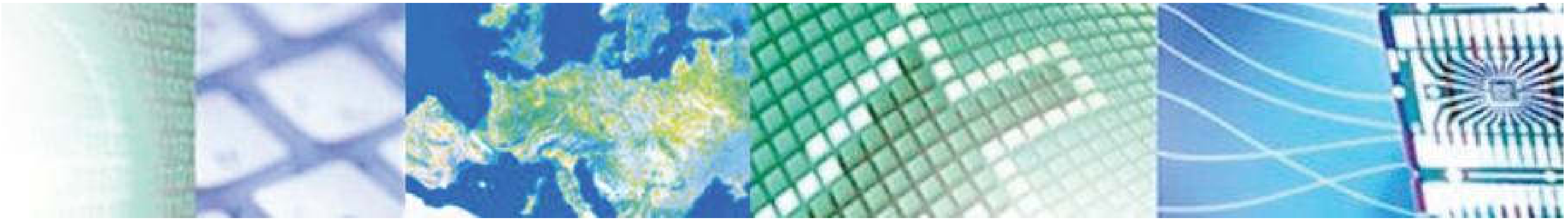


CHIST-ERA Projects Seminar
Topic Human Language Understanding – Grounded Language Learning

Katrien Beuls

Paris, April 12, 2018

FUNDING OPPORTUNITIES from the
FUTURE & EMERGING TECHNOLOGIES scheme

Introduction: Topic description

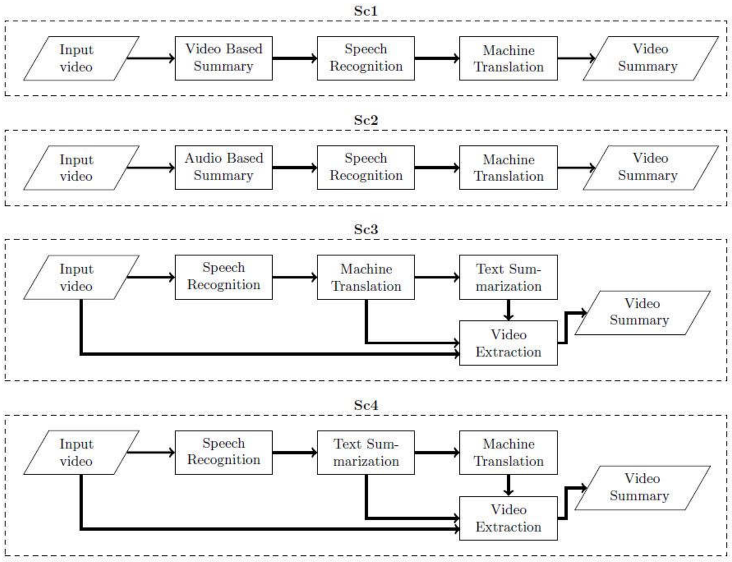
- **The goal:**
  - ✓ Ground language learning in the perceptual, emotional and sensorimotor experience of the system
- **Why:**
  - ✓ To model high-level, semantic & pragmatic knowledge in a robust way, from varied data, considering situational context
- **How:**
  - ✓ Multidisciplinary approach: combine human language processing with related fields such as developmental robotics or cognitive science.
  - ✓ Evaluation
    - **Well defined metrics and protocols to measure progress**

Introduction: 6 Projects of the topic

- **AMIS** (AGH Poland, DEUSTO Spain, LIA France, LORIA France)
- **ATLANTIS** (VUB Belgium, LATTICE France, OFAI Austria, SONY France, UPF Spain)
- **IGLU** (UMONS Belgium, Lille1 & INRIA Bordeaux France, UNIZAR Spain, KTH Sweden, MILA Umontréal & Usherbrooke Québec)
- **M2CR (**CVC Barcelona Spain, LIUM Le Mans France, MILA UMontréal Québec)
- **MUSTER** (ETH Zurich (CH), KU Leuven (BE), University of the Basque Country (SP), Sorbonne Universite (FR))
- **ReGROUND** (KU Leuven Belgium, Koç University Turkey, Örebro University Sweden)

Access Multilingual Information opinionS (AMIS)

- Partners: **LORIA** (France), AGH (Poland), DEUSTO (Spain), LIA (France)
- Challenge:
  - ✓ Understanding a foreign video by summarizing



Sc1: Input video → Video Based Summary → Speech Recognition → Machine Translation → Video Summary

