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 dataset, 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 translations from English to German and English to French. We also introduce a simple heuristic to quantify the extent of the ambiguity of a word from the distribution of its translations and use it to select subsets of the MLT Dataset which are difficult to translate. These form a 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 Resources

1. Introduction

Multimodal Machine Translation 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 last two editions of the Conference on Machine Translation (WMT16, WMT17) (Specia et al., 2016; Elliott et al., 2017). Within the Conference on Machine Translation, the task is defined as: Given an image and its description in the source language, the objective is to translate the description into a target language, where this process can be supported by information from the image, as depicted in Figure 1.

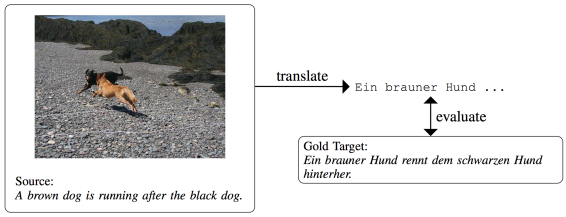


Figure 1: Multimodal Machine Translation Shared Task

One of the main motivations to introduce multimodality in Machine Translation 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 (Figure 2) could help a Multimodal Machine Translation system disambiguate the correct sense of the word *seal* and translate accordingly.

Disambiguation of word senses, popularly known as Word Sense Disambiguation or Lexical Disambiguation, 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 based on a pre-defined sense inventory. A review of approaches to Word Sense Disambiguation can be found in Navigli (2009) and Raganato et al. (2017).

(a) Ein Mann hält ein Siegel

(b) Ein Mann hält einen Seehund

Figure 2: Two different translations of "A man is holding a seal" depending on the visual context

In standard Word Sense Disambiguation, words are disambiguated based on their textual context. However, in a multimodal setting we could also disambiguate words using visual context. This modified version of Word Sense Disambiguation that uses visual context instead of textual context is called Visual Sense Disambiguation.

In monolingual work, Visual Sense Disambiguation has previously been attempted for ambiguous nouns like the word 'bank' which could refer to a financial institution or a river bank (Barnard et al., 2003; Loeff et al., 2006; Saenko and Darrell, 2009; Chen et al., 2015). Recently, Visual Sense Disambiguation 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 Machine Translation, including Multimodal Machine

Translation, 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 Multimodal Machine Translation, 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 Word Sense Disambiguation within Multimodal Machine Translation framework by:

1. Generating a language resource of ambiguous words and its translations together with visual and textual contexts. We call this the Multimodal Lexical Translation Dataset (MLT Dataset).
2. Introducing a new task - Multimodal Lexical Translation (MLT) - where the objective is to translate a single word into the target language given an image (visual context) and a sentence in the source language (textual context).
3. Demonstrating a simple way to evaluate lexical disambiguation within Multimodal Machine Translation using the MLT Dataset.

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

2. Language Resource - MLT Dataset

The MLT Dataset is a collection of 4-tuples of the form:

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

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 MLT Dataset

We make use of the Multi30K dataset (Elliott et al., 2016; Elliott et al., 2017), an extension of the Flickr30K dataset (Young et al., 2014), 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 a description of the image in the source language (English) and $\mathbf{y}_i$ is a translation of the description in the target language (German and French) by human translators (i is an integer index ranging from 1 to 31,014). From this sentence-level dataset, we extract the ambiguous words and their lexical translations 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] (Riedl and Biemann, 2016). Since we are not interested in distinguishing 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 in the later step.

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 corpus comprising of the Europarl parallel corpus[6] (Koehn, 2005) in addition to the Multi30K parallel corpus for the 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 (POS) tags (using the NLP tool in footnote 4). Next, we remove all English words that get aligned to a single 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. These retained alignments are converted into a dictionary format where 'Keys' are the potentially ambiguous English words and 'Values' are the words in the target language that get aligned to it. For instance, in English-French language direction we have cases like:

four → *quart, quartequatre*
woods → *forêt, bois*
western → *occidental, western*
hat → *casque, casquette, chapeau, haut, bonnet, couvre, képi, béret*

One dictionary from the word alignments of each language pair is built independently, i.e. one for English-German and one for English-French.

