Cross-lingual Visual Verb Sense Disambiguation

Spandana Gella[*] and **Desmond Elliott**[†] and **Frank Keller**[*]
[*]School of Informatics, University of Edinburgh
[†]Department of Computer Science, University of Copenhagen
{spandana.gella,frank.keller}@ed.ac.uk, de@di.ku.dk

Abstract

Recent work has shown that visual context improves cross-lingual sense disambiguation for nouns. We extend this line of work to the more challenging task of cross-lingual *verb sense* disambiguation, introducing the MultiSense dataset of 9,504 images annotated with English, German, and Spanish verbs. Each image in MultiSense is annotated with an English verb and its translation in German or Spanish. We show that cross-lingual verb sense disambiguation models benefit from visual context, compared to unimodal baselines. We also show that the verb sense predicted by our best disambiguation model can improve the results of a text-only machine translation system when used for a multimodal translation task.

1 Introduction

Resolving lexical ambiguity remains one of the most challenging problems in natural language processing. It is often studied as a word sense disambiguation (WSD) problem, which is the task of assigning the correct sense to a word in a given context (Kilgarrif, 1998). Word sense disambiguation is typically tackled using only *textual context*; however, in a multimodal setting, *visual context* is also available and can be used for disambiguation. Most prior work on visual word sense disambiguation has targeted noun senses (Barnard and Johnson, 2005; Loeff et al., 2006; Saenko and Darrell, 2008), but the task has recently been extended to verb senses (Gella et al., 2016, 2019).

Resolving sense ambiguity is particularly crucial for translation tasks, as words can have more than one translation, and these translations often correspond to word senses (Carpuat and Wu, 2007; Navigli, 2009). As an example consider the verb *ride*, which can translate into German as *fahren* (ride a bike) or *reiten* (ride a horse). Recent work on multimodal machine translation has partly addressed

Source: Three guys **riding** on an elephant.
Target: Drei Männer **reiten** auf einem Elefanten.
Output: Drei Jungs **fahren** auf einem Elefanten.

Figure 1: An example of a verb sense translation error (shown in **bold red**) by the English-German neural translation system of Sennrich et al. (2017).

lexical ambiguity by using visual information, but it still remains unresolved especially for the part-of-speech categories such as verbs (Specia et al., 2016; Shah et al., 2016; Hitschler et al., 2016; Lala and Specia, 2018). Prior work on cross-lingual WSD has been limited in scale and has only employed textual context (Lefever and Hoste, 2013), even though the task should benefit from visual context, just like monolingual WSD.

Visual information has been shown to be useful to map words across languages for bilingual lexicon induction. For this, images are used as a pivot between languages or visual information is combined with cross-lingual vector spaces to learn word translations across languages (Bergsma and Van Durme, 2011; Kiela et al., 2015; Vulic et al., 2016). However, as with other grounding or word similarity tasks, bilingual lexicon induction has so far mainly targeted nouns and these approaches was shown to perform poorly for other word categories such as verbs. Recent work by Gella et al. (2017) and Kádár et al. (2018) has shown using image as pivot between languages can lead to better multilingual multimodal representations and can have successful applications in crosslingual retrieval and

Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: Short Papers, 1998–2004. Minneapolis, 2019.

multilingual image retrieval.

In this paper, we introduce the MultiSense dataset of 9,504 images annotated with English verbs and their translations in German and Spanish. For each image in MultiSense, the English verb is translation-ambiguous, i.e., it has more than one possible translation in German or Spanish. We propose a series of disambiguation models that, given an image and an English verb, select the correct translation of the verb. We apply our models on MultiSense and find that multimodal models that fuse textual context with visual features outperform unimodal models, confirming our hypothesis that cross-lingual WSD benefits from visual context.

