The Discovery of Spatial Knowledge from Images and Language

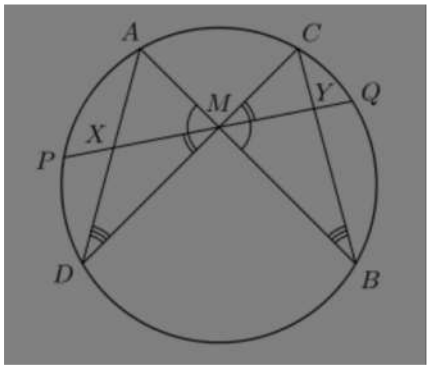
Marie-Francine Moens

Joint work with Guillem Collell, Parisa Kordjamshidi and Thierry Deruyttere

Department of Computer Science, KU Leuven

EKAW 2018, Nancy, France, 16-11-2018

Study of space



- In antiquity the study of space emerged among the ancient Babylonians and Greeks and led to Euclidean geometry
- The next breakthrough was probably the development of analytic geometry by René Descartes and the projective geometry by Girard Desargues in the 17th century
- In the 19th century non-Euclidean geometries were developed extending the concept of space beyond what could be intuited through everyday perception

[Chilton Language, Cognition, Space 2010]

Study of space

- Neuroscientist John O'Keefe contributed pioneering work on mammalian spatial cognition: three-dimensional Euclidean construction is inherent to the human nervous system

[O'Keefe 1999]

- The human experience of space includes knowledge relating to size, shape, location and distribution of entities in a three-dimensional environment

[Evans Language, Cognition, Space 2010]

Cognitive maps

- The human brain constructs spatial or cognitive maps (Edward C. Tolman 1948), which:
  - Facilitate navigation
  - Are a prerequisite to experience objects and their motions
  - Allow us to perceive places independent of the entities and objects that places and locations occupy
  - Humans - like many other organisms - can compute distances and other spatial relations between distant places such as directions without having to physically experience the spatial relationship
  - Three-dimensional Euclidean space is imposed on perceptual experience
- Map-like representations of the environment are constructed by humans and other species: from a neuroscience viewpoint it is **still unclear what the nature of these representations is**

[Evans Language, Cognition, Space 2010]

Language and space

- Language enables humans to communicate about space
- Understanding the spatial meaning of language is important: both in:
  - Human-human communication

  - Human-machine communication

  => both for humans and machines
- **However, language is incomplete, vague and ambiguous**

Language and space

- Humans usually have no trouble of understanding the spatial meaning of language:
  - Rely on huge experience of grounding language meaning in visual and other perceptual experiences

Language and space

- Given that we perceive our surrounding world at 25 frames per second, a child of 3 years who is awake 8 hours a day has been exposed to about 700 million visual frames (not counting the first 4 months), from which it learns

Language and space

- Can processing imagery help automated language understanding and especially in recovering the spatial meaning of a language utterance?

Spatial inference

- Humans acquire additional insights by reasoning:
  - **Logical reasoning**: deductive, inductive, abductive
  - To solve unknown situations in case no teacher is available, humans rely on **case-based reasoning** and solve the problem following the solution of an analogous problem that they have experienced, where a key component in the analogy is the shared relationships found in both problems [Gentner *Cogn. Scienc.* 2010]
  - Humans are also very proficient in **geometric reasoning** estimating positions, angles and routes while mentally making computations in 3-D Euclidean spaces, and can even adapt the computations to 4-D spaces [Alflalo & Graziano *Journ. of Exp. Psych,* 2008]

Spatial inference

- So, beyond language processing and computer vision in which spaces do machines best reason with regard to acquiring spatial knowledge?

Overview

- Qualitative versus quantitative understanding of spatial language
- Qualitative understanding of spatial language
- The role of imagery in quantitative understanding of spatial language
- In which spaces to reason ?

Overview

- Qualitative versus quantitative understanding of spatial language
- Qualitative understanding of spatial language
- The role of imagery in quantitative understanding of spatial language
- In which spaces to reason ?

Qualitative versus quantitative spatial understanding

- **Qualitative spatial representations and reasoning:**
  - Translation of spatial language to discrete symbolic representations
  - Provides a calculus which allows a machine to represent and reason with spatial entities and their attributes

- **Quantitative spatial representations and reasoning:**
  - Translation of spatial language to continuous representations, involving geometric spaces and reasoning in these spaces

[Cohn & Renz 2008]

Overview

- Qualitative versus quantitative understanding of spatial language
- Qualitative understanding of spatial language
- The role of imagery in quantitative understanding of spatial language
- In which spaces to reason ?

Spatial information extraction

- **Spatial role labeling**
  - Recognizing the spatial relation between two objects
  - Including recognizing the role of argument objects (spatial roles), such as trajector and landmark
  - Including recognizing attributes of the spatial relation
- Ultimate goal: **predicting the position of objects** in a 2D or 3D space

Spatial information extraction

- Spatial role labeling (SpRL): considers the extraction of a set of generic spatial roles and relations which includes, but is not limited to:
  - The role of **trajector,** which is defined as an entity whose location or translocation is described in a sentence
  - The role of **landmark** which is defined as an entity by which we describe the location of the trajector
  - The role of **spatial indicator** which is a linguistic signal that indicates the presence of a spatial relationship between a trajector and landmark (not always present)
  - The **attributes** of the spatial relation: e.g., direction, orientation, connectivity

Annotation schemes: spatial semantics

- SemEval 2012
- SemEval 2013
- SpaceEval 2015 (as part of SemEval 2015): ISOSpace in SemEval 2015: the spatial language specification language
- CLEF 2017: use of ISOSpace, also images as context
- Several small annotated corpora available

