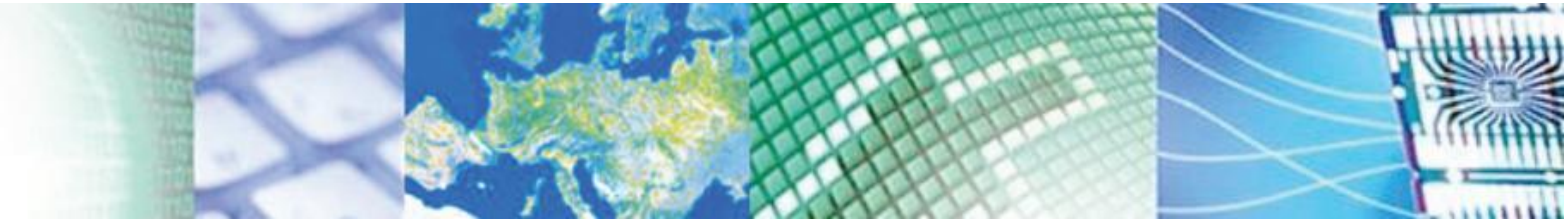


chist-era

CHIST-ERA Projects Seminar 2019
Call: HLU
Human Language Understanding

Presenter: Benjamin Piwowarski (CNRS/LIP6, Sorbonne Université)
Slides: prepared by Stephen McGregor (LATTICE, CNRS, France)

Bucharest, April 4, 2019

Programme co-funded by the
EUROPEAN UNION

Introduction of the Topic

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 and cognitive science.

- ❑ **Evaluation**

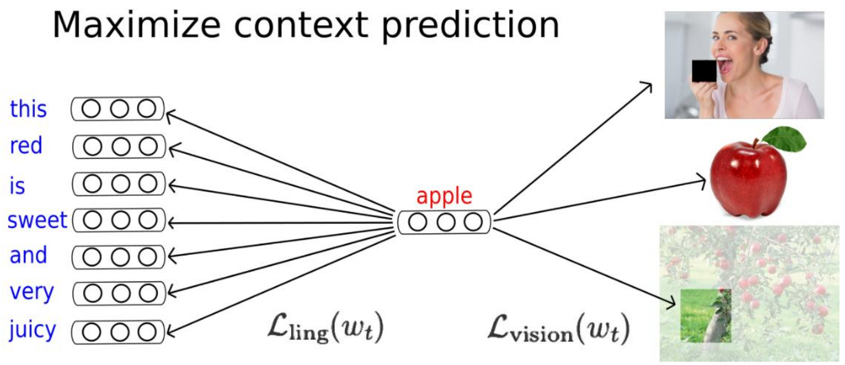
  Well defined metrics and protocols to measure progress.

chist-era

MUSTER

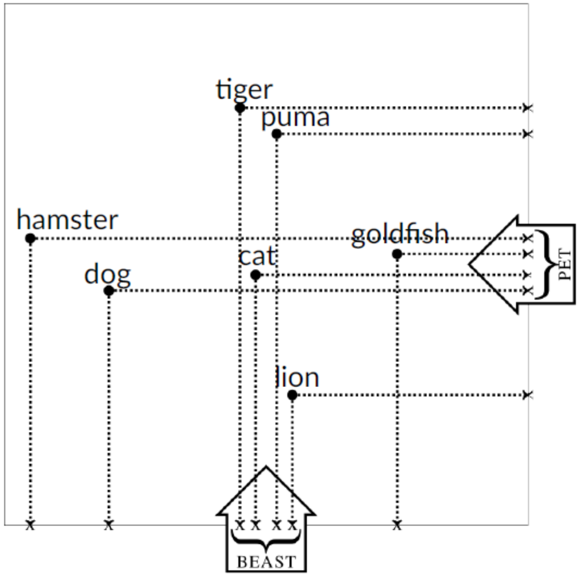
KU Leuven (Be), ETH Zurich (Ch), SU – Paris (Fr), U. Basque Country (Spain)

