Analysing the Correlation between Lexical Ambiguity and Translation Quality in a Multimodal Setting using WordNet

Ali Hatami, Paul Buitelaar and Mihael Arcan
Insight SFI Research Centre for Data Analytics,
Data Science Institute, National University of Ireland Galway
firstname.lastname@insight-centre.org

Abstract

Multimodal Neural Machine Translation is focusing on using visual information to translate sentences in the source language into the target language. The main idea is to utilise information from visual modalities to promote the output quality of the text-based translation model. Although the recent multimodal strategies extract the most relevant visual information in images, the effectiveness of using visual information on translation quality changes based on the text dataset. Due to this, this work studies the impact of leveraging visual information in multimodal translation models of ambiguous sentences. Our experiments analyse the Multi30k evaluation dataset and calculate ambiguity scores of sentences based on the WordNet hierarchical structure. To calculate the ambiguity of a sentence, we extract the ambiguity scores for all nouns based on the number of senses in WordNet. The main goal is to find in which sentences, visual content can improve the text-based translation model. We report the correlation between the ambiguity scores and translation quality extracted for all sentences in the English-German dataset.

1 Introduction

In recent years, Neural Machine Translation (NMT) model is widely used in translation tasks and represents remarkable performance in terms of fluency and precision compared with the previous generations of machine translation. Recurrent Neural Network (RNN)-based NMT with Attention mechanism has found broad application in different fields of NLP tasks such as machine translation. The transformer model as a Self-attention based model has been introduced by Google in 2017 as a new architecture for NMT (Vaswani et al., 2017). The self-attention mechanism uses cross-lingual attention that allows the input words to interact with each other (self) and find out which one should pay more attention to (attention). In addition to the mechanism of cross-lingual attention, the transformer model uses a stacked self-attention layer that follows with a point-wise feed-forward component. Recently many studies in machine translation have been increasingly focusing on using visual content well as textual to improve the translation quality. Therefore, Multimodal Neural Machine Translation (MNMT) as a subarea of NMT has been introduced to use visual information extracted from other modalities such as speech, image or video to translate a sentence in a source language into the target language.

MNMT is an area of research that plays an important role in machine translation tasks since multimodal resources have been increasingly used in deep learning techniques. MNMT tries to extend the ability of the NMT models by taking visual context such as images as an additional input to better translate the source text. The main idea behind this is that the textual context does not provide sufficient information for the text-based NMT model in some situations to translate ambiguous sentences (ambiguous terms or grammatical gender). Due to this, visual information can enrich text-best NMT systems by adding extra information to disambiguate the input words and provide correct translations on the target side.

One of the main ideas of using multimodality in Machine Translation is that visual information can help the textual context to find the correct sense of ambiguous words in the translation process of the source sentence. For example, the word "track" in the English sentence "A man is performing a trick on a track" is an ambiguous word and could have at least two different translations in German – (1) "*Ein Mann führt einen Trick auf einer Strecke aus*", and (2) "*Ein Mann führt einen Trick auf einem Bahngleis aus*". Given the word "track", the context does not provide enough information to disambiguate and translate it correctly. Therefore, multimodal resources such as images can guide the

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop, pages 89 - 95

translation system to select the correct sense based on the visual information. Word Sense Disambiguation (WSD) is widely studied in different natural language processing tasks. WSD analyses given the context of an ambiguous word to assign the correct sense based on a pre-defined sense net for words. Visual Sense Disambiguation (VSD) as a modified version of WSD use visual context instead of textual to disambiguate words. Although disambiguation of word sense can be done directly by Machine Translation models, research on Multimodal Machine Translation more focuses on analysing of contributions of each modality to disambiguate words in the translation process.

In this work, we focus on identifying ambiguous sentences and leverage therefore the WordNet hierarchical structure to calculate an ambiguity score for each sentence. This is then used to study a correlation between ambiguity and translation evaluation scores. Analysing the lexical ambiguity and translation quality allowed us to identify sentences that are more challenging in the translation process and most likely visual content can help the text-based NMT to translate sentences more accurate.

2 Related Work

Multimodal Machine Translation is a new trend in machine translation tasks that aims to create multimodal frameworks to use information from visual modality as well as text context (Specia et al., 2016). Different practices were used for the visual part of the MMT framework. The common approach is to extract visual information by using Convolutional Neural Networks (CNN) and then integrate this information with textual features (Yao and Wan, 2020). Many MMT models were developed based on the Transformer approach. The transformer approach extracts the relationships between words in the source and target sentences by using a multihead self-attention mechanism (Vaswani et al., 2017)

In some studies, the global image features are used in the encoder beside word sequences to use both types of features in the decoding stage (Huang et al., 2016) or used to initialise the hidden parameters of the encoder and decoder in RNN (Calixto and Liu, 2017). (Caglayan et al., 2017) use elementwise multiplication to initialise hidden states of encoder/decoder in the attention-based model. (Zhou et al., 2018) links visual and corresponding text semantically by using a visual attention mechanism.