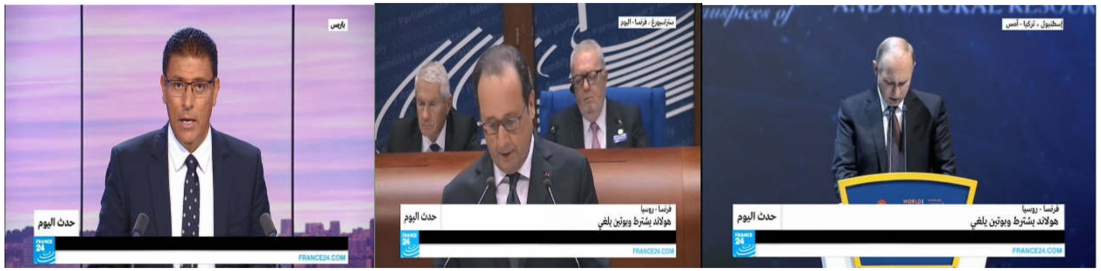
Sc2: Input video → Audio Based Summary → Speech Recognition → Machine Translation → Video Summary

Sc3: Input video → Speech Recognition → Machine Translation → Text Summarization → Video Extraction → Video Summary

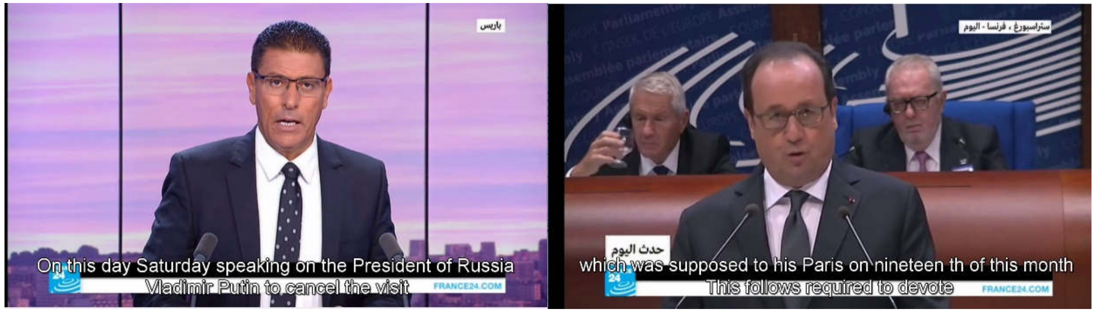
Sc4: Input video → Speech Recognition → Text Summarization → Machine Translation → Video Extraction → Video Summary

Different Architectures for AMIS

Arabic Source Video



A summarized Video subtitled in English



On this day Saturday speaking on the President of Russia Vladimir Putin to cancel the visit

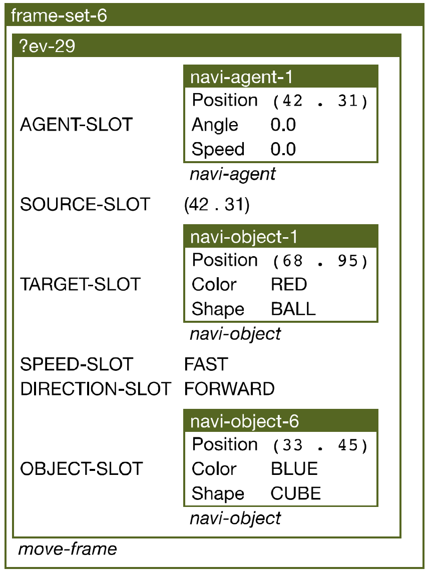
which was supposed to his Paris on nineteen th of this month This follows required to devote

Artificial language understanding in robots (ATLANTIS)

Synthesize the major transitions in the emergence of languages using agent-based computational models

- **Object reference scenario**: draw attention to objects and/or their properties and spatial relations
- **Spatial reference scenario**: describe a particular path of movement

