Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6776–6785
Marseille, 20-25 June 2022

`MultiSubs`: A Large-scale Multimodal and Multilingual Dataset

Josiah Wang[1], **Josiel Figueiredo**[2], **Lucia Specia**[1,3]
[1]Imperial College London, UK
[2]Federal University of Mato Grosso, Brazil
[3]University of Sheffield, UK
josiah.wang@imperial.ac.uk, josiel@ic.ufmt.br, l.specia@imperial.ac.uk

Abstract

This paper introduces a large-scale multimodal and multilingual dataset that aims to facilitate research on grounding words to images in their contextual usage in language. The dataset consists of images selected to unambiguously illustrate concepts expressed in sentences from movie subtitles. The dataset is a valuable resource as (i) the images are aligned to text fragments rather than whole sentences; (ii) multiple images are possible for a text fragment and a sentence; (iii) the sentences are free-form and real-world like; (iv) the parallel texts are multilingual. We also set up a fill-in-the-blank game for humans to evaluate the quality of the automatic image selection process of our dataset. Finally, we propose a fill-in-the-blank task to demonstrate the utility of the dataset, and present some baseline prediction models. The dataset will benefit research on visual grounding of words especially in the context of free-form sentences, and can be obtained from `https://doi.org/10.5281/zenodo.5034604` under a Creative Commons licence.

Keywords: Multimodality, Visual Grounding, Multilinguality, Multimodal Dataset

1. Introduction

"Our experience of the world is multimodal – we see objects, hear sounds, feel texture, smell odours, and taste flavours" (Baltrušaitis et al., 2019). In order to understand the world around us, we need to be able to interpret such multimodal signals together. Learning and understanding languages is not an exception: humans make use of multiple modalities when doing so. In particular, words are generally learned with visual (among others) input as additional modality. Research on computational models of language grounding using visual information has led to many interesting applications, such as Image Captioning (Vinyals et al., 2015), Visual Question Answering (Antol et al., 2015) and Visual Dialog (Das et al., 2017).

Various multimodal datasets comprising images and text have been constructed for different applications. Many of these are made up of images annotated with text labels, and thus do not provide a context in which to apply the text and/or images. More recent datasets for image captioning (Chen et al., 2015; Young et al., 2014) go beyond textual labels and annotate images with sentence-level text. While these sentences provide a stronger context for the image, they suffer from one primary shortcoming: Each sentence 'explains' an image given as a whole, while most often focusing on only some of the elements depicted in the image. This makes it difficult to learn correspondences between elements in the text and their visual representation. Indeed, the connection between images and text is multifaceted, *i.e.* the former is not strictly a representation of the latter, thus making it hard to describe a whole image in a single sentence or to illustrate a whole sentence with a single image. A tighter, local correspondence between images and text segments is therefore needed in order to learn better groundings between words and images. Additionally, the texts are limited to very specific domains (image descriptions), while the images are also constrained to very few and very specific object categories or human activities; this makes it very hard to generalise to the diversity of possible real-world scenarios.

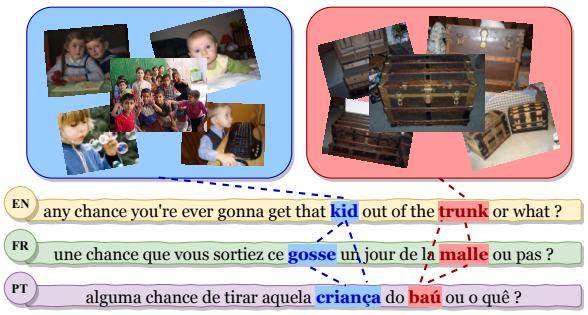

Figure 1: An example instance from our proposed large-scale multimodal and multilingual dataset. `MultiSubs` comprises predominantly conversational or narrative texts from movie subtitles, with text fragments illustrated with images and aligned across languages.

