Images as Context in Statistical Machine Translation*

Iacer Calixto[1], Teófilo de Campos[2] and Lucia Specia[3]

[1]University of Wolverhampton, Stafford St, Wolverhampton WV1 1SB, UK
[2]CVSSP, University of Surrey, Guildford GU2 7XH, UK
[3]DCS, University of Sheffield, 211 Portobello St, Sheffield S1 4DP, UK

1 Problem statement

This paper reports ongoing experiments towards exploiting the use of images to provide additional context for statistical machine translation (SMT). We investigate whether this contextual information can be helpful in targeting two well-known challenges in machine translation: ambiguity (incorrect translation of words that have multiple senses) and out-of-vocabulary words (words left untranslated).

As a motivating example, consider Figure 1, which depicts a news headline extracted from the BBC News website[1] and its incorrect translation into Portuguese, as generated by the Google Translate online service.

[en] Approval for *pup* rescue
A tired *seal pup* which refused to leave an oil industry ship 100 miles out to sea is recovering before being released back into the wild.

[pt]Aprovação para resgate *filhote de cachorro*
Um *selo* cansado, que se recusou a sair de um navio indústria petrolífera 100 milhas para o mar está se recuperando antes de ser liberado de volta na natureza.

Figure 1: Example of incorrect translation of a news headline due to ambiguity.

The word *seal* is ambiguous, with at least two possible translations into Portuguese: *selo* (*stamp*) and *foca* (marine animal). In this short context, *seal pup* should have been translated as *filhote de foca* (*young seal*), but it has been translated as *selo*. Moreover, the word *pup* in the title has been incorrectly translated as *filhote de cachorro* (*young dog*), when it should also have been translated as *filhote de foca*. Our hypothesis is that cases like these, where ambiguous words cannot be correctly disambiguated because their textual context is very short, could benefit from the use of images provided along with the textual information to improve the overall translation quality.

Our goal in this project was to investigate this hypothesis by selecting and inspecting English short texts accompanied by images and their machine translation. As the first attempt to combine visual and textual cues for machine translation, this short project has only scratched the surface of a number of complex questions:

1. Can visual information help solve textual issues of ambiguity and unknown words in translation?
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?

In order to answer these questions, we automatically built a dataset containing (i) images from Wikipedia, (ii) their captions in English, (iii) their machine translations into Portuguese, Spanish, German or French, (iv) their "reference" (human) translation as found in Wikipedia, (v) a similar image retrieved from ImageNet using standard computer vision methods, and (vi) keywords from the WordNet synset associated with the retrieved image. We are in the process of evaluating a sample of this dataset in order to answer questions 1-2 above.

2 Dataset

The textual part of the dataset contains a collection of English captions from Wikipedia which had a corresponding (human) translation in at least one of our four languages of interest: French, German, Portuguese, and Spanish. These captions were extracted by parsing and post-processing the May-2012 dump of the Wikipedia database. Taking pairs of captions in English and each of these four languages, we removed descriptions with more than 80 words (normally difficult to translate), and sentences in which the difference in length between the source and the translation (reference) was greater than 30% (most likely versions as opposed to translations). The next step was to machine translate the English captions into French, German, Portuguese, and Spanish. We used a standard, state-of-the-art Moses phrase-based SMT system trained on

*The authors acknowledge the EPSRC Vision and Language Network (EP/H018557/1) for the support through a pump-priming grant. Additionally, TdC received support from EPSRC grant EP/F069421/1 (ACASVA) and IC received a partial travel grant from the PASCAL2 Network of Excellence (EU). IC is currently with the Centre Tesnière - Université de Franche-Comté, Besançon, France.

[1]BBC news 29 November 2011 `http://www.bbc.co.uk/news/uk-scotland-north-east-orkney-shetland-15940058`

parallel texts from the European Parliament.[2] The source-translation pairs of captions were then automatically filtered to keep cases with potential translation problems, but sufficient translation quality: we used the METEOR metric[3] to measure the similarity between the human and automatic translations, and kept only the cases with a METEOR score larger than 0.1 (translation has sufficient quality), but smaller than 0.9 (translation is not a perfect). Finally, we checked the subset of WordNet synsets included in ImageNet (see below) to select only (source) captions that contained at least one content word from those synsets. The statistics of the final dataset are as follows:

Language pair	Number of captions
English-French	57,646
English-German	114,402
English-Portuguese	9,161
English-Spanish	29,786

Given the selected pairs of source and machine translated captions, the Wikipedia images associated to the English captions were extracted from the Wikipedia database dump. The next step was to find similar images to the ones from Wikipedia which are linked to textual information that we believe can be useful to solve ambiguities and out-of-vocabulary words. This was done using ImageNet, a dataset that was built by searching the web using keywords obtained from nouns in WordNet. We took the subset of $1,000$ synsets of the ILSVRC2010 challenge dataset,[4] with $1,043,415$ images, all of which were used in our training set.