Execute motion frames



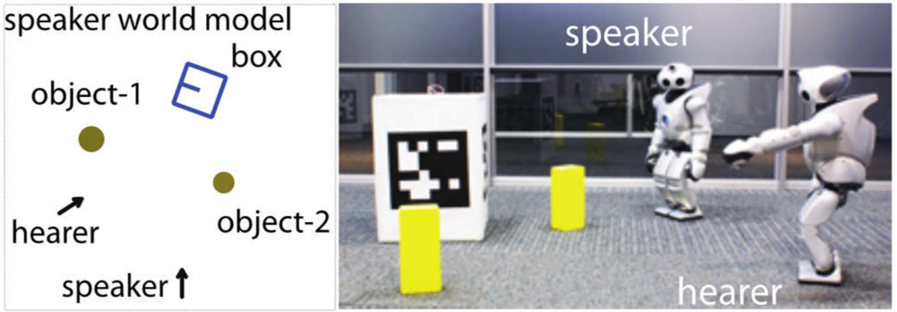
frame-set-6

?ev-29

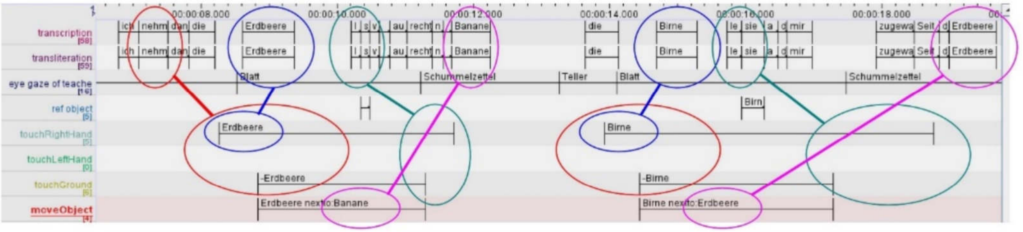
AGENT-SLOT	navi-agent-1 Position (42 . 31) Angle 0.0 Speed 0.0 *navi-agent*
SOURCE-SLOT	(42 . 31)
TARGET-SLOT	navi-object-1 Position (68 . 95) Color RED Shape BALL *navi-object*
SPEED-SLOT	FAST
DIRECTION-SLOT	FORWARD
OBJECT-SLOT	navi-object-6 Position (33 . 45) Color BLUE Shape CUBE *navi-object*

move-frame

Grounded language games



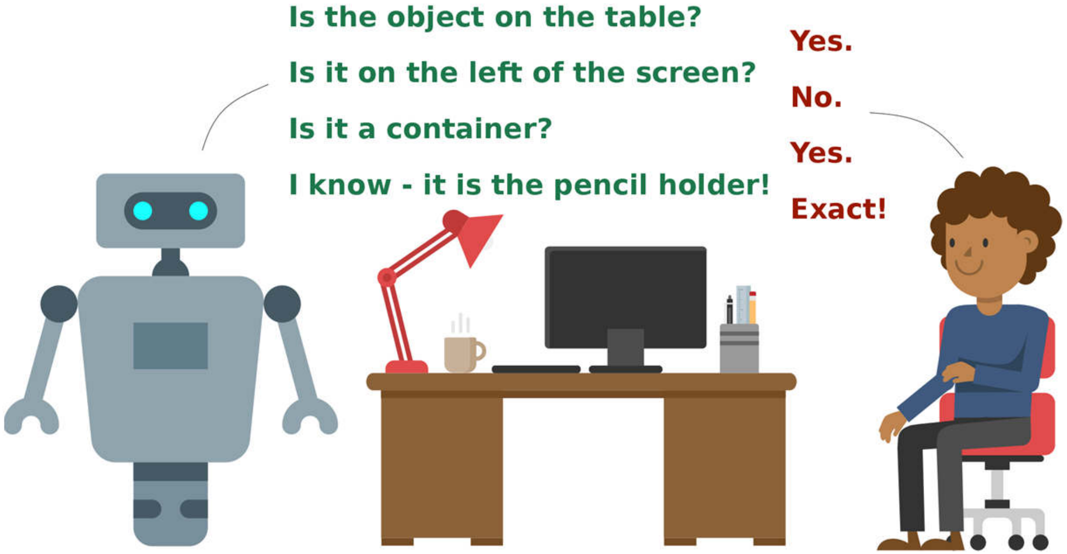
Multi-modal task description corpus



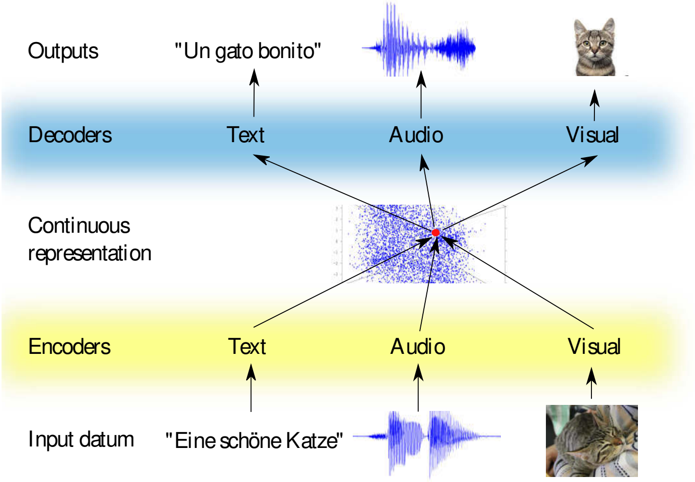
Interactive Grounded Language Understanding (IGLU)

Language learning and grounding through dialogue and interaction in multimodal environments.

