Natural Language Processing (NLP) Group Overview

Andrei Popescu-Belis

www.idiap.ch/nlp

Idiap Research Institute – Innovation Day
September 1st, 2016
Objectives of the NLP group

• Semantic and discourse processing

APPLIED TO

• Statistical machine translation

AND TO

• Information indexing and retrieval, multimedia recommendation
Current NLP group members

• Senior researcher & EPFL lecturer
  
  **Andrei Popescu-Belis** (since 2007)
  (École Polytechnique 1995, PhD LIMSI Paris-Sud 1999)

• Postdocs
  
  **Quang Luong** (PhD U of Grenoble 2014)
  **Nikos Pappas** (PhD EPFL/Idiap 2016)

• PhD students
  
  **Xiao Pu** (MSc RWTH Aachen 2014)
  **Lesly Miculicich** (MSc U of Fribourg 2016)
Past members of the NLP group

- **Postdocs**
  - Najeh Hajlaoui (2011-2013)
  - Chidansh Bhatt (2012-2014)
  - Parvaz Mahdabi (2014-2016)

- **PhD students graduated**
  - Majid Yazdani (2009-2013)
  - Thomas Meyer (2010-2014)
  - Nikos Pappas (2012-2016)

- **Interns**
  - Catherine Gasnier  (MSc EPFL)
  - Quoc Anh Le  (MSc U of Namur)
  - Jeevanthi Liyanapathirana  (PhD student U of Copenhagen)
  - Lukas Matena  (MSc Brno U of T)
  - Kexing Li  (MSc EPFL)
  - Sharid Loaiciga  (PhD student UniGe)
Overview

- Machine Translation (MT)
- **Highlight**: exploiting text-level constraints

- Information Retrieval and Recommender Systems
- **Highlights**
  - just-in-time document retrieval
  - multimedia navigation
  - sentiment analysis of user comments
  - emotion-based recommender
Machine translation

How can semantic and discourse processing improve statistical machine translation?

– Goals
  • disambiguate discourse connectives, pass labels to MT
  • predict verb tense translation prior to MT
  • probabilistic co-reference and pronoun resolution for MT decoding and/or post-editing of nouns (X. Pu, L. Miculicich, N.Q. Luong)

– Projects: with the Universities of Geneva, Zurich, Utrecht, Edinburgh
  • COMTIS (SNSF Sinergia, 2010-13): Improving the coherence of machine translation output by modeling intersentential relations
  • MODERN (SNSF Sinergia, 2013-17): Modeling discourse entities and relations for coherent machine translation
  • SUMMA (EU, 2016-19): Scalable understanding of multilingual media
Discourse in MT: an example

- Improving the translation of certain lexical items that require knowledge of a wider context than state-of-the-art SMT models
Highlight: improvements of statistical machine translation (1/4)

- When automatically labeling EN discourse connectives
  - e.g. since | CAUSAL vs. since | TEMPORAL (Moses factored model)
  - especially when using in-domain data

<table>
<thead>
<tr>
<th>Languages</th>
<th>Test set</th>
<th>System</th>
<th>BLEU</th>
<th>Δ</th>
<th>p</th>
<th>ACT</th>
<th>Δ</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>EN/FR</td>
<td>nt2012</td>
<td>baseline</td>
<td>26.1</td>
<td></td>
<td></td>
<td>56.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>25.8</td>
<td>-0.3</td>
<td>**</td>
<td>57.68</td>
<td>1.40</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>nt2010</td>
<td>baseline</td>
<td>24.4</td>
<td></td>
<td></td>
<td>68.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>24.3</td>
<td>-0.1</td>
<td>**</td>
<td>68.60</td>
<td>0.48</td>
<td>*</td>
</tr>
<tr>
<td></td>
<td>nt2008+sy2009</td>
<td>baseline</td>
<td>28.9</td>
<td></td>
<td></td>
<td>61.36</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>29.2</td>
<td>0.3</td>
<td>*</td>
<td>60.94</td>
<td>-0.42</td>
<td>*</td>
</tr>
<tr>
<td>EN/DE</td>
<td>nt2012</td>
<td>baseline</td>
<td>11.8</td>
<td></td>
<td>n/s</td>
<td>62.28</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>11.8</td>
<td>0.0</td>
<td>n/s</td>
<td>65.08</td>
<td>2.80</td>
<td>**</td>
</tr>
<tr>
<td></td>
<td>nt2010</td>
<td>baseline</td>
<td>15.0</td>
<td></td>
<td></td>
<td>62.42</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>15.0</td>
<td>0.0</td>
<td>n/s</td>
<td>69.28</td>
<td>6.86</td>
<td>***</td>
</tr>
<tr>
<td></td>
<td>nt2008+sy2009</td>
<td>baseline</td>
<td>13.0</td>
<td></td>
<td></td>
<td>71.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>labeled connectives</td>
<td>13.1</td>
<td>0.1</td>
<td>n/s</td>
<td>70.30</td>
<td>-0.76</td>
<td>n/s</td>
</tr>
</tbody>
</table>
Highlight: improvements of statistical machine translation (2/4)

- When automatically predicting the **tense, aspect and mode** of EN/FR verb phrase translations

<table>
<thead>
<tr>
<th>French tense</th>
<th>System</th>
<th>TAM Incorrect ≠ ref</th>
<th>TAM Correct = ref</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imparfait</td>
<td>Baseline</td>
<td>82</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>Predicted</td>
<td>42</td>
<td>23</td>
</tr>
<tr>
<td></td>
<td>Oracle</td>
<td>13</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>121</td>
<td></td>
</tr>
</tbody>
</table>
Highlight: improvements of statistical machine translation (3/4)