Despite successfully using multimodal information in MMT, recent studies show that most of the information in the image is not related to the text while the translation process and when there is limited textual information, visual content plays more important for the translation model (Caglayan et al., 2019). The studies use visual features by focusing on relative importance among different modalities. (Lala et al., 2018) introduced a multimodal cross-lingual word sense disambiguation model based on Multimodal Lexical Translation Dataset (MLTD) (Lala and Specia, 2018) to generate contextually correct translations for the ambiguous words. MLTD includes a list of words of the source language with multiple translations in the training set of Multi30k. (Ive et al., 2019) introduced a translate-and-refine mechanism by using images in a second stage decoder to refine the text-based NMT model in the ambiguous words listed in MLT dataset. (Calixto et al., 2019) use a latent variable model to extract the multimodal relationships between modalities. Recent methods try to reduce the noise of visual information and select visual features related to the text. (Yao and Wan, 2020) use a multimodal transformer-based self-attention to encode relevant information in images. To capture various relationships, (Yin et al., 2020) propose a graph-based multimodal fusion encoder.

3 Experimental Setup

This section provides insights on the dataset used in this work, neural architectures and the translation evaluation metric BLEU.

3.1 Multi30K Dataset

Multi30K (Elliott et al., 2016) is an extended version of the Flickr30K dataset that includes images and paired descriptions expressed by one English sentence and translated sentences in multiple languages. Firstly, the German translation was added to the dataset (Young et al., 2014) and then it extended to French and Czech (Elliott et al., 2017) (Barrault et al., 2018). Many recent models in MNMT have focused on Multi30K as it provides an image for each sentence in English and three translation directions, i.e. in German, French and Czech. In this study, the evaluation dataset of Multi30k contains 1,000 instances.

3.2 Text-based NMT

OpenNMT (Klein et al., 2018) is used to train the text-based NMT model on a general En-De dataset. The model used a 6-layer transformer mechanism for both the encoder and decoder stage. We trained the model for 50,000 steps on a general dataset and set the parameters of the model to the original implementations of OpenNMT.

As the text-based NMT system cannot leverage the visual information, and to ensure a broad lexical and domain coverage of our text-based NMT system, we merged existing parallel for the English-German language pair from the OPUS web page[1] into one parallel corpus, i.e., Europarl (Koehn, 2005), DGT (Steinberger et al., 2014), EMEA, KDE4, OpenOffice (Tiedemann, 2009), OpenSubtitles2012 (Tiedemann, 2012), and randomly selected 10 million sentences for our training step.

3.3 Doubly-attentive MNMT

For the visual side, we used the model that proposed in (Zhao et al., 2020) to apply semantic image region features[2] for MNMT. This model is based on the Doubly-attentive mechanism (Calixto and Liu, 2017) to integrate visual and textual features by applying 100 semantic image features with a dimension of 2,048 at each time step. The hidden state dimension of the visual model is 500 for both 2-layer GRU encoder and 2-layer GRU decoder. The work also set the dimension of the source word embedding to 500, batch size to 400, beam size to 5, text dropout to 0.3, and image region dropout to 0.5. After training the model for 25 epochs using stochastic gradient descent with ADADELTA (Zeiler, 2012) and a learning rate of 0.002, the model of epoch 16 has been selected based on comparing BLEU scores of the final models.

3.4 Evaluation Metric

We report the automatic evaluation based on BLEU for the automatic evaluation. BLEU (Papineni et al., 2002) is calculated for individual translated segments (n-grams) by comparing them with a dataset of reference translations. For this work we use the *sacrebleu*[3] library (Post, 2018).

[1] https://opus.nlpl.eu/

[2] https://github.com/Zhao-Yuting/MNMT-with-semantic-regions

[3] https://github.com/mjpost/sacrebleu

3.5 Princeton WordNet

Princeton WordNet (Fellbaum, 1998) is a manually created resource that has been used in many different tasks and applications across linguistics and natural language processing. WordNet's hierarchical structure makes it a useful tool for many semantic applications and it also plays a vital role in various deep learning approaches (Rychalska et al., 2016).

