KU LEUVEN

Human Language Understanding: General Picture and Approaches

Marie-Francine Moens

Joint work with Steven Bethard, Guillem Collell, Quynh Do and Oswaldo Ludwig

Department of Computer Science, https://liir.cs.kuleuven.be

KU Leuven, Belgium

What is human language understanding?

Bill MacCartney and Christopher Potts
Course on Natural Language Understanding, Stanford University, 2016

What is understanding?

To understand a statement is to:

- determine its truth (with justification)
- calculate its entailments
- take appropriate action in light of it
- translate it into another language
- …

Overview

1. When and how did it all start?
2. What is possible now?
3. What is still needed?
4. What are important properties that models should have?
5. Imagery to empower automated language understanding
6. Conclusions

1. When and how did it all start?

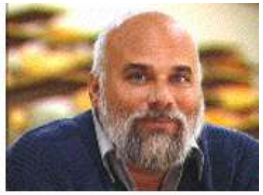
ELIZA

- 1964-1966: Joseph Weizenbaum (MIT)
- Limited interpretation of a dialogue (with a psycho-analyst)
- Use of pre-coded patterns of language

  - http://jerz.setonhill.edu/if/canon/eliza.htm
  - http://www.inspiratron.org/Eliza/

Conceptual dependency theory

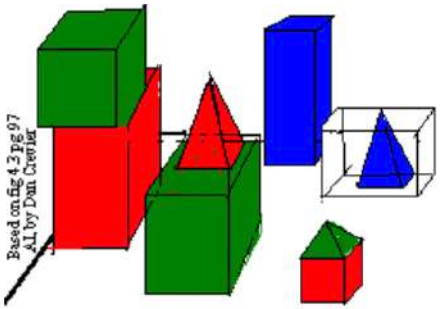
- Late 1960-1977: Roger Schank (Yale University):
- Models that:
  - Use pre-coded patterns of language
  - Recognize a number of primitives and predicates in language
  - Sentence analysis based on recognition of action and its anticipated arguments (*picture producers*)
  - Discourse analysis based on script of anticipated actions



SHRDLU

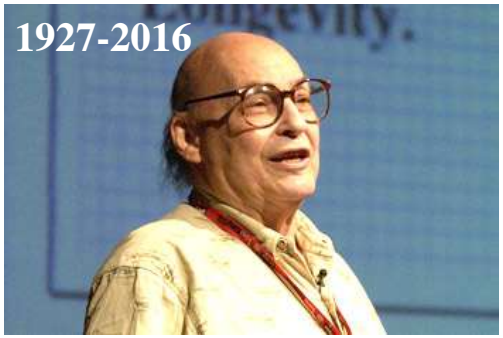
- 1972: Terry Winograd: SHRDLU (MIT)
  - Scripts, plans, goals, and semantic roles for processing natural language commands with pre-coded patterns
  - One of the first question-answering systems

Person: Pick up a big red block.
Computer: OK.
Person: Grasp the pyramid.
Computer: I don't understand which pyramid you mean.
Person (changing their mind): Find a block which is taller than the one you are holding and put it into the box.
Computer: By "it", I assume you mean the block which is taller than the one I am holding.
Computer: OK.
Person: What does the box contain?
Computer: The blue pyramid and the blue block.
Person: What is the pyramid supported by?
Computer: The box.
Person: How many blocks are not in the box?
Computer: Four of them.
Person: Is at least one of them narrower than the one which I told you to pick up?
Computer: Yes, the red cube.



Frames

- 1975: Marvin Minsky (MIT): *father of artificial intelligence*:
  - Pre-coded patterns to analyze text are represented as frame-based knowledge representations
  - Possible ordered in scripts



2. What is possible now?

Human language understanding

- Recognizing actions and states and the actors that play a role in the events, which includes semantic role labeling, i.e., recognizing "who" does "what", "where", "when" and "how"
- Coreferent resolution, i.e., identifying coreferring expressions in a discourse (e.g., that refer to the same entity)
- Recognizing temporal and spatial relations between actions and entities, and identification of other relationships
- Detecting modality, i.e., the factual status of a statement, which can involve negation, possibility and obligational entailments of statements

Human language understanding

- Often entails **making inferences** :
  - With information found in the discourse
  - With background knowledge that speaker/writer and audience possess: world, commonsense and domain knowledge
  - With other contextual knowledge

Today's language understanding systems

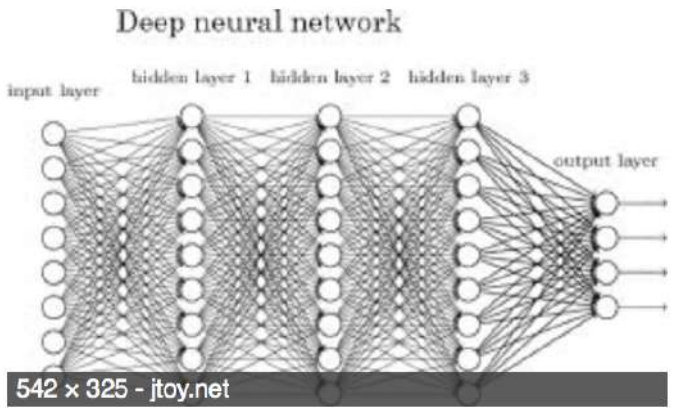
- **Linguistic pipelines** composed of preprocessing, morpho-syntactic analysis, semantic and discourse processing (semantic roles, negation detection, coreference resolution, temporal and spatial information extraction, etc.) and final mapping to formal representation
  - Individual systems are trained on annotated text corpora
  - Output of one system serves as input (features) for the next system in the pipeline
- Disadvantages:
  - Errors made earlier in the pipeline propagate
  - Dependent on developed language resources (annotated corpora)

Today's language understanding systems

- **End-to-end systems** implementing deep learning (often realized by a deep neural network):
  - Input = text or some distributional semantic representation of text
  - Output = formal representation (e.g., into a sequence of labels)

  - Deep layers model the linguistic patterns

- Disadvantage:
  - Many annotated training data needed, mitigated in multitask settings where a task can be trained with its own data



Today's language understanding systems

- **Hybrid systems**: e.g.,
  - Leveraging (uncertain) linguistic annotations/features as well as training from raw input
    - Input = text and linguistic annotations
    - Output = formal representation (e.g., into a sequence of labels)
  - Inference to compute best fusion of results of individual systems and mapping to outputs
- Advantage:
  - Less training data needed
  - Some recovery from errors

Today's language understanding systems

- Inferencing in natural language understanding is **traditionally realized by logical reasoning with symbolic representations** that are **logical** in nature [Bunt et al. *Computing Meaning* 2014]
- Recently **statistical inference** has emerged with **representations** that are **Bayesian** or **algebraic** in nature, and that can be composed (e.g., by simple additions of vectors): but very little is known …

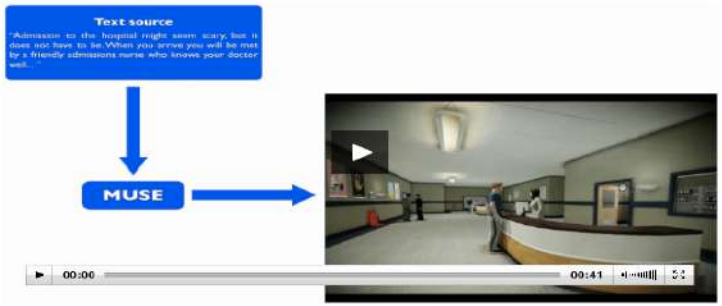
Today's understanding systems

- **Output of a language understanding system** =
  - Other discrete symbolic representation, often referred to as semantic parsing:
    - Semantic labels
    - Logical expressions
    - Task specific commands or instructions in a programming language

    [Liang & Potts *Ann. Rev. Ling*. 2015]
  - Continuous representation:
    - Numerical representations such as vectors and matrices that capture the meaning of an utterance or of a larger discourse unit
  - **Translation to other natural language or to other modality**

An example of language understanding seen as machine translation

- = bringing text to life = ultimate test of machine understanding of language
  - Render children's stories (use case in this presentation) and patient education guidelines as 3D-virtual worlds: http://www.muse-project.eu/



MUSE project

- MUSE: **Machine Understanding** for interactive StorytElling
- Algorithms for translating text into virtual worlds, 9/2012-11/2015, EU FP7-296703 (FET-open call)

Teesside University

Institut "Jožef Stefan"

