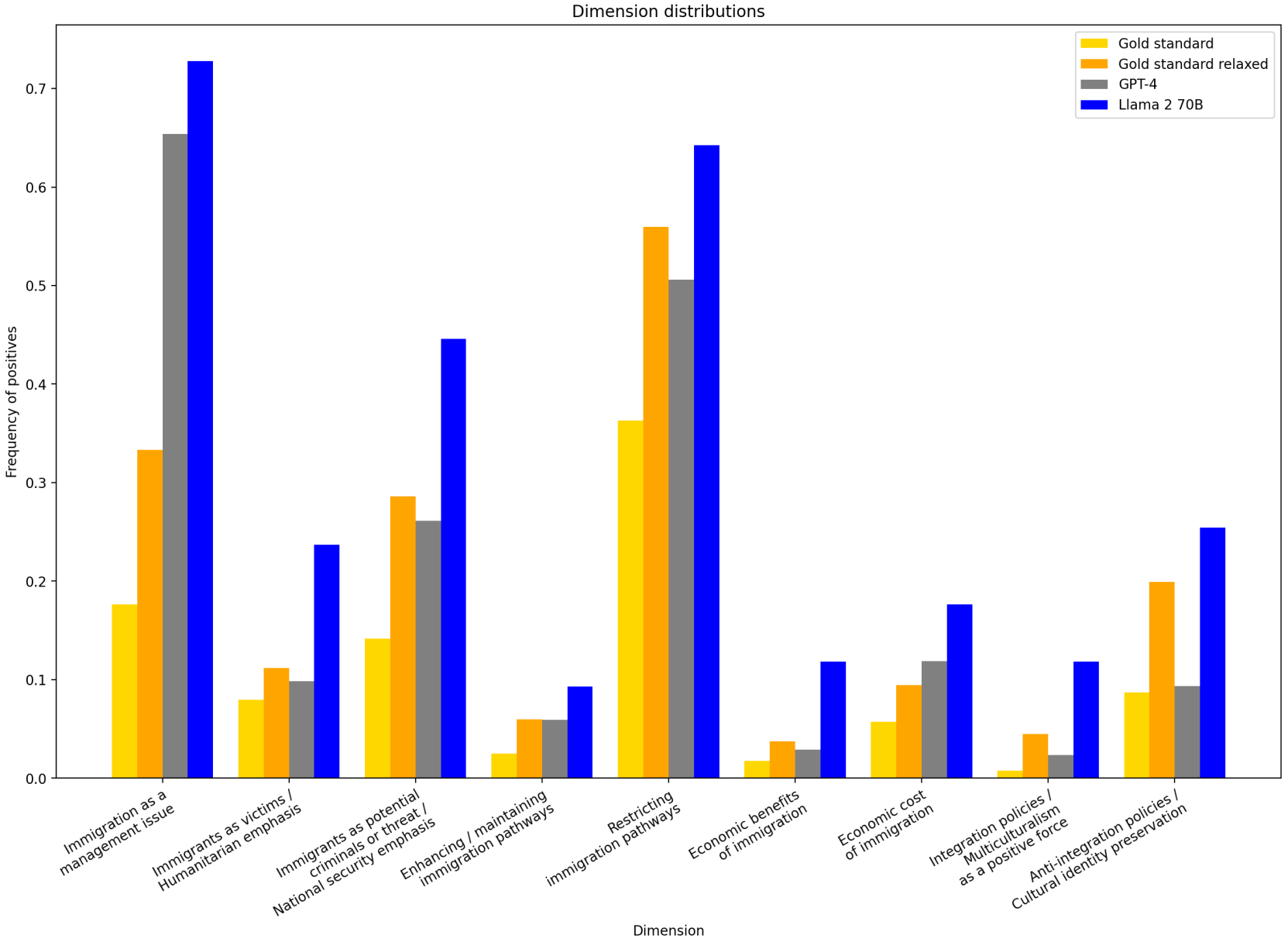
Enhancing / Maintaining Immigration Pathways vs
Restricting Immigration Pathways

Economic benefits of Immigration vs
Economic cost of Immigration

Integration Policies / Multiculturalism as a positive force vs
Anti Integration Policies / Cultural Identity Preservation


Number of positives
0
20
40
60
80
100
120
140
17.66%
7.96%
14.18%
2.49%
36.32%
1.74%
5.72%
0.75%
8.71%
Immigration as a management issue
Immigrants as victims / Humanitarian emphasis
Immigrants as potential criminals or threat / National security emphasis
Enhancing / maintaining immigration pathways
Restricting immigration pathways
Economic benefits of immigration
Economic cost of immigration
Integration policies / Multiculturalism as a positive force
Anti-integration policies / Cultural identity preservation
Dimension


Immigrants as victims / Humanitarian emphasis
1.00
0.75
0.50
0.25
0.00
2023-06-01
2023-06-14
2023-06-27
2023-07-08
2023-07-18
2023-08-04
2023-08-15
2023-08-28
Date
Immigrants as potential criminals or threat / National security emphasis
1.00
0.75
0.50
0.25
0.00
2023-06-01
2023-06-14
2023-06-27
2023-07-08
2023-07-18
2023-08-04
2023-08-15
2023-08-28
Date
Enhancing / maintaining immigration pathways
1.00
0.75
0.50
0.25
0.00
2023-06-01
2023-06-14
2023-06-27
2023-07-08
2023-07-18
2023-08-04
2023-08-15
2023-08-28
Date
Restricting immigration pathways
1.00
0.75
0.50
0.25
0.00
2023-06-01
2023-06-14
2023-06-27
2023-07-08
2023-07-18
2023-08-04
2023-08-15
2023-08-28
Date


Dimension distributions
Gold standard
Gold standard relaxed
GPT-4
Llama 2 70B
Frequency of positives
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Immigration as a management issue
Immigrants as victims / Humanitarian emphasis
Immigrants as potential criminals or threat / National security emphasis
Enhancing / maintaining immigration pathways
Restricting immigration pathways
Economic benefits of immigration
Economic cost of immigration
Integration policies / Multiculturalism as a positive force
Anti-integration policies / Cultural identity preservation
Dimension

Discussion

- Approach is promising, in particular because it goes into a direction of a multi-dimensional analysis of the perspectives expressed in the media about a topic, without introducing (human) bias
- Preliminary evaluation shows that agreement among human annotators is comparable to agreement between humans and LLM
- However, there are still issues related to producing a robust gold standard
- Current work focusing on improving quality of gold standard and fine-tuning open LLMs to improve performance


KMi
Knowledge Media Institute

Family of News Classification Ontologies

The following four OWL ontologies have been completed and are publicly available

- NCO: http://data.open.ac.uk/ontology/newsclassification#
  - Realises the formal framework in an OWL ontology
  - Imports SKOS and the W3C Time Ontology
  - Use of *punning* to model reification
  - Use of *property chains* to represent the various axioms in the model
- NCO_ex: http://data.open.ac.uk/ontology/ncoexamples#
  - Provides concrete examples of how to use NCO to characterise topics in news items
- NCO-IPTC: http://data.open.ac.uk/ontology/nco-iptc#
  - Imports the IPTC taxonomy of news codes into NCO
- News2D0: http://www.ontologydesignpatterns.org/ont/news/news2d0.owl
  - Provides a full alignment of NCO with the Dolce D0 ontology