Spatial role labeling

SemEval 2012

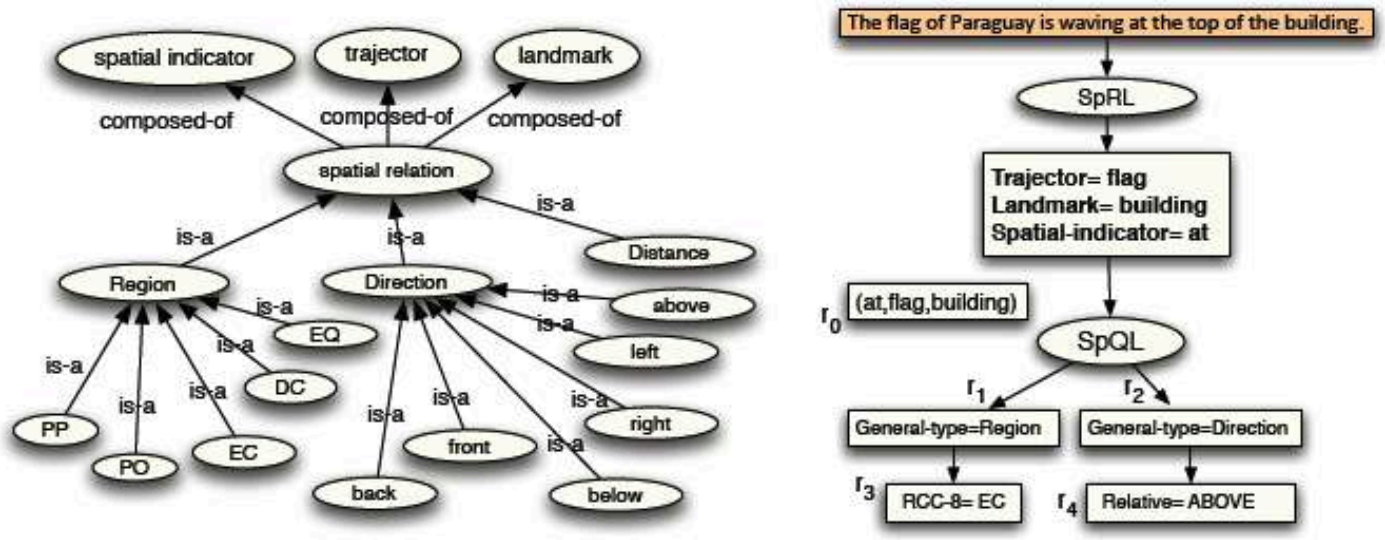


Figure 1. (a) The spatial ontology. (b) Example sentence and the recognized spatial concepts.

The goal is to jointly assign the labels of the ontology to a text item

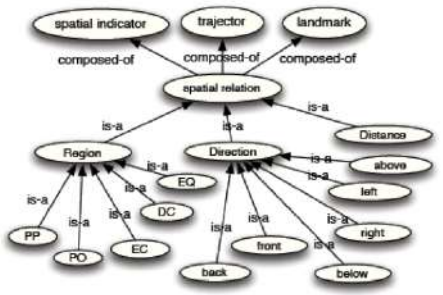
SpaceEval 2015

- Locations: regions, spatial objects, geographic and geopolitical places
- Entities participating in spatial relations
- Paths: routes, lines, turns, arcs
- Topological relations: *in*, *connected*.
- Direction and orientation: *North*, *down*
- Time and space measurements: *20 miles away*, *for two hours*
- Object properties: intrinsic orientation, dimensionality.
- Frames of reference: absolute, intrinsic, relative
- Motion: tracking objects over time

http://alt.qcri.org/semeval2015/task8/

Spatial role labeling

- Joint or global learning – **structured learning** ≠ local learning of independent classifiers or pipelining of classifiers:
  - 1 classification model for the global structure
  - Output is = structure (here spatial ontology)



[Kordjamshidi et al. Journal of Web Semantics 2015]

Input

- Object to which the classification model is applied: e.g., sentence (in our case), paragraph, full document, ...
- Is usually composed of different input components: single words, phrases, ... depending on the type of text snippet to which a label will be assigned

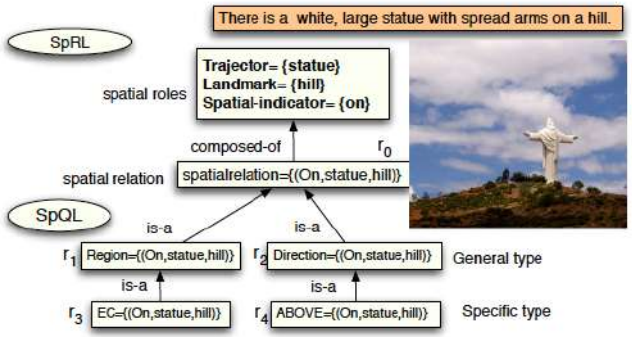
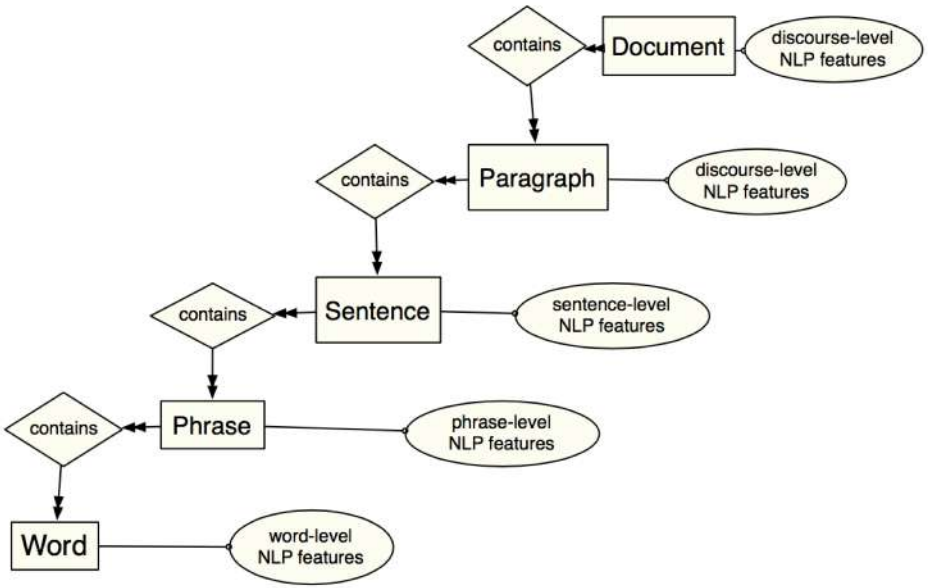