In this paper we propose `MultiSubs`, a new **large-scale multimodal and multilingual dataset** that facilitates research on grounding words to images in the *context* of their corresponding sentences (Figure 1). In contrast to previous datasets, ours ground words not only to images but also to their contextual usage in language, potentially giving rise to deeper insights into real-world human language learning. More specifically, (i) text fragments and images in `MultiSubs` have a tighter *local* correspondence, facilitating the

learning of associations between text fragments and their corresponding visual representations; (ii) the images are more general and diverse in scope and not constrained to particular domains, in contrast to image captioning datasets; (iii) multiple images are possible for each given text fragment and sentence; (iv) the text comprises a grammar or syntax similar to free-form, real-world text; and (v) the texts are multilingual and not just monolingual or bilingual. Starting from a parellel corpus of movie subtitles (Section 3), we propose a **crosslingual multimodal disambiguation method** to illustrate text fragments by exploiting the parallel multilingual texts to disambiguate the meanings of words in the text (Figure 2) (Section 4). To the best of our knowledge, this has not been previously explored in the context of text illustration. We also evaluate the quality of the dataset and illustrated text fragments via human judgment by casting it as a game (Section 6). To demonstrate a multimodal application using `MultiSubs`, we further propose a **fill-in-the-blank** task for automatic models to predict a missing word from a sentence, with or without image(s) of the word as clues (Section 7). We also present simple baseline models to automatically predict the missing word from the sentence.

The dataset can be obtained from `https://doi.org/10.5281/zenodo.5034604` under a Creative Commons licence.

2. Related work

Most existing multimodal grounding datasets consist of images/videos annotated with noun labels[1] (Deng et al., 2009; Lin et al., 2014). The main applications of these datasets include multimedia annotation/indexing/retrieval (Snoek and Worring, 2005) and object recognition/detection (Lin et al., 2014; Russakovsky et al., 2015). They also enable research on grounded semantic representation or concept learning (Baroni, 2016; Beinborn et al., 2018). Besides nouns, other work and datasets focus on labelling and recognising actions (Gella and Keller, 2017) and verbs (Gella et al., 2016). These works, however, are limited to single word labels independent of a contextual usage.

Recently multimodal grounding work has been moving beyond textual labels to include free-form sentences or paragraphs. Various datasets were constructed for these tasks, including image and video descriptions (Bernardi et al., 2016; Aafaq et al., 2018), news articles (Feng and Lapata, 2010; Hollink et al., 2016; Ramisa et al., 2018), cooking recipes (Marín et al., 2018), among others. These datasets, however, ground whole images to the whole text, and making it difficult to identify correspondences between text fragments and elements in the image. In addition, the text also does not explain all elements in the images.

Apart from monolingual text, there has also been work on multimodal grounding on multilingual text. One primary application of such work is in bilingual lexicon induction using visual data (Kiela et al., 2015), where the task is to find words in different languages sharing the same meaning. Hewitt *et al*. (Hewitt et al., 2018) developed a large-scale dataset to investigate bilingual lexicon learning for 100 languages. However, this dataset is limited to single word tokens; no textual context is provided with the words. Beyond word tokens, there are also multilingual datasets that are provided at sentence level, primarily extended from existing image description/captioning datasets (Elliott et al., 2016; Miyazaki and Shimizu, 2016). Schamoni *et al*. (Schamoni et al., 2018) also introduce a dataset with images from Wikipedia and their captions in multiple languages; however, the captions are not parallel across languages. These datasets are either very small or use machine translation to generate texts in a different language. More importantly, they are literal descriptions of images gathered for a specific set of object categories or activities and written by users in a constrained setting (*A woman is standing beside a bicycle with a dog*). Like monolingual image descriptions, whole sentences are associated with whole images. This makes it hard to ground image elements to text fragments. In contrast, our dataset grounds text fragments (words, phrases) to whole images that are not domain specific.

3. Corpus and text fragment selection

`MultiSubs` is based on the OpenSubtitles 2016 (OPUS) corpus (Lison and Tiedemann, 2016), which is a large-scale dataset of movie subtitles in 65 languages obtained from OpenSubtitles (OpenSubtitles, 2019). We use a subset by restricting the movies[2] to five categories that we believe are potentially more 'visual': adventure, animation, comedy, documentaries, and family. The mapping of IMDb identifiers (used in OPUS) to their corresponding categories are obtained from IMDb's official list (IMDb, 2019). Most of the subtitles are conversational (dialogues) or narrative (story narration or documentaries).

The subtitles are further filtered to only a subset of English subtitles that has been aligned in OPUS to subtitles from at least one of the top 30 non-English languages in the corpus. This resulted in 45,482 movie instances overall with ≈38M English sentences. The number of movies ranges from 2,354 to 31,168 for the top 30 languages.