- □ **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 and sentence representations leveraging images (context, appearance, spatial information)
  - ✓ Multimodal tasks (e.g. visual sentence similarity, query-biased video summary, visual QA)
  - ✓ Study of the properties of multimodal representations
- □ **Valorization**
  - ✓ 22 publications
  - ✓ 4 Datasets produced for evaluating the quality of representations
  - ✓ Tools (dataset manager, annotations, benchmarks, and models)



Maximize context prediction
this
red
is
sweet
and
very
juicy
apple
$\mathcal{L}_{\text{ling}}(w_t)$
$\mathcal{L}_{\text{vision}}(w_t)$

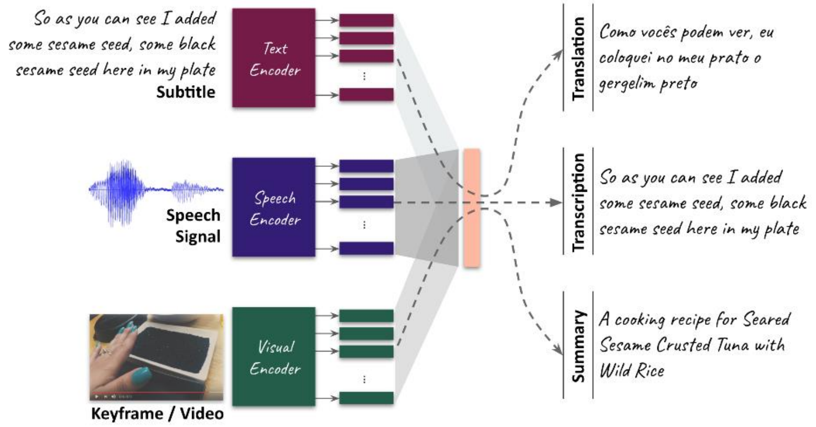
- **Our project has explored the way that agents acquire flexible, composable linguistic representations from the earliest stages of development.**
  - We have developed a framework for the context-specific projection of word meaning.
  - We have applied this framework to image classification tasks and modelling linguistic phenomena such as semantic type coercion.
  - We have gathered data on humans interacting with language learning robots and trained models to learn from this data.
  - We have run simulations of the way semantic representations can begin to emerge from interactions between basic agents without recourse to internal representations.



M2CR: Multimodal Multilingual Continuous Representations for HLU

❑ Goal

- ✓ Design a unified DL architecture
- ✓ Address major HLU tasks
- ✓ Multiple languages and modalities



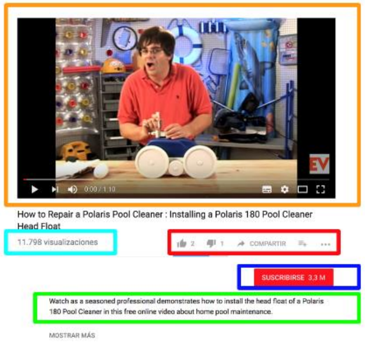
So as you can see I added some sesame seed, some black sesame seed here in my plate
Subtitle
Text Encoder
Speech Signal
Speech Encoder
Keyframe / Video
Visual Encoder
Translation
Como vocês podem ver, eu coloquei no meu prato o gergelim preto
Transcription
So as you can see I added some sesame seed, some black sesame seed here in my plate
Summary
A cooking recipe for Seared Sesame Crusted Tuna with Wild Rice

❑ Achievements:

- ✓ End-to end multimodal neural MT, ASR and SLU systems
- ✓ Image to image translation
- ✓ Multi-task learning with multiple modalities
- ✓ Open source datasets and toolkit: nmtpytorch

❖ **Partners:** CVC (Barcelona, Spain), LIUM (Le Mans, France), MILA (Montreal, Québec)

HOW2 dataset



How to Repair a Polaris Pool Cleaner : Installing a Polaris 180 Pool Cleaner Head Float