Figure 1 An input example structure represented as a document and its NLP features at different layers (document, paragraph, sentence, ...) independent from the output representation and its elements.



Output

- Output variables = labels in the structure

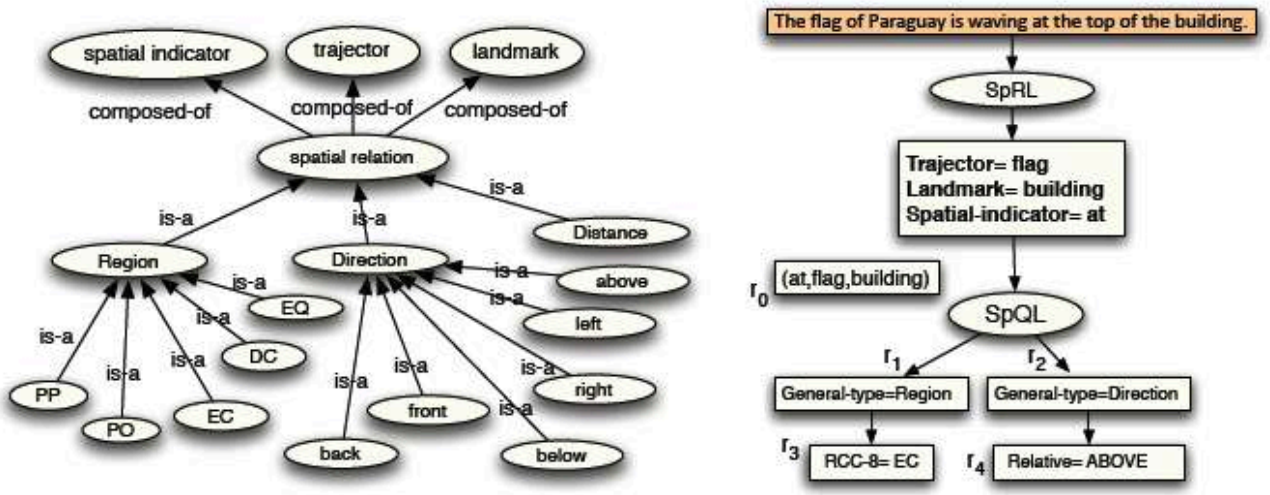


Figure 1. (a) The spatial ontology. (b) Example sentence and the recognized spatial concepts.

[Kordjamshidi et al. Journal of Web Semantics 2015]

Spatial role labeling

- Spatial role labeling: linear structured classifier: we learn the weights of each feature function, but use a linear scoring function g, with assignment of a label structure with maximum score:

$$g(X, Y; W) = \langle W, f(x, y) \rangle$$

$$\hat{Y} = \arg\max_{Y_l} g(X, Y; W)$$

$g(X, Y; W)$ is a linear function in terms of the combined feature representations associated with each candidate input component x and an output label y

Training of the model

- Given N training examples: $E = \{(X^{(i)}, Y^{(i)}) \in \mathcal{X} \times \mathcal{Y} : i = 1..N\}$
- Objective: the score of a training example with a correct labeling $Y^{(i)}$ should be larger or equal than the score of a training example of an incorrect labeling $Y \in \mathcal{Y}$ + some cost :

$$g(X^{(i)}, Y^{(i)}; W) \geq g(X^{(i)}, Y; W) + \Delta(Y^{(i)}, Y), \quad \forall Y \in \mathcal{Y}$$

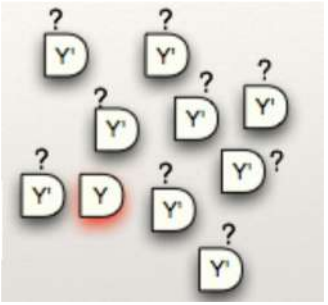
- A violation :

$$g(X^{(i)}, Y; W) - g(X^{(i)}, Y^{(i)}; W) + \Delta(Y^{(i)}, Y) > 0$$

Training of the model

- To improve the efficiency of the training: training is done only with the most violated Y for each X

$$\arg\max_{Y \in \mathcal{Y}} \left(g\left(X^{(i)}, Y; W\right) - g\left(X^{(i)}, Y^{(i)}; W\right) + \Delta\left(Y^{(i)}, Y\right) \right)$$