2.1.4. Human Filtering

Finally, each dictionary (English-German and English-French) from the Automatic Filtering step is given to human annotators for a final inspection and filtering. Human

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

[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/

annotators are native speakers of French/German who are also fluent in English. They were asked to

1. filter out instances which they believe did not have multiple senses, e.g. 'western → *occidental, western*'

2. filter out target words which they believe are not translations of the source word in any context, such as *haut* in the example 'hat → *casque, casquette, chapeau, haut, bonnet, couvre, képi, béret*'

The annotators were given the freedom to use any other resource, such as bilingual dictionaries, existing translation tools, etcetera that may help them filter the dictionaries. After the final filtering and inspection, for each (Key, Value) pair in the dictionaries we retrieve the visual and textual contexts from the Multi30K dataset to complete the MLT Dataset language resource.

2.2. Dataset Statistics and Examples

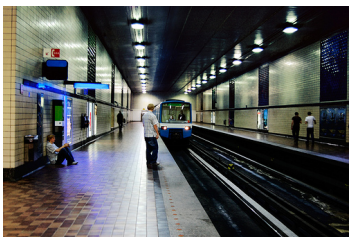
We extracted 1108 words in English which are ambiguous in either German or French or both (i.e. they have multiple translations in the target language). Each of these words can appear in multiple sentences, thus resulting in a total of 98,647 MLT datapoints.
Both English-German and English-French MLT language resources are made freely available[7] under the Creative Commons Attribution Non Commercial ShareAlike 4.0 International license.

English - French
We extracted 661 words in English which are ambiguous in French with 2.98 translations per word (on average) and 22.73 instances per translation (on average) totaling to 44,779 MLT datapoints.

English - German
We extracted 745 words in English which are ambiguous in German with 4.09 translations per word (on average) and 17.69 instances per translation (on average) totaling to 53,868 MLT datapoints. A couple of examples are shown below.

Examples

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



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

3. Multimodal Lexical Translation Task

Once we have the MLT Dataset, 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 (or disambiguate its sense) using its

1. Textual context only, i.e. source sentence $\mathbf{x}$ only

2. Visual context only, i.e. image $\mathbf{v}$ only

3. Both Textual and Visual contexts $(\mathbf{x}, \mathbf{v})$

These three versions can help identify the relative importance of the textual context and visual context when translating (or disambiguating) an ambiguous word.

3.1. Evaluating Machine Translation Systems

MLT can be used to evaluate (Multimodal) Machine Translation systems in their ability to correctly translate ambiguous words. Consider a MLT datapoint $(x, y, \mathbf{x}, \mathbf{v})$ of the form in Equation 1. A Multimodal Machine Translation 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 a translation of the source sentence in the target language. A straightforward evaluation strategy is to simply check if the correct lexical translation y of the ambiguous word x, as given in the reference translation, is also found in the system's output $S(\mathbf{x}, \mathbf{v})$[8]. When run across all the examples in the MLT dataset, we can then count the number of times a system translated ambiguous words correctly to compute its accuracy in the task. We call this the *MLT accuracy*. More elaborate metrics and metrics that also consider variants of the words in the reference, will be developed and tested in future. For now, we use this simple accuracy measure and demonstrate a potential application of the MLT Dataset.

[7] https://github.com/sheffieldnlp/mlt

[8] For consistency, the system's outputs undergo the same preprocessing steps in Section 2.1.1.

3.1.1. Evaluating Machine Translation Systems

In Elliott et al. (2017), the Multimodal and Text-only Machine Translation systems submitted to the shared task were evaluated and ranked using the Meteor metric and Human scoring.

The Meteor metric (Denkowski and Lavie, 2014) calculates a sentence-level similarity score between 0 and 100 between the system output and the reference (human) translation, where 0 means no similarity and 100 means 'perfect' similarity. Similarity is computed as a function of the proportion of words that can be aligned between the system and human translations, allowing for different types of alignments (e.g. lemma, synonym). The overall Meteor score of a system is the mean of the sentence-level scores over the test set.