We aim to select text fragments that are potentially 'visually depictable', and which can therefore be illustrated with images. We start by chunking the English subtitles[3] to extract nouns, verbs, compound nouns, and simple adjectivial noun phrases. The fragments are

[1] Other modalities include speech, audio, etc., but we focus our discussion only on images and text in this paper.

[2] We use the term 'movie' to cover all types of shows such as movies, TV series, and mini series.

[3] PoS tagger from spaCy v2: `en_core_web_md` from `https://spacy.io/models/en`.

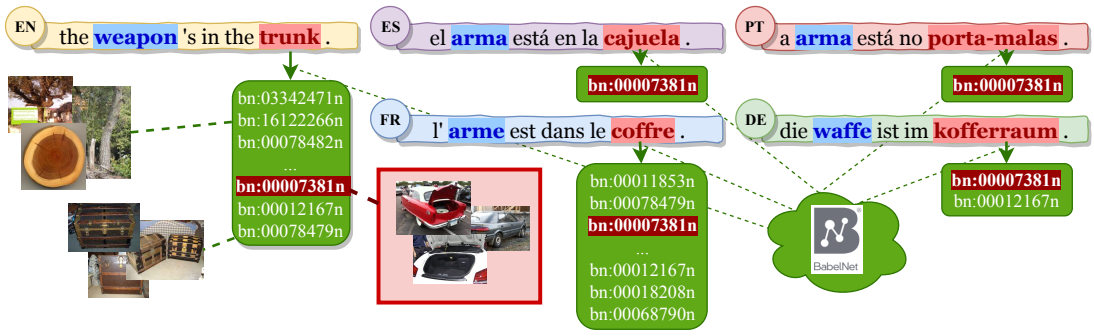

Figure 2: Overview of the MultiSubs construction process. Starting from parallel corpora, we selected 'visually salient' English words (*weapon* and *trunk* in this example). We automatically align the words across languages (e.g. *trunk* to *cajuela*, *coffre* etc.), and queried BabelNet with the words to obtain a list of synsets. In this example, *trunk* in English is ambiguous, but *cajuela* in Spanish is not. We thus disambiguated the sense of *trunk* by finding the intersection of synsets across languages (bn:00007381n), and illustrate *trunk* with images associated with the intersecting synset, as provided by BabelNet.

ranked by imageability scores obtained via bootstrapping from the MRC Psycholinguistic database (Paetzold and Specia, 2016); for multi-word phrases we average the imageability score of each individual word, assigning a zero score to each unseen word. We retain text fragments with an imageability score of at least 500, which is determined by manual inspection of a subset of words. After removing fragments occurring only once, the output is a set of 144,168 unique candidate fragments (more than 16M instances) across ≈11M sentences.

4. Illustration of text fragments

Our approach for illustrating MultiSubs obtains images for a subset of text fragments: *single word nouns*. Such nouns occur substantially more often in the corpus and are thus more suitable for learning algorithms. Additionally, single nouns (*dog*) make it more feasible to obtain good representative images than longer phrases (*a medium-sized white and brown dog*). This filtering step results in 4,099 unique English nouns occurring in ≈10.2M English sentences.

We aim to obtain images that illustrate the correct *sense* of these nouns in the context of the sentence. For that, we propose a novel approach that exploits the aligned multilingual subtitle corpus for sense disambiguation using BabelNet (Navigli and Ponzetto, 2012) (Section 4.1), a multilingual sense dictionary. Figure 2 illustrates the process.

MultiSubs is designed as a subtitle corpus illustrated with *general* images. Taking images from the video from where the subtitle comes is not possible since we do not have access to the copyrighted materials. In addition, there are no guarantees that the concepts mentioned in the text would be depicted in the video.

4.1. Cross-lingual sense disambiguation

The key intuition to our proposed text illustration approach is that an ambiguous English word may be unambiguous in the parallel sentence in the target language. For example, the correct word sense of *drill* in an English sentence can be inferred from a parallel Portuguese sentence based on the occurrence of the word *broca* (the machine) or *treino* (training exercise).

Cross-lingual word alignment. We experiment with up to four *target* languages in selecting the correct images to illustrate our candidate text fragments (nouns): **Spanish** (**ES**) and **Brazilian Portuguese** (**PT**), which are the two most frequent languages in OPUS; and **French** (**FR**) and **German** (**DE**), both commonly used in existing Machine Translation (MT) and Multimodal Machine Translation (MMT) research (Elliott et al., 2017). For each language, subtitles are selected such that (i) each is aligned with a subtitle in English; (ii) each contains at least one noun of interest.