- Training = finding the weights W that minimize these violations:

$$\text{minimize } l(W) = \sum_{i=1}^{N} \max_{Y \in \mathcal{Y}} \left(g\left(X^{(i)}, Y; W\right) - g\left(X^{(i)}, Y^{(i)}; W\right) + \Delta\left(Y^{(i)}, Y\right) \right)$$

e.g., Hamming distance between ground truth labeling and wrong labeling

Training of the model

- Training is done with a structured support vector machines (SSVM) or structured perceptron
- Training with the most violated constraints/outputs (Y) per training example:
  - Using domain specific constraints in integer linear programming formulation
  - Using ILP solver to find most violated Y given spatial constraints

Spatial inference or reasoning during training

Constraints

$$sp_i + nsp_i = 1$$

$$sp_i tr_j + sp_i lm_j + sp_i nrol_j = 1$$

$$sp_i tr_j - sp_i \leq 0, \quad sp_i lm_j - sp_i \leq 0$$

$$sp_i - \sum_j (sp_i tr_j) \leq 0, \quad sp_i - \sum_j (sp_i lm_j) \leq 0$$

$$sp_i tr_j + sp_i lm_j \leq 1$$

$$\sum_i (sp_i tr_j) \leq 1, \quad \sum_i (sp_i lm_j) \leq 1$$

$$sp_i tr_j lm_k r_{\gamma'} - sp_i tr_j lm_k r_\gamma \leq 0, \quad \forall \gamma < \gamma' \quad \gamma, \gamma' \in \mathcal{H}$$

$$\sum_{\gamma \in \mathcal{H}_{leafs}} r_\gamma(x_i, x_j, x_k) \geq r_0(x_i, x_j, x_k)$$

$$\sum_{\gamma \in QSR_k} sp_i tr_j lm_k r_\gamma \leq 1, \quad \forall h, \quad \forall QSR_h \subset \mathcal{H}_{leafs}$$

Constraints are linear and variables take the form of integers

Constraints are applied: during training and during testing

Target	Precision	Recall	F1	Annotated	Positive candidates	Negative candidates
sp	0.881	0,942	0,909	1466	1437	1992
sp.tr	0.752	0.622	0.678	1693	1640	20133
sp.lm	0.853	0.815	0.832	1196	1161	24123
r_0	0.526	0.533	0.529	1703	1619	–

Table 5. L+I (LISpRL): Local training, global prediction for single label *sp*, linked labels *sp.tr*, *sp.lm* and producing r_0 using rule 16, BSVM.

SemEval-2012 corpus

Class	Precision	Recall	F
Region	0.667	0.541	0.594
Direction	0.602	0.544	0.57
Distance	0.633	0.409	0.477
EQ	0.9	0.7	0.6
DC	0.383	0.304	0.330
EC	0.571	0.442	0.486
PO	0.85	0.464	0.458
PP	0.577	0.48	0.521
BELOW	0.6	0.55	0.49
LEFT	0.449	0.292	0.331
RIGHT	0.372	0.537	0.359
BEHIND	0.602	0.563	0.573
FRONT	0.558	0.508	0.525
ABOVE	0.654	0.485	0.513
W.Avg	0.593	0.493	0.527

Table 9. Pipeline SpRL and SpQL (EtoE-pipe): AvGSPerc.

[Kordjamshidi et al. *Journal of Web Semantics* 2015]

Relation extraction in the biomedical domain

- [GE] Genia Event Extraction for NFkB knowledge base
- [CG] Cancer Genetics
- [PC] Pathway Curation
- [GRO] Corpus Annotation with Gene Regulation Ontology
- [GRN] Gene Regulation Network in Bacteria
- [BB] Bacteria Biotopes (semantic annotation by an ontology)

- Challenging !

Participant	Rank	Recall	Precision	F1
	1	0.28	0.82	0.42
	2	0.36	0.46	0.40
	3	0.21	0.38	0.27
	4	0.04	0.19	0.06

Extraction of localization relations between bacteria and habitats.

Participant	Rank	Recall	Precision	F1
	1	0.12	0.18	0.14
	2	0.04	0.12	0.06
Structured SVM		0.31	0.17	0.22

Bacteria habitat categorization through the MBTO-Habitat ontology.

http://2013.bionlp-st.org/BioNLP

[Kordjamshidi et al. BMC Bioinformatics 2015]

Qualitative representations

- When an autonomous system has to naturally, fast and correctly interact with its users about the shared context that they all observe, we still need to translate the visual context and the natural language to another symbolic language, which is error-prone [e.g., Kordjamshidi & Moens 2015], then perform some symbolic or qualitative reasoning, and finally translate the symbolic language to action controls in a real world physical space
- Over the years many symbolic representation languages that use a limited symbolic vocabulary were developed, many of which follow first-order logic as underlying knowledge representation formalism:
  - Yet another human language - albeit usually less complex -, and could be prone to ambiguity and redundancy [Ritter et al. 2006]
  - Their primary goal is to facilitate reasoning about the world, rather than taking action in it [Davis 1993]

Overview

- Qualitative versus quantitative understanding of spatial language
- Qualitative understanding of spatial language
- The role of imagery in quantitative understanding of spatial language
- In which spaces to reason ?

Quantitative 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 in the physical space that capture this knowledge and potentially make inferences with it?

