Using Discourse Structure as Textual Context for Statistical Machine Translation: the COMTIS Project

Andrei Popescu-Belis
Idiap Research Institute

September 11, 2012

META-NET Workshop: Machine Translation and Multimodal Contexts
Special session at ICANN 2012, Lausanne, Switzerland, Sept. 11-14, 2012
COMTIS: SNF Sinergia project

• “Improving the coherence of machine translation output by modeling intersentential relations”

• Idiap + two groups at the University of Geneva

• Three-year project: March 2010 – February 2013
  – likely extended (Aug. 2013) and continued (Aug. 2014)
### Motivation

<table>
<thead>
<tr>
<th>1. Connective</th>
<th>2. Pronoun</th>
<th>3. Verb Tense</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The matrix</strong></td>
<td><strong>has been</strong></td>
<td><strong>four times,</strong></td>
</tr>
<tr>
<td></td>
<td><strong>since</strong></td>
<td><strong>it</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>was</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>too large.</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1. Connective</th>
<th>2. Pronoun</th>
<th>3. Verb Tense</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>La matrice</strong></td>
<td><strong>a été</strong></td>
<td><strong>quatre fois,</strong></td>
</tr>
<tr>
<td></td>
<td><strong>depuis qu'</strong></td>
<td><strong>il</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>a été</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>trop grand.</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>1. Connective</th>
<th>2. Pronoun</th>
<th>3. Verb Tense</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><strong>car</strong></td>
<td><strong>elle</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>était</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>trop grande.</strong></td>
</tr>
</tbody>
</table>

Current machine translation systems: red

COMTIS considers the context of longer-range dependencies: translation in green
Main idea of the project

• MT can be improved by giving it information about context across sentences
  – text-level information or, rather, discourse-level
  – might involve intra-sentence dependencies as well

• How to use context
  – extract contextual features from source text
  – design MT systems that can use these features
How to proceed?

1. Define the contextual phenomena to target
   • linguistic analysis, relevance to MT, tractability

2. Create data for training and evaluation
   • also for corpus linguistics

3. Build classifiers for each phenomenon
   • not necessarily perfect, but useful for MT

4. Adapt MT systems to use classifiers’ output

5. Evaluate the improvement of MT
What to model and annotate

languages: EN, FR, DE, IT
What makes discourse discourse?

• “Text-level” exists because texts are generally coherent
  – coherence is ensured by cohesion markers

• What cohesion markers might help SMT in COMTIS?
  – tense: modeling (PhD at UniGe)
  – connectives: modeling + annotation + classifiers (PhD at Idiap)
  – pronouns: annotation + post-editing (intern/PhD at Idiap)
  – lexical choice: later
  – style / register: later

• Cohesion markers are long-range or intersentential
  – current SMT systems translate sentence-by-sentence
  – some commercial rule-based MT consider text-level domains
Examples (1)

Discourse connectives

- SOURCE: Why has no air quality test been done on this particular building since we were elected? (Europarl)
  - Ref: Comment se fait-il qu'aucun test de qualité de l'air n'ait été réalisé dans ce bâtiment depuis notre élection?
  - SMT: Pourquoi aucun test de qualité de l'air a été réalisé dans ce bâtiment car nous avons été élus ?

- SOURCE: While no-one wants to see public demonstration, I have to say I understand the anxiety and share their concern. (Europarl)
  - Ref: Alors que personne ne veut voir de manifestations publiques, je dois dire que je comprends leur anxiété et que je partage leur inquiétude.
  - SMT: Bien que personne ne veut voir la démonstration publique, je dois dire que je comprends l'inquiétude et de partager leurs préoccupations.
Examples (2)

- **Tense**
  - SOURCE: Grandmother drank three cups of coffee a day.
    - Ref: Grand-maman *buvait* trois tasses de café par jour.
    - SMT: Grand-mère *a bu* trois tasses de café par jour.
  - SOURCE: Je me lèве à cinq heures depuis 20 ans.
    - Ref: I *have been waking up* at five o’clock for the last 20 years.
    - SMT: I *get up* at five in the last 20 years.