3.6 Correlation Coefficients

The correlation coefficient is a measure to determine the relationship between two variables (Janse et al., 2021). In correlated data, the change in the magnitude of one variable leads to a change in the magnitude of another variable either in the same or in the opposite directions. Pearson product-moment correlation is a typical type of correlation for a linear relationship between two continuous variables. The range of the correlation coefficient is between -1 and +1, where 0 shows that there is no correlation between the two variables. The correlation coefficient near +1 and -1 shows a strong, same or opposite, correlation respectively. The equation for the correlation coefficient is:

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

where $\bar{x}$ and $\bar{y}$ are the sample means of array *X* and *Y* respectively.

4 Methodology

In this section, we explain our methodology to calculate the ambiguity scores for each sentence based on the hierarchical structure of WordNet. To find a meaningful relationship between ambiguity and translation quality, we analyse the correlation functions between different ambiguity scores and the translation evaluation metric BLEU. Our focus in this work is on the inherited structure of English nouns in WordNet. Each noun in WordNet can be defined as a set *W* of pairs *(w,s)* where *w* is a word in that language and a sense *s* is possible set of meanings (synonyms or *synsets*) for the word *w*. Table 1 shows all *synset* entries (11) for the noun *track* in WordNet. The inherited structure in WordNet is a hierarchical structure to organise the semantic relations of *synsets*. Furthermore, synsets in WordNet have different hierarchical structures from each other including *hyponymy* and *hypernymy*. Figure 1 shows the WordNet inherited structure of *synset* entries for the word *track*. *Entity*

path, track, course	a line or route along which something travels or moves
lead, track, trail	evidence pointing to a possible solution
track	a pair of parallel rails providing a runway for wheels
racetrack, racecourse, raceway, track	a course over which races are run
cut, track	a distinct selection of music from a recording or a compact disc
track, caterpillar track, caterpillar tread	an endless metal belt on which tracked vehicles move over the ground
track, data track	one of the circular magnetic paths on a magnetic disk that serve ... for writing and reading data
track	a groove on a phonograph recording
track, rail, rails, runway	a bar or pair of parallel bars of rolled steel making the railway along which railroad ... can roll
track, cart track, cartroad	any road or path affording passage especially a rough one
track, running	the act of participating in an athletic competition involving running on a track

Table 1: Synset entries (11) for the word *track* in the Princeton WordNet.

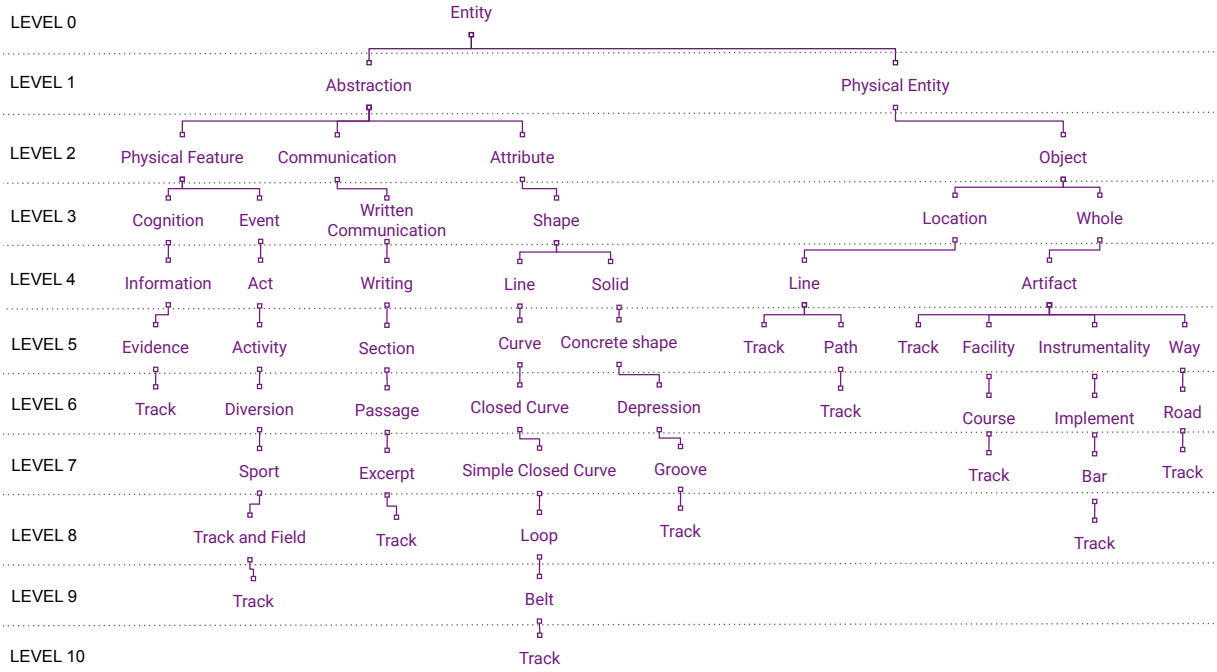

Figure 1: Hierarchical structure of the WordNet entry *track*.