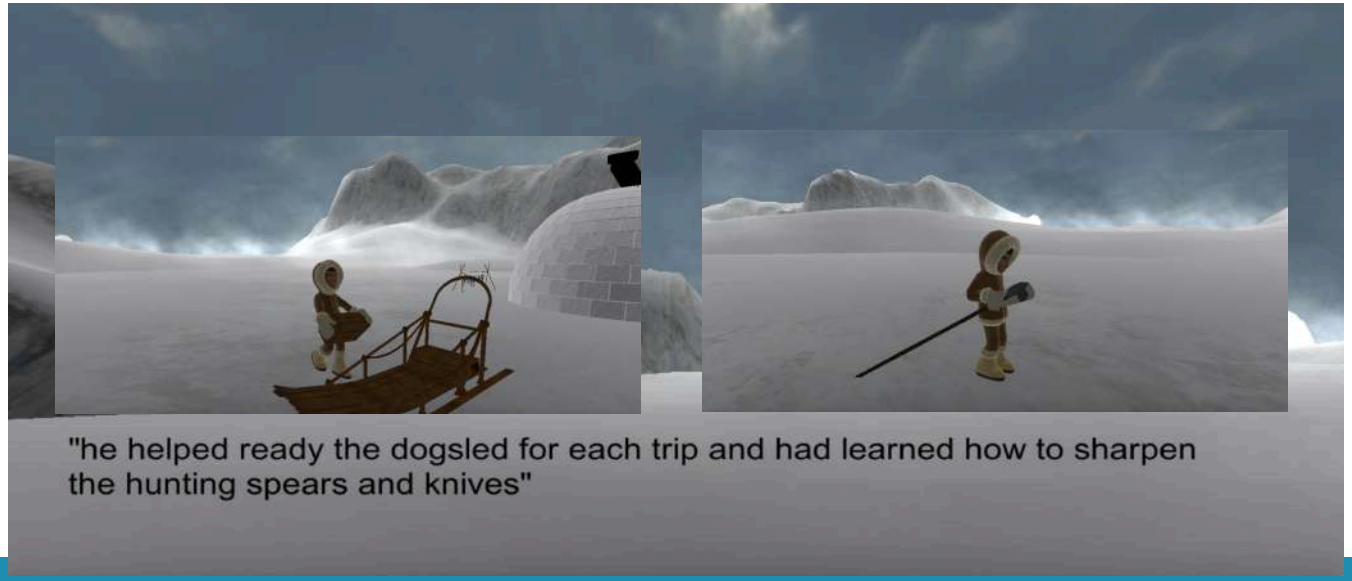
Universiteit Leiden

HAS HAUTE AUTORITÉ DE SANTÉ

KU LEUVEN

MUSE = test case of language understanding

- Bringing text to life
  - Render children's stories as 3D-virtual worlds: http://www.muse-project.eu/



http://roshi.cs.kuleuven.be/muse_demon/#/children-story

The MUSE approach

- Translate the sentences of the children's story to a knowledge representation that steers the graphical engine
  - **Deep learning ?** but problem of lack of training data to build end-to-end system
  - Deep translation was restricted to finding related/equivalent word patterns, but still reliance on common natural language processing subtasks
  - => Hybrid of a linguistic pipeline and some deep learning

EXAMPLE OF THE MAPPING OUTPUT

"He practiced using a spear and even knew how to cut up animals"					
semantic frame #1					
action	**char−subj**	**char−obj**	**obj/item**	**tool**	**direction**
to practice	tuk	none	none	spear	none
semantic frame #4					
action	**char−subj**	**char−obj**	**obj/item**	**tool**	**direction**
to cut	tuk	none	animals	knife	none

Symbolic representation in first-order logic

Unity scripts

```
using UnityEngine;
using System.Collections;

public class ExampleBehaviourScript : MonoBehaviour
{
    void Update()
    {
        if (Input.GetKeyDown(KeyCode.R))
        {
            GetComponent<Renderer> ().material.color = Color.red;
        }
        if (Input.GetKeyDown(KeyCode.G))
        {
            GetComponent<Renderer>().material.color = Color.green;
        }
        if (Input.GetKeyDown(KeyCode.B))
        {
            GetComponent<Renderer>().material.color = Color.blue;
        }
    }
}
```

Rendering

The MUSE approach

- **Lack of annotated training examples** (we only used two unrelated annotated stories)
- => goal is to combine all evidences to realize the mapping to the knowledge representation: use of a Bayesian framework for uncertainty reasoning
- **Bayesian framework** allows to add linguistic annotations, raw text features + any available world, commonsense and domain knowledge

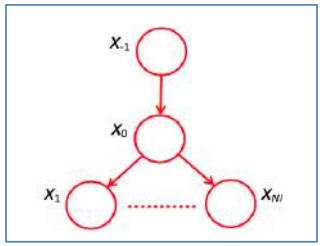
The MUSE approach

- At the **sentence level** recognition of:
  - Actions/events and their semantic roles (actor, patient, instrument, …)
- At the **discourse level** recognition of:
  - Coreferent noun phrases
  - Temporal relations between actions
  - Spatial relations between objects

Mapping to knowledge representation that steers the virtual world =

semantic parsing of the narrative

Inference with Bayesian network



[Do Thi et al. *EBLP* 2016]

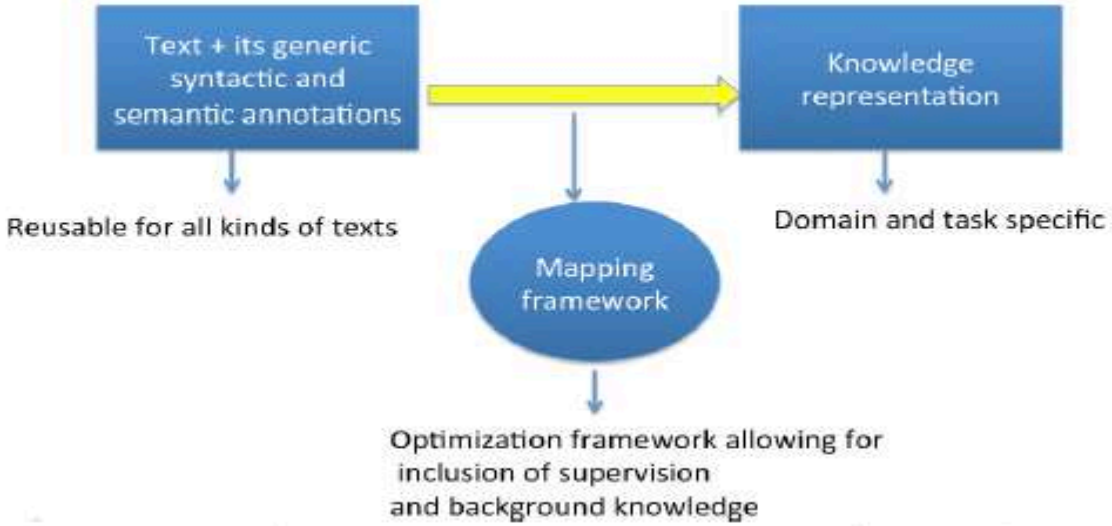


Figure 1: The general architecture for translating text in a knowledge representation.

[MUSE Scientific Report 2015]

Mapping to a knowledge representation

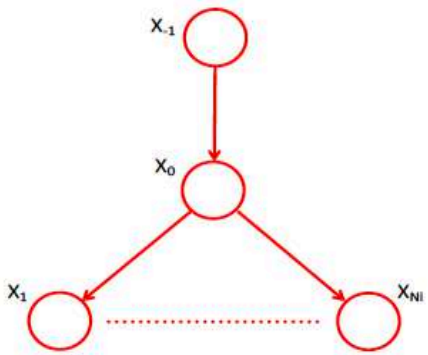


Fig. 2. Graphical model of the adopted statistical framework, in which X_{-1} represents the action of the previous semantic frame, X_0 the action of the current frame, and $X_1 \ldots X_{N_i}$ their arguments.

$$P\left(X_q = x_{(q,i)} \mid Pa_q, f; \theta_q\right) = \frac{e^{\theta_q \phi_q\left(x_{(q,i)}, Pa_q, f\right)}}{\sum_{h=1}^{|S_q|} e^{\theta_q \phi_q\left(x_{(q,h)}, Pa_q, f\right)}}$$

$$\arg\max_{X_{-1}, X_0, \ldots, X_{N_i}} P\left(X_{-1}, X_0, \ldots, X_{N_i} \mid f; \theta_{-1}, \ldots, \theta_{N_i}\right)$$

[Ludwig et al. *IEEE Trans. on Comp. Intell. and AI in Games* 2017]

ϕ_q = vector of features, modeling outputs of the semantic role labeler, coreference resolver, etc., sometimes in the form of a constraint rule

θ_q = parameters to be learned

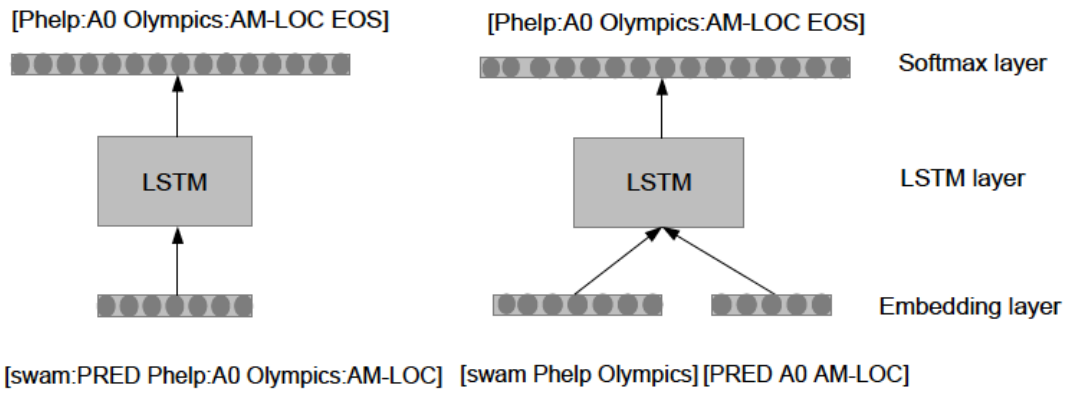
Implicit semantic role labeling

- Content is often left **implicit in a story**, but inferred by readers
- "Tuk thought a lot about the day that he would hunt his first animal."