For English and each target language, we trained fast_align (Dyer et al., 2013) on the *full* set of parallel sentences (regardless of whether the sentence contains a candidate fragment) to obtain alignments between words in both languages (symmetrised by the intersection of alignments in both directions). This generates a dictionary which maps English nouns to words in the target language. We filter this dictionary to remove pairs with infrequent target phrases (under 1% of the corpus). We also group words in the target language that share the same lemma[4].

Sense disambiguation. A noun being translated to different words in the target language does not necessarily mean it is ambiguous. The target phrases may simply be synonyms referring to the same concept. Thus, we further attempt to group synonyms on the target side, while also determining the correct word sense by looking at the aligned phrases across *multilingual* corpora.

For word senses, we use BabelNet (Navigli and Ponzetto, 2012), which is a large semantic network and

[4]We used the lemmas provided by spaCy.

multilingual encyclopaedic dictionary covering many languages and unifies other semantic networks. We query BabelNet with the English noun and its possible translations in each target language from our automatically aligned dictionary. The output (queried separately per language) is a list of BabelNet synset IDs matching the query.

To help us identify the correct sense of an English noun for a given context, we use the aligned word in the parallel sentence in the target language for disambiguation. We compute the intersection between the BabelNet synset IDs returned from both queries. For example, the English query *bank* could contain the synsets *financial-bank* and *river-bank*, and the Spanish query for the corresponding translation *banco* only returns the synset *financial-bank*. In this case, the intersection of both synset sets allows us to decide that *bank*, when translated to *banco*, refers to its *financial-bank* sense. Therefore, we can annotate the respective parallel sentence in the corpus with the correct sense. Where multiple synset IDs intersect, we take the union of all intersecting synsets as possible senses for the particular alignment. This potentially means that (i) the term is ambiguous and the ambiguity is carried over to the target language; or (ii) the distinct BabelNet synsets actually refer to the same or similar sense, as BabelNet unifies word senses from multiple sources automatically. We name this dataset ***intersect***$_1$.

If the above is only performed for one language pair, this single target language may not be sufficient to disambiguate the sense of the English term, as the term might be ambiguous in both languages (e.g. *coffre* is also ambiguous in Figure 2). This is particularly true for closely related languages such as Portuguese and Spanish. Thus, we propose exploiting *multiple* target languages to further increase our confidence in disambiguating the sense of the English word. Our assumption is that more languages will eventually allow the correct context of the word to be identified.

More specifically, we examine subtitles containing parallel sentences for up to four target languages. For each English phrase, we retain instances with at least one intersection between the synset IDs across all N languages, and discard if there is no intersection. We name these datasets ***intersect***$_N$, which comprise sentences that have valid synset alignments to at least N languages. Note that ***intersect***$_{N+1} \subseteq$ ***intersect***$_N$.

Table 1 shows the final dataset sizes of ***intersect***$_N$.

Image selection. The final step to constructing `MultiSubs` is to assign at least one image to each disambiguated English term, and by design the term in the aligned target language(s). As BabelNet generally provides multiple images for a given synset ID, we illustrate the term with all Creative Commons images associated with the synset.

	$N=1$	$N=2$	$N=3$	$N=4$
ES	2,159,635	1,083,748	335,484	45,209
PT	1,796,095	1,043,991	332,996	45,203
FR	1,063,071	641,865	305,817	45,217
DE	384,480	250,686	131,349	45,214

Table 1: Number of sentences for the ***intersect***$_N$ subset of `MultiSubs`, where N is the minimum number of target languages used for disambiguation. The slight variation in the final column is due to differences in how the aligned sentences are combined or split in OPUS across languages.

	tokens	**types**	**avg length**	**singletons**
EN	27,423,227	152,520	12.70	2,005,874
ES	25,616,482	245,686	11.86	2,012,476
EN	23,110,285	138,487	12.87	1,685,102
PT	20,538,013	205,410	11.43	1,687,903
EN	13,523,651	104,851	12.72	1,012,136
FR	12,956,305	149,372	12.19	1,004,304
EN	4,670,577	62,138	12.15	364,656
DE	4,311,350	123,087	11.21	364,613

Table 2: Token/type statistics on the sentences of ***intersect***$_1$ `MultiSubs`.

