CHIST-ERA Projects Seminar
Day 2, Cross Topics

D2K

Presenter: Rob Gaiaauskas
Transcriber: Anselmo Peñas
Content: D2K Project Members

Bern, April 29th, 2016
What is Data 2 Knowledge?

- Give semantics to data
- Make data meaningful and useful/manipulable in higher level tasks
  - Make explicit hidden information
  - Find patterns, predictions
  - Useful for decision making, analysis and communication
  - Give structure, and link to other data and processes
  - Human in the loop as source, target or both
## Introduction: Projects of the topic

<table>
<thead>
<tr>
<th>Title</th>
<th>Kind of data / field</th>
<th>Problem addressed</th>
<th>Methods</th>
<th>Achievements / releases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Camomile</td>
<td>Audiovisual</td>
<td>Person identification in broadcast TV</td>
<td>Active Learning</td>
<td>Framework Evaluation benchmarks &amp; annotated data</td>
</tr>
<tr>
<td>Reframe</td>
<td>Agnostic to data type</td>
<td>Transfer / reuse models between changing contexts</td>
<td>Reframing of Machine Learning models</td>
<td>Methodology and integration of tools in a platform</td>
</tr>
<tr>
<td>uComp</td>
<td>Textual Social media</td>
<td>Extraction of factual and affective knowledge</td>
<td>Human computation. Natural Language Processing (NLP)</td>
<td>Multilingual content repository, annotated corpora, human computation engine, game applications</td>
</tr>
<tr>
<td>ViSen</td>
<td>Image &amp; Text</td>
<td>Natural description of visual data</td>
<td>Deep learning for joint NLP and computer vision models</td>
<td>Annotated data, evaluation benchmarks, prototypes</td>
</tr>
</tbody>
</table>
Upcoming challenges and needs

- Dealing with human-created information is challenging
  - Deep / complex representations
  - Ambiguous information
  - Implicit information
  - Noise / approximate information
  - Contradictory
  - Differing perspectives
  - Long tail problem
  - Cultural and social diversity
Upcoming challenges and needs

- Learning to correlate information from multiple sources and use one to interpret the other is an important challenge.
  - Within and across modes and languages

- Examples:
  - Eye gaze/gesture and spoken language
  - News reports on political events and financial market movements
  - Body sensor outputs and patient self-reporting of condition
  - Climate data and social media discussion
  - Multi-lingual accounts of the same events, topics, ...
  - ...

Upcoming challenges and needs

- **Learning to scale not only in amount of data but also in its growing complexity**
  - Big Data is more than map-reduce ...
  - Not Big data, but the large amount of dimensions involved in human activity

- **Security and privacy issues working with human data**
  - Social media
  - Image and videos
  - Biometrics
  - ...
Upcoming challenges and needs

- Moving beyond
  target = knowledge as fixed/pre-defined meaning representation
to
  target = knowledge as ability to interpret data appropriately in
context of use/with understanding of human goals and intentions
(aka **pragmatic knowledge**)
Possible roadmap

Roadmap from Hope to Lands End

…
Possible roadmap

- D2K – should learn roadmap (K) from natural data (D) – 😊
- There are many roadmaps ....
- Need to move more to unsupervised methods
  - To address the annotation bottleneck
- Need to move more to joint models
  - To address the multisource corelation/co-interpretation challenge
- Need to understand how to reuse models
  - On-line learning, beyond active-learning...
- Need to increase robustness of models across different domains
Possible roadmap

❖ **Goal?: Data 2 Actionable Intelligence**
  ✓ Produce systems able to generate sound interpretations from
    ▪ previously unseen data, including
      • human-generated data in intentional contexts
      • Data from multiple inter-relatable sources
    ▪ in unseen contexts

❖ **Mechanism?**
  ✓ Shared task challenge
  ✓ Beyond the Turing Test – embodied agent in real settings that behave indistinguishably from human agent?
  ✓ How to arrive at challenge that pushes field without being AI complete
Role of the CHIST-ERA support

❖ D2K is a broad area
  ✓ But some projects are close to each other
  ✓ How to strengthen interaction among projects?
    ▪ Requires previous action from CHIST-ERA organization
  ✓ Could help facilitate sharing of platforms and data

❖ Other calls related to D2K
  ✓ HLU, IUI
    ▪ How to share/get/exchange/evaluate results from previous related calls?
      • Shared tasks /benchmarks are a useful tools
  ✓ Thematic corners for posters and demos
Role of the CHIST-ERA support

- Promoting synergies within Data 2 Knowledge
  - Prior need for data
    - Need of data annotation
    - Need of annotated data delivery
    - Use of Open Data from e.g. public administrations
      - CHIST-ERA could select domains of Open Data
      - Sponsor annotation of data for benchmarking
      - Create calls to work in that domain with the same data (but diverse research topics and agendas)
    - DARPA model?
    - Other option is additional funds to shared tasks among projects selected inside the same call
Role of the CHIST-ERA support

▶ Challenges

✓ How to share? ELRA/ELDA?
✓ How to keep data and platforms alive?
Questions