If we would visualize his thought:
the location is not made explicit :
Outside snow fields in the artic world

Implicit semantic role labeling

- Improving implicit semantic role labeling by predicting semantic frame arguments in texts



2 models that implement an encoder –decoder architecture trained on large corpus with noisy SRL annotations

[Do Thi et al. *IJCNLP* 2017]

Implicit semantic role labeling

The task is to infer and instantiate the missing semantic roles given a number of possibilities from the discourse

	P	R	F1
Gerber and Chai (2010)	44.5	40.4	42.3
Laparra and Rigau (2013)	47.9	43.8	45.8
Schenk and Chiarcos (2016)	33.5	39.2	36.1
Model 1	48.0	38.2	42.6
Model 2	**52.6**	41.0	**46.1**

Table 1: Implicit role labeling evaluation.

Use WordNet and manually annotated iSRL data

Use WordNet, named entity annotations, and manual semantic category mappings

[Do Thi et al. *IJCNLP* 2017]

Through the deep learning representations we are able to transfer valuable knowledge

Other inferences

- **Language** of stories is often more **abstract** than the details needed to render the actions in a virtual world:
  - To a certain extent - but not completely - solved by word representations and dependency on previous action in the story

Fig. 3. Two-dimensional PCA projection for the vectorial representation of the words "take" and "care", beyond some low-level action instances, in red, and the vectorial composition $p \circ g(\texttt{take care}) = p \circ g(\texttt{take}) + p \circ g(\texttt{care})$ in blue.

3. What is still needed?

MUSE :

- Can we learn representations that capture world and commonsense knowledge ?

- Can we learn representations that are better suited for translating to other modality?

The role of world and commonsense knowledge

- To truthfully render the content of a text in a virtual world: a large amount of world and commonsense knowledge is needed because content is:
  - Not made explicit
  - Communicated in a more abstract way

He helped to make ready the dog sled for each trip

Representation learning

- Over decades symbolic representation languages that use a limited symbolic vocabulary were developed, many of which follow first-order logic as underlying knowledge representation formalism:
  - A primary goal is to facilitate reasoning about the world, rather than taking action in it (Davis 1993)
  - They form yet another human language - albeit usually less complex -, and are prone to ambiguity and redundancy (Ritter et al. 2006)

Representation learning

- Qualitative symbolic representations of language are not scalable:
  - Impossible to learn models for all types of objects, attributes, relationships
  - Many different “label” structures
  - Poses practical problems of scalability
  - Still need to be translated to the real-world physical space in many real-world applications

4. What are important properties that models should have?

What Does It Mean to Understand Language?

TERRY WINOGRAD
Stanford University

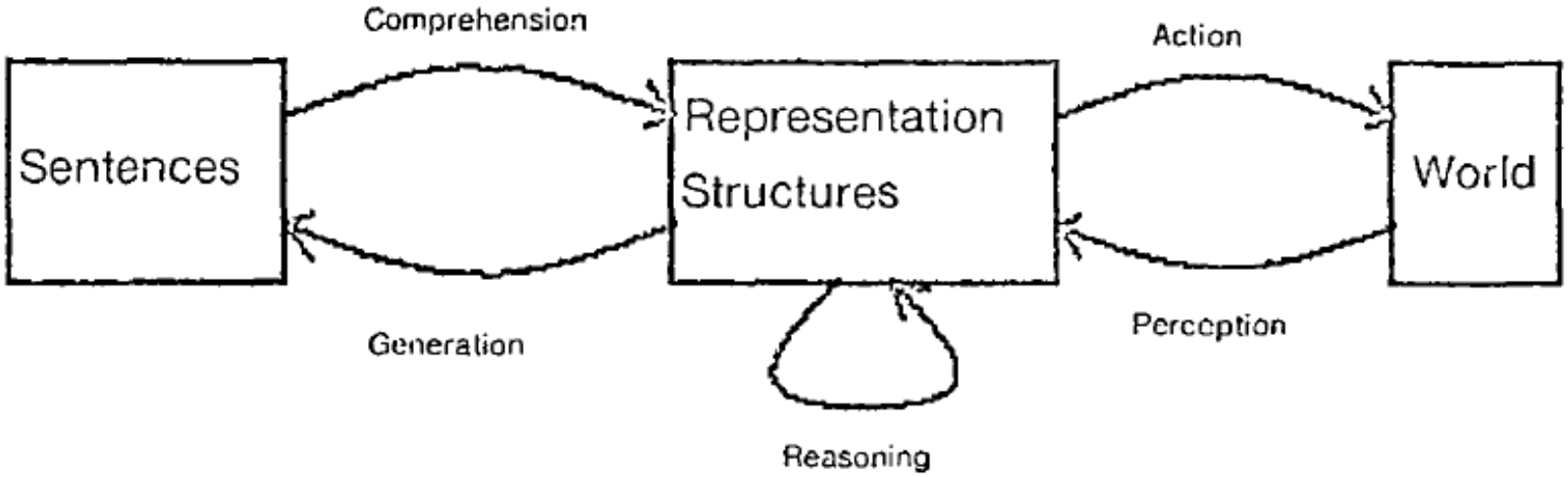


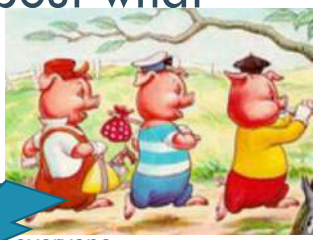
Figure 1. Basic AI model of language understanding.

[Winograd 1980]

Learning anticipatory representations from language and visual data

- Cognitive and neuroscience studies: human brain uses **anticipation** to perform tasks very efficiently [Friston *Nature* 2010, Vernon *Artif. Cogn. Syst.* 2014 p. 2]:
  - Based on lifelong **verbal and perceptual experiences** a human anticipates what events might occur in his or her environment
  - When humans read texts :
    - A reader's comprehension system continuously makes predictions about what information will be presented next in the text

[Kurby & Zacks *Cogn. Neurosc. of Nat. Lang. Use* 2015, Lambon Ralph et al. *Neuroscience* 2017]



All summer long, they roamed through the wood
plains,playing games and having fun. None we
three little pigs, and they easily made friends
everyone.
Wherever they went, they were given a w
welcome, but as
summer drew to a close, they realized that folk were drifting back
to their usual jobs, and preparing for winter.

Learning anticipatory representations from language and visual data

- Neuroscientists demonstrate the existence in the brain of: [Handjaras et al. NeuroImage 2016, Lambon Ralph et al. *op. cit.* 2017]
  - Large scale modality independent conceptual representations
  - Small scale modality dependent and category specific representations
  - Storage of anticipated events as **template structures** that are independent of the various distinctive contexts that might be encountered, providing a basis for conceptual generalization

Grounding representations in perception

- When humans imagine events:
  - Multiple forms of **mental imagery** exist: e.g., [Moulton & Kosslyn *Philos. Trans. Roy. Soc.* 2009]
    - Object-based (of shapes, colors),
    - Spatial (e.g., of locations)
    - Auditory
    - Motoric

Requirements of representations usable by a machine

- Reflecting the real world:
  - By nature relational: involves some **degree of scene reconstruction** with key elements (actions, people, objects, settings and their relations)
- Invariant to:
  - Paraphrasing of language
  - Physical permutations that do not change the meaning of its language description
- At different levels of abstraction:
  - Group objects that behave similarly
  - Compositional in nature
  - Predictive/anticipatory

[Moens Arg. & Computation 2018]

Requirements of representations usable by a machine

- Compositional in nature:
  - Allowing to **represent and constrain** more complex scenes or discourses
  - Allowing to **infer** implicit or obfuscated information
  - Allowing to deal with **recursion** in language grammar

- Predictive:
  - Allowing fast parsing

[Moens Arg. & Computation 2018]

Machine learning requirements

- Initial representations can be learned from large datasets (e.g., perception paired with language)
  - Supervised in sense that language describes image
  - Weakly supervised

- Incremental learning:
  - Representations can be adjusted to fit a task
  - With little or no supervision

Evaluation of HLU

What is understanding?

To understand a statement is to:

- determine its truth (with justification)
- calculate its entailments
- take appropriate action in light of it
- translate it into another language
- …

=> **Among others, visualization of language**

=> **Inference with quantitative representations**

=> **Translation to events in a real physical space**

=> **Translation to other modality**

[MacCartney & Potts 2016]

We might need novel evaluation metrics besides the existing narrowly focused task-based metrics

Still a large program to realize in language understanding by an intelligent machine

5. Imagery to empower automated language understanding

The role of world and commonsense knowledge

Going back to translation of language to events happening in a virtual world:

He learned to sharpen a hunting spear:

where is the spear located in relation to the body of the actor, concerning the sharpen action?