- Goal-oriented visual and dialogue tasks (GuessWhat?!).
- Evaluation frameworks of language-learning cognitive agents for dialogue (HoME – 3D multimodal simulator) and incremental learning (Multimodal Human Robot Interaction dataset).
- Integration of developed algorithms on multiple humanoïd robotic platforms.


Is the object on the table?
Yes.
Is it on the left of the screen?
No.
Is it a container?
Yes.
I know - it is the pencil holder!
Exact!

M2CR: Multimodal Multilingual Continuous Representations for HLU

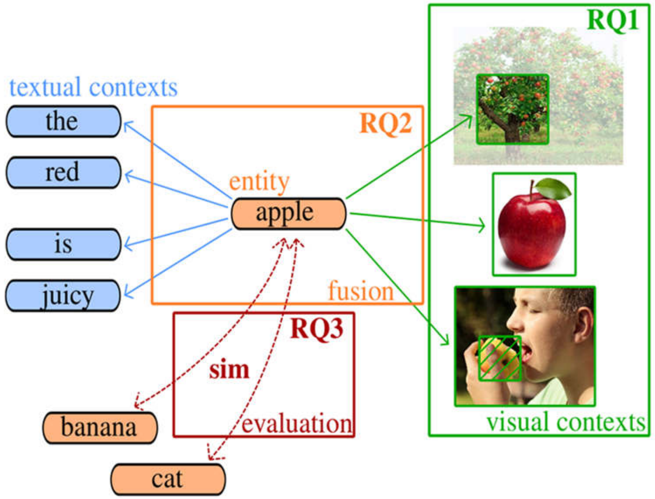


- **Goal**
  - ✓ Design a unified deep architecture
  - ✓ Address major HLU tasks
  - ✓ Multiple languages and modalities
- **Achievements:**
  - ✓ Pure neural MT, ASR and SLU systems
  - ✓ Image to image translation (sharing encoders and decoders)
  - ✓ Multimodal Machine Translation and Image description
- **Ongoing:**
  - ✓ Multi-task systems using multiple modalities
- **Next:**
  - ✓ Corpus targeting specific (linguistic or visual) aspects (e.g. gender agreement)
  - ✓ Integrate encoders and decoders into a single NN
- **Partners:** CVC (Barcelona, Spain), LIUM (Le Mans, France), MILA (Montreal, Québec)

MUSTER

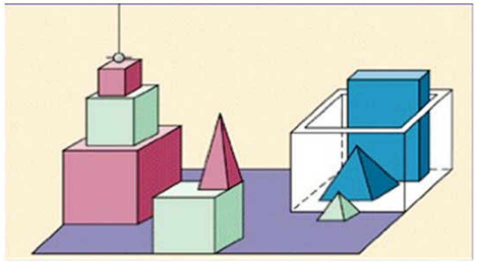
KU Leuven (Be), ETH Zurich (Ch), SU – Paris (Fr), U. Basque Contry (Spain)

- **MUSTER – Mu**ltimodal processing of **S**patial and **T**emporal Exp**R**essions
  - ✓ Multi-modal embeddings for text (word & sentence level)
  - ✓ Understanding & evaluation for various HLU tasks
- **Main results so far**
  - ✓ Multimodal word representations leveraging images (context, appearance, spatial information)
  - ✓ Multimodal tasks (e.g. visual sentence similarity, query-biased video summary)
  - ✓ Study of the properties of multimodal representations
- **Valorisation**
  - ✓ 13 publications
  - ✓ 4 Datasets produced for evaluating the quality of representations
  - ✓ Tools (annotations, benchmarks, and models)

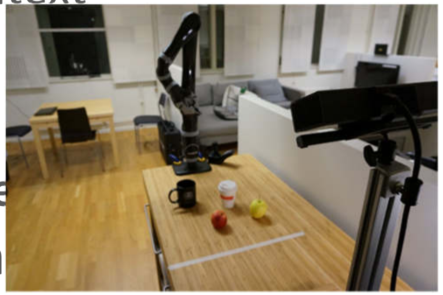


Relational Symbol Grounding through Affordance learning (ReGround)