Representations of spatial knowledge

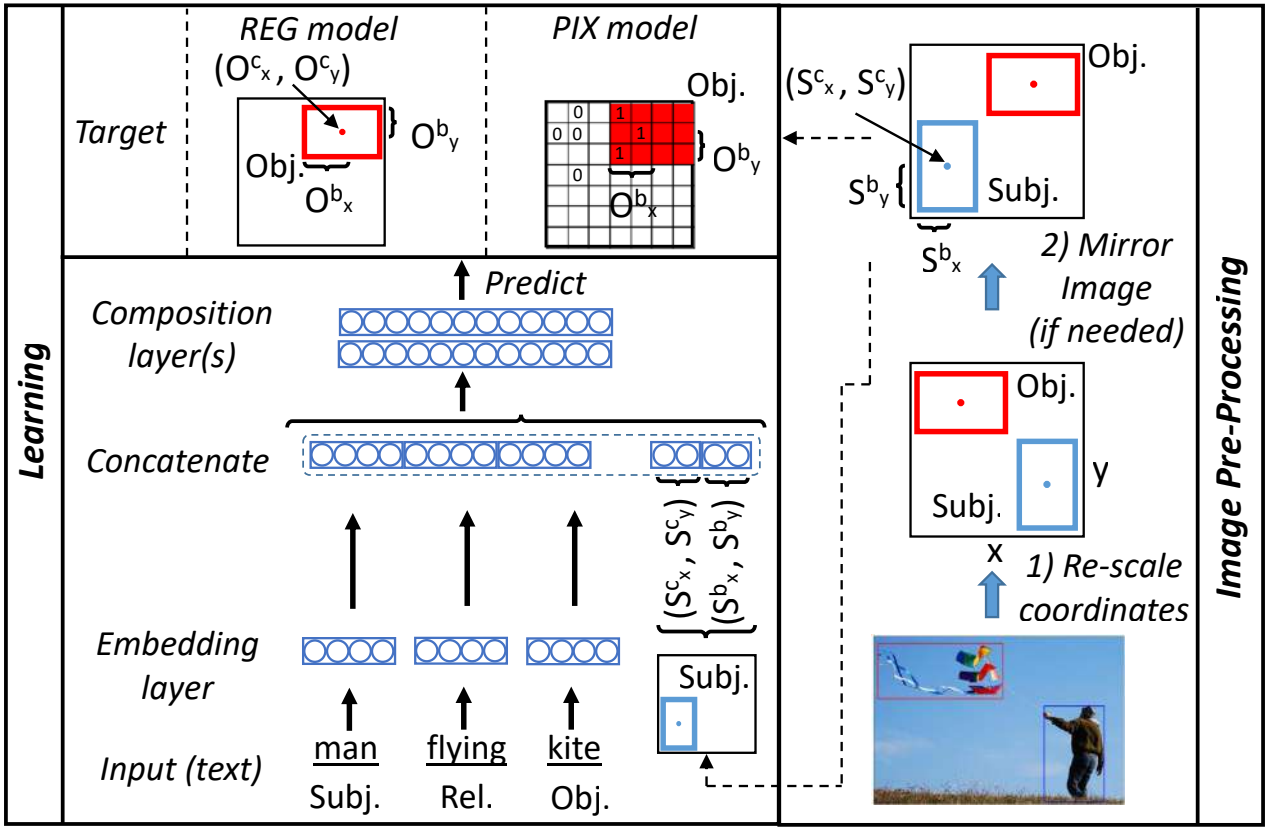
- Focus on spatial understanding of language and representing language with spatial templates = regions of acceptability of two objects under a spatial relationship
- Prior work restricts spatial templates to language that **explicitly** uses spatial cues (e.g., "glass *on* table") [Logan and Sadler *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 et al. *AAAI* 2018, Collell & Moens *TACL* 2018]
-

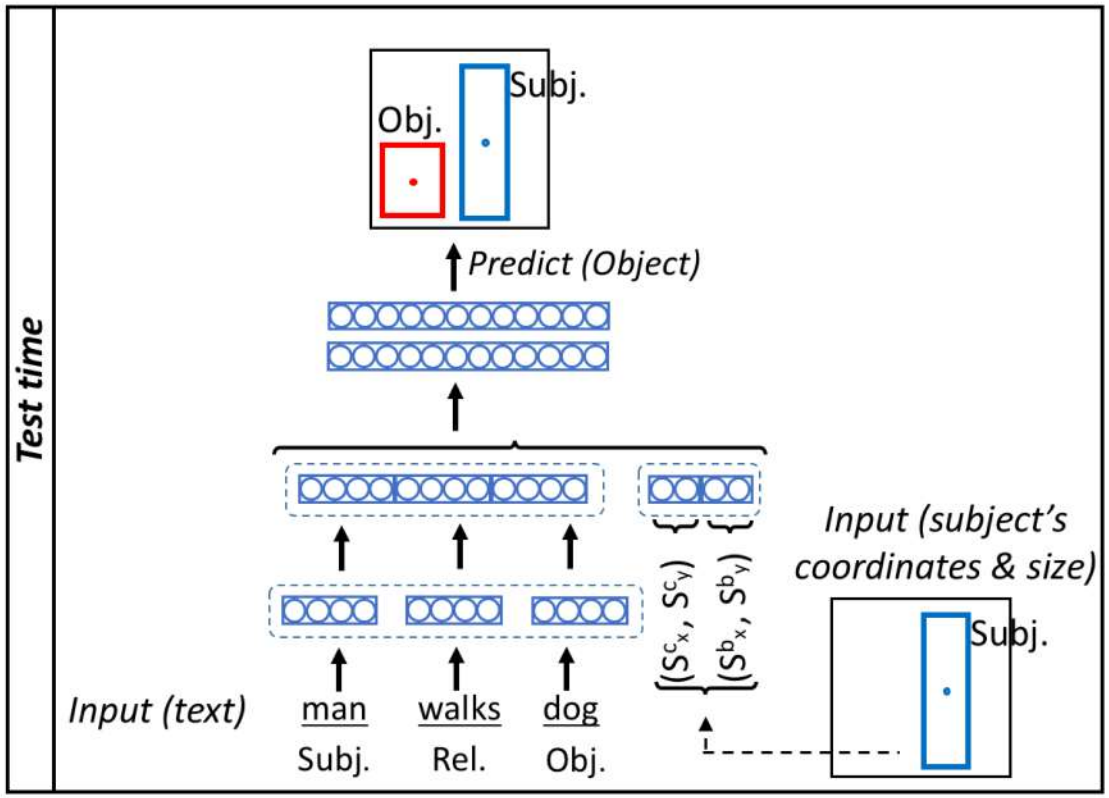
- We propose the task of:
  - Given a structured text input of the form (Subject, Relationship, Object) = (S,R,O)
  - Predict the 2D relative spatial arrangement of two objects (output)
- Train the task in a supervised setting:
  - Training set of image-text pairs, where the size and location of bounding 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]