5. MultiSubs statistics and analysis

Table 1 shows the number of sentences in `MultiSubs`, according to their degree of intersection. On average, there are 1.10 illustrated words per sentence in `MultiSubs`, where about 90-93% sentences contain one illustrated word per sentence (depending on the target language). The number of images for each BabelNet synset ranges from 1 to 259, with an average of 15.5 images (excluding those with no images).

Table 2 shows some statistics of the sentences in `MultiSubs`. `MultiSubs` is substantially larger and less repetitive than Multi30k (Elliott et al., 2016) (≈300k tokens, ≈11-19k types, and only ≈5-11k singletons), even though the sentence length remains similar.

Figure 3 shows an example of how multilingual corpora is beneficial for disambiguating the correct sense of a word and subsequently illustrating it with an image. The top example shows an instance from ***intersect***$_1$, where the English sentence is aligned to only one target language (French). In this case, the word **sceau** is ambiguous in BabelNet, covering different but mostly related senses, and in some cases is noisy (terms are obtained by automatic translation). The bottom example shows an example where the English sentence is aligned to four target languages, which came to a consensus on a single BabelNet synset (and illustrated with the correct image). A manual inspection of a randomly selected subset of the data to assess our automated

bn:00070012n (seal wax), bn:00070013n (stamp), bn:00070014n (sealskin), ...
EN: stamp my heart with a **seal** of love !
FR: frapper mon cœur d' un **sceau** d' amour !

---

bn:00021163n (animal)
EN: even the **seal** 's got the badge .
ES: que hasta la **foca** tiene placa .
PT: até a **foca** tem um distintivo .
FR: même l' **otarie** a un badge .
DE: sogar die **robbe** hat das abzeichen .

Figure 3: Example of using multilingual corpora to disambiguate and illustrate a phrase.

they knew the gods put dewdrops on plants in the night.
sabiam que os deues punham orvalho nas plantas à noite

today we are announcing the closing of 11 of our older plants.
hoje anunciamos o encerramento de 11 das fábricas mais antigas.

Figure 4: Example disambiguation in the EN-PT portion of `MultiSubs`. In both cases, *plants* were correctly disambiguated using 4 languages.

disambiguation procedure showed that $\boldsymbol{intersect}_4$ is of high quality. We found many interesting cases of ambiguities, some of which are shown in Figure 4.

6. Human evaluation

To quantitatively assess our automated cross-lingual sense disambiguation cleaning procedure, we collect human annotations to determine whether images in `MultiSubs` are indeed useful for predicting a missing word in a fill-in-the-blank task.
We set up the annotation task as *The Gap Filling Game* (Figure 5). In this game, users are given three attempts at guessing the exact word removed from a sentence from `MultiSubs`. In the first attempt, the game shows only the sentence (along with a blank space for the missing word). In the second attempt, the game additionally provides one image for the missing word as a clue. In the third and final attempt, the system shows all images associated with the missing word. At each attempt, users are awarded a score of 1.0 if the word they entered is an exact match to the original word, or otherwise a partial score (between 0.0 and 1.0) computed as the cosine similarity between pre-trained CBOW word2vec (Mikolov et al., 2013) embeddings of the predicted and the original word. Each 'turn' (one sentence) ends when the user enters an exact match or after he or she has exhausted all three attempts, whichever occurs first. The score at the second and third attempts are multiplied by a *penalty factor* (0.90 and 0.80 respectively) to encourage users to

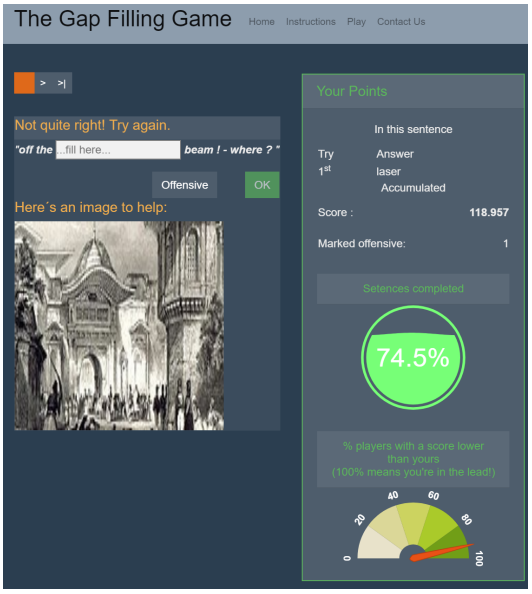

Figure 5: A screenshot of *The Gap Filling Game*, used to evaluate our automated cleaning procedure, as an upperbound to how well humans can perform the task without images, and to evaluate whether images are actually useful for the task.

guess the word correctly as early as possible. A user's score for a single turn is the maximum over all three attempts, and the final cumulative score per user is the sum of the score across all annotated sentences. This final score determines the winner and runner-up at the end of the game (after a pre-defined cut-off date), both of whom are awarded an Amazon voucher each. Users are not given an exact 'current top score' table during the game, but are instead provided the percentage of all users who has a lower score than the user's current score.