- When automatically enforcing **consistent noun translations**
  - machine learning to decide if two occurrences of the same source noun must be translated with an identical noun, and if yes, which one
  - results on German-English and Chinese-English
    - closed up to 80% of the BLEU gap between baseline MT and oracle classifier
    - syntactic features more useful than semantic ones, esp. for ZH/EN

---

**Example 1**

*Source*: nach eingührung dieser politik [...] die politik auf dem gebiet der informationstechnik [...]  
*Reference*: once the policy is implemented [...] the information technology policy [...]  
*MT*: after introduction of policy [...] the politics in the area of information technology [...]  

**Example 2**

*Source*: 欺诈性旅行或身份 证件 系指有下列情形之一的任何旅行或身份 证件  
*Reference*: Fraudulent travel or identity document; shall mean any travel or identity document  
*MT*: 欺诈性 travel or identity papers. 系指 have under one condition; any travel, or identity document
Highlight: improvements of statistical machine translation (4/4)

• When the antecedent (i.e. the “referent”) of a pronoun is recognized and passed to the MT
  – combine pronoun resolution and MT
  – pronoun-aware decoder and language model

⇒ DEMO BY N.Q. LUONG
Outputs

- Launched ACL DiscoMT 2013 and EMNLP DiscoMT 2015 workshops
- First rank at the pronoun-focused translation task at DiscoMT 2015
- Publications
  - conference proceedings: AMTA, ACL student session, EAMT, LREC, SIGdial, WMT, CICLing, ACL/EACL workshops
  - journals: IEEE/ACM Trans. on Audio Speech and Language Processing, Computer Speech & Language; Dialogue & Discourse
- Data and software
  - Discourse connective annotation on Europarl
  - Parallel verb phrase annotation on Europarl
  - Directional corpora from Europarl (e.g., original EN)
  - ACT metric: accuracy of connective translation
  - Automatic discourse connective labeler
  - APT metric: accuracy of pronoun translation
Multimedia information recommendation

How can semantic analysis improve information retrieval and recommendation?

– Topics
  • learning semantic similarity in networked data
  • multimedia segmentation, recommendation, summarization
  • sentiment analysis for lecture recommendation (*N. Pappas*)

– Projects: with the Universities of Fribourg and Edinburgh,
  IBM Haifa, Klewel SA, Faveeo SA, Unono.net
  • coordination: AROLES (SNSF, 2012-14), REMUS (Hasler Foundation, 2014), PIRMIN (Armasuisse, 2011), HYBRID (TheArk)
Highlight: just-in-time document retrieval in meetings

- Build multiple topically-separated implicit queries from ASR output, then merge results

“The virtual secretary”
Diverse keyword extraction for building just-in-time implicit queries

• For a short dialogue fragment (e.g. 30-90 sec.)
• New algorithm that rewards
  – representativeness of keywords
  – diversity of the keyword set
• Evaluation by human subjects (Mechanical Turk)
  – results of new algorithm preferred over others

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Compared methods (m₁ vs. m₂)</th>
<th>Relevance (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>m₁</td>
</tr>
<tr>
<td>Fisher</td>
<td>D(.75) vs. TS</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>TS vs. WF</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>WF vs. D(.5)</td>
<td>95</td>
</tr>
<tr>
<td>AMI</td>
<td>D(.75) vs. TS</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>TS vs. WF</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>WF vs. D(.5)</td>
<td>78</td>
</tr>
</tbody>
</table>

+ robustness to ASR noise
Highlight: multimedia navigation based on segment recommendation

- Using the metadata from multimedia recordings enhanced with NLP and semantic similarity
Highlight: sentiment analysis of user comments for recommendation

- New collaborative filtering models (neighborhood) with
  - users’ explicit one-class ratings (favorites)
  - the sentiments of users’ comments (computed by a SotA classifier)
- Several combination models, including learned ones
- More comments improve average precision and recall at $N (1 \leq N \leq 50)$ on three datasets: TED lectures, Vimeo videos, Flickr pictures
Highlight: emotion-based recommender

Multi-aspect sentiment profiles from transcript or from comments
Outputs in multimedia retrieval and recommendation

• First rank at the Grand Challenge on Lecture Segmentation and Recommendation at ACM Multimedia 2013

• First rank on the Hyperlinking task at MediaEval 2013

• Publications
  – conferences: ACL, Coling, EMNLP, SIGIR, ACM Multimedia, ICMR, NLDB

• Data and software
  – TED metadata for recommendation tasks
  – Software for visiting probability over networks
  – AREX dataset: spontaneous queries over AMI meetings
  – Software for multiple instance learning of aspects
  – Software for diverse keyword extraction from conversations
Objectives for the next 4 years

• **Machine translation**
  – demonstrate merits of text-level features for MT
    • conclude on co-reference in MT: noun phrases, pronouns
    • find new methods to combine text-level knowledge sources
    • integrate text-level features into neural MT
  – federate community and continue shared tasks

• **Multimedia retrieval and recommendation**
  – break new ground on explicit document modeling using DNNs
  – apply sub-topic-level semantic models to big multilingual data
  – improve usability of prototypes, continue technology transfer
Thank you!

Merci!

Grazie!

Vielen Dank!

謝謝！

 شكرا! 

...