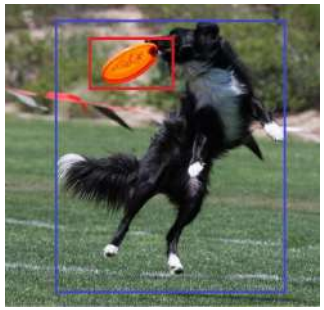




[Collell et al. *UCL Commonsense* 2017]

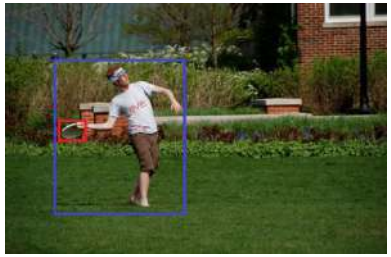
Experimental set-up

- 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

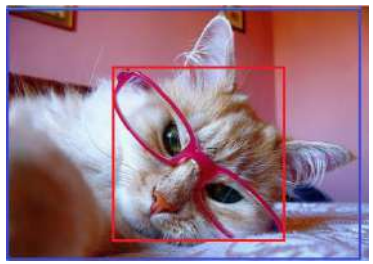


dog, catches, frisbee

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

		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]

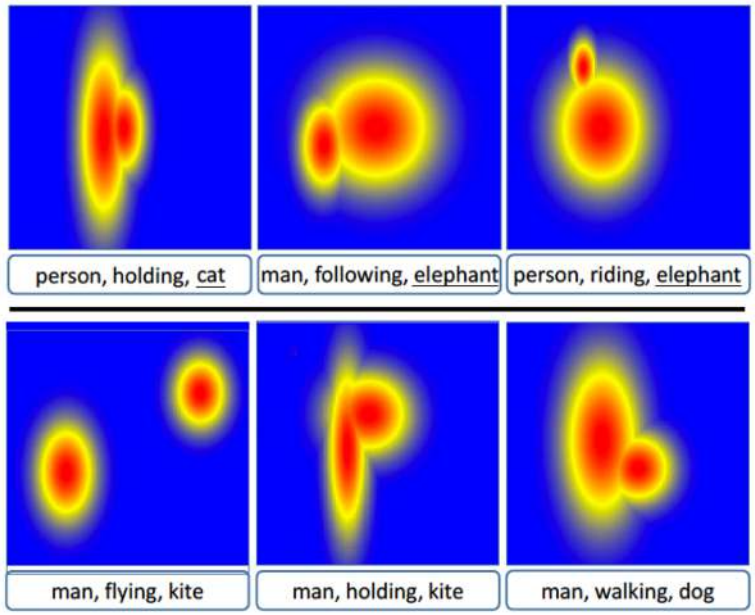
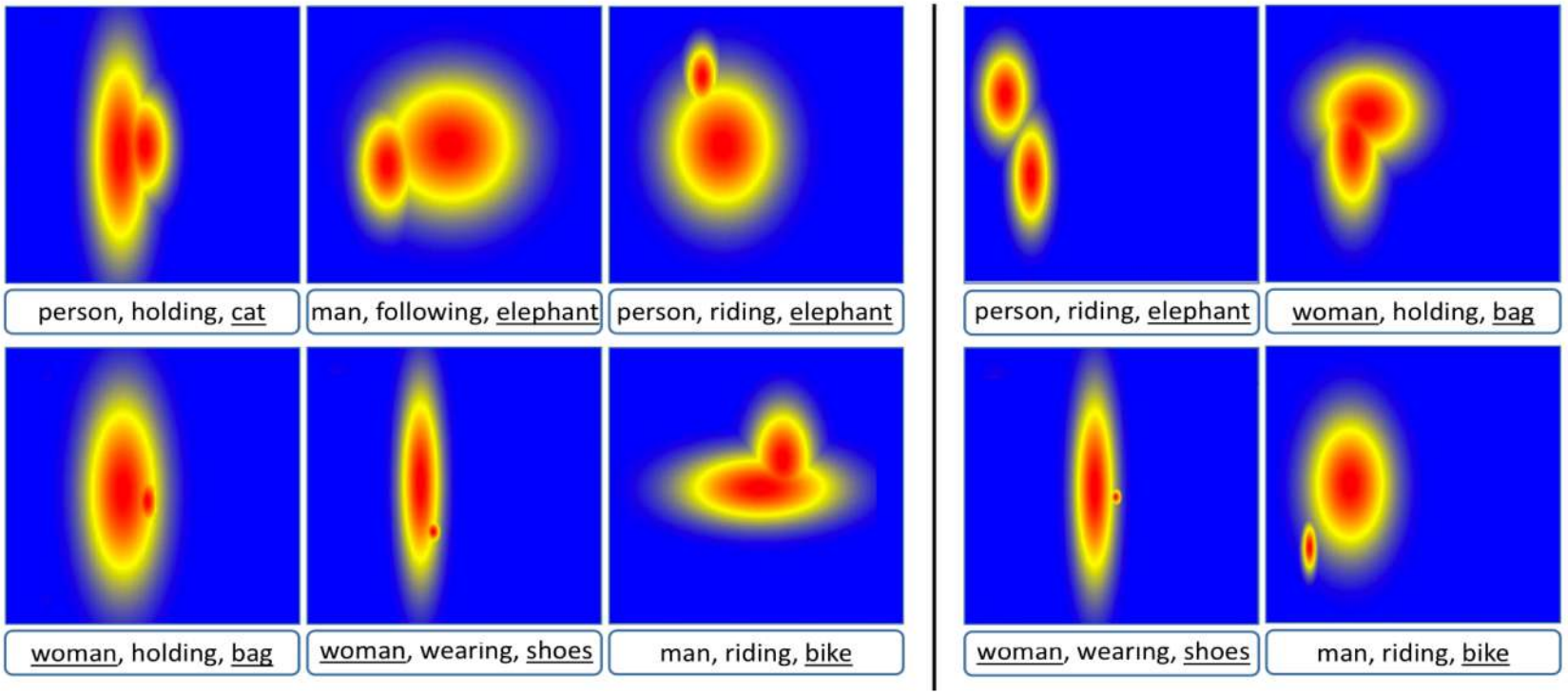