For the human annotations, we also introduce the $intersect_0$ dataset where the words are not disambiguated, i.e. images from all matching BabelNet synsets are used. This is to evaluate the quality of our automated filtering process. Annotators are allocated 100 sentences per batch, and are able to request for more batches once they complete their allocated batch. Sentences are selected at random. To select one image for the second attempt, we select the image most similar to the majority of other images of the synset, by computing the cosine distance of each image's ResNet152 pool5 feature (He et al., 2016) against all remaining images in the synset, and averaged the distance across these images.

Users are allowed to complete as many sentences as they like. The annotations were collected over 24 days in December 2018, and participants are primarily staff and student volunteers from the University of Sheffield, UK. 238 users participated in our annotation, resulting in 11,127 annotated instances (after filtering out invalid annotations).

Results of human evaluation Table 3 shows the results of human annotation, comparing the proportion of instances correctly predicted by annotators at different attempts: (1) no image; (2) one image; (3) many images; and also those that fail to be correctly predicted after three attempts. We consider a prediction correct if the predicted word is an exact match to the original word. Overall, out of $11,127$ instances, 21.89% of instances were predicted correctly with only the sentence as context, 20.49% with one image, and 15.21% with many images. The annotators failed to guess the remaining 42.41% of instances. Thus, we can estimate a human upper bound of 57.59% for correctly predicting missing words in the dataset, regardless of the cue provided. Across different $intersect_N$ splits, there is an improvement in the proportion of correct predictions as N increases, from 54.55% for $intersect_0$ to 60.83% for $intersect_4$. We have also tested sampling each split to have an equal number of instances $(1,598)$ to ensure that the proportion is not an indirect effect of imbalance in the number of instances; we found the proportions to be similar.

A user might fail to predict the exact word, but the word might be semantically similar to the correct word (e.g. a synonym). Thus, we also evaluate the annotations with the cosine similarity between word2vec embeddings of the predicted and correct word. Table 4 shows the average word similarity scores at different attempts across $intersect_N$ splits. Across attempts, the average similarity score is lowest for attempt 1 (text-only, 0.36), compared to attempts 2 (one image) and 3 (many images) – 0.48 and 0.49 respectively. Again, we verified that the scores are not affected by the imbalanced number of instances, by sampling equal number of instances across splits and attempts. We also observe a generally higher average score as we increase N, albeit marginal.

Figure 6 shows a few example human annotations, with varying degrees of success. In some cases, textual context alone is sufficient for predicting the correct word. In other cases, like in the second example, it is difficult to guess the missing word purely from textual context. In this case, images are useful.

We conclude that the task of filling in the blanks in `MultiSubs` is quite challenging even for humans, where only 57.59% instances were correctly guessed. This inspired us to introduce fill-in-the-blank task to evaluate how well automatic models can perform the same task, with or without images as cues (Section 7).

he was one of the best pitchers in **baseball** .
baseball (1.00)

uh , you know , i got to fix the **sink** , catch the game .
car (0.06), *sink* (1.00)

i saw it at the **supermarket** and i thought that maybe you would have liked it .
market (0.18) *shop* (0.50), *supermarket* (1.00)

It's mac , the night **watchman** .
before (0.07), *police* (0.31), *guard* (0.26)

Figure 6: Example annotations from our human experiment, with the masked word **boldfaced**. Users' guesses are *italicised*, with the word similarity score in brackets. The first example was guessed correctly without any images. The second was guessed correctly after one image was shown. The third was only guessed correctly after all images were shown. The final example was not guessed correctly after all three attempts.

7. Fill-in-the-blank task

The objective of the fill-in-the-blank task is to predict a word that has been removed from a sentence in `MultiSubs`, given the masked sentence as *textual context* and optionally one or more images depicting the missing word as *visual context*. Formally, given a sequence $S=\{w_1,\ldots,w_{t-1},w_t,w_{t+1},\ldots,...w_T\}$ of length T, where w_t is unobserved while the others are observed, the task is to predict w_t given S and optionally one or more images $\{I_1,I_2,\ldots,I_K\}$.

