Multimodal Lexical Translation

Chiraag Lala and Lucia Specia
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
{clala1, l.specia}@sheffield.ac.uk

Abstract

Inspired by the tasks of Multimodal Machine Translation and Visual Sense Disambiguation we introduce a task called Multimodal Lexical Translation (MLT). The aim of this new task is to correctly translate an ambiguous word given its context - an image and a sentence in the source language. To facilitate the task, we introduce the MLT datasets, where each data point is a 4-tuple consisting of an ambiguous source word, its visual context (an image), its textual context (a source sentence), and its translation that conforms with the visual and textual contexts. The dataset has been created from the Multi30K corpus using word-alignment followed by human inspection for English to German and English to French language directions. These datasets form a very valuable multimodal and multilingual language resource with several potential uses including evaluation of lexical disambiguation within (Multimodal) Machine Translation systems.

Keywords: Multimodal Machine Translation, Visual Sense Disambiguation, Multimodal Multilingual Language Resource

1. Introduction

Multimodal Machine Translation (MMT) is the task of translating text using information in other modalities (such as images) as auxiliary cues. It has been recently framed as a shared task as part of the two last editions of the Conference on Machine Translation (WMT16, WMT17) (Specia et al., 2016; Elliott et al., 2017). Within WMT, the task is is defined as such: Given an image and its description in the source language, the objective is to translate the description into the target language, where this process can be supported by information from the image, as depicted in Figure 1.

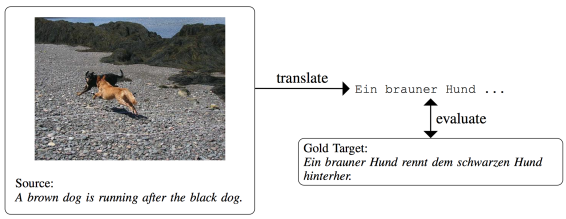

Figure 1: Multimodal Machine Translation Shared Task

One of the main motivations to introduce multimodality in Machine Translation (MT) is the intuition that information from other modalities could help find the correct sense of ambiguous words in the source sentence, which could potentially lead to more accurate translations. For example, the English sentence "A man is holding a seal" could have at least two different translations in German depending on the sense of the word *seal* - (1) "Ein Mann hält ein Siegel", and (2) "Ein Mann hält einen Seehund". The images (Figures 2 and 3) could help a MMT system disambiguate the correct sense of the word *seal* and translate accordingly.
Disambiguation of word senses, known as Word Sense Disambiguation (WSD), is a widely studied natural language processing task. Given an ambiguous word and its context, the objective is to assign the correct sense of the word

Figure 2: A man is holding a seal (Ein Mann hält ein Siegel)

Figure 3: A man is holding a seal (Ein Mann hält einen Seehund)

based on a pre-defined sense inventory. An review of approaches to WSD can be found in (Navigli, 2009; Navigli, 2012; Raganato et al., 2017). In standard WSD, words are disambiguated based on their textual context. However, in a multimodal setting we could disambiguate words using visual context. This modified version of WSD that uses visual context instead of textual context is called Visual Sense Disambiguation (VSD). In monolingual work, VSD has previously been attempted for ambiguous nouns like the word 'bank' which could refer to the financial institution or the river bank (Barnard et al., 2003; Loeff et al., 2006; Saenko and Darrell, 2008; Chen et al., 2015). Re-

cently, VSD has also been attempted for ambiguous verbs like the word 'play' which could refer to playing a musical instrument or playing a sport (Gella et al., 2016).
In MT, including MMT, disambiguation of word sense happens implicitly. For instance, in the same example "A man is holding a seal", we would come to know whether the system disambiguated the correct sense of the word *seal* only indirectly from the translation produced by the system. The corresponding translation of the word *seal* in the target language (*Siegel* or *Seehund* in German) acts as a "sense label". Further, in MMT, we would like know which modality (visual or textual) contributed to the disambiguation and to what extent.
The main contribution of this paper is to facilitate the study of sense disambiguation within MMT framework by:

1. Generating a language resource of ambiguous words and its translations together with visual and textual contexts. We call it the Multimodal Lexical Translation Dataset (MLTD).