He helped to make ready the dog sled for each trip:

what actions does this involve in an arctic environment?

Could we learn world or commonsense knowledge from visual data?

- MUSTER project: MUltimodal processing of Spatial and TEmporal expRession:
- **How can computer vision improve language understanding?**

http://www.chistera.eu/projects/muster

[Elliott & Keller *EMNLP* 2013]

A man is riding a bike down the road.
A car and trees are in the background.

KU LEUVEN

ETH zürich

UNIVERSITE PIERRE & MARIE CURIE LA SCIENCE A PARIS

Universidad del País Vasco Euskal Herriko Unibertsitatea

2016-2019

MUSTER: Learning from visual data

- Learn world or commonsense knowledge from visual data
- **Deep learning** offers a joint methodology for processing language and other modalities
- Deep learning : to learn temporal and spatial knowledge from visual data or from visual data aligned with textual data, and integrate this knowledge in suitable representations for language understanding
- Only the acquisition of a fraction of world knowledge needed for language understanding, but very useful for language understanding

MUSTER project

- *Q1. How to automatically create text representations in the form of single-word and multi-word embeddings that integrate perceptual knowledge in the representations of objects, actions, their spatial and temporal relations?* **Problem of multimodal representation construction**

- *Q2. How to use the novel improved semantic representations (i.e., embeddings) to improve machine understanding of human language?* **Problem of multimodal representation integration and usage**

Multimodal representations of words

- Multimodal representations of words obtained by mapping of the textual word embeddings of the word to its visual image: proven usefulness in word similarity tasks
- Integrating vision and language in a single multimodal representation:
  - Learning of a language-to-vision mapping
  - Output of the mapping = vector representation of the imagined concept
  - = cognitively plausible way of building representations, consistent with the inherently reconstructive and associative nature of human memory

[Collell et al. *AAAI* 2017]

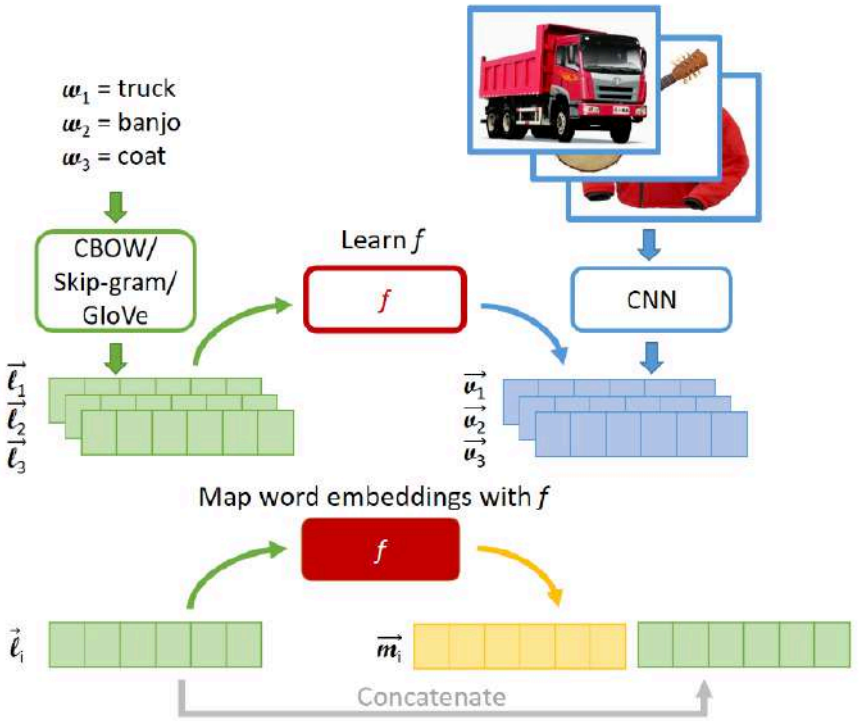


Figure 1: Overview of our model. The *imagined* representations are the outputs of a text-to-vision mapping.

[Collell et al. *AAAI* 2017]

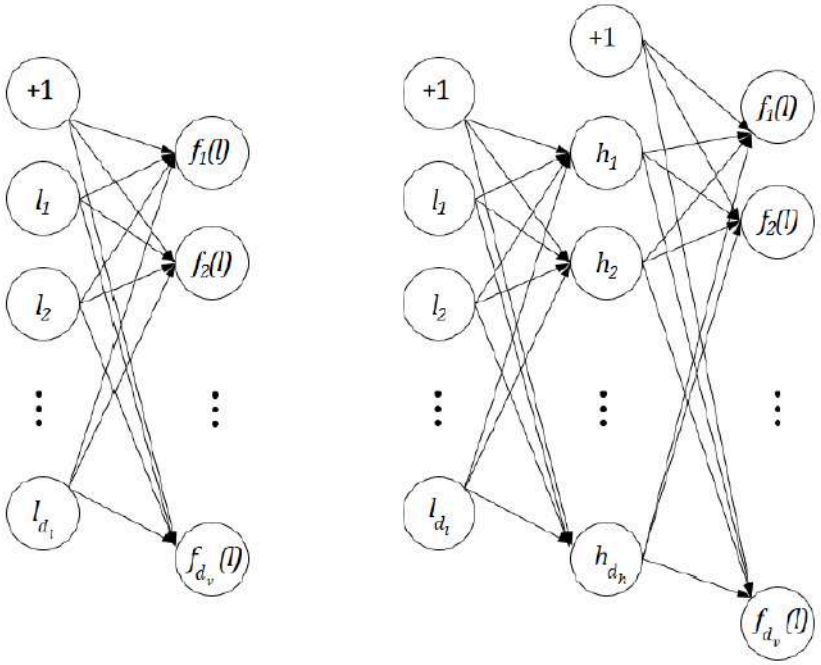


Figure 2: Architecture of the linear (left) and neural network (right) mappings.

[Collell et al. *AAAI* 2017]

	Wordsim353			MEN			SemSim			VisSim			Simlex999		
	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS
Silberer & Lapata 2014	-	-	-	-	-	-	0.7	-	-	0.64	-	-	-	-	-
Lazaridou et al. 2015	-	-	-	0.75	0.76	-	0.72	0.72	-	0.63	0.63	-	0.4	**0.53**	-
Kiela & Bottou 2014	-	0.61	-	-	0.72	-	-	-	-	-	-	-	-	-	-
GloVe	**0.712**	0.632	**0.705**	0.805	0.801	0.801	0.753	0.768	0.701	0.591	0.606	0.54	0.408	0.371	**0.429**
CNN_{avg}	-	0.448	-	-	0.593	-	-	0.534	-	-	0.56	-	-	0.406	-
CONC	-	0.606	-	-	0.8	-	-	0.734	-	-	0.651	-	-	0.442	-
MAP_{NN}	0.443	0.534	0.391	0.703	0.761	0.68	0.729	0.732	0.718	**0.658**	**0.659**	**0.655**	0.322	0.451	0.296
MAP_{lin}	0.402	0.539	0.366	0.701	0.774	0.674	0.738	0.738	0.74	0.646	0.644	0.651	0.322	0.412	0.286
$MAP\text{-}C_{NN}$	0.687	0.644	0.673	**0.813**	**0.82**	**0.806**	0.783	**0.791**	0.754	0.65	0.657	0.626	0.405	0.404	0.417
$MAP\text{-}C_{lin}$	0.694	**0.649**	0.684	0.811	0.819	0.802	**0.785**	**0.791**	**0.764**	0.641	0.647	0.623	**0.41**	0.388	0.422
# inst.	353	63	290	3000	795	2205	6933	5238	1695	6933	5238	1695	999	261	738

	Wordsim353-rel			Wordsim353-sim			SimVerb-3500		
	ALL	VIS	ZS	ALL	VIS	ZS	ALL	VIS	ZS
GloVe	**0.644**	0.759	**0.619**	**0.802**	0.688	**0.783**	0.283	0.32	0.282
CNN_{avg}	-	0.422	-	-	0.526	-	-	0.235	-
CONC	-	0.665	-	-	0.664	-	-	0.437	-
MAP_{NN}	0.33	0.606	0.267	0.536	0.599	0.475	0.213	**0.513**	0.21
MAP_{lin}	0.28	0.553	0.243	0.505	0.569	0.477	0.212	0.338	0.21
$MAP\text{-}C_{NN}$	0.623	0.778	0.589	0.769	0.696	0.745	**0.286**	0.49	0.284
$MAP\text{-}C_{lin}$	0.629	**0.797**	0.601	0.781	**0.698**	0.766	**0.286**	0.371	**0.285**
# inst.	252	28	224	203	45	158	3500	41	3459

[Collell et al. *AAAI* 2017]