Human scoring was carried out in Elliott et al. (2017) using bilingual Direct Assessment (Graham et al., 2017), where the assessors were asked to evaluate the semantic relatedness between the system outputs and the source sentence (not the reference translation) given the image. The assessors gave a sentence-level score between 0 and 100, where 0 indicates that the meaning of the source sentence is not preserved in the system output, and 100 means that the meaning is 'perfectly' preserved. The sentence-level scores were standardized according to each individual assessor's overall mean and standard deviation score. The overall Human score of a system was computed as the mean of the standardized sentence-level scores over the test set.

We evaluate the participating systems of the Multimodal Machine Translation Task using the MLT Accuracy. The MLT Accuracy of a system measures the proportion of ambiguous words in the test set that are correctly translated by the system. The ambiguous words in the test set and its lexical translations are obtained from the MLT Dataset. We extract ambiguous words from the official test sets – Multi30K 2017 test set and Ambiguous COCO test set (Elliott et al., 2017) – for our evaluation. The MLT Accuracy is measured in percentages on a scale of 0 to 100.

The performance of all submissions to the Multi30K 2017 test set is shown in Table 1 for English to German, and Table 2 for English to French. The performance of all submissions to the Ambiguous COCO test set is shown in Table 3 for English to German, and Table 4 for English to French. For the Ambiguous COCO test set, no human evaluation was performed.

3.1.2. System ranking correlation

We observe that our evaluation of MLT Accuracy is mostly consistent with Meteor and human scores. To measure the extent of this consistency, we computed the Spearman's rank correlation coefficient ρ_s and Pearson's Correlation Coefficient ρ_p between MLT and Meteor or MLT and Human in Tables 1, 2, 3, and 4. The results are as follows.

For EN-DE on Multi30K 2017 test set (Table 1)
$\rho_s(\text{MLT}, \text{Meteor}) = 0.94 \quad \rho_s(\text{MLT}, \text{Human}) = 0.90$
$\rho_p(\text{MLT}, \text{Meteor}) = 0.99 \quad \rho_p(\text{MLT}, \text{Human}) = 0.78$

For EN-FR on Multi30K 2017 test set (Table 2)
$\rho_s(\text{MLT}, \text{Meteor}) = 0.93 \quad \rho_s(\text{MLT}, \text{Human}) = 0.54$

System	**MLT** ↑	Meteor ↑	Human ↑
NICT_1_NMTrerank_C	**75.49**	53.9	70.3
LIUMCVC_NMT_C	**74.70**	53.8	65.1
LIUMCVC_MNMT_C	**73.78**	54.0	77.8
DCU-ADAPT_MultiMT_C	**71.54**	50.5	68.1
UvA-TiCC_IMAGINATION_U	**70.75**	53.5	74.1
UvA-TiCC_IMAGINATION_C	**70.75**	51.2	59.7
CUNI_NeuralMonkeyTextualMT_U	**69.96**	51.0	68.1
CUNI_NeuralMonkeyMultimodalMT_U	**69.30**	50.2	60.6
OREGONSTATE_2NeuralTranslation_C	**68.64**	50.6	54.4
CUNI_NeuralMonkeyTextualMT_C	**67.72**	49.2	54.2
CUNI_NeuralMonkeyMultimodalMT_C	**64.95**	47.1	55.9
OREGONSTATE_1NeuralTranslation_C	**64.82**	48.9	53.3
SHEF_ShefClassProj_C	**60.74**	43.4	49.4
SHEF_ShefClassInitDec_C	**60.47**	44.5	46.6
AFRL-OHIOSTATE-MULTIMODAL_U	**23.06**	20.2	36.6

Table 1: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Multi30K 2017 test set for English to German.

System	**MLT** ↑	Meteor ↑	Human ↑
NICT_1_NMTrerank_C	**82.50**	72.0	79.4
LIUMCVC_NMT_C	**81.34**	70.1	60.5
LIUMCVC_MNMT_C	**81.23**	72.1	71.2
DCU-ADAPT_MultiMT_C	**81.00**	70.1	74.1
OREGONSTATE_2NeuralTranslation_C	**78.68**	68.3	65.4
OREGONSTATE_1NeuralTranslation_C	**75.78**	67.2	60.8
CUNI_NeuralMonkeyTextualMT_C	**75.55**	67.0	61.9
CUNI_NeuralMonkeyMultimodalMT_C	**74.97**	67.2	74.2
SHEF_ShefClassInitDec_C	**73.70**	62.8	54.7
SHEF_ShefClassProj_C	**72.42**	61.5	54.0

Table 2: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Multi30K 2017 test set for English to French.