2. Introducing a new task - Multimodal Lexical Translation (MLT) - and demonstrating a simple way to evaluate lexical disambiguation within MMT using the MLTD datasets.

We build this resource for English to German and English to French translations.

2. Language Resource - MLTD

MLTD is a collection of 4-tuples of the form:

$$\{(\mathbf{v}_i, \mathbf{x}_i, x_i, y_i)\}_{i=1}^{n} \quad (1)$$

where x_i is an ambiguous[1] word, $\mathbf{x}_i$ is its textual context (a source sentence), $\mathbf{v}_i$ is its visual context (an image), and y_i is its translation that conforms with both the textual and visual contexts.

2.1. Generating the Dataset

We make use of the Multi30K dataset (Elliott et al., 2016; Elliott et al., 2017), which consists of 31,014 triples of the form $(\mathbf{v}_i, \mathbf{x}_i, \mathbf{y}_i)$ where $\mathbf{v}_i$ is an image, $\mathbf{x}_i$ is an English description of the image and $\mathbf{y}_i$ is a translation of the description in the target language (German and French) by professional translators (and i is an integer index ranging from 1 to 31,014). From this sentence-level dataset, we extract ambiguous words using the following steps:

Pre-processing → Word Alignment → Automatic Filtering → Human Filtering

2.1.1. Pre-processing

Sentences in all languages are lowercased and tokenized using scripts from the Moses toolkit[2] (Koehn et al., 2007). German sentences, which can contain compound words like 'sonnenblumenkerne' (sunflower seeds), are split/decompounded using pre-computed model of SEmantic COmpound Splitter (SECOS)[3] (Martin Riedl, 2016). Since we are not interested in morphological variants of the words, we also lemmatized[4] all sentences in the respective languages, which reduced vocabulary size and led to better word alignment.

2.1.2. Word Alignment

After the pre-processing step, the word tokens in the Multi30K parallel corpus are aligned using Fast Align[5] (Dyer et al., 2013). Fast Align generates asymmetric word alignments depending on which language in the parallel corpus is treated as the source. We generate both alignments - 'forward' (where English is treated as the source language) and 'reverse' (where German or French is treated as the source language). To learn better word alignments, we train Fast Align on a larger parallel corpus comprising of the Europarl parallel corpus[6] (Koehn, 2005) in addition to the Multi30K parallel corpus for English-German and English-French language pairs separately. The Europarl corpus also undergoes the same pre-processing steps in Section 2.1.1. before word alignment.

2.1.3. Automatic Filtering

In this step we remove all the word alignments having stop words and select only those alignments which are to be found in both 'forward' and 'reverse' directions. In addition, we filter out the alignments between words with different Part-Of-Speech tags (using the NLP tool in footnote 4). Next, we remove all English words that get aligned to just one word in the target language across the entire Multi30K corpus, retaining only the potentially ambiguous English words, i.e. those aligned to multiple words in the target language.

2.1.4. Human Filtering

Finally, the set of each potentially ambiguous English word and all the words in the target language that get aligned to them is presented to human annotators in a 'dictionary' format. For instance, a French annotator is asked to inspect cases like

hat → casque, casquette, chapeau, haut, bonnet, couvre, képi, béret

Human annotators are native speakers of French and German who are also fluent in English. They are asked to filter out target words which are not translations of the source word in any possible context, such as *haut* in the above example. After the final filtering we retrieve the visual and textual contexts from Multi30K to complete the language resource.

2.2. Dataset Statistics and Examples

Currently the datasets are under development undergoing the human filtering step. All the previous steps have been completed. Based on our findings thus far, we expect to

[1] We use the term 'ambiguous' for those words in the source language that have multiple translations in the target language in a given dataset, representing different 'senses'.