Cross-lingual WSD also has a clear application in machine translation. Determining the correct sense of a verb is important for high quality translation output, and sometimes text-only translation systems fail when the correct translation would be obvious from visual information (see Figure 1). To show that cross-lingual visual sense disambiguation can improve the performance of translation systems, we annotate a part of our MultiSense dataset with English image descriptions and their German translations. There are two existing multimodal translation evaluation sets with ambiguous words: the Ambiguous COCO dataset (Elliott et al., 2017) contains sentences that are "possibly ambiguous", and the Multimodal Lexical Translation dataset is restricted to predicting single words instead of full sentences (Lala and Specia, 2018). This type of resource is important for multimodal translation because it is known that humans use visual context to resolve ambiguities for nouns and gender-neutral words (Frank et al., 2018). MultiSense contains sentences that are known to have ambiguities, and it allows for sentence-level and verb prediction evaluation. Here, we use the verbs predicted by our visual sense disambiguation model to constrain the output of a neural translation system and demonstrate a clear improvement in Meteor, BLEU, and verb accuracy over a text-only baseline.

2 MultiSense Dataset

Images Paired with Verb Translations The MultiSense dataset pairs sense-ambiguous English verbs with images as visual context and contextually appropriate German and Spanish translations. Table 1 shows examples of images taken from MultiSense with their Spanish and German translations. To compile the dataset, we first chose a set of English verbs which had multiple translations into German and Spanish in Wiktionary, an online dictionary. Then we retrieved 150 candidate images from Google Images using queries that included the target English verb. We constructed the verb phrases by extracting the 100 most frequent phrases for each verb from the English Google syntactic n-grams dataset (Lin et al., 2012), which we then manually filtered to remove redundancies, resulting in 10 phrases per verb. Examples of verb phrases for *blow* include *blowing hair*, *blowing a balloon*, and *blowing up a bomb*. We filtered the candidate images using crowdworkers on Amazon Mechanical Turk, who were asked to remove images that were irrelevant to the verb phrase query. Overall pairwise agreement for this image filtering task was 0.763. Finally, we employed native German and Spanish speakers to translate the verbs into their language, given the additional visual context.

This resulted in a dataset of 9,504 images, covering 55 English verbs with 154 and 136 unique translations in German and Spanish, respectively. We divided the dataset into 75% training, 10% validation and 15% test splits.

Sentence-level Translations We also annotated a subset of MultiSense with sentence-level translations for English and German. This subset contains 995 image–English description–German translation tuples that can be used to evaluate the verb sense disambiguation capabilities of multimodal translation models. We collected the data in four-steps: (1) crowdsource English descriptions of the images using the gold-standard MultiSense verb as a prompt; (2) manually post-edit the English descriptions to ensure they contain the correct verb; (3) crowdsource German translations, given the English descriptions, the German gold-standard MultiSense verb, and the image; (4) manually post-edit the German translations to ensure they contain the correct verb. Figure 1 shows an example of an image paired with its English description and German translation.

3 Verb Sense Disambiguation Modeling

We propose three models for cross-lingual verb sense disambiguation, based on the visual input, the textual input, or using both inputs. Each model is trained to minimize the negative log probability of predicting the correct verb translation.



Spanish	mandar	hinchar	explotar
German	zublasen	aufblasen	detonieren

Table 1: Images for the English verb *blow* annotated with translations in Spanish and German. The images correspond to the uses of *blowing with a hair dryer* and *blowing a balloon*, and *blowing up a bomb*.

3.1 Unimodal Visual Model

Visual features have been shown to be useful for learning semantic representations of words (Lazaridou et al., 2015), bilingual lexicon learning (Kiela et al., 2015), and visual sense disambiguation (Gella et al., 2016), amongst others. We propose a model that learns to predict the verb translation using only visual input. Given an image $\mathbf{I}$, we extract a fixed feature vector from a Convolutional Neural Network, and project it into a hidden layer $\mathbf{h}_v$ with the learned matrix $\mathbf{W}_i \in \mathbb{R}^{h \times 512}$ (Eqn. 1). The hidden layer is projected into the output vocabulary of v verbs using the learned matrix $\mathbf{W}_o \in \mathbb{R}^{h \times v}$, and normalized into a probability distribution using a softmax transformation (Eqn. 2).