In order to retrieve similar images from ImageNet, a baseline bag-of-visual-words approach was used. We took the implementation provided by ImageNet which is based on dense feature extraction using SIFT at multiple scales and pooling using hard voting with the visual vocabulary built with K-means ($K = 1,000$). Therefore, for each image, a 1000-dimensional feature vector was generated. Classification into one of the $1,000$ classes (synsets) was done using linear SVM with ℓ_1 regularisation.[5] Each Wikipedia image is thus classified into one of the $1,000$ ImageNet synsets. For evaluation purposes, a sample image from the selected synset was randomly chosen. This approach is not the state-of-the-art for image classification, but it provides a solid baseline.

Using this approach, the image of the *seal* in the headline of Figure 1 was classified as *poodle* (ImageNet's synset *n02113799*). This suboptimal result is mostly due to the fact that our subset of ImageNet does not contain a synset for *seal*. This result can nevertheless still be useful as it indicates that the *seal* in the image is a mammal, rather than an object (stamp).

In order to answer the research questions listed in the previous section, we are collecting human judgements based on an online form as the one in Figure 2. The form shows the source caption and its machine translation, and asks the annotator to indicate the number of out-of-vocabulary words and number of words incorrectly translated due to ambiguity. With these questions we expect to have a more reliable assessment on the quality of the automatic translations as compared to the METEOR-based filter, i.e., whether there is room for improvement in translation. The form also asks the annotator to judge whether both the original (Wikipedia) image and the image retrieved from ImageNet could be useful to solve those problems, without suggesting any details on how this would be implemented. Finally, taking the content words from the WordNet synset corresponding to the image retrieved, the form asks whether these could be useful to solve the same two problems.

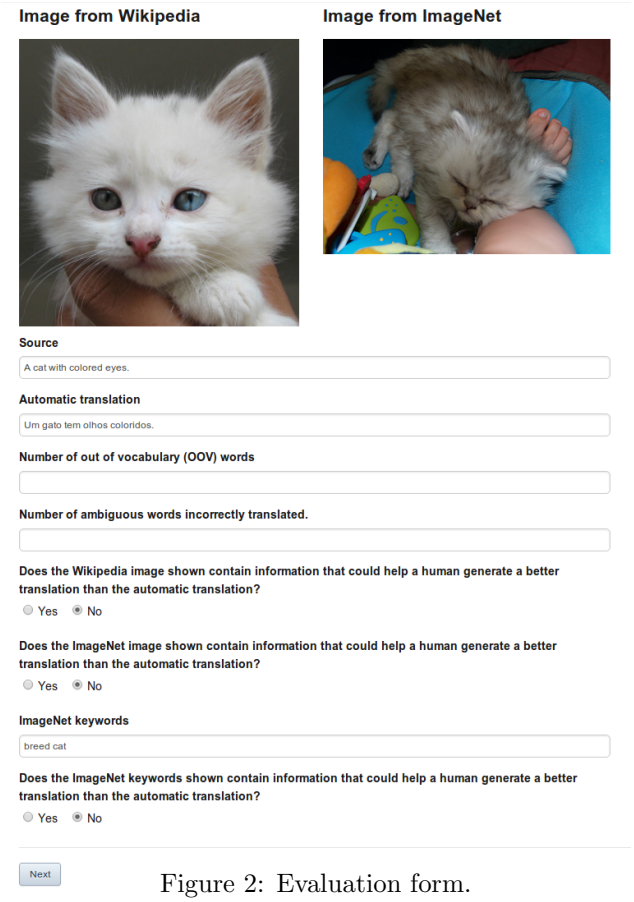

Figure 2: Evaluation form.

We note that this evaluation does not aim at assessing methods to use these sources of information as part of the translation process, but rather assessing whether exploiting these sources is worthwhile. Some interesting directions that we will investigate are the use of keywords in the target language to bias language models towards certain lexical choices (disambiguation) and the use of keywords in the source language to replace out-of-vocabulary words by synonyms or related words that are known to the translation model.

The dataset is freely available for download at `http://www.dcs.shef.ac.uk/~lucia/resources.html`. We expect this dataset to provide fertile ground for insights to be used in future research on the combination of textual and visual information for translation.

[2] `http://www.statmt.org/wmt11/baseline.html`
[3] `http://www.cs.cmu.edu/~alavie/METEOR/`
[4] `http://www.image-net.org/challenges/LSVRC/2010/`
[5] `http://www.csie.ntu.edu.tw/~cjlin/liblinear/`.