This task is similar to the human annotation (Section 6). Thus, we use the statistics from human evaluation as an estimated human upperbound for the task. We observe that this task is challenging even for humans who successfully predicted the missing word for only 57.59% of instances, regardless of whether they use images as contextual cue.

7.1. Dataset and settings

We blank out each illustrated word of a sentence as a fill-in-the-blank instance. If a sentence contains multiple illustrated nouns, we replicate the sentence and generate a blank per sentence for each noun, treating each as a separate instance.

The number of validation and test instances is fixed at $5,000$ each. These comprise sentences from $intersect_4$, which we consider to be the cleanest subset of `MultiSubs`. The validation and test sets are made more challenging by (i) uniformly sampling nouns from $intersect_4$ to increase their diversity; (ii) sampling an instance for each possible BabelNet sense of a sampled noun; this increases the semantic (and visual) variety for each word (*e.g.* sampling both the financial institution sense and the river sense of the noun 'bank'). The training set comprises all remaining instances.

We sample one image at random from the corresponding synset to illustrate each sense-disambiguated noun.

	Correct at attempt				Total
	1	2	3	Failed	
$intersect_0$	611 (18.75%)	660 (20.26%)	503 (15.44%)	1484 (45.55%)	3258
$intersect_1$	534 (21.86%)	481 (19.69%)	378 (15.47%)	1050 (42.98%)	2443
$intersect_2$	462 (22.35%)	408 (19.74%)	303 (14.66%)	894 (43.25%)	2067
$intersect_3$	432 (24.53%)	388 (22.03%)	260 (14.76%)	681 (38.67%)	1761
$intersect_4$	397 (24.84%)	343 (21.46%)	248 (15.52%)	610 (38.17%)	1598
all	2436 (21.89%)	2280 (20.49%)	1692 (15.21%)	4719 (42.41%)	11127

Table 3: Distribution across different attempts by humans in the fill-in-the-blank task.

	Average scores for attempt		
	1	2	3
$intersect_0$	0.33 (3258)	0.47 (2647)	0.47 (1987)
$intersect_1$	0.36 (2443)	0.47 (1909)	0.49 (1428)
$intersect_2$	0.37 (2067)	0.48 (1605)	0.48 (1197)
$intersect_3$	0.38 (1761)	0.51 (1329)	0.50 (941)
$intersect_4$	0.39 (1598)	0.50 (1201)	0.52 (858)
all	0.36 (11127)	0.48 (8691)	0.49 (6411)

Table 4: Average word similarity scores of human evaluation of the fill-in-the-blank task.

Our preliminary analysis showed that, in most cases, an image tends to correspond to only a single word label. This makes it less challenging for an image classifier which simply performs an exact matching of a test image to a training image, as the same image is repeated frequently across instances of the same noun. To circumvent this problem, we ensured that the images in the validation and test sets are both disjoint from the images in the training set. This is done by reserving 10% of all unique images for each synset in the validation and test sets respectively, and removing all these images from the training set. Our final training set consists of $4,277,772$ instances with $2,797$ unique masked words. The number of unique words in the validation and test set is 496 and 493 respectively, signifying their diversity.

7.2. Evaluation metrics

The models are evaluated using two metrics: (i) accuracy; (ii) average word similarity. The **accuracy** measures the proportion of correctly predicted words (exact token match) across test instances. The **word similarity** score measures the average semantic similarity across test instances between the predicted word and the correct word. For this paper, the cosine similarity between word2vec embeddings is used. Our evaluation script can be found on `https://github.com/josiahwang/multisubs-eval`

7.3. Baseline models

We present results of several baseline models that only use only the blanked out sentence as input, without using any image cues. The baseline models are (i) a **random** baseline that predicts a random target word from the full training set; (ii) a **random-multinomial** baseline that randomly samples a target word based on its frequency distribution in the full training set; (iii) a classic n**-gram** model with back-off. The n-gram model learns the most frequent target word from the full training set given the previous $n-1$ context words; the context window is iteratively reduced if the context is not found. In the case of $n = 1$, the model always predicts the most frequent blanked-out word (*man* for our dataset). We report results for $n \leq 9$ (the predictions do not change after $n = 9$).