$$\mathbf{h}_v = \mathbf{W}_i \cdot \texttt{CNN}(\mathbf{I}) + \mathbf{b}_i \quad (1)$$

$$\mathbf{y} = \texttt{softmax}(\mathbf{W}_o \cdot \mathbf{h}_v + \mathbf{b}_o) \quad (2)$$

3.2 Unimodal Textual Model

Each image in MultiSense is associated with the query phrase that was used to retrieve it. Given a query phrase with N words, we embed each word as a d-dimensional dense vector, and represent the phrase as the average of its embeddings E. We then project the query representation into a hidden layer with the learned matrix $\mathbf{W}_q \in \mathbb{R}^{h \times d}$ (Eqn. 3). The hidden layer is projected into an output layer and normalized to a probability distribution, in the same manner as the unimodal visual model.

$$\mathbf{h}_q = \mathbf{W}_q \cdot \left(\frac{1}{N} \sum_{i}^{N} \mathrm{E}[w_i] \right) + \mathbf{b}_q \quad (3)$$

3.3 Multimodal Model

We also propose a multimodal model that integrates the visual and textual features to predict the correct verb sense. In our multimodal model, we concatenate the inputs together before projecting them into a hidden layer with a learned matrix $\mathbf{W}_h \in \mathbb{R}^{h \times (512+h)}$ (Eqn. 4). We follow the same steps as the unimodal models to project the multimodal hidden layers into the output label space.

$$\mathbf{h}_{early} = \mathbf{W}_h \cdot [\texttt{CNN}(\mathbf{I}) \,;\, \mathbf{h}_q] + \mathbf{b}_h \quad (4)$$

	Chance	Majority	Text	Image	MM
German	0.7	2.8	49.1	52.1	**55.6**
Spanish	0.7	4.0	52.7	50.3	**56.0**

Table 2: Cross-lingual verb sense disambiguation accuracy of our unimodal models and the multimodal model. We also show the performance of a random chance baseline and a majority label baseline.

4 Verb Disambiguation Experiments

Our experiments are designed to determine whether the integration of textual and visual features yields better cross-lingual verb sense disambiguation than unimodal models.

4.1 Setup and Evaluation

We embed the textual queries using pre-trained $d = 300$ dimension word2vec embeddings (Mikolov et al., 2013). We represent images in the visual model using the features extracted from the 512D `pool5` layer of a pre-trained ResNet-34 CNN (He et al., 2016). All our models have a $h = 128$ dimension hidden layer. The German models have an output vocabulary of $v = 154$ verbs, and the Spanish models have a vocabulary of $v = 136$ verbs. All of our models are trained using SGD with mini-batches of 16 samples and a learning rate of 0.0001.

We evaluate the performance of our models by measuring the accuracy of the predicted verb against the gold standard. We also compare against chance and majority label baselines. Our preliminary experiments show that with better visual representation we achieve better acccuracy scores similar to others who observed better visual representation contributes to better downstream tasks such as image description (Fang et al., 2015), multimodal machine translation (Specia et al., 2016) and representation learning (Kádár et al., 2018).

4.2 Results

We present the results in Table 2. The chance and majority label baselines perform very poorly. The unimodal textual model performs better than the

Source	A large herd of sheep is **blocking** the road.	A woman smiles as she **brushes** her long, dark hair.
Target	Eine große Herde Schafe **blockiert** die Straße.	Eine Frau lächelt während sie sich ihre dunklen langen Haare **bürstet** .
Baseline	Eine große Herde Schafe **kriecht** die Straße entlang.	Eine Frau lächelt , als sie ihren langen und dunklen Haaren **putzt** .
+WSD	Eine große Herde Schafe **blockieren** die Straße.	Eine Frau lächelt , als sie ihr lange , dunklen Haaren **bürsten** .