- **Main ideas**
  - ✓ Associate symbols in language with referents in an environment
  - ✓ Goal: From Winograd’s SHRDLU to the real world, here kitchen environment
- **Distinctive features**
  - ✓ multi-modal input (perception and language)
  - ✓ take into account the context & environment; multiple objects and their relationships
  - ✓ build on a notion of affordance from robotics
    - potential actions in objects in the environmental context
- **Results so far**
  - ✓ Anchoring + relational affordances
    - **Link perception of physical world to object propertie**
  - ✓ Identification of position in NLP (symbol groundin



Put the blue pyramid on the block in the box



Bring me the tea pot and the sugar

- **Partners:** KU Leuven (Belgium), Koç University (Turkey), Örebro University (Sweden)

Introduction: Projects of the topic

	Multi-modal	Multi-lingual	Physical Robots	Actions	Relations	Open Data	Systematic Evaluation
AMIS	✓	✓				✓	✓
Atlantis	✓		✓	✓	✓	✓	✓
IGLU	✓		✓	✓	✓	✓	✓
M2CR	✓	✓				✓	✓
Muster	✓			✓	✓	✓	✓
ReGround	✓		✓	✓	✓	✓	✓

What we learned so far

- At the beginning we were looking to very wide solutions and approaches but realized that we had to focus on more specific goal oriented tasks
- CHIST-ERA HLU created a new research community [HLU-master class 10-11 April 2018, ~ 10 workshops + others planned]
- 3 years is short we would like to find a way to keep the HLU community alive.

Produced databases

- AMIS: Video database, 3 languages, 300 hours (100 per langage)
- ATLANTIS: Manual annotation of multimodal task description
- IGLU: 3 databases and 1 3D multimodal simulator
- M2CR: 1 multilingual, multi-modal (image and text descriptions in 4 languages)
- MUSTER: Dataset on spatial similarity for word pairs, Web-image dataset (2.5 M images), visual Word Sense Disambiguation, Visual semantic textual similarity
- ReGROUND: 2 artificial data generators for instruction following (infinite)

Upcoming challenges and needs (1/3)

- **Scientific challenges & needs**
  - ✓ How to combine low level neural approaches with higher level reasoning?
  - ✓ Many current techniques need large amounts of data: how to address this challenge?
    - **investigate techniques that only need few data ...**
      - e.g: unsupervised / weakly supervised learning
  - ✓ How to improve the transfer between modalities across different contexts?
  - ✓ Evaluation is always a challenge

chist-era

Upcoming challenges and needs (2/3)

- **Scientific challenges & needs**
  - ✓ How to connect data to actions?
    - **use data for action**
      - e.g: anticipation, planning
    - **use action for data acquisition**
      - e.g: sensor planning, perception focus
  - ✓ How to evaluate an intelligent interactive system in real situations?
    - **what are the performance metrics?**
    - **how can we define benchmarks?**

chist-era

Upcoming challenges and needs (3/3)

❖ **Organizational challenges & needs – provide resources for:**

- ✓ Data
  - **create bodies of annotated data**
- ✓ Evaluation
  - **shared test facilities, standard challenges, evaluation campaigns**
- ✓ Common platform (hardware/software)
  - **affordable, maintenable**
- ✓ Exchange data, software components, knowledge
  - **Make public the outputs of the inititiative**
    - Workshops, summer school, …
    - HLU Website listing the facilities developed by the projects
      - **e.g. data sharing, platforms produced by the partners**

chist-era

Remaining challenges and needs - SWOT

- **Strengths**
  - ✓ Large multimodal datasets are starting to be available usually developed by communities or large groups
  - ✓ Deep Learning as a common framework for different modalities is convenient
- **Weaknesses**
  - ✓ Deep Learning is not enough
  - ✓ Relevant multi-modal data not yet always available
  - ✓ Copyright and distribution of corpora
- **Opportunities**
  - ✓ Needs from industry – hopefully
  - ✓ Common practices, component and data reuse, availability of data, software, computation
  - ✓ Common representations for multi-modal phenomena
  - ✓ Better perception and actuation hardware (robots)
- **Threats**
  - ✓ Ethics – use of technology by companies
  - ✓ Research driven by some companies

chist-era

Remaining challenges and needs

- **Challenges**
  - ✓ **Combining neural implementations and high-level reasoning**
  - ✓ Transfer knowledge from different contexts
  - ✓ Incremental learning with limited data
  - ✓ Integrating different approaches
- **In progress**
  - ✓ Evaluation of the different systems that have been built in the projects

Questions

Questions ?