MODERN: Modeling Discourse Entities and Relations for Coherent Machine Translation

Andrei POPESCU-BELIS\textsuperscript{1}, Jacqueline EVERS-VERMEUL\textsuperscript{2}, Mark FISHEL\textsuperscript{3}, Cristina GRISOT\textsuperscript{2}, Martin GROEN\textsuperscript{1}, Jet HOEK\textsuperscript{1}, Sharid LOAGIC\textsuperscript{2}, Ngoc Quang LUONG\textsuperscript{1}, Laura MASCARELL\textsuperscript{1}, Thomas MEYER\textsuperscript{1}, Lesly MICULICICH\textsuperscript{1}, Jacques MOESCHLER\textsuperscript{3}, Xiao PU\textsuperscript{1}, Annette RIOS\textsuperscript{1}, Ted SANDERS\textsuperscript{1}, Martin VOLK\textsuperscript{1}, Sandrine ZUFFEREY\textsuperscript{2}

\textsuperscript{1} Idiap Research Institute, Martigny, Switzerland
\textsuperscript{2} University of Geneva, Department of Linguistics
\textsuperscript{3} University of Zürich, Institute of Computational Linguistics
\textsuperscript{4} Utrecht Institute of Linguistics, The Netherlands

**Efficient when working on the project**

 Motivation

- MT is efficient, has good coverage, is quite intelligible, but it always translates by sentence and uses local features → it does not propagate information across sentences or even clauses
- Such information is crucial for the correct and coherent translation of complex sentences or entire texts
- Referring information: noun phrases, pronouns
- Verbs: tense, mode, aspect
- Discourse relations, as signaled by discourse connectives
- Other features, not targeted here: style, register, politeness
- This information is not (yet) accurately captured or used by mainstream MT systems (SMT or RBMT)

Discourse connectives

- **Goal:** model the human translation of connectives, annotate parallel data, disambiguate source-side connectives automatically, use the result for SMT, and measure the improvement on EN > FR, DE, IT, AN
- **Cross-lingual studies of discourse connectives**
- **Automatic disambiguation and use for SMT with Moses**
- **Automatic selection of connectives to translate implicitly**
- Additional goal: model the implicature of discourse connectives, i.e. conveying a discourse relation without using an explicit discourse connective
- **Cross-lingual modeling**
- **Automatic alignment, disambiguation and MT of EN/FR verbs**
- **Noun phrases**
- Goal: improve pronoun translation by addressing translation divergencies, in particular, for EN > FR, the possible translations of it (as it, elle, es, ce, etc.) and they (as ils, elles, etc.) — participated in the DiSCoMT ’15 and WMT ’16 shared tasks
- **Modeling uncertainty on coreference resolution and definition of a coreference-aware decoder**
- **Defining a pronoun language model to re-rank SMT output**

1. **Linguistic analyses**
   - Cohesion markers for MT
   - Cross-lingual features for discourse connectives
2. **Corpus data and annotation**
   - Define tagset
   - Locate problematic examples
   - Execute annotation and deliver data
3. **Automatic labeling**
   - Build and test classifiers using surface features
4. **SMT of labeled texts**
   - Phrase-based SMT for labeled texts
   - Factorized SMT models using labels
5. **Evaluation**
   - Define metrics of coherence
   - Performance of past systems
   - Apply metrics

Method for collaboration: linguistics, NLP, and MT

Support and related initiatives

- Swiss National Science Foundation, Sinergia program, grant n. 147653
- FINN-Nansen, Swedens Research Council, NansenRF, SUBGRID.
- Continues COMITIS SNSF Sinergia project grant n. 127510
- Resources available at www.idiap.ch/dataset
- Contributed to the creation of the Discosmt workshops on Discourse and Machine Translation (in conjunction with ACL 2013 and EMNLP 2015)
- Involved in the TextLink EU COST network: Structuring Discourse in Multilingual European

Verb tenses

- **Goal:** model verb tense, mode and aspect; infer features for translation; study the EN > FR dityency of Simple Past translated e.g. as imperfect, present, or past tense, depending on context
- **Cross-language modeling of verb TMA:** linguistics and pragmatics
- **Annotation of narrativity, disambiguation and use for MT**
- Belis A. (2013). Annotation through translation spotting on discourse connectives
- Annotation of narrativity, disambiguation and use for MT
- **Automatic alignment, disambiguation and MT of EN/FR verbs**

Pronouns

- **Goal:** improve pronoun translation by addressing translation divergencies, in particular, for EN > FR, the possible translations of it (as it, elle, es, ce, etc.) and they (as ils, elles, etc.) — participated in the DiSCoMT ’15 and WMT ’16 shared tasks
- **Modeling uncertainty on coreference resolution and definition of a coreference-aware decoder**
- **Defining a pronoun language model to re-rank SMT output**

Noun phrases

- **Goal:** improve the consistency of noun phrase translation, in particular for repeated nouns, including compounds, for EN > FR, FR > EN
- **Defining coreference resolution and definition of a coreference-aware decoder**
- **Automatic alignment, disambiguation and MT of EN/FR verbs**

Example (two clauses)

<table>
<thead>
<tr>
<th>Connectives</th>
<th>French</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>since</td>
<td>a</td>
<td>since</td>
</tr>
<tr>
<td>has been</td>
<td>quatre</td>
<td>been</td>
</tr>
<tr>
<td>reduced</td>
<td>fois</td>
<td>four</td>
</tr>
<tr>
<td>too large</td>
<td></td>
<td>too</td>
</tr>
</tbody>
</table>

Current machine translation systems: red

Using longer-range dependencies: green