Images as Context in Statistical Machine Translation

Iacer Calixto
University of Wolverhampton

Lucia Specia
University of Sheffield

Teófilo de Campos
University of Surrey
Images as Context in Statistical Machine Translation

- Outline
  - Quality of translations in Statistical Machine Translation;
  - Research questions;
  - Goals;
  - Dataset;
  - Image processing pipeline;
  - Dataset examples;
  - Results;
  - Remarks.
Quality of translations in Statistical Machine Translation

- Learn how to translate based on previously seen examples

- Some known problems:
  - ambiguity (AMB);
  - low quality when training is done on a different domain;
  - words left untranslated or out-of-vocabulary (OOV) words.
Research questions

1) Can images help solve the problem of ambiguous and out-of-vocabulary words?

2) Can computer vision techniques help retrieve textual information that complements the original context?

3) In which ways can textual cues extracted from images be used in SMT systems?
Goals

- Create a dataset composed of:
  - short texts;
  - potentially useful images;
  - keywords derived from these images.

- Conduct an evaluation on a sample of the dataset and answer research question (1).

- Provide a basis for answering the other research questions (2) and (3).
Dataset

- Images from Wikipedia and their textual captions in English
- Their machine translations (Moses) into Pt, Es, De and Fr
  - Moses SMT Toolkit + Europarl corpus for training
- Their human (reference) translations as found in Wikipedia
- Related images retrieved from ImageNet
- Keywords from WordNet synset associated with the retrieved image

- Translations are filtered in a number of ways to keep only the medium quality translations
- # sentences: En-Fr 57,646; En-De 114,402; En-Pt 9,161; En-Es 29,786.
Image processing pipeline

- Bag of Visual Words baseline method was used to retrieve images from ImageNet
- Implementation provided by ImageNet based on dense feature extraction using SIFT at multiple scales and pooling using hard voting with visual vocabulary built with K-means (k=1000)
- 1000-dimensional feature vector generated for each image
- we used linear SVM with L1 regularisation to classify each Wikipedia image into one ImageNet synset
- for evaluation purposes, image representative of the synset was randomly chosen
Dataset examples

- Human – Casa de Santos Dumont: estante de madeira para livros (alguns claramente danificados).

- 1 OOV word
- 3 ambiguous words
Dataset examples

- The “Pico do Jaraguá” in the West zone of São Paulo (city)
- Human – O pico do Jaraguá (na zona Oeste de são Paulo)
- MT – A “do pico fazer Jaraguá” no Ocidente zona de S. Paulo (City)
- • 1 OOV word
- • 2 ambiguous words
Evaluation

• Non-expert, bilingual speakers evaluated English-Portuguese automatically translated sentences.

• How useful are Wikipedia and ImageNet images, and ImageNet keywords for translation.

• Results:
  - 5.03% of sentences evaluated for En-Pt (355);
  - 23.04% have one or more OOV words (average of 1.51 words per sentence);
  - 43.72% have one or more AMB words incorrectly translated (average of 1.38 words per sentence).
Evaluation

% of 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>75.39%</td>
</tr>
<tr>
<td>ImageNet image</td>
<td>9.16%</td>
</tr>
<tr>
<td>ImageNet keyword</td>
<td>6.81%</td>
</tr>
</tbody>
</table>

% of sentences with one or more OOV or one or more AMB word 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 keyword</td>
<td>6.64%</td>
</tr>
</tbody>
</table>
Evaluation

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

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

% of sentences with either +1 OOV but 0 AMB or +1 AMB or 0 OOV for which visual cues help:

<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 keyword</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 translation (OOV and AMB)
  - model trained on different domain + short textual context
- Wikipedia images can be useful in providing context to MT in 79%-84% of problematic cases
- To a certain degree, ImageNet images can also be useful (10%-13% of problematic cases)
  - only 1000 synsets from ImageNet were used, many query images had objects that did not appear in the training set;
  - the simple BoW method used for image representation is a baseline and can be improved;
- Dataset will be available for download soon at: http://www.dcs.shef.ac.uk/~lucia/resources.html
Thank you!

Questions... ?

Iacer Calixto
calixto.iacer@gmail.com

Lucia Specia
l.specia@sheffield.ac.uk

Teófilo de Campos
t.decampos@surrey.ac.uk