Figure 2: Predictions by the model that leverages word embeddings (*EMB*). **Top**: Predictions in unseen words (underlined). **Bottom:** Predictions in unseen *triplets*.

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

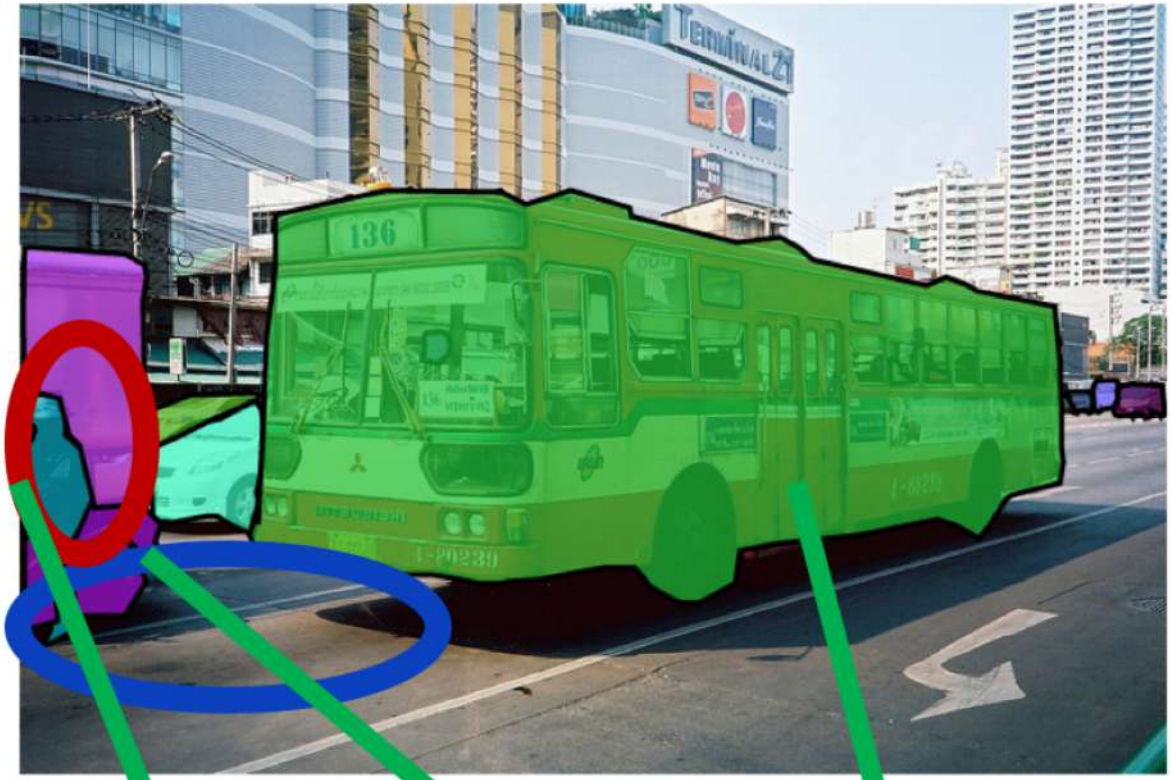
MACCHINA project (KU Leuven 2018-2022)

Many real life situations benefit from **communication in natural language with a machine about a shared visual environment**: e.g., conversation with a self-driving car

[nuScenes dataset]

Oh, I know him, the guy wearing the yellow hat. Stop next to him!

