Images as Context in Statistical Machine Translation

Iacer Calixto†, Teo de Campos§ and Lucia Specia‡
†University of Wolverhampton, calixto.iacer@gmail.com
‡University of Sheffield, t.decamps@surrey.ac.uk
§University of Surrey, l.specia@surrey.ac.uk

Statistics Machine Translation

• Systems like “Google Translate”
• Reasonable quality on documents, low quality:
  - When short textual context is available - ambiguities cannot be resolved
  - When trained on different domains - unknown words left untranslated (OOV)

Research questions

• Can images help solve the problem of ambiguous and out-of-vocabulary words?
• Can computer vision techniques help retrieve textual information that complements the original context?
• In which ways can textual cues extracted from images be used in SMT systems?

Goals

• Compile a dataset composed of short texts and potentially useful images and keywords derived from these images
• Evaluate a sample of such dataset to answer whether images could help solve the problem of ambiguous and out-of-vocabulary words
• Provide a basis for answering the other research questions

Dataset

• Moses toolkit to build an SMT system based on Europarl data
• Translations filtered in a number of ways to keep medium quality translations
• Content:
  - Images from Wikipedia
  - Their captions in English
  - Their machine translations (Moses) into Portuguese, Spanish, German or French
  - Their “reference” (human) translation as found in Wikipedia
  - Related images retrieved from ImageNet using a standard computer vision method using bags of visual words
  - Keywords from the WordNet synset associated with the retrieved image

Dataset examples


Human – Casa de Santos Dumont: estante de madeira para livros (alguns claramente danificados).
MT – Santos Dumont Assembleia (Petrópolis, realizada no Rio de Janeiro, o Brasil): Estado de madeira danificada livros bookcase.

• 1 OOV word
• 2 ambiguous words

Evaluation results

• Non-expert, bilingual speakers evaluated English-to-Portuguese automatically translated sentences
• How useful the Wikipedia & ImageNet images, and ImageNet keywords are for translation
• Results:
  - 5.03% of sentences evaluated for English-Portuguese (355)
  - 23.04% have 1+ OOV words
  - Average of 1.51% OOV words per sentence
  - 43.72% have 1+ ambiguous (AMB) words incorrectly translated
  - Average of 1.38% AMB words per sentence
• % of the sentences for which images/keywords are useful:

<table>
<thead>
<tr>
<th>Helps in translation</th>
<th>% of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia image</td>
<td>75.99%</td>
</tr>
<tr>
<td>ImageNet image</td>
<td>9.16%</td>
</tr>
<tr>
<td>ImageNet keywords</td>
<td>6.81%</td>
</tr>
</tbody>
</table>

% of the sentences with 1+ OOV or 1+ AMB for which images/keywords are useful:

<table>
<thead>
<tr>
<th>Helps in translation</th>
<th>% of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia image</td>
<td>78.20%</td>
</tr>
<tr>
<td>ImageNet image</td>
<td>10.43%</td>
</tr>
<tr>
<td>ImageNet keywords</td>
<td>6.44%</td>
</tr>
</tbody>
</table>

% of the sentences for which more than one image/keyword combination is useful:

<table>
<thead>
<tr>
<th>Helps in translation</th>
<th>% of sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia image +</td>
<td>10.07%</td>
</tr>
<tr>
<td>ImageNet image</td>
<td></td>
</tr>
<tr>
<td>ImageNet keywords</td>
<td>7.99%</td>
</tr>
</tbody>
</table>

% of the sentences for which visual cues help for:

• Sentences with 1+ OOV, but 0 AMB
• Sentences with 1+ AMB, but 0 OOV

<table>
<thead>
<tr>
<th>Helps in translation</th>
<th>1+ OOV but 0 AMB</th>
<th>1+ AMB but 0 OOV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia image</td>
<td>84.09%</td>
<td>79.67%</td>
</tr>
<tr>
<td>ImageNet image</td>
<td>13.64%</td>
<td>10.57%</td>
</tr>
<tr>
<td>ImageNet keywords</td>
<td>4.55%</td>
<td>9.76%</td>
</tr>
</tbody>
</table>

Remarks

• State of the art SMT systems produce a large number of incorrect translations (OOV and AMB)
  - Model trained on a different domain – fairly standard scenario in MT
• Wikipedia’s images can be useful in providing context to MT in 79%–84% of problematic cases
  - To a certain degree, ImageNet images can be also useful (10%–13% of problematic cases)
• Only a subset of 1000 synsets were used from ImageNet, many query images had objects that did not appear in the training set
• The simple BoW method used for image representation is a baseline that can be improved
• Dataset to be released with the following expected number of bilingual sentences (and images):
  - English-Portuguese – 9,239;
  - English-Spanish – 29,786;
  - English-French – 57,646;
  - English-German – 114,402;
• http://www.dcs.shef.ac.uk/~lucia/resources.html