(level 0), is the root node for all *synset* entries in WordNet. Each path between the root node and a *synset* entry has a different length that shows the different abstraction level. For example of the word *track*, `min_length` has a path length of 4, with six unique abstract concepts (*Information, Act, Writing, Line, Solid, Artifact*). On the other hand, `min_length-1` at the path length of 3, has six concepts as well, i.e. *Cognition, Event, Written Communication, Shape, Location, Whole*. The number of all *synsets* for *track* in WordNet is 11. After extracting this information for each word, we use the `sum` and `multiply` functions on all nouns of a sentence to calculate the overall ambiguity score (see example in Table 2 for the sentence *Dog runs at a track*). We normalised these scores by dividing them by the number of content words (nouns with more than one synset in WordNet) of the sentence to minimise the effect of sentence length on our experiments.

5 Results

This section provides the results of our experiments. After calculating ambiguity and BLEU scores (NMT, MNMT) for each sentence in the test set, we analysed the correlation coefficients between ambiguity and translation quality scores to find a meaningful relationship between them. To better analyse the correlation between the sentence ambiguity and translation quality, we grouped them into sets of 50 sentences (resulting in 20 groups) after ranking them by the ambiguity score. The corpus BLEU scores for NMT and MNMT on the evaluation dataset in En-De are 30.66 and 35.80 respectively.

Table 3 illustrates the correlation score (see Sec-

Approach	# of Concepts	# Nouns	Ambiguity
Sum(synsets)	7 + 11	2	9.0
Sum(min_length)	7 + 10	2	8.5
Sum(min_length-1)	6 + 6	2	6.0
Multiply(synsets)	7 * 11	2	38.5
Multiply(min_length)	7 * 10	2	35.0
Multiply(min_length-1)	6 * 6	2	18.0

Table 2: Examples of calculating the ambiguity score based on the number of concepts of each word, i.e. *dog* and *track*, at the certain hierarchical level, normalised with the set of nouns in the sentence.

Approach	NMT	MNMT
Sum(Synsets)	0.3987	0.3841
Sum(min_length)	0.2226	0.0445
Sum(min_length-1)	0.1017	-0.0453
Multiply(Synsets)	-0.5511	-0.6744
Multiply(min_length)	-0.5846	-0.6020
Multiply(min_length-1)	-0.5292	-0.6039

Table 3: Correlation between the calculated ambiguity scores and BLEU metric for NMT and MNMT on 20 groups.

tion 3.6), ambiguity scores and the BLEU evaluation metric for the approaches used to calculate the ambiguity scores of the sentences. As seen in the table, the best correlations for NMT and MNMT are obtained by the `Multiply(min_length)` and `Multiply(Synsets)` approaches respectively. Due to this, we focused on the `Multiply` approaches and provide graphs, which illustrate the correlation between the ambiguity and translation quality.

As seen in Figure 2 the ambiguity score calculated by the WordNet hierarchy correlates with the translation quality, i.e., if the ambiguity of a sentence is high, the translation quality in terms of BLEU is low. On the other hand, if the ambiguity of a sentence is low, the translation quality in terms of the BLEU metric improves. This can be seen for all methods used to calculate the ambiguity, i.e. `synsets`, `min_length`, `min_length-1`. In addition to that, the graphs also illustrate the better performance of the MNMT system (orange points) compared to the text-based NMT system (blue points).