[2] https:/github.com/moses-smt/mosesdecoder

[3] https://github.com/riedlma/SECOS

[4] http://staffwww.dcs.shef.ac.uk/people/A.Aker/activityNLPProjects.html

[5] https://github.com/clab/fast_align

[6] http://www.statmt.org/europarl/

extract 800 to 1000 unique ambiguous English words for each language direction. Also, for each of those words we expect to have, on average, 3 to 4 translations per word. Further, for each translation, we expect to have, on average, 12 to 18 instances/data points having the textual and visual contexts. In total we expect the MLTD size to be between 28,000 and 70,000 4-tuples for each language direction. Upon completion, the MLTD language resource will be made freely available under the Creative Commons Attribution Non Commercial ShareAlike 4.0 International license.

2.2.1. English–German

So far, we have extracted 376 unique ambiguous English words with 3.94 translations per word (on average) and 17.02 instances per translation (on average) totaling to 25,196 MLTD datapoints. Example data points:

1. Visual Context $\mathbf{v}_1$: (see Figure 4)
   Textual Context $\mathbf{x}_1$: "a few people are waiting in a subway, with an arriving car in the distance."
   Ambiguous Word x_1: *subway*
   Translation y_1: *bahnstation*

2. Visual Context $\mathbf{v}_2$: (see Figure 5)
   Textual Context $\mathbf{x}_2$: "pedestrians bombard a city street covered in consumerism, including signs for burger king, mcdonalds, subway, and heineken."
   Ambiguous Word x_2: *subway*
   Translation y_2: *subway*

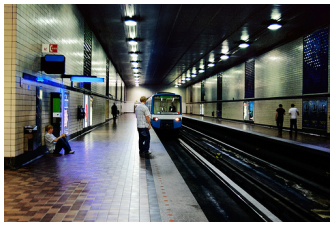

Figure 4: Multimodal Lexical Translation data point 1

Figure 5: Multimodal Lexical Translation data point 2

2.2.2. English–French

So far, we have extracted 267 unique ambiguous English words with 3.16 translations per word (on average) and 13.09 instances per translation (on average) totaling to 11,045 MLTD data points. Examples are similar to the above ones for English to German.

3. Multimodal Lexical Translation Task

Once we have MLTD, which is of the form in Equation 1, then at least three versions of MLT task can be defined. Given an ambiguous word x, translate it using its
(1) Textual context, i.e. source sentence $\mathbf{x}$ only
(2) Visual context, i.e. image $\mathbf{v}$ only
(3) Both Textual and Visual contexts $(\mathbf{x}, \mathbf{v})$

3.1. Evaluating MMT Systems

The MLT task can be used to evaluate MMT systems (and text-only MT systems too) in their ability to correctly translate ambiguous words. Consider a MLTD data point $(\mathbf{v}, \mathbf{x}, x, y)$. Now, an MMT system S can take two arguments as inputs - the source sentence $\mathbf{x}$ and the image $\mathbf{v}$ - and generate an output $S(\mathbf{x}, \mathbf{v})$ which is just a translation of the source sentence in the target language. A very straightforward evaluation strategy is to simply check if the correct lexical translation y of the ambiguous word x is also found in the MMT system's output $S(\mathbf{x}, \mathbf{v})$. When run across all the data points in MLTD, we can count the number of times MMT systems translated an ambiguous word correctly and compute the accuracy. More elaborate metrics, other than simple accuracy, will be developed and tested in future. For now, we use this accuracy measure and demonstrate an application of MLTD.
In Elliott et al. (2017), the MMT systems submitted to the shared task were evaluated and ranked using Meteor metric and human scoring. In this paper, we evaluate the same systems using the MLT task described above and compute MLT Accuracy (the accuracy of translating an ambiguous word correctly). The official test set – Multi30K 2017 test set – which was used for evaluation in Elliott et al. (2017) is used to extract MLTD data points for our evaluation. Performance of all the submissions is shown in Table 1 for English to German, and Table 2 for English to French.