11.798 visualizaciones

2 1 COMPARTIR

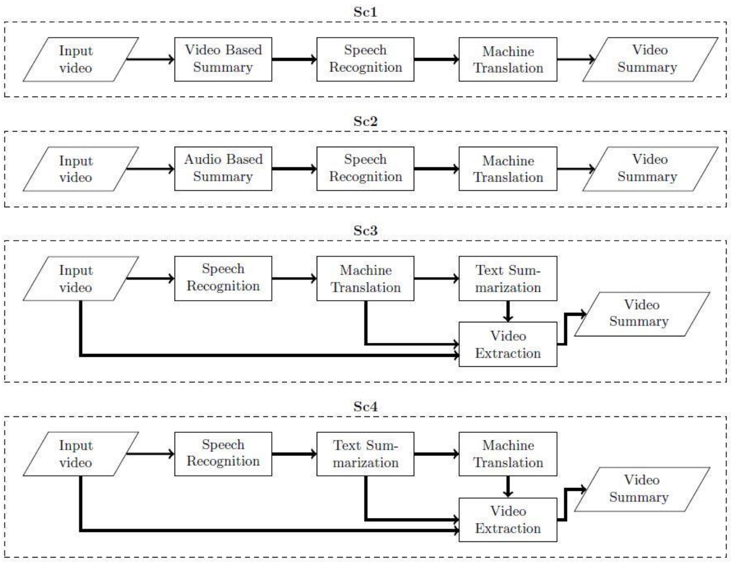
SUSCRIBIRSE 3,3 M

Watch as a seasoned professional demonstrates how to install the head float of a Polaris 180 Pool Cleaner in this free online video about home pool maintenance.

MOSTRAR MÁS

AMIS: Access Multilingual Information opinionS

- 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

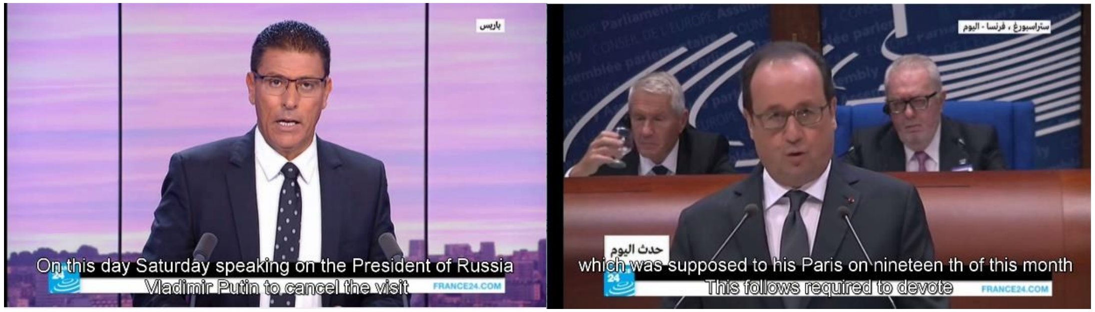
Sc4

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

Different Architectures for AMIS

Arabic Source Video

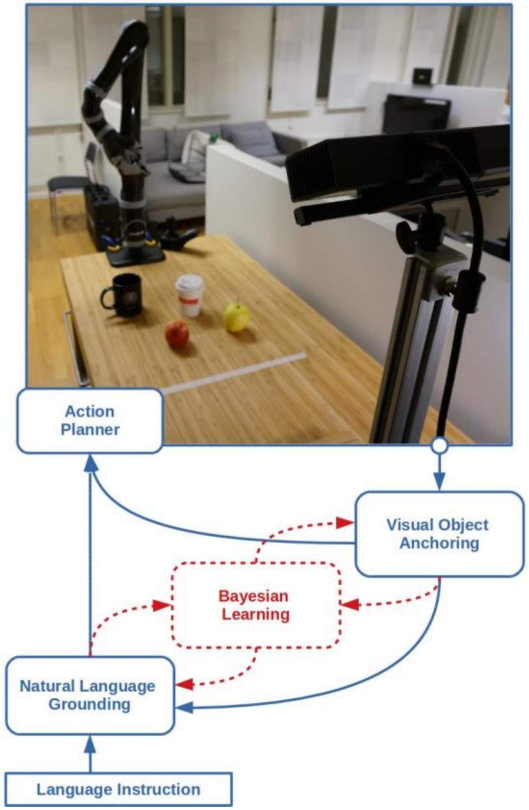
A summarized Video subtitled in English


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

ReGROUND: Relational Symbol Grounding through Affordance Learning

- ❑ **Main ideas:**
  - ❑ Associate symbols in language with referents in an environment
  - ❑ From Winograd’s SHRDLU to the real world
- ❑ **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
- ❑ **Results (so far):**
  - ❑ Anchoring + Probabilistic Reasoning
  - ❑ Resolving Inconsistencies between Language
  - ❑ and Perception