System	**MLT** ↑	Meteor ↑
DCU-ADAPT_MultiMT_C	**68.50**	46.8
LIUMCVC_NMT_C	**68.24**	48.9
NICT_1_NMTrerank_C	**67.19**	48.5
UvA-TiCC_IMAGINATION_C	**67.19**	45.8
LIUMCVC_MNMT_C	**66.40**	48.8
CUNI_NeuralMonkeyMultimodalMT_U	**65.35**	45.6
UvA-TiCC_IMAGINATION_U	**64.30**	48.1
CUNI_NeuralMonkeyTextualMT_U	**63.78**	46.0
OREGONSTATE_1NeuralTranslation_C	**62.99**	46.5
OREGONSTATE_2NeuralTranslation_C	**62.99**	45.7
CUNI_NeuralMonkeyTextualMT_C	**62.99**	43.8
CUNI_NeuralMonkeyMultimodalMT_C	**57.74**	42.7
SHEF_ShefClassProj_C	**55.64**	40.0
SHEF_ShefClassInitDec_C	**54.33**	40.7

Table 3: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Ambiguous COCO test set for English to German.

System	**MLT** ↑	Meteor ↑
LIUMCVC_MNMT_C	**77.55**	65.9
NICT_1_NMTrerank_C	**77.55**	65.6
DCU-ADAPT_MultiMT_C	**76.42**	64.1
LIUMCVC_NMT_C	**75.28**	63.4
OREGONSTATE_2NeuralTranslation_C	**74.83**	63.8
CUNI_NeuralMonkeyTextualMT_C	**74.83**	62.5
CUNI_NeuralMonkeyMultimodalMT_C	**74.83**	62.5
OREGONSTATE_1NeuralTranslation_C	**70.75**	61.6
SHEF_ShefClassProj_C	**68.93**	57.0
SHEF_ShefClassInitDec_C	**68.48**	57.3

Table 4: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Ambiguous COCO test set for English to French.

$\rho_p(\text{MLT}, \text{Meteor}) = 0.94 \quad \rho_p(\text{MLT}, \text{Human}) = 0.68$

For EN-DE on Ambiguous COCO test set (Table 3)
$\rho_s(\text{MLT}, \text{Meteor}) = 0.80 \quad \rho_p(\text{MLT}, \text{Meteor}) = 0.90$

For EN-FR on Ambiguous COCO test set (Table 4)
$\rho_s(\text{MLT}, \text{Meteor}) = 0.95 \quad \rho_p(\text{MLT}, \text{Meteor}) = 0.96$

Ranking of the systems using MLT Accuracy differs only slightly from the ranking using Meteor or human scores, and the top performing systems are often the same. On further inspection, we notice that the high correlation of MLT Accuracy and Meteor is mainly due to words with skewed distributions for their translations.

For instance, the ambiguous word *lean* has the following translations in German in our training set - *lehnen* (to be leaning), *schlank* (slim), *stützen* (support), and *beugen* (bend). However, in the training set *lehnen* occurs 137 times while the rest of the lexical translations combined occur only 16 times. Such a skewed distribution makes the word *lean* virtually unambiguous (or less ambiguous) compared to the cases when the distribution is more uniform over the translations. We propose a way to deal with such skewed distributions using a simple heuristic we call the 'ambiguity score'.

3.1.3. Ambiguity Score

On inspecting the generated MLT Dataset, we noticed that ambiguity of words is not a simple concept to define and measure. Some words appear to be more ambiguous than others based on the distribution of their translations in the training set, while other words, like *lean* (see Section 3.1.2.), appear less ambiguous in the training set due to the skewed distribution of its translations. We propose to quantify the extent of ambiguity using a simple heuristic that looks at the distribution of the translations.