System	**MLT** ↑	Meteor ↑	Human ↑
LIUMCVC_MNMT_C	**0.65**	54.0	77.8
NICT_1_NMTrerank_C	**0.64**	53.9	70.3
LIUMCVC_NMT_C	**0.64**	53.8	65.1
UvA-TiCC_IMAGINATION_U	**0.63**	53.5	74.1
DCU-ADAPT_MultiMT_C	**0.61**	50.5	68.1
UvA-TiCC_IMAGINATION_C	**0.60**	51.2	59.7
CUNI_NeuralMonkeyTextualMT_U	**0.60**	51.0	68.1
OREGONSTATE_2NeuralTranslation_C	**0.59**	50.6	54.4
CUNI_NeuralMonkeyMultimodalMT_U	**0.58**	50.2	60.6
OREGONSTATE_1NeuralTranslation_C	**0.58**	48.9	53.3
CUNI_NeuralMonkeyTextualMT_C	**0.57**	49.2	54.2
CUNI_NeuralMonkeyMultimodalMT_C	**0.54**	47.1	55.9
SHEF_ShefClassInitDec_C	**0.50**	44.5	46.6
SHEF_ShefClassProj_C	**0.48**	43.4	49.4
AFRL-OHIOSTATE-MULTIMODAL_U	**0.20**	20.2	36.6

Table 1: Performance of systems submitted to MMT Shared Task 2017 for English to German

System ranking correlation We observe that our evaluation of MLT Accuracy (although using partially complete MLT dataset only) is consistent with Meteor and human

System	**MLT** ↑	Meteor ↑	Human ↑
LIUMCVC_NMT_C	**0.68**	70.1	60.5
NICT_1_NMTrerank_C	**0.67**	72.0	79.4
LIUMCVC_MNMT_C	**0.65**	72.1	71.2
DCU-ADAPT_MultiMT_C	**0.64**	70.1	74.1
OREGONSTATE_1NeuralTranslation_C	**0.64**	67.2	60.8
CUNI_NeuralMonkeyTextualMT_C	**0.63**	67.0	61.9
OREGONSTATE_2NeuralTranslation_C	**0.62**	68.3	65.4
SHEF_ShefClassProj_C	**0.60**	61.5	54.0
CUNI_NeuralMonkeyMultimodalMT_C	**0.59**	67.2	74.2
SHEF_ShefClassInitDec_C	**0.57**	62.8	54.7

Table 2: Performance of systems submitted to MMT Shared Task 2017 for English to French

scores. To demonstrate this, we computed the Pearson Correlation Coefficient on the values in Table 1 to get:
$Correlation(\text{MLT}, \text{Meteor}) = 0.9$
$Correlation(\text{MLT}, \text{Human}) = 0.7$
Ranking of the systems using MLT differs slightly from the ranking using Meteor or human scores, but the top performing systems are the same.

Constrained vs. unconstrained For the teams that submitted both constrained and unconstrained models (those using additional external data for training): Unconstrained models show considerable improvement (at least 3%) over their constrained counterparts.

Multimodal vs. text-only For teams that submitted both multimodal and text-only systems: CUNI's multimodal system performs worse than its text-only system. In other cases, the contribution of multimodality is not evident as far as MLT is concerned.

Qualitative example Consider the following MLTD data point $(\mathbf{v}, \mathbf{x}, x, y)$ where $\mathbf{v}$ = Figure 6, $\mathbf{x}$ = "a man in an orange hat starring at something.", x = *hat*, and y = *hut*. While most MMT systems translated *hat* to *hut* in German, a few translated differently. 'CUNI_NeuralMonkeyMultimodalMT_C' translated it as *kappe*, 'OREGONSTATE_1NeuralTranslation_C' translated it as *mütze*, and 'SHEF_ShefClassInitDec_C' translated it as *kopfbedeckung*. All these words are referring to the same object but have slightly different senses. *Kappe* refers to the modern caps with shade extending out from front side only, usually worn in sports. *Hut* refers to the hat with edges/small extensions coming off from all sides and usually worn in summer. *Mütze* refers to differently designed hats for the winter. *Kopfbedeckung* means a headgear which could refer to any kind of object worn on the head. From the image (see Figure 6), it can be seen that the hat looks a bit unusual because of its colour and brand logo on it. Perhaps that could be a reason why some systems chose another translation instead of *hut*.