Table 3: The visual verb sense predictions ("blockieren", "bürsten") successfully constrains the decoder to predict the correct sense of the verb ("block", "brush") in the German translation **(+WSD)**. The incorrect verb in the baseline translation is shown in **bold red**.

looking
for directions

Model	German	Spanish
Textual	schauen	mirar
Visual	tragen	**buscar**
MM	**suchen**	**buscar**

Figure 2: Examples of the Top-1 predictions of our unimodal and multimodal models. Only the early fusion multimodal model predicts the correct verb sense for both languages (shown in bold).

	Meteor	BLEU	VAcc
Baseline NMT	38.6	17.8	22.9
+ Predicted Verb	40.0	18.5	49.5
+ Oracle Verb	40.4	19.1	77.7
Caglayan et al.	46.1	25.8	29.3
Helcl & Libovický	42.5	22.3	25.1

Table 4: Translation results: Meteor and BLEU are standard text-similarity metrics; verb accuracy (VAcc) counts how often the model proposal contains the gold standard German verb.

unimodal visual model for German verb sense disambiguation, but we find the opposite for Spanish unimodal verb sense disambiguation. However, the early fusion multimodal model outperforms the best unimodal model for both German and Spanish. This confirms that cross-lingual verb sense disambiguation benefits from multimodal supervision compared to unimodal supervision.

4.3 Discussion

We analyzed the outputs of our models in order to understand where multimodal features helped in identifying the correct verb translation and the cases where they failed. In Figure 2, we show an example that illustrates how varying the input (textual, visual, or multimodal) affects the accuracy of the verb prediction. We show the top verb predicted by our models for both German and Spanish. The top predicted verb using text-only visual features is incorrect. The unimodal visual features model predicts the correct Spanish verb but the incorrect German verb. However, when visual information is added to textual features, models in both the languages predict the correct label.

5 Machine Translation Experiments

We also evaluate our verb sense disambiguation model in the challenging downstream task of multimodal machine translation (Specia et al., 2016). We conduct this evaluation on the sentence-level translation subset of MultiSense. We evaluate model performance using BLEU (Papineni et al., 2002) and Meteor scores (Denkowski and Lavie, 2014) between the MultiSense reference description and the translation model output. We also evaluate the verb prediction accuracy of the output against the gold standard verb annotation.

5.1 Models

Our baseline is an attention-based neural machine translation model (Hieber et al., 2017) trained on the 29,000 English-German sentences in Multi30k (Elliott et al., 2016). We preprocessed the text with punctuation normalization, tokenization, and lowercasing. We then learned a joint byte-pair-encoded vocabulary with 10,000 merge operations to reduce sparsity (Sennrich et al., 2016).

Our approach uses the German verb predicted by the unimodal visual model (Section 3.1) to constrain the output of the translation decoder (Post and Vilar, 2018). This means that our approach does not directly use visual features, instead it uses the output of the visual verb sense disambiguation model to guide the translation process.

We compare our approach against two state-of-the-art multimodal translation systems: Caglayan et al. (2017) modulate the target language word embeddings by an element-wise multiplication with a learned transformation of the visual data; Helcl and Libovický (2017) use a double attention model that learns to selectively attend to a combination of the source language and the visual data.

5.2 Results

Table 4 shows the results of the translation experiment. Overall, the Meteor scores are much lower than on the Multi30k test sets, where the state-of-the-art single model scores 51.6 Meteor points compared to 46.1 Meteor we obtained. This gap is most likely due evaluating the models on an out-of-domain dataset with out-of-vocabulary tokens. Using the predicted verb as a decoding constraint outperforms the text-only translation baseline by 1.4 Meteor points. In addition, the translation output of our model contains the correct German verb 27% more often than the text-only baseline model. These results show that a multimodal verb sense disambiguation model can improve translation quality in a multimodal setting.

We also calculated the upper bound of our approach by using the gold standard German verb as the lexical constraint. In this oracle experiment we observed a further 0.4 Meteor point improvement over our best model, and a further 27% improvement in verb accuracy. This shows that: (1) there are further improvements to be gained from improving the verb disambiguation model, and (2) the OOV rate in German means that we cannot achieve perfect verb accuracy.