- **Pronouns**
  - SOURCE: The European commission must make good these omissions as soon as possible. *It* must also cooperate with the Member States ...
    - SMT: *La commission européenne doit réparer ces omissions dès que possible. *Il* doit également coopérer avec les états membres ...
Some achievements from the first two years
Modeling verb tense

• How to label verb tenses to ensure that they are coherently translated?
  – depends on the language pair
  – must be tractable for NLP
    • existing linguistic theories of tense are complex
  – what features are useful to compute labels?

• A model for the translation of EN simple past into FR (mainly *passé simple* vs. *imparfait*) has been proposed and justified (theoretically and empirically)
  – pilot annotation of resources, more needed for training
Translating EN simple past

• Proposed label: ‘narrative’ vs. ‘non-narrative’
  – must be assigned globally, at the text level

• Proposed impact of label on MT
  – simple past ‘narrative’ \(\rightarrow\) passé simple (or composé)
  – simple past ‘non-narr.’ \(\rightarrow\) imparfait

• How to assign this label automatically?
  – we don’t know yet, but will look at training data

• This is really a simplified view
  – more labels, e.g. ‘subjective’ or not
  – more EN and FR tenses
Modeling and annotating discourse connectives

• Existing theories and annotated resources (mainly EN)
  – PDTB: complex hierarchy of possible senses of connectives
    • difficult to annotate, not necessarily relevant to MT

• In COMTIS, annotation through **translation spotting**
  – annotators only identify the human translation of each connective in a parallel corpus (Europarl)
  – for each connective type, observed translations are clustered into *a posteriori* “senses” relevant to MT
    • compact set of labels, cheaper to annotate
    • done for English/French, English/German/Italian, Arabic in progress

• Example
  – PDTB: *while* has 21 possible composite labels
  – COMTIS: *while* signals either a contrast, a concession, or has a temporal meaning (durative, temporal, or causal)
## Annotations of connectives in COMTIS

<table>
<thead>
<tr>
<th>Lexical items</th>
<th>A priori senses</th>
<th>A posteriori senses</th>
<th>N.S.</th>
<th>F.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>EN CONNECTIVE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>although</td>
<td>contrast, concession</td>
<td>contrast, concession</td>
<td>197</td>
<td>183</td>
</tr>
<tr>
<td>even though</td>
<td>contrast, concession</td>
<td>contrast, concession</td>
<td>212</td>
<td>191</td>
</tr>
<tr>
<td>since</td>
<td>temporal, causal</td>
<td>temporal, causal, causal_known_relation, causal_new_relation, causal_other</td>
<td>423</td>
<td>423</td>
</tr>
<tr>
<td>though</td>
<td>contrast, concession</td>
<td>contrast, concession</td>
<td>200</td>
<td>155</td>
</tr>
<tr>
<td>while</td>
<td>contrast, concession,</td>
<td>contrast, concession, contrast_and_concession, temporal, temporal_durative, temporal_punctual, temporal_causal</td>
<td>499</td>
<td>294</td>
</tr>
<tr>
<td></td>
<td>comparison, temporal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FR CONNECTIVE</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dans la mesure où</td>
<td>causal, explanation</td>
<td>causal_and_explanation</td>
<td>175</td>
<td>150</td>
</tr>
<tr>
<td>pourtant</td>
<td>contrast_and_concession</td>
<td>contrast_and_concession</td>
<td>312</td>
<td>250</td>
</tr>
<tr>
<td>alors que</td>
<td>contrast, temporal</td>
<td>contrast, temporal, temporal_and_con-trast</td>
<td>423</td>
<td>366</td>
</tr>
<tr>
<td>bien que</td>
<td>concession</td>
<td>concession_and_contrast</td>
<td>55</td>
<td>51</td>
</tr>
</tbody>
</table>