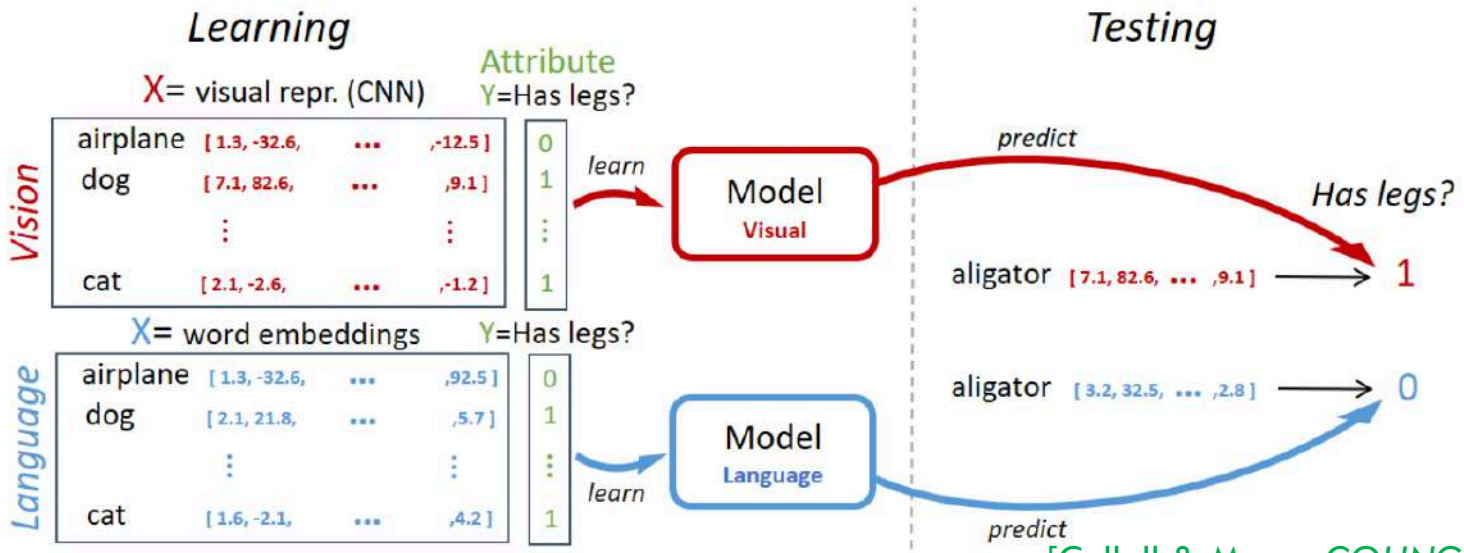
Table 1: Spearman correlations between model predictions and human ratings. For each test, ALL correspond to the whole set of word pairs, VIS to those pairs for which we have both visual representations, and ZS denotes its complement, i.e., zero-shot words. Boldface indicates the best results per column and # inst. the number of word pairs in each region (ALL, VIS, ZS). We notice that comparison methods are not available for test sets in the second row. Additionally, the VIS subset of the compared methods is only approximated, as the authors do not report the exact evaluated instances.

- Other realizations of learning meaning representations of words in MUSTER:
  - [Zablocki et al. AAAI 2018]: multimodal word embeddings
  - [Artetxe et al. AAAI 2018] : bilingual word embeddings

Representations of object attributes

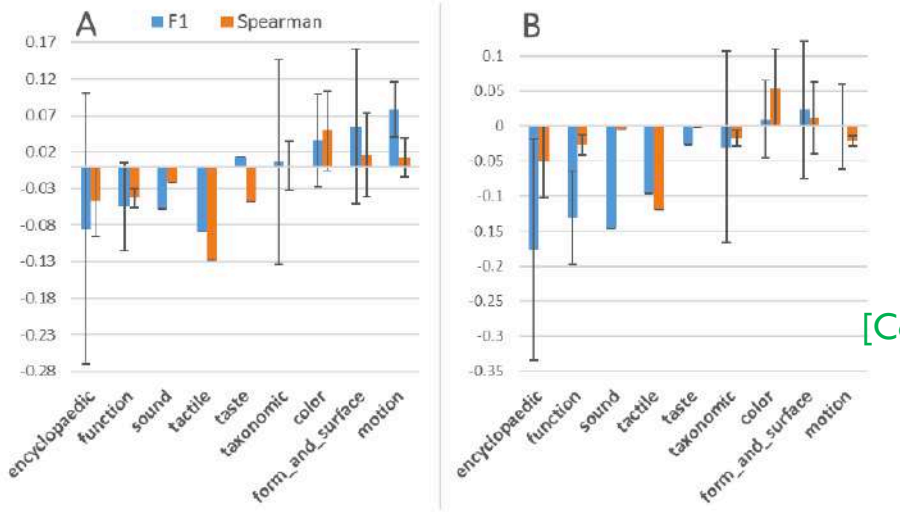
- Vision and language provide complementary information that, properly combined, can potentially yield more complete concept representations
- Study of visual and linguistic representations:
  - Which attributes are generally better captured by either the vision or by the language modality?
  - What type of attributes or semantic knowledge are better encoded by each modality?

[Collell & Moens *COLING* 2016]



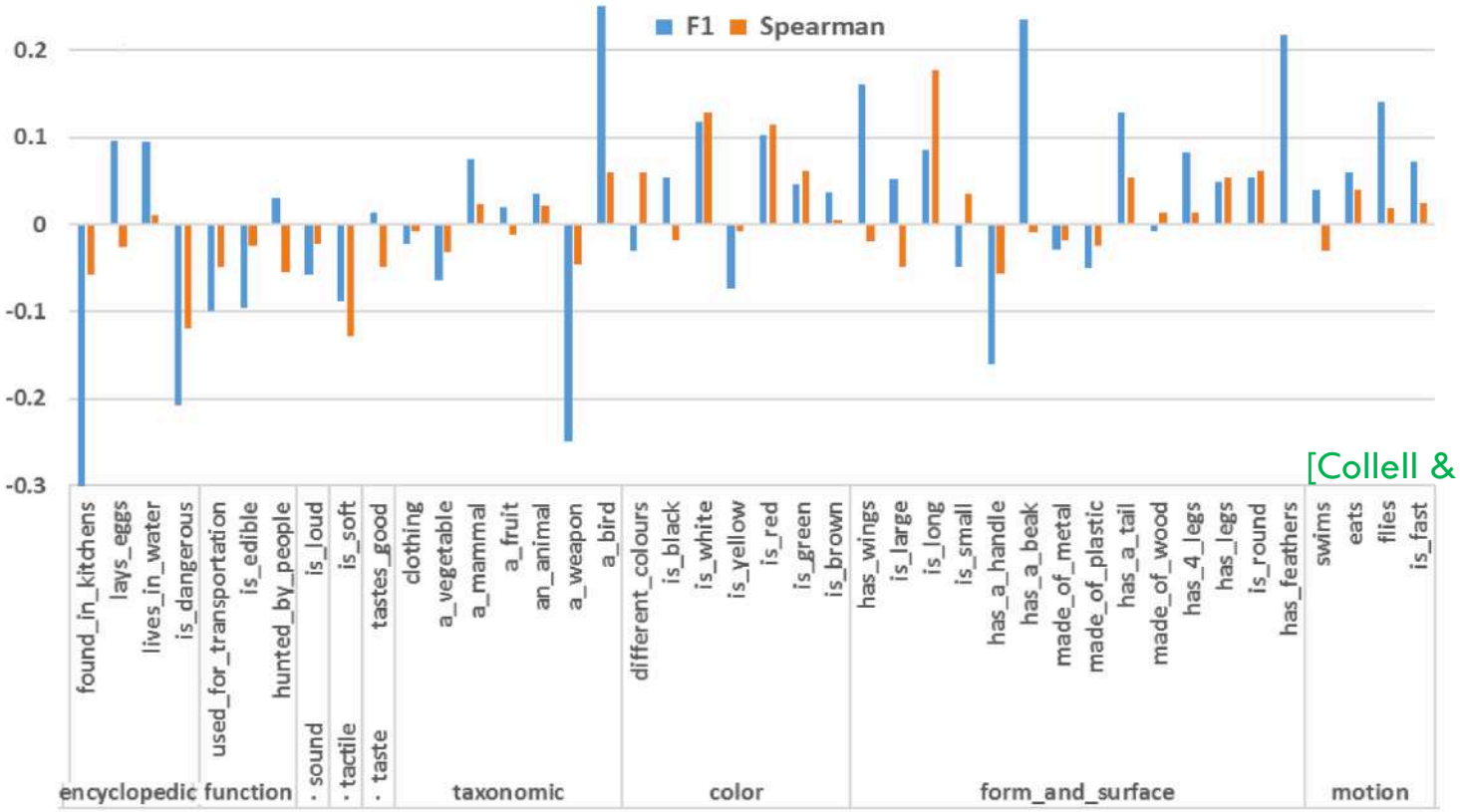
[Collell & Moens COLING 2016]

Figure 1: Overview of our experimental setting. Attributes are learned from the embeddings of each modality (left side), and afterwards new concepts are classified on whether the attribute is present or not (classification) or to which degree the attribute is present (regression). For clarity, we omitted the regression problem since its setting is identical to classification except for a continuous output $\mathcal{Y}$ instead of 0/1.