4. Future Work

Currently, the MLT evaluation using counts and accuracy is too simplistic and has its limitations. First of all, it is based on exact matching of the surface-form of the gold standard lexical translation with the corresponding word in the system generated translation. Thus, any other correct translation that does not consist our gold standard lexical translation will be considered an error. Secondly, no partial credit is given to synonymous words (although it may have different sense). For instance, in the example discussed in Section 3.1. all systems translated *hat* into some hat. Maybe not the correct kind of hat, but some hat nevertheless. To overcome these limitations, our future work will be focused on developing more elaborate scoring for MLT evaluation. Finally, we are interested in understanding whether the disambiguation happening within the system is due to the textual context or the visual context, both or none. For this, we propose to use the same MMT system to translate the MLT data in four different ways. Recall from equation 1, MLT data is of the form $(\mathbf{v}, \mathbf{x}, x, y)$. Given a MMT System S, we compare four kinds of output with the reference.

Figure 6: Multimodal Lexical Translation Datapoint used in the evaluation of submissions to MMT Shared Task 2017

$$S(\mathbf{x}, \mathbf{v}) \sim S(\mathbf{x}, \mathbf{0}) \sim S(x, \mathbf{v}) \sim S(x, \mathbf{0}) \sim y \quad (2)$$

where $\mathbf{0}$ refers to absence of visual context (no image). Such a comparison should help measure a system's ability of making use of different modalities. Thus, in addition to the *inter-system* comparisons that already happen in the evaluation of MMT system, in future we will work on these *intra-system* comparisons of equation 2.

5. Conclusion

We introduced the MLTD language resource and the process of generating it from Multi30K using word alignments followed by human filtering. 25,196 MLTD data points for English to German and 11,045 MLTD data points for English to French have been generated so far. These numbers are expected to rise to anywhere between 28,000 and 70,000. Different versions of MLT task were introduced. We demonstrated the use of MLT task to evaluate MMT systems' ability to translate ambiguous words correctly. For this, submissions to the MMT Shared task were evaluated using our simple MLT accuracy metric. This metric, in spite of its limitations, was found to be consistent with Meteor and human scoring used in Elliott et al. (2017). We observed that unconstrained MMT models perform better than their constrained counterparts under the MLT accuracy metric. To sum up, the multimodal multilingual MLTD is a useful language resource that can facilitate, among many things, the study of lexical disambiguation with MMT systems.

6. Acknowledgements

This work is supported by the MultiMT project (H2020 ERC Starting Grant No. 678017). The authors also thank Mareike Hartmann, Julia Ive, Fred Blain, Pranava Madhyastha, and Josiah Wang.

7. Bibliographical References

Barnard, K., Johnson, M., and Forsyth, D. (2003). Word sense disambiguation with pictures. In *Proceedings of the HLT-NAACL Workshop on Learning Word Meaning from non-Linguistic Data*, pages 1–5.

Chen, X., Ritter, A., Gupta, A., and Mitchell, T. (2015). Sense discovery via co-clustering on images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5298–5306.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL-HLT*.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. pages 70–74.

Elliott, D., Frank, S., Barrault, L., Bougares, F., and Specia, L. (2017). Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, Copenhagen, Denmark.

Gella, S., Lapata, M., and Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of NAACL-HLT*, pages 182–192.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., et al. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of ACL*, pages 177–180.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of 10th Machine Translation Summit (MT Summit)*, pages 79–86.

Loeff, N., Alm, C. O., and Forsyth, D. A. (2006). Discriminating image senses by clustering with multimodal features. In *Proceedings of COLING/ACL*, pages 547–554, Sydney, Australia.

Martin Riedl, C. B. (2016). Unsupervised compound splitting with distributional semantics rivals supervised methods. In *Proceedings of NAACL-HLT*, pages 617–622, San Diego, CA, USA.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2):10.

Navigli, R. (2012). A quick tour of word sense disambiguation, induction and related approaches. *SOFSEM 2012: Theory and practice of computer science*, pages 115–129.

Raganato, A., Delli Bovi, C., and Navigli, R. (2017). Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1167–1178, Copenhagen, Denmark.

Saenko, K. and Darrell, T. (2008). Unsupervised learning of visual sense models for polysemous words. In *Proceedings of NIPS*, pages 1393–1400, Vancouver, British Columbia, Canada.

Specia, L., Frank, S., Sima'an, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation*, pages 543–553.