Muster

Multimodal processing of Spatial and Temporal Expressions

Participants

- Belgium - KU Leuven — NLP
  - Marie-Francine Moens, Guile Collell Talleda
- Switzerland – ETH Zurich — Vision
  - Luc Van Gool
- France - University Pierre and Marie Curie- Paris — Machine Learning
  - Patrick Gallinari, Benjamin Piwowarski
- Spain – University of the Basque Country — NLP
  - Aitor Soroa, Eneko Agirre

Objectives

- Focus
  - Exploit visual and perceptual input (images and videos) coupled with textual modality for learning **multi-modal** semantic **representations for the recognition of objects and actions, and their spatial and temporal relations.**
  - Ground language in perception (visual inputs) and extract representations of meaning tied to the physical world
    - Language has been used to help image/video understanding e.g. visual annotations, sentence generation for images, video
    - The goal of the project is the opposite: use visual modality to help understand language
- Methodology
  - language **representation learning** based on text and visual modalities
- Output
  - New pilot framework for joint representation learning from text and vision data tailored for spatial and temporal language processing.
  - Evaluation on a series of semantic tasks (i.e., semantic textual similarity and disambiguation, spatial role labeling, zero-shot learning, temporal action ordering)

Objectives

- *Two fundamental questions*
  - *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 multi-modal representation construction**.

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

Background - Temporal and spatial information reasoning in NLP

- Active research area in AI for quite a time, more recent for NLP
- Temporal/ spatial reasoning:
  - Representation formalism
  - Extraction of information
  - Inference over extracted information
- SemEval evaluation campaign
  - TempEval
    - Cross Document Event Ordering
    - QA TempEval
    - Clinical TempEval
  - SpaceEval
  - Methodology
    - Most often statistical classifiers on top of linguistic handcrafted features
    - Well understood areas (entities), challenging tasks (relations)

SemEval Exemples

- QA TempEval

Texte	O'SMACH, Cambodia (AP)_ The top commander of a Cambodian resistance force said Thursday he has sent a team to recover the remains of a British mine removal expert kidnapped and presumed killed by Khmer Rouge guerrillas almost two years ago. Gen. Nhek Bunchhay, a loyalist of ousted Cambodian Prime Minister Prince Norodom Ranariddh, said in an interview with The Associated Press at his hilltop headquarters that he hopes to recover the remains of Christopher Howes within the next two weeks. Howes had been working for the Britain-based Mines Advisory Group when he was abducted with his Cambodian interpreter Houn Hourth in March 1996. There were many conflicting accounts of his fate...
Questions	IS ei48 AFTER ei47—Did Howes flee from Cambodia after the coup d'etat? (YES)—yes IS ei65 AFTER ei66—Was Howes killed after his capture? (YES)—yes IS ei83 AFTER ei81—Was the coup d'etat after Pol Pot was arrested? (YES)—yes

SpaceEval

1. [Arriving$_{m1}$] [in$_{ms1}$] the [town of Juanjui$_{pl1}$], near the [park$_{pl2}$], [I$_{se1}$] learned that my map had lied to me.

```
<MOTION id=m1 extent=''Arriving''
motion_type=PATH motion_class=REACH
motion_sense=LITERAL>
<MOTION_SIGNAL id=ms1 extent=''in''
motion_signal_type=PATH>
<PLACE id=pl1 extent=''town of
Juanjui'' form=NAM countable=TRUE
dimensionality=AREA>
<PLACE id=pl2 extent=''park'' form=NAM
countable=TRUE dimensionality=AREA>
<SPATIAL_ENTITY id=se1 extent=''I''
form=NOM countable=TRUE
dimensionality=VOLUME>
<MOVELINK id=mvl1 trigger=m1
goal=pl1 mover=se1 goal_reached=TRUE
motion_signalID=ms1>
```

Background - Temporal and spatial information reasoning in vision

- Object + relation recognition in images
- Modeling the temporal structure of videos, action description
- Example: Learning semantic relationships for better action retrieval in images (Ramanathan et al. 2015) Stanford-Google
  - Action recognition: large set of actions - > sparse supervision
  - Infer relations between actions to enrich supervision
  - Jointly learn action recognition and relationship extraction through embeddings

Figure 1. Given a query, such as "Person interacting with panda" (a) standard models for action recognition treat every action independently, while (b) our method identifies the relation between actions, and uses these relations to extrapolate labels for images of related actions. In this example, "person interacting with panda" is implied-by "person feeding panda", and mutually exclusive of "Person feeding a calf". Hence, the images of these actions could also be used to train a model for "person interacting with panda". The green and the red boxes indicate the positive and negative examples considered by the methods for training the model.