6 Conclusions

We introduced the MultiSense dataset of 9,504 images annotated with an English verb and its translation in Spanish and German. We proposed a range of cross-lingual visual sense disambiguation models and showed that multimodal models that fuse textual and visual features outperform unimodal models. We also collected a set of image descriptions and their translations, and showed that the output of our cross-lingual WSD system boosts the performance of a text-only translation system on this data. MultiSense is publicly available at `https://github.com/spandanagella/multisense`

Acknowledgements

DE was supported by an Amazon Research Award. This work was supported by the donation of a Titan Xp GPU by the NVIDIA Corporation.

References

Kobus Barnard and Matthew Johnson. 2005. Word sense disambiguation with pictures. *Artificial Intelligence*, 167(1-2):13–30.

Shane Bergsma and Benjamin Van Durme. 2011. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, page 1764.

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*.

Marine Carpuat and Dekai Wu. 2007. Improving statistical machine translation using word sense disambiguation. In *EMNLP-CoNLL*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

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*.

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, VL@ACL 2016*.

Hao Fang, Saurabh Gupta, Forrest N. Iandola, Rupesh K. Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015*, pages 1473–1482.

Stella Frank, Desmond Elliott, and Lucia Specia. 2018. Assessing multilingual multimodal image description: Studies of native speaker preferences and translator choices. *Natural Language Engineering*, 24(3):393—413.

Spandana Gella, Frank Keller, and Mirella Lapata. 2019. Disambiguating visual verbs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):311–322.

Spandana Gella, Mirella Lapata, and Frank Keller. 2016. Unsupervised visual sense disambiguation for verbs using multimodal embeddings. In *Proceedings of 2016 Conference of the North American Chapter of the Association for Computational Linguistics*.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Jindřich Helcl and Jindřich Libovický. 2017. Cuni system for the wmt17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 450–457.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2017. Sockeye: A Toolkit for Neural Machine Translation. *arXiv preprint arXiv:1712.05690*.

Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal pivots for image caption translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*.

Ákos Kádár, Desmond Elliott, Marc-Alexandre Côté, Grzegorz Chrupala, and Afra Alishahi. 2018. Lessons learned in multilingual grounded language learning. In *Proceedings of the 22nd Conference on Computational Natural Language Learning, CoNLL 2018*, pages 402–412.

Douwe Kiela, Ivan Vulic, and Stephen Clark. 2015. Visual bilingual lexicon induction with transferred convnet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*.

Adam Kilgarrif. 1998. Senseval: An exercise in evaluating word sense disambiguation programs. In *Proc. of the first international conference on language resources and evaluation*.

Chirag Lala and Lucia Specia. 2018. Multimodal lexical translation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2018)*.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 153–163.

Els Lefever and Véronique Hoste. 2013. Semeval-2013 task 10: Cross-lingual word sense disambiguation. *Proc. of SemEval*, pages 158–166.

Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, and Slav Petrov. 2012. Syntactic annotations for the google books ngram corpus. In *Proceedings of the ACL 2012 system demonstrations*, pages 169–174.

Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. 2006. Discriminating image senses by clustering with multimodal features. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

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

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*.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics NAACL-HLT 2018*, pages 1314–1324.

Kate Saenko and Trevor Darrell. 2008. Unsupervised learning of visual sense models for polysemous words. In *Advances in Neural Information Processing Systems*, pages 1393–1400.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. 2017. The university of edinburgh's neural MT systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, WMT 2017*, pages 389–399.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Kashif Shah, Josiah Wang, and Lucia Specia. 2016. Shef-multimodal: Grounding machine translation on images. In *Proceedings of the First Conference on Machine Translation, WMT*, pages 660–665.

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*, pages 543–553.

Ivan Vulic, Douwe Kiela, Stephen Clark, and Marie-Francine Moens. 2016. Multi-modal representations for improved bilingual lexicon learning. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 188.