“Turn after the motorcycle” [nuScenes dataset]

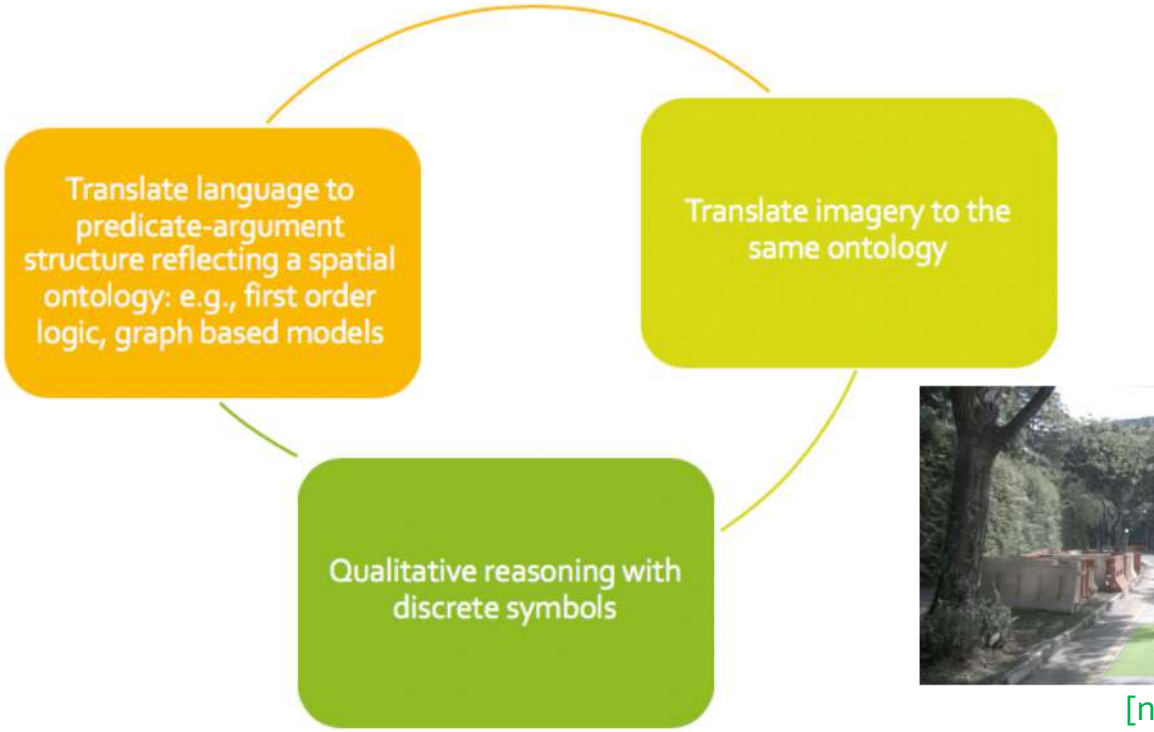


a man riding a motorcycle in front of an orange bus

Overview

- Qualitative versus quantitative understanding of spatial language
- Qualitative understanding of spatial language
- The role of imagery in quantitative understanding of spatial language
- In which spaces to reason ?

Reasoning in language space



[nuScenes dataset]

Reasoning in language space

- Reasoning in the language space is useful in processing human-human communications:
  - Retrieval of textual information based on textual query
  - In machine translation between languages
  - In machine translation between natural language and a programming language (e.g., SQL, programming code
  - Inference with abstract concepts
  - …

  - But not well suited for inference in a concrete physical world

Reasoning in physical space

- As seen above translate language to concrete coordinates 2D or 3D physical space
- Imagery can be easily mapped to same 2D or 3D physical space
- Quantitative reasoning with continuous values in Euclidean space with mathematical operators
- Probably faster and less error-prone

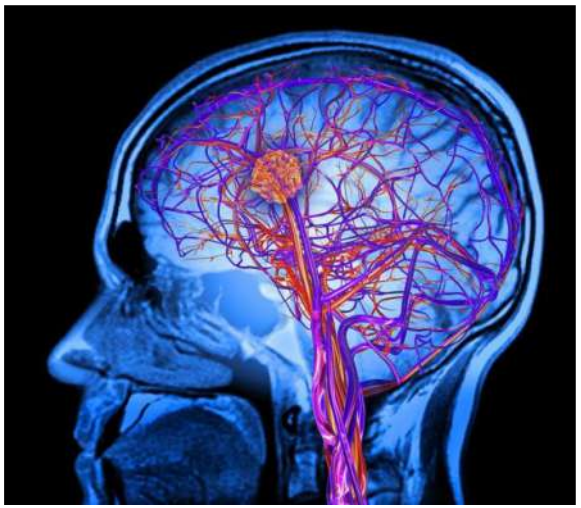
In line with Vapnik's principle to never solve a problem which is more general than the one that we are interested in when possessing a restricted amount of information for solving it [Vapnik *Statistical Learning Theory* 1998]

Reasoning in physical space

- Reasoning in the physical is useful in processing human-machine communications: e.g.,
  - Communications with robots and autonomous vehicles: inference of additional spatial information
  - In machine translation between natural language and a programming language (e.g., numeric arguments of commands)
- When to reason in the language space and when in the physical space is an interesting research question

Reasoning in representation space that mimics the human brain ?

- Representations that generate the mappings of language to 2D or 3D spaces contain the spatial information in a dense form
- Eventually quantitative reasoning with these ???
- Inspired by the human brain?
- Possibly computations in non-Euclidean geometric spaces ???



Conclusions

- We focused on spatial knowledge acquisition
- Both qualitative and quantitative calculi have a long tradition
- Learning to map natural language to 2D or 3D physical spaces is novel: the imagery jointly processed with language helps to accomplish this goal
- Still many questions unsolved, demanding future research

CALCULUS : ERC Advanced Grant, 2018-2023

Commonsense and Anticipation enriched Learning of Continuous representations sUpporting Language UnderStanding

MACCHINA project (KU Leuven 2018-2022)

Some references

- 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. & 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., 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.
- Kordjamshidi, P. & Moens, M.-F. (2015). Global Machine Learning for Ontology Population. *Journal of Web Semantics,* 30, 3-21.
- Kordjamshidi, P., Roth, D. & Moens, M.-F. (2015). Structured Learning for Spatial Information Extraction from Biomedical Text: Bacteria Biotopes. *BMC Bioinformatics* 16: 129.

Questions?