Consider a word en in English with *n* different translations in German $de_1, de_2, ..., de_n$. Let $freq(de_i|en)$ denote the number of times the word de_i occurs as a translation of en in the training set. Also, without loss of generality, arrange the translations in decreasing order of frequency, i.e. $freq(de_1|en) > freq(de_2|en) > ... > freq(de_n|en)$. Then we define ambiguity score of en as:

$$Ambiguity(en) = \frac{\sum_{i=2}^{n} freq(de_i|en)}{freq(de_1|en)} \qquad (2)$$

Using the above formulation, an ambiguity score of zero signifies unambiguous words and closer to zero signifies low ambiguity. The higher the ambiguity score, the more difficult it is to translate/disambiguate the source word correctly. Thus, to increase the difficulty of the MLT Dataset we can filter out words with ambiguity scores below a certain threshold. To demonstrate this, we set an ambiguity threshold of 0.2 and filter out all those words in the MLT Dataset with ambiguity score below this threshold. We then evaluate the WMT 2017 participating systems using MLT accuracy on this difficult version of the MLT Dataset (denoted as $\text{MLT}_{0.2}$). The results are shown in Tables 5, 6, 7 and 8:

System	**$\text{MLT}_{0.2}$** ↑	Meteor ↑	Human ↑
NICT_1_NMTrerank_C	**69.08**	53.9	70.3
LIUMCVC_NMT_C	**69.08**	53.8	65.1
LIUMCVC_MNMT_C	**68.10**	54.0	77.8
DCU-ADAPT_MultiMT_C	**65.75**	50.5	68.1
UvA-TiCC_IMAGINATION_U	**65.36**	53.5	74.1
CUNI_NeuralMonkeyTextualMT_U	**63.99**	51.0	68.1
UvA-TiCC_IMAGINATION_C	**63.01**	51.2	59.7
OREGONSTATE_2NeuralTranslation_C	**62.23**	50.6	54.4
CUNI_NeuralMonkeyMultimodalMT_U	**61.64**	50.2	60.6
CUNI_NeuralMonkeyTextualMT_C	**61.25**	49.2	54.2
OREGONSTATE_1NeuralTranslation_C	**59.69**	48.9	53.3
CUNI_NeuralMonkeyMultimodalMT_C	**58.71**	47.1	55.9
SHEF_ShefClassProj_C	**56.16**	43.4	49.4
SHEF_ShefClassInitDec_C	**53.82**	44.5	46.6
AFRL-OHIOSTATE-MULTIMODAL_U	**18.40**	20.2	36.6

Table 5: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Multi30K 2017 $\text{MLT}_{0.2}$ subset for English to German.

System	**$\text{MLT}_{0.2}$** ↑	Meteor ↑	Human ↑
NICT_1_NMTrerank_C	**71.43**	72.0	79.4
LIUMCVC_MNMT_C	**71.17**	72.1	71.2
DCU-ADAPT_MultiMT_C	**69.61**	70.1	74.1
LIUMCVC_NMT_C	**69.09**	70.1	60.5
OREGONSTATE_2NeuralTranslation_C	**65.71**	68.3	65.4
OREGONSTATE_1NeuralTranslation_C	**63.38**	67.2	60.8
CUNI_NeuralMonkeyTextualMT_C	**60.78**	67.0	61.9
SHEF_ShefClassInitDec_C	**60.00**	62.8	54.7
CUNI_NeuralMonkeyMultimodalMT_C	**59.48**	67.2	74.2
SHEF_ShefClassProj_C	**58.44**	61.5	54.0

Table 6: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Multi30K 2017 $\text{MLT}_{0.2}$ subset for English to French.