*Total: 1,246

*Total: 817*
Automatic labeling of connectives

• Classification problem: for each discourse connective
  – given features extracted from the text
  – determine its most probable label ("sense")
  – using MaxEnt, decision trees, etc.
  • trained on manually labeled data (PDTB or COMTIS)
  • tested on unseen data or plugged into an SMT system

• Features
  – standard
    • token, capitalization, POS tag, parent syntactic class, punctuation
    • first/last word/POS of previous/current clause
  – novel
    • similarity/antonymy for word pairs in the two clauses (WordNet)
    • features related to temporal relations (Tarsqi Toolkit)
    • candidate translation from a baseline SMT system
## Example of results on PDTB

<table>
<thead>
<tr>
<th>Connective</th>
<th>Number of occurrences and senses</th>
<th>F1 Score</th>
<th>F1 Score</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training set: total and per sense</td>
<td>PT</td>
<td>PT+</td>
</tr>
<tr>
<td>after</td>
<td>507 456 As, 51 As/Ca</td>
<td>0.66</td>
<td>1.00</td>
</tr>
<tr>
<td>although</td>
<td>267 135 Cs, 118 Ct, 14 Cp</td>
<td>0.60</td>
<td>0.66</td>
</tr>
<tr>
<td>however</td>
<td>176 121 Ct, 32 Cs, 23 Cp</td>
<td>0.33</td>
<td>1.00</td>
</tr>
<tr>
<td>indeed</td>
<td>69 37 Cd, 24 R, 3 Ca, 3 E, 2 I</td>
<td>*0.50</td>
<td>*0.50</td>
</tr>
<tr>
<td>meanwhile</td>
<td>117 66 Cj/S, 16 Cd, 16 S, 14 Ct/S, 5 Ct</td>
<td>0.32</td>
<td>0.53</td>
</tr>
<tr>
<td>nevertheless</td>
<td>26 15 Ct, 11 Cs</td>
<td>0.44</td>
<td>0.66</td>
</tr>
<tr>
<td>nonetheless</td>
<td>12 7 Cs, 3 Ct, 2 Cp</td>
<td>*1.00</td>
<td>*1.00</td>
</tr>
<tr>
<td>rather</td>
<td>10 6 R, 2 Al, 1 Ca, 1 Ct</td>
<td>*0.00</td>
<td>*0.00</td>
</tr>
<tr>
<td>since</td>
<td>166 75 As, 83 Ca, 8 As/Ca</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>still</td>
<td>114 56 Cs, 51 Ct, 7 Cp</td>
<td>0.60</td>
<td>0.66</td>
</tr>
<tr>
<td>then</td>
<td>145 136 As, 6 Cd, 3 As/Ca</td>
<td>0.83</td>
<td>1.00</td>
</tr>
<tr>
<td>while</td>
<td>631 317 Ct, 140 S, 79 Cs, 41 Ct/S, 36 Cd, 18 Cp</td>
<td>0.93</td>
<td>0.96</td>
</tr>
<tr>
<td>yet</td>
<td>80 46 Ct, 25 Cs, 9 Cp</td>
<td>*0.5</td>
<td>*1.00</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>2,320</strong></td>
<td><strong>0.57</strong></td>
<td><strong>0.75</strong></td>
</tr>
</tbody>
</table>

Integration with MT

• How to train SMT to use labeled connectives?
• Several methods have been studied
  – replace in the system’s phrase table all unambiguous occurrences of the connective with the labeled connective
  – train the system on manually or on automatically labeled data (e.g., while becomes while_Temporal)
  – combine contextual features into factored MT models
  – train over multiplied data in proportion to the label prob.

• Also: use a modified SMT system only when the connective labeler is confident enough
Sample results

• Modified phrase table
  – tested on ~10,000 instances of connectives (5 types)
  – 34% improved, 20% degraded, 46% unchanged [SAMPLE]

• Trained on manually labeled data
  – 26% improved, 8% degraded, 66% unchanged [SAMPLE]

• Trained on automatically labeled data
  – 18% improved, 14% degraded, 68% unchanged [SAMPLE]
  – smaller improvement, but cheaper and larger data

• Thresholding based on labeler’s confidence
  – experimented with two connectives until now
  – improvement of 0.2-0.4 BLEU points (quite significant)
What about global output quality?