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

Produced Datasets

- ❑ AMIS: Video database, 3 languages, 300 hours (100 per language)
- ❑ 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, visual Word Sense Disambiguation, Visual semantic textual similarity, How To instructions
- ❑ ReGROUND: 2 artificial data generators for instruction following (infinite)

Individual Achievements

- ❑ **AMIS** has created a system that can translate and summarize video from a source language to a target language.
- ❑ **ATLANTIS** has developed a framework for the representation of how meaning comes about in context and achieved positive results on experiments employing this framework.
- ❑ **REGROUND** has combined language grounding, object anchoring, and reasoning in a principled fashion through probability calculus.
- ❑ **MUSTER** has made advances in learning continuous multi-modal representations and studying their properties.
- ❑ **M2CR** created data and deep learning models to train systems for multi-modal and multi-lingual HLU tasks.

Major Achievements and Outputs

- ❑ **Last year, we expressed a desire to continue and extend collaboration on this project.**
  - ❑ We have organized seminars and workshops
  - ❑ Published open-source data & tools
- ❑ **Last year, we noted a need for additional time to accomplish our project objectives.**
  - ❑ We understand more than ever how ambitious the goals associated with grounded language learning are.

Outcomes from Last Year

- ❑ **How to model the transfer between modalities across different contexts:**
  - ❑ We have explored mapping between and combining data of various modalities, with positive results
    ex. for using the simulation of environmental affordance to perform mappings
- ❑ **How to evaluate system performance:**
  - ❑ Designing tasks where meaningful evaluation is possible
    ex. tasks with a tangible physical outcome.
  - ❑ Subjective evaluations of entire systems and programs.

Outcomes from Last Year

- ❑ **How to connect data to actions:**
  - ❑ We have designed experiments involving moving from sub-symbolic data to concrete actions in the world.
- ❑ **How to capture linguistic flexibility from the earliest stages of development:**
  - ❑ We have designed experiments in which semantic representations emerge from the physiognomy of simplistic language learning agents.

Topic Challenges and Needs

- ❑ **We've identified a number of specific topics that are relevant across multiple components within this project:**
  - ❑ Affordances in grounded language learning;
  - ❑ Embodiment and language learning agents;
  - ❑ Identifying and modelling potentially multi-modal context;
  - ❑ Designing 'the right task' for the question being asked;
  - ❑ Generalization from event-specific training—avoiding the learning of bias.

Role of the CHIST-ERA Support

- **Helpful features of CHIST-ERA**
  - The ability to gather a variety of researchers with different views on a single topic has been beneficial.
  - Periodic reporting and gatherings have facilitated exchanges of ideas within and across teams.
- **Things we might look for from CHIST-ERA in the future**
  - More opportunities for meetings with partners between the big annual events, particular smaller scale meetings between sub-groups within the project: could part of the core budget be directed toward this?

chist-era

Events Organised by Project Partners

- ❑ AMIS: Special session on Accessing Multilingual Information and Opinions (AMIS) at MISSI 2018 (https://missi.pwr.edu.pl/2018/).
- ❑ ATLANTIS: Symposium on Language Learning for Artificial Agents (L2A2) at AISB 2019 (www.l2a2.github.io/symposium)
- ❑ M2CR: JHU workshop << Grounded seq. To seq. Transduction>>
- ❑ M2CR: Multimodal Machine Translation at WMT 2016-2018 http://statmt.org/wmt18/multimodal-task.html
- ❑ M2CR: ICML Workshop: « The HOW2 challenge » https://srvk.github.io/how2-challenge/

Events Organised by Project Partners

- M2CR: IWSLT: Multimodal Spoken Language Translation (in preparation, to be announced)
- M2CR: Using the HOW2 dataset
- M2CR: Daghstul Seminar https://www.dagstuhl.de/no_cache/en/program/calendar/semhp/?semnr=19021
- Overall: HLU Mastercall https://chistera-hlu.sciencesconf.org/

Questions

Questions ?