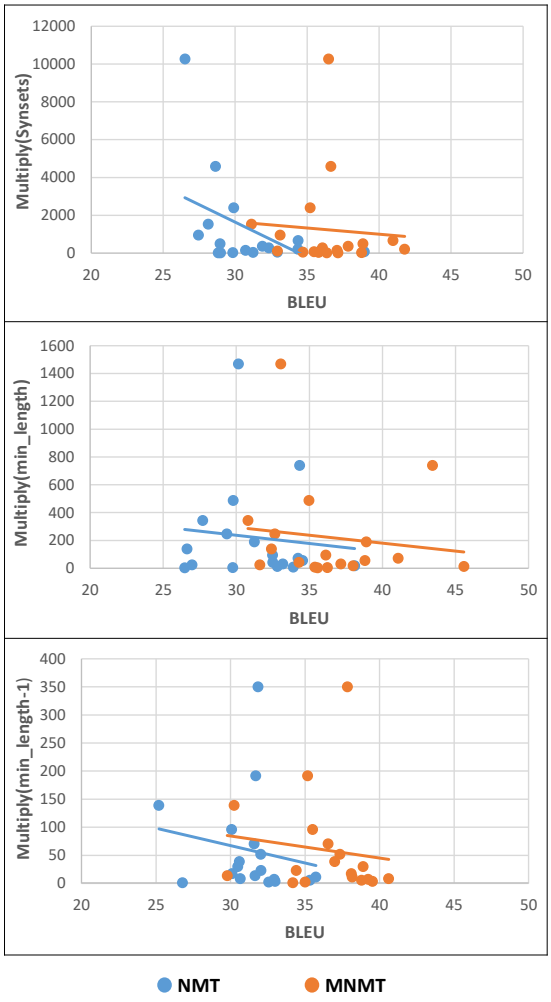

Figure 2: Correlation representation between the Multiply approach's ambiguity scores and the BLEU metric for NMT and MNMT on 20 groups.

6 Conclusion

Recent studies in Multimodal Machine Translation focused on using visual information to improve the quality of translation tasks. One of the main challenges for the translation systems is to find a correct translation in terms of the context used. Despite the progress of research in this area, the performance of multimodal translation systems is more related to the quality of visual content which is used along with textual dataset. In this study, we analysed different approaches to calculate the ambiguity of the sentence to find a correlation between sentence ambiguity and the translation quality in terms of the BLEU metric. We tested different approaches to calculate the ambiguity and observed that multiplying the number of entries at the minimum length level of the WordNet hierarchy for each noun provided the best correlation to the evaluation metric for each sentence. Within our future work, we plan to consider the frequency and further linguistic features of WordNet synsets. In addition to that, we plan to leverage the Polylingual Wordnet (Arcan et al., 2019), a large multilingual WordNet in more

than 20 European languages, to calculate the lexical ambiguity beyond English. Furthermore, we plan the incorporation of ImageNet (Deng et al., 2009), which has an image dataset organised according to the WordNet hierarchy.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission. We would like to thank the anonymous reviewers for their insights on this work.

References

Mihael Arcan, John P. McCrae, and Paul Buitelaar. 2019. Polylingual wordnet. *CoRR*, abs/1903.01411.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark. Association for Computational Linguistics.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark. Association for Computational Linguistics.

Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multi-modal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6392–6405, Florence, Italy. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 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, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based multimodal neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Julia Ive, Pranava Madhyastha, and Lucia Specia. 2019. Distilling translations with visual awareness. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6525–6538, Florence, Italy. Association for Computational Linguistics.

Roemer J Janse, Tiny Hoekstra, Kitty J Jager, Carmine Zoccali, Giovanni Tripepi, Friedo W Dekker, and Merel van Diepen. 2021. Conducting correlation analysis: important limitations and pitfalls. *Clinical Kidney Journal*, 14(11):2332–2337.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander Rush. 2018. OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*. AAMT.

Chiraag Lala, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia. 2018. Sheffield submissions for WMT18 multimodal translation shared task. In *Proceedings of the Third Conference on Machine*

Translation: Shared Task Papers, pages 624–631, Belgium, Brussels. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Barbara Rychalska, Katarzyna Pakulska, Krystyna Chodorowska, Wojciech Walczak, and Piotr Andruszkiewicz. 2016. Samsung Poland NLP team at SemEval-2016 task 1: Necessity for methods to measure semantic similarity. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 614–620.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Ralf Steinberger, Mohamed Ebrahim, Alexandros Poulis, Manuel Carrasco-Benitez, Patrick Schlüter, Marek Przybyszewski, and Signe Gilbro. 2014. An overview of the european union's highly multilingual parallel corpora. *Language Resources and Evaluation*, 48(4):679–707.

Jorg Tiedemann. 2009. News from OPUS – A Collection of Multilingual Parallel Corpora with Tools and Interfaces. In *Advances in Natural Language Processing*, volume V, chapter V, pages 237–248. Borovets, Bulgaria.

Jörg Tiedemann. 2012. Character-based pivot translations for under-resourced languages and domains. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 141–151, Avignon, France.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online. Association for Computational Linguistics.

Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *ACL*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Yuting Zhao, Mamoru Komachi, Tomoyuki Kajiwara, and Chenhui Chu. 2020. Double attention-based multimodal neural machine translation with semantic image regions. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 105–114, Lisboa, Portugal. European Association for Machine Translation.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium. Association for Computational Linguistics.