• It depends on how we measure it
  – traditional BLEU metric: n-gram automatic comparison of a candidate text with one or more reference translations

• COMTIS
  – contextual factors are not frequently determinant
  – so impact on BLEU should be small (and it is)
  – goal is at least not to decrease BLEU scores

• Need for specific automatic metrics
  – still reference-based, but sensitive to sparse phenomena
Towards new evaluation metrics

• Goal: automatic procedure to count how many connectives were correctly translated

• *ACT metric: Accuracy of Connective Translation*
  – given a source sentence with a discourse connective \( C \)
  – use automatic alignment to find out:
    • how \( C \) is translated in the reference translation(s)
    • how \( C \) is translated in the candidate translation
  – compare the two translations of \( C \)
    • identical / “synonymous” / incompatible / missing

• ACT empirically tested: within 1-5% of human ratings
  – can also be used to spot litigious sentences, which are given to human assessment (10-20% of all sentences)
Recent results, on WMT10 data

- **Factored models (with Moses SMT)**
  - source factors: POS tags, labeled connectives (DL), or both (POS+DL)
  - phrase-based or hierarchical SMT models

- **Non-factored models: multiplied data based on labels’ probabilities**

<table>
<thead>
<tr>
<th>Translation model</th>
<th>SMT system</th>
<th>BLEU</th>
<th>$ACT_a$</th>
<th>$ACT_{a5}$</th>
<th>$ACT_m$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factored phrase-based</td>
<td>POS + DL</td>
<td>22.19</td>
<td>70.7</td>
<td>86.1</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td>DL</td>
<td>21.69</td>
<td>70.0</td>
<td>85.2</td>
<td>80.7</td>
</tr>
<tr>
<td></td>
<td>POS</td>
<td>22.26</td>
<td>67.9</td>
<td>81.2</td>
<td>76.4</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>21.71</td>
<td>65.0</td>
<td>77.8</td>
<td>73.6</td>
</tr>
<tr>
<td>Factored hierarchical</td>
<td>DL</td>
<td>19.20</td>
<td>67.9</td>
<td>78.5</td>
<td>77.1</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>19.31</td>
<td>63.6</td>
<td>74.8</td>
<td>74.3</td>
</tr>
<tr>
<td>Phrase-based with label probabilities</td>
<td>LPD</td>
<td>21.60</td>
<td>69.4</td>
<td>82.0</td>
<td>78.5</td>
</tr>
<tr>
<td></td>
<td>Baseline</td>
<td>21.30</td>
<td>68.8</td>
<td>81.1</td>
<td>79.2</td>
</tr>
</tbody>
</table>
Wrap up
1. Linguistic analyses
   Cohesion markers for MT
   Features for classification
   Cross-linguistic perspective

2. Corpus data and annotation
   Select corpora
   Define tagset and guidelines
   Locate problematic examples
   Execute annotation and deliver data

3. Automatic labeling of cohesion markers
   Classifiers w. contextual features
   Use of synchronous parsing
   Dependencies across classifiers

4. SMT of labeled texts
   Phrase-based SMT for labeled texts
   SMT using synchronous parsing
   Synchronous parsing SMT with labels

5. Evaluation
   Define metrics of coherence
   How to use test suites
   Performance of past systems
   Apply metrics

COMTIS
Perspectives

• Make progress on all tasks
  – more resources, better integration with MT, process new phenomena, improve evaluation

• Towards a proof-of-concept
  – text-level processing is efficient enough for MT
  – it can be efficiently combined with MT

• Check [www.idiap.ch/comtis](http://www.idiap.ch/comtis) for more details
Some COMTIS publications