[Collell & Moens COLING 2016]

Figure 4: Averages of performance difference per attribute type. For each attribute type (e.g., taxonomic, taste, etc.), the bar indicates the average performance difference of its set of attributes. Plot A shows performance difference between VIS_{avg} and GloVe and B between VIS_{max} and GloVe. As in Fig. 3, positive bars indicate better performance of visual embeddings and negative bars otherwise. Error bars show standard error.



[Collell & Moens *COLING* 2016]

Figure 3: Difference of performance between VIS_{avg} minus GloVe. Attributes are shown on the horizontal axis and grouped by their type. Positive bars indicate better performance of visual embeddings and negative bars otherwise. Results with VIS_{max} are omitted as they exhibit almost identical patterns as VIS_{avg}, yet slightly worse.

Representations of spatial knowledge

A girl rides a horse

- Where is the horse located, where is the girl located in relation to the horse?
- Can we build suitable representations that caption this knowledge and potentially make inferences with it?

Representations of spatial knowledge

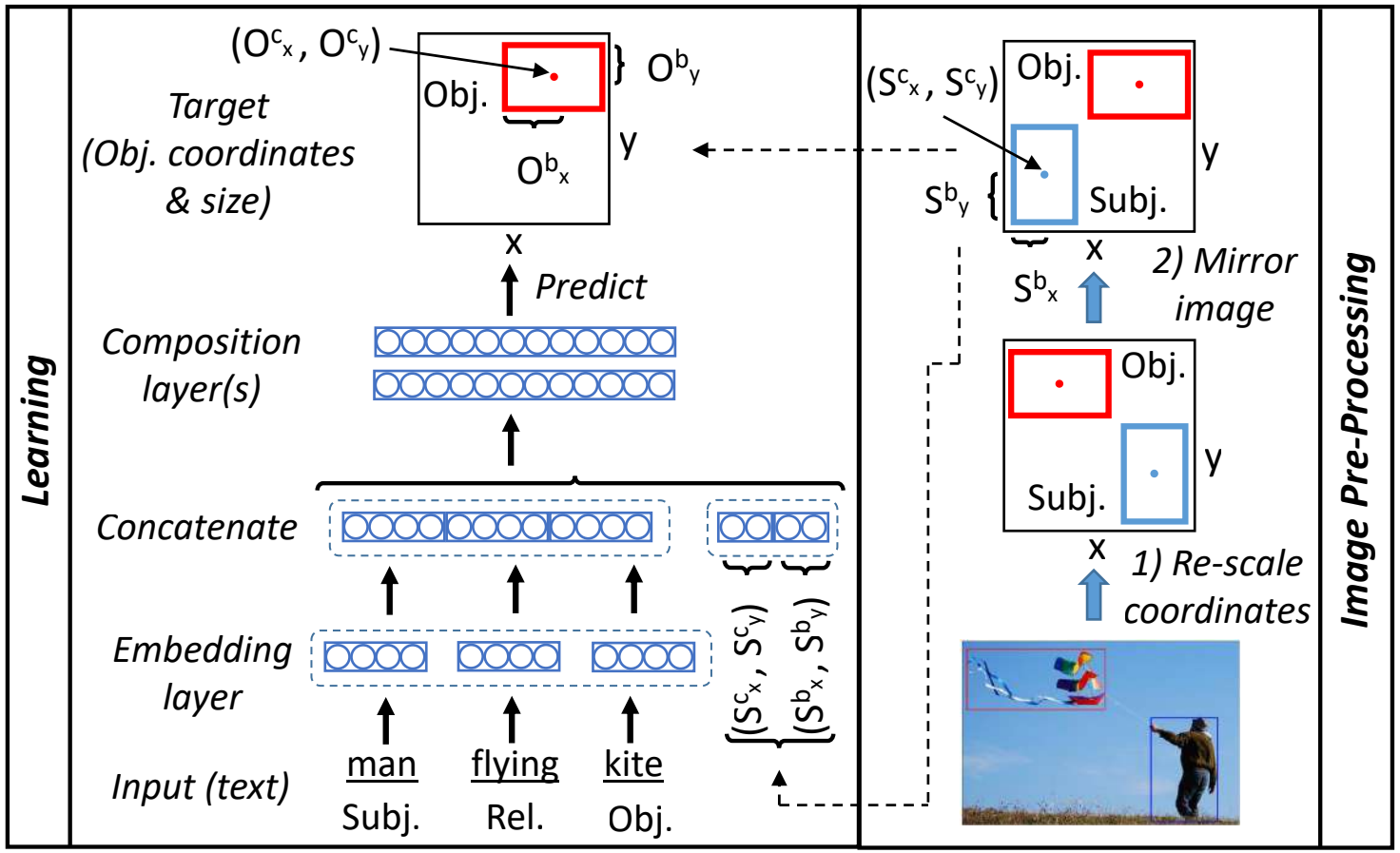
- Focus on spatial understanding of language and representing language with **spatial templates** = regions of acceptability of two objects under a spatial relationship
- Prior work restricts spatial templates to language that **explicitly** uses spatial cues (e.g., "glass *on* table") [Logan and Sadler *Language, Speech, and Communication* 1996, Moratz and Tenbrink *Spatial Cognition and Computation* 2006, Malinowski and Fritz arXiv 2014]
- We extend this concept to **implicit** spatial language, i.e., those relationships (generally actions) for which the spatial arrangement of the objects is only implicitly implied (e.g., "man *riding* horse") => requires significant commonsense spatial understanding [Collell & Moens *TACL* 2018]
-

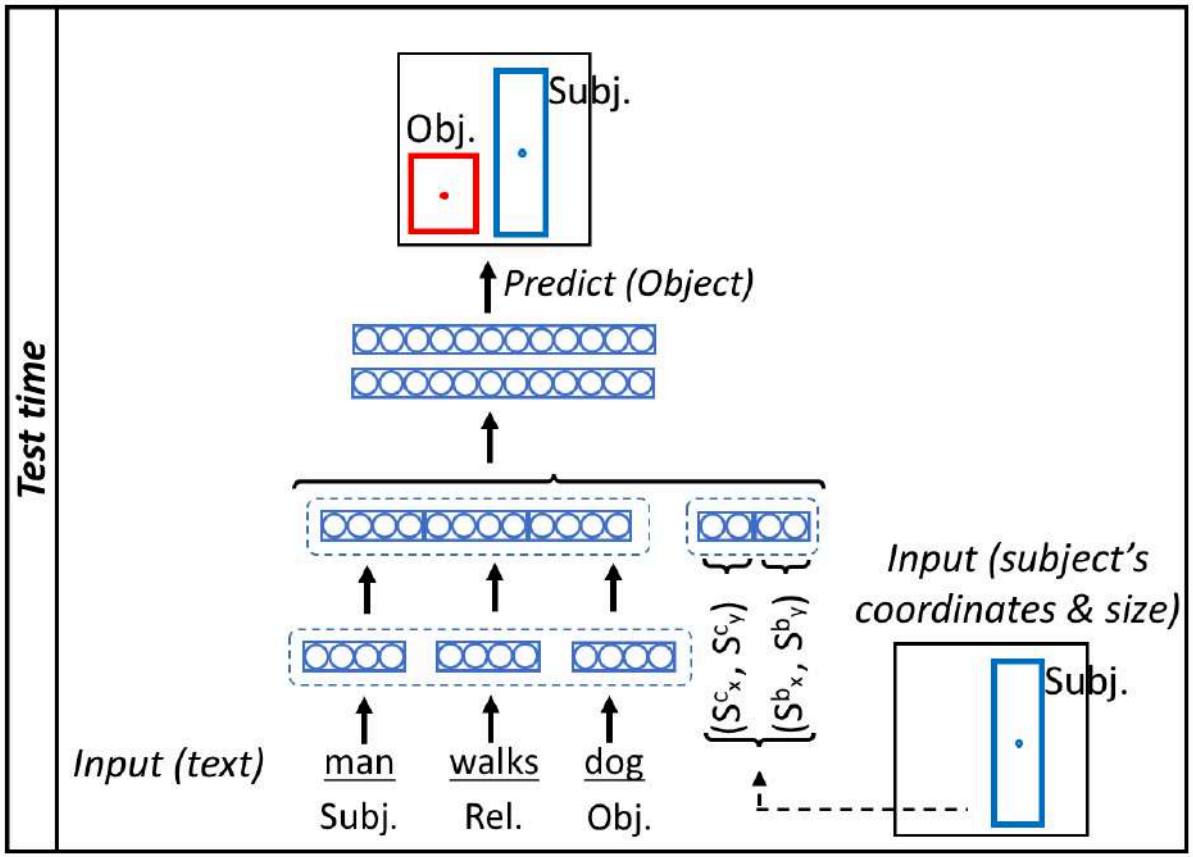
- We propose the task of:
  - Given a structured text input of the form (Subject, Relationship, Object) = (S,R,O)
  - Predict the 2D relative spatial arrangement of two objects (output)
- Train the task in a supervised setting:
  - Training set of image-text pairs, where the size and location of bounding boxes of objects in images serve as ground truth
- = a spatial "question-answering" task where the question consists in a spatial commonsense query such as *where is the* "man" *located with respect to a* "horse" *when a* "man" *is* "feeding" *the* "horse"?
- The answer is a 2D "imagined" representation in contrast with a sentence/word as typically done in question-answering tasks