Background – representation learning for image and text

- Representation learning
  - Language models
  - Text + image for image annotation
  - Text + image for caption generation: Encoder – Decoder with CNN + recurrent NN
  - e.g. Neural image caption generator

  (Vinyals et al. 2015)

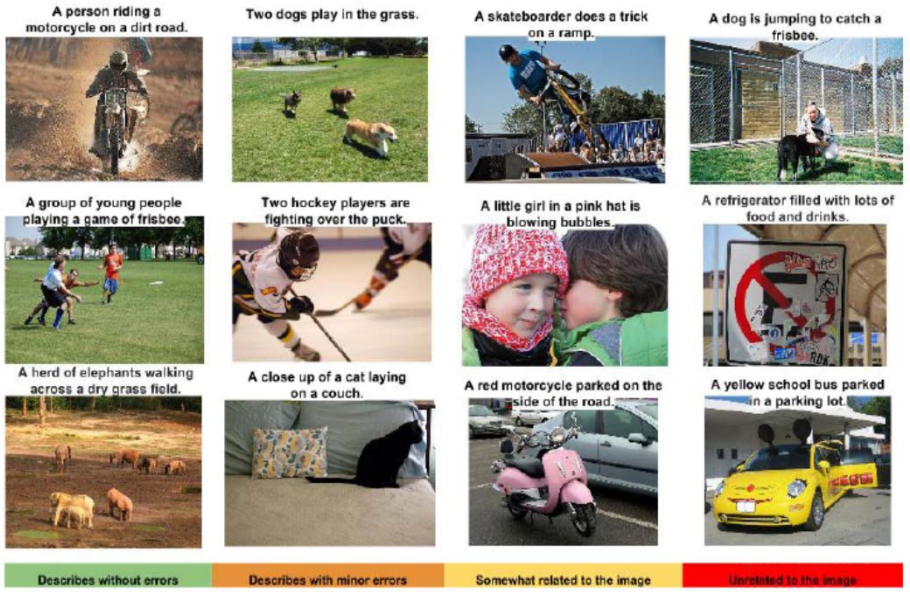

Figure 5. A selection of evaluation results, grouped by human rating.

Project objectives

- **Objective 1 : How to build improved structured semantic representations for nouns and verbs**, with an emphasis on verbs describing actions, motions, and nouns that are often used to refer to world objects, perpetrators of actions, locations, trajectors, landmarks, etc. Combine different spurces: text, KB
- **Objective 2 : How to build structured semantic representations targeting spatial language phenomena**, relying on aligned textual and visual input in the form of static images with textual descriptions.
- **Objective 3 : How to build structured semantic representations targeting temporal language phenomena**, relying on aligned textual and visual input in the form of captioned and/or annotated videos.
- **Objective 4 : How to integrate improved multi-modal semantic representations into novel systems for processing of spatial and temporal language.**
- **Objective 5 : How to evaluate these novel multi-modal spatial and temporal processing systems in a spectrum of spatial and temporal HLU tasks.**

Impact

- Major innovation
  - extend semantic representations obtained from text by using contextual visual knowledge in spatial and temporal language processing.
- *Impact on human language understanding:*
  - introduce a **new paradigm in processing of linguistic phenomena related to objects, actions, space and time in language**.
- *Impact on machine learning for semantic processing*
  - **new data-driven algorithms and representations** for **deep semantic understanding of human language**.
- *Impact on multilingual modeling*
  - the methodology is data-driven and language-independent, and potentially applicable to other languages as well as in multilingual settings.