System	**$\text{MLT}_{0.2}$** ↑	Meteor ↑
DCU-ADAPT_MultiMT_C	**65.59**	46.8
LIUMCVC_NMT_C	**64.16**	48.9
UvA-TiCC_IMAGINATION_C	**63.44**	45.8
NICT_1_NMTrerank_C	**60.93**	48.5
UvA-TiCC_IMAGINATION_U	**60.93**	48.1
LIUMCVC_MNMT_C	**59.50**	48.8
OREGONSTATE_2NeuralTranslation_C	**59.14**	45.7
CUNI_NeuralMonkeyMultimodalMT_U	**59.14**	45.6
CUNI_NeuralMonkeyTextualMT_U	**57.71**	46.0
OREGONSTATE_1NeuralTranslation_C	**56.99**	46.5
CUNI_NeuralMonkeyTextualMT_C	**56.63**	43.8
CUNI_NeuralMonkeyMultimodalMT_C	**51.61**	42.7
SHEF_ShefClassInitDec_C	**50.54**	40.7
SHEF_ShefClassProj_C	**49.10**	40.0

Table 7: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Ambiguous COCO $\text{MLT}_{0.2}$ subset for English to German.

System	**$\text{MLT}_{0.2}$** ↑	Meteor ↑
LIUMCVC_MNMT_C	**66.83**	65.9
NICT_1_NMTrerank_C	**65.83**	65.6
OREGONSTATE_2NeuralTranslation_C	**64.82**	63.8
DCU-ADAPT_MultiMT_C	**64.32**	64.1
LIUMCVC_NMT_C	**63.32**	63.4
CUNI_NeuralMonkeyTextualMT_C	**61.81**	62.5
CUNI_NeuralMonkeyMultimodalMT_C	**61.81**	62.5
OREGONSTATE_1NeuralTranslation_C	**58.29**	61.6
SHEF_ShefClassProj_C	**55.78**	57.0
SHEF_ShefClassInitDec_C	**53.77**	57.3

Table 8: Performance of systems submitted to the Multimodal Shared Task at WMT 2017 on Ambiguous COCO $\text{MLT}_{0.2}$ subset for English to French.

The first thing to notice is that for every system the MLT accuracy drops when evaluated on $MLT_{0.2}$ (Tables 5, 6, 7 and 8) as compared to the full MLT Dataset (Tables 1, 2, 3 and 4). This shows that by setting an ambiguity threshold we are extracting ambiguous words which are more difficult to translate/disambiguate.
In general, for any threshold τ, we can extract a subset MLT_τ of the MLT Dataset consisting only of words with ambiguity score $\geq \tau$. In other words, the threshold τ can be used to regulate the difficulty of the MLT Dataset. Also, system rankings change as threshold τ changes. This can help in error analysis and identify the strengths and weaknesses of the systems.

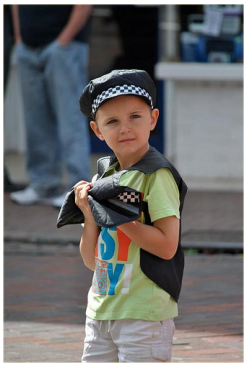
3.1.4. Analysis

According to MLT accuracy, for the teams that submitted both constrained and unconstrained models (those using additional external data for training), unconstrained models show improvement over their constrained counterparts in most cases (see Tables 1 and 5). For teams that submitted both multimodal and text-only systems, the role of multimodality is not evident as far as MLT Accuracy is concerned: sometimes multimodal systems perform better and sometimes text-only systems perform better. However, Human scores show that overall multimodal systems tend to be better than the text-only counterparts. MLT Accuracy fails to show this maybe because in its current form the matching performed between reference and system words is still too simplistic and does not take synonyms into account.

(a) hut

(b) kappe

(c) mütze



(d) kopfbedeckung

Figure 3: Different kinds of 'hats' translated into German differently based on the visual context in Multi30K corpus

Qualitative example In our English-German MLT Dataset, the word *hat* has been identified ambiguous because professional translators have translated it differently depending on the textual and visual contexts. Sometimes it has been translated as *hut* which refers to hats with edges/extensions coming off from all sides and usually worn in summer (see Figure 3a). Sometimes it has been translated as *kappe* which refers to the modern caps with shade extending out from front side only, usually worn in sports (see Figure 3b). Sometimes it has been translated as *mütze* which refers to differently designed hats usually worn in winters (see Figure 3c) and sometimes as *kopfbedeckung* which means a headgear which could refer to any kind of object worn on the head (see Figure 3d).
Now consider the following MLT data point from the 2016 test set whose textual context $\mathbf{x}$ was translated by the systems submitted to the Multimodal Shared Task:

Ambiguous Word x: *hat*
Lexical Translation y: *hut*
Textual Context $\mathbf{x}$: *"a man in an orange hat starring at something."*
Visual Context $\mathbf{v}$:

While most systems translated *hat* to *hut* in this example, 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 as seen earlier in Figure 3. From the image - visual context $\mathbf{v}$ - it can be seen that this hat looks a bit unusual because of its colour, texture and brand logo on it. Perhaps, that could be a reason why some systems chose another translation instead of *hut*. This can be considered as an instance where ambiguity is being introduced by the image.

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 form of the correct translation that does not appear in our gold standard lexical translation will be considered an error. This could be partly addressed by performing morphological analysis during the matching process. Secondly, no partial credit is given to synonymous words. This is a more difficult issue to address as synonyms can also have different senses. For instance, in the *hat* example discussed in the previous Section 3.1.4. all systems translated *hat* into some form of hat, which could be considered synonymous to some extent. Maybe not the correct kind of hat. Our future work will be focused on developing more elaborate scoring for MLT evaluation.

Additionally, we are interested in understanding whether the disambiguation happening within the system is due to the textual context, the visual context, both or none. For this, we propose to use a (Multimodal) Machine Translation system to translate the MLT data in four different ways. Recall from equation 1, MLT data is of the form $(x, y, \mathbf{x}, \mathbf{v})$. Given a Machine Translation System S, we should be able to compare four kinds of output with the reference.

$$S(\mathbf{x}, \mathbf{v}) \sim S(\mathbf{x}, \mathbf{0}) \sim S(x, \mathbf{v}) \sim S(x, \mathbf{0}) \sim y \quad (3)$$

where $\mathbf{0}$ refers to absence of visual context (no image) and $\sim$ refers to comparison of the different outputs. 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 was demonstrated in section 3.1.1., in future we will work on these *intra-system* comparisons depicted in equation 3.

5. Conclusion

We introduced the Multimodal Lexical Translation language resource and the process of generating it from Multi30K using word alignments followed by human filtering. 53,868 MLT data points for English to German and 44,779 MLT data points for English to French have been generated.

Different versions of MLT tasks were also introduced. We demonstrated the use of MLT task to evaluate Multimodal and Text-only Machine Translation systems' ability to translate ambiguous words correctly. For this, submissions to the WMT 2017 Multimodal Shared task were evaluated using a 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). Further, we introduced a simple heuristic to quantify ambiguity based on the distribution of the translations in the training set and demonstrated its use to extract more ambiguous subset of the MLT Dataset which was found to be more difficult to translate/disambiguate.

We observed that in most cases unconstrained (Multimodal) Machine Translation models perform better than their constrained counterparts in terms of MLT accuracy. The contribution of multimodality in machine translation has not yet proved evident in terms of MLT. The qualitative example of *hat* in Section 3.1.4. showed us that the MLT Dataset can be useful to compare different Machine Translation Systems.

We believe the multimodal multilingual MLT Dataset is a useful language resource that can facilitate, among many things, the study of lexical disambiguation within Multimodal and Text-only Machine Translation systems.

6. Acknowledgements

This work is supported by the MultiMT project (H2020 ERC Starting Grant No. 678017). The authors also thank Mareike Hartmann, Charles Escudier, Julia Ive, Frédéric 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 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2003) 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.

Denkowski, M. and Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2013)*, pages 644–649.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, 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*, pages 215–233.

Gella, S., Lapata, M., and Keller, F. (2016). Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, pages 182–192.

Graham, Y., Baldwin, T., Moffat, A., and Zobel, J. (2017). Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

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 the 45th annual meeting of the Association for Computational Linguistics*, 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 the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 547–554.

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

Raganato, A., Bovi, C. D., 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 1156–1167.

Riedl, M. and Biemann, C. (2016). Unsupervised compound splitting with distributional semantics rivals supervised methods. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 617–622.

Saenko, K. and Darrell, T. (2009). Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems*, pages 1393–1400.

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.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.