General approach

- Neural network approach (simple feedforward neural network):
  - **Input**: triplet of words, optionally size of subject
  - **Embedding layer**: aim is to generalize over unseen words by using embedding look-up (e.g., Glove [Pennington et al. *EMNLP* 2014])
  - **Concatenation** of the triplet embedding and possibly size of subject
  - **Composition layer**: to build a compositional representation
  - **Output layer**: coordinates and size of predicted object (i.e., the bounding box)
  - **Objective function**: mean squared error loss

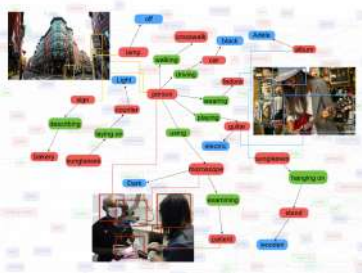
[Collell et al. *AAAI* 2018]



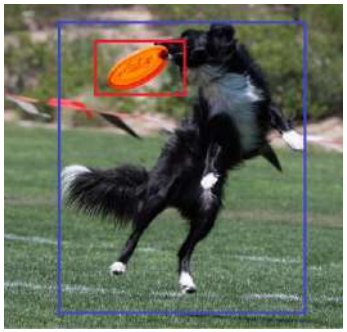


[Collell et al. *AAAI* 2018]

Experimental set-up

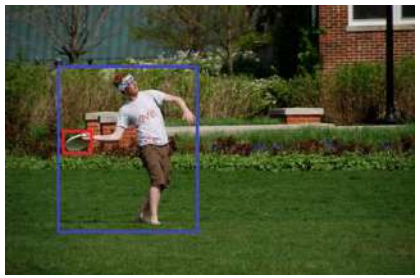


- Source of annotated images:
  - Visual Genome data set [Krishna et al. *CVPR* 2016]
  - 108K images with 1,5M human-annotated (Subject, Relationship, Object) instances with bounding boxes for Subject and Object

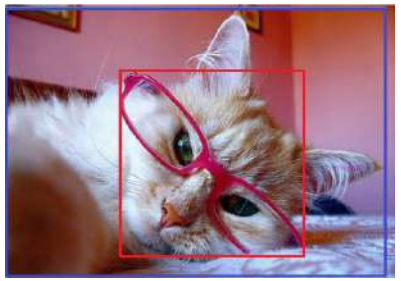


dog, catches, frisbee

boy, feeds, giraffe



man, throws, frisbee



cat, wears glasses

Experimental set-up

- Keep triplets for which pretrained word embeddings are available:
  - Implicit spatial relationships: 378K instances: 2,183 unique relationships and 5,614 unique objects
  - Explicit spatial relationships: 852K instances, 31 unique spatial prepositions and 6,749 unique objects
- Evaluation metrics:
  - Mean Squared Error (MSE) between predicted and true object center and size
  - Coefficient of Determination (R^2) between the predicted and true object center and size
  - Pearson Correlation (r) between the predicted and true object x and y coordinates
  - Accuracy and F1

Quantitative evaluation

- 10-fold cross-validation and results averaged over the 10 folds:

		MSE	R^2	acc_y	$F1_y$	r_x	r_y
Implicit	*EMB*	**0.008**	0.705	0.756	0.755	0.894	0.834
	RND	**0.008**	0.691	0.750	0.750	0.891	0.826
	1H	**0.008**	**0.717**	**0.762**	**0.762**	**0.896**	**0.842**
	ctrl	0.054	-1.000	0.522	0.521	0.000	-0.001
Explicit	*EMB*	0.013	0.586	0.768	0.770	0.811	0.823
	RND	0.013	0.580	0.767	0.769	0.808	0.815
	1H	**0.012**	**0.604**	**0.778**	**0.780**	**0.815**	**0.828**
	ctrl	0.060	-1.000	0.633	0.630	0.000	0.000

Table 1: Results on **implicit** and **explicit** relations.

EMB: Glove embeddings as input
RND: Random embeddings as input
1H: 1-hot encodings as input
Ctrl: control method that outputs random normal predictions

[Collell et al. *AAAI* 2018]

Quantitative evaluation

- 10-fold cross-validation and results averaged over the 10 folds:

		Extrapolated						No extrapolated					
		MSE	R^2	acc_y	$F1_y$	r_x	r_y	MSE	R^2	acc_y	$F1_y$	r_x	r_y
Triplets	*EMB*	**0.006**	0.749	0.786	0.789	0.904	0.871	0.008	0.711	0.758	0.759	0.894	0.839
	RND	0.007	0.727	0.767	0.771	0.899	0.861	0.008	0.701	0.757	0.757	0.893	0.832
	1H	**0.006**	**0.764**	**0.792**	**0.795**	**0.906**	**0.880**	**0.007**	**0.724**	**0.768**	**0.768**	**0.897**	**0.846**
	ctrl	0.053	-1.097	0.515	0.505	0.000	0.001	0.054	-1.016	0.521	0.521	-0.001	-0.001
Words	*EMB*	**0.010**	**0.635**	**0.747**	**0.747**	**0.879**	**0.793**	**0.008**	0.708	0.760	0.760	0.895	0.836
	RND	0.015	0.424	0.602	0.597	0.853	0.606	**0.008**	0.694	0.755	0.755	0.892	0.828
	1H	0.015	0.424	0.595	0.587	0.861	0.611	**0.008**	**0.721**	**0.766**	**0.766**	**0.897**	**0.845**
	ctrl	0.054	-1.022	0.519	0.518	-0.001	0.000	0.054	-1.003	0.520	0.520	0.001	0.000

Table 2: Results on the Extrapolated **Triplets** (top) and Extrapolated **Words** (bottom) sets (see Sect. 4.2). Right tables show results in the same sets without imposing extrapolation conditions, i.e., allowing to see all combinations/words during training.

[Collell et al. *AAAI* 2018]

Qualitative evaluation

[Collell & Moens *UCL Commonsense* 2017]

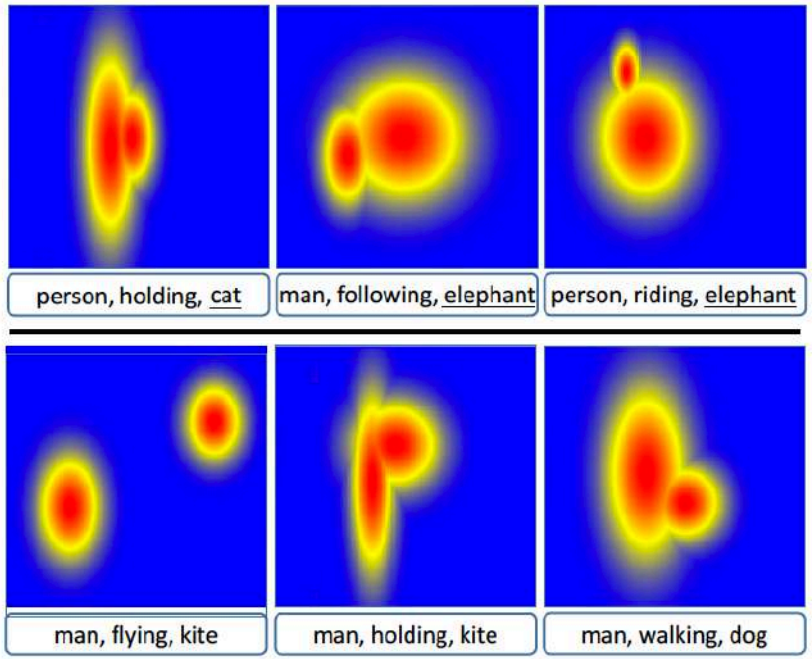
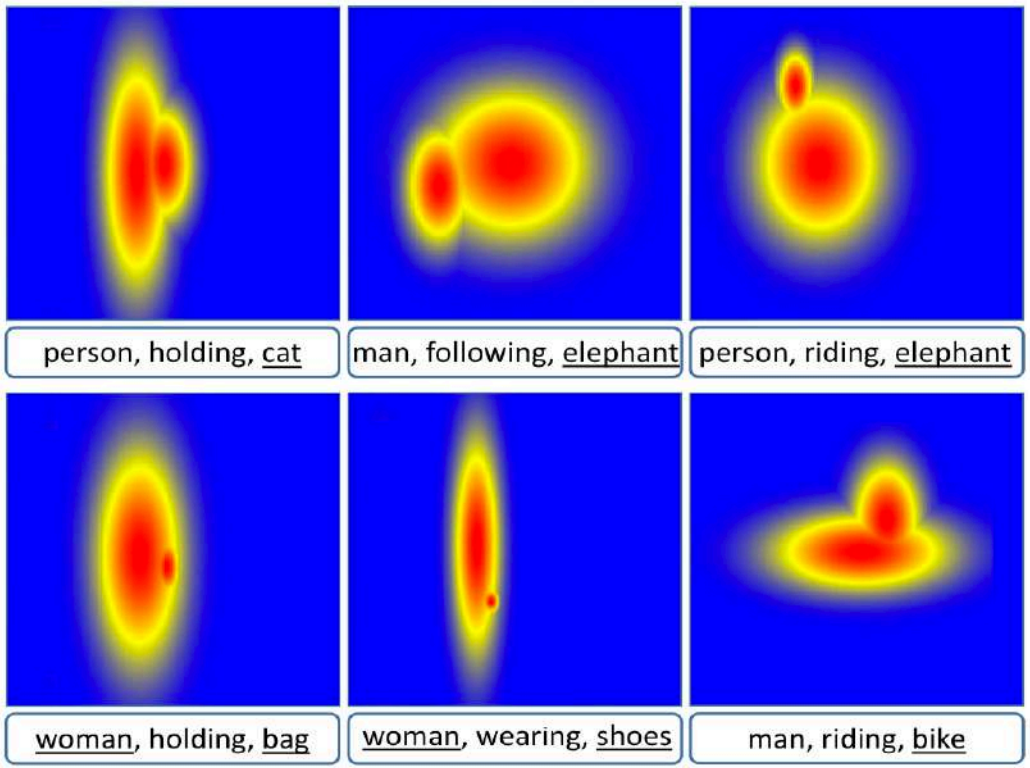


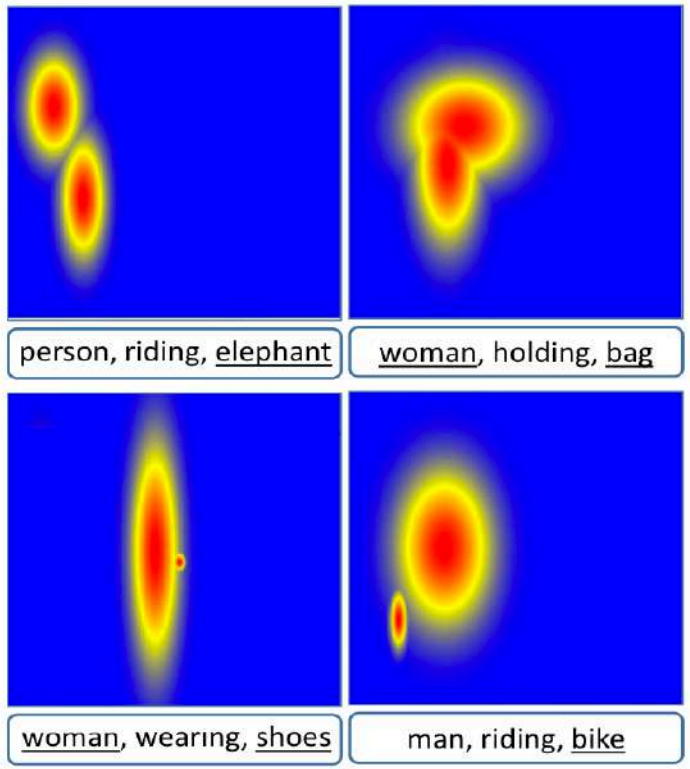
Figure 2: Predictions by the model that leverages word embeddings (*EMB*). **Top**: Predictions in unseen words (underlined). **Bottom:** Predictions in unseen *triplets*.

Qualitative evaluation

[Collell & Moens UCL Commonsense 2017]



Model: Initialized with distributional word embeddings



Model: Initialized with random word embeddings

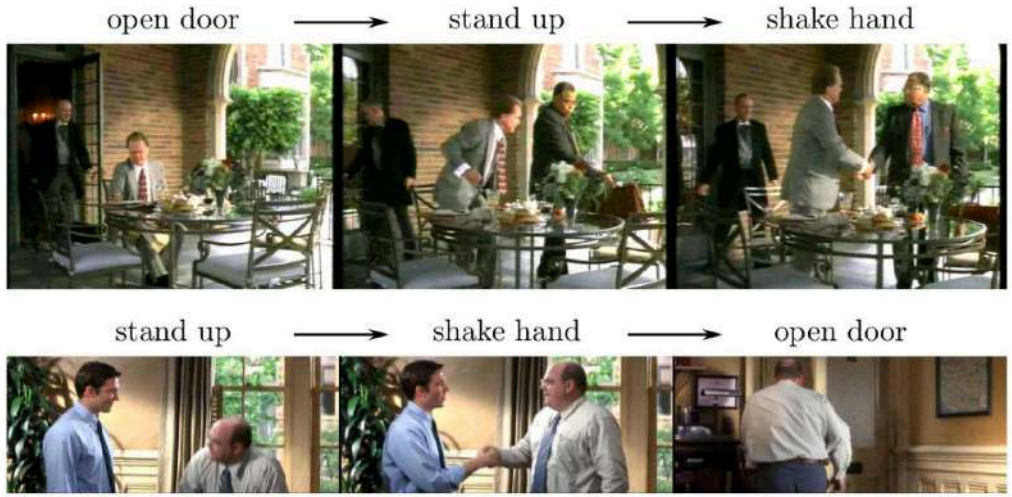
- Our work can easily be expanded to predicting relative 3D spatial arrangements of objects from language input given that suitable training data are available
- Our work has potential for real-time language understanding in a visual context:
  - Language communication to robots, machines, self-driving cars, …
  - Translation of spatial language to geometric space opens possibilities of fast **quantitative reasoning in such a space**, which can complement qualitative symbolic representations and reasoning
- Our work is a step towards opening the black box of neural models applied to language processing by visualizing the interpreted content

Representations of temporal knowledge

Tuk and his father tied the last things to the sled and then set off

- Entails that:
  - Sled dogs are hooked to the sled
  - Tuk and his father sit down on the sled
  - The sled dogs pull the sled
- Knowledge of scripts:
  - Traditionally learned as language models from text [e.g., Jans et al. *EACL* 2012, Pichotta and Mooney *AAAI* 2016]

MUSTER wants to learn this temporal knowledge from visual data



[MUSTER project proposal]

Figure 2: An example of actions along with their annotations available from the dataset published in [BOJA14]. One aim of MUSTER will be exploiting such annotations for building multi-modal representations of actions (MUSTER objective 3) which could then be integrated into systems for temporal recognition and ordering of actions in text (MUSTER objective 4) which will in turn be evaulated in temporal HLU tasks (MUSTER objective 5).

First step in this direction: [Vasudevan et al. ACM Multimedia 2017]

CALCULUS : ERC Advanced Grant, 2018-2023

Commonsense and Anticipation enriched Learning of Continuous representations sUpporting Language UnderStanding

European Research Council

Established by the European Commission

6. Conclusions

Conclusions

- The goal of natural language understanding is already there a long time
- Today's neural network based approaches – whether deep or not so deep – have learned us an interest in continuous representations of meaning:
  - Offer a methodology to jointly process language and vision
- In this talk: how imagery helps in building richer meaning representations and integrate commonsense knowledge
- More needs to be done:
  - To fast parse language
  - To fast and incrementally learn
  - To make compositions and inferences with the representations that we have learned

For more details …

- Collell Talleda, G. & Moens, M.-F. (2016). Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations. In *Proceedings of the 26th International Conference on Computational Linguistics* (COLING 2016) (pp. 2807-2817). ACL.
- Collell Talleda, G., Zhang, T. & Moens, M.-F. (2017). Imagined Visual Representations as Multimodal Embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (AAAI 17) (pp. 4378-4383). AAAI.
- Collell, G. & Moens, M.-F. (2017). Learning Visually Grounded Common Sense Spatial Knowledge for Implicit Spatial Language. In *Proceedings of the 13th International Symposium on Commonsense Reasoning, University College London*. CEUR.
- Collell, G. & Moens, M.-F. (2018). Learning Representations Specialized in Spatial Knowledge: Leveraging Language and Vision. *Transactions of the Association for Computational Linguistics* (*TACL*), 6, 133-144.
- Collell, G., Van Gool, L. & Moens, M.-F. (2018). Acquiring Common Sense Spatial Knowledge through Implicit Spatial Templates. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence* (*AAAI 2018*). AAAI.
- Do Thi, Q.N., Bethard, S. & Moens, M.-F. (2016). Visualizing the Content of a Children's Story in a Virtual World. In *Proceedings of Uphill Battles in Language Processing Scaling Early Achievements to Robust Methods*, Workshop held at EMNLP 2016.
- Do Thi, Q.N., Bethard, S. & Moens, M.-F. (2017). Improving Implicit Semantic Role Labeling by Predicting Semantic Frame Arguments. In *Proceedings of the 8th International Joint Conference on Natural Language Processing* (IJCNLP 2017) (pp. 90-99). ACL.
- Ludwig, O., Do Thi, Q.N., Smith, C., Cavazza, M. & Moens, M.-F. (2017). Learning to Extract Action Descriptions from Narrative Text. *IEEE Transactions on Computational Intelligence and AI in Games* 10 (1), 15-28.
- Moens, M.-F. (2018). Argumentation Mining: How Can a Machine Acquire Common Sense and World Knowledge? *Argument and Computation,* 9 (1), 1-14.