	Accuracy (%)	**Word similarity**
random	0.00	0.10
random-multinomial	0.03	0.12
1-gram	1.07	0.17
2-gram	8.74	0.22
3-gram	16.03	0.31
4-gram	23.67	0.38
5-gram	27.35	0.41
6-gram	29.28	0.43
7-gram	30.07	0.43
8-gram	30.32	0.44
9-gram	30.35	0.44

Table 5: Accuracy and word similarity scores for our baseline (text-only) models on the fill-in-the-blank task, evaluated on the test subset and trained on the full training set.

Table 5 presents the baseline results on the test subset. As expected, randomly guessing the blank word does not get the system far. It is useful to note that the word similarity score has a lower-bound of 0.10. Always guessing *man* (1-gram) is slightly better than randomly guessing, although the accuracy is still low at 1%. Surprisingly, the simple n-gram models with back-off actually perform well, with an accuracy of 23.67% for 4-grams and up to 30.35% for 9-grams. The word similarity scores show a similar trend, with a maximum score of 0.44 with 9-grams.

We hope the baseline models will be useful to spur further research to investigate more complex prediction models, as well as those that incorporate image cues as input.

8. Conclusions

We introduced MultiSubs, a large-scale multimodal and multilingual dataset aimed at facilitating research on grounding words to images in the context of their corresponding sentences. The dataset consists of a parallel corpus of subtitles in English and four other languages, and selected words are illustrated with one or more images in the context of the sentence. This provides a tighter local correspondence between text and images, allowing the learning of associations between text fragments and their corresponding images. The structure of the text is also less constrained than existing multilingual and multimodal datasets, making it more representative of multimodal grounding in real-world scenarios.

Human evaluation in the form of a fill-in-the-blank game showed that the task is quite challenging, where humans failed to guess a missing word 42.41% of the time, and could correctly guess only 21.89% of instances without any images. We also proposed a fill-in-the-blank task to demonstrate the utility of the dataset, and presented some text-only baseline predicted models, with the best model achieving 30.55% accuracy. An extended version of this paper (Wang et al., 2021) contains further experiments using more complex models (spoiler: they do not achieve better results), and also proposes another task called lexical translation.

We plan to further develop MultiSubs to annotate more phrases with images, and to improve the quality and quantity of images associated with the text fragments. MultiSubs will benefit research on visual grounding of words especially in the context of free-form sentences.

9. Acknowledgements

This work was supported by a Microsoft Azure Research Award for Josiah Wang. It was also supported by the MultiMT project (H2020 ERC Starting Grant No. 678017), and the MMVC project, via an Institutional Links grant, ID 352343575, under the Newton-Katip Celebi Fund partnership. The grant is funded by the UK Department of Business, Energy and Industrial Strategy (BEIS) and Scientific and Technological Research Council of Turkey (TÜBİTAK) and delivered by the British Council. Lucia Specia also received support from the Air Force Office of Scientific Research (under award number FA8655-20-1-7006).

10. Bibliographical References

Aafaq, N., Gilani, S. Z., Liu, W., and Mian, A. (2018). Video description: A survey of methods, datasets and evaluation metrics. *CoRR*, abs/1806.00186.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2425–2433, Santiago, Chile, December. IEEE.

Baltrušaitis, T., Ahuja, C., and Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, February.

Baroni, M. (2016). Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Beinborn, L., Botschen, T., and Gurevych, I. (2018). Multimodal grounding for language processing. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2325–2339, Santa Fe, NM, USA, August. Association for Computational Linguistics.

Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., and Plank, B. (2016). Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. (2015). Microsoft COCO captions: Data collection and evaluation server. *CoRR*, abs/1504.00325v2.

Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M. F., Parikh, D., and Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 1080–1089, Honolulu, HI, USA, July. IEEE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 248–255, Miami, FL, USA, June. IEEE.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of IBM Model 2. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 644–648, Atlanta, Georgia, June. Association for Computational Linguistics.

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, Berlin, Germany, August. Association for Computational Linguistics.

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, Copenhagen, Denmark. Association for Computational Linguistics.

Feng, Y. and Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics, pages 91–99, Los Angeles, CA, USA, June. Association for Computational Linguistics.

Gella, S. and Keller, F. (2017). An analysis of action recognition datasets for language and vision tasks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 64–71, Vancouver, Canada. Association for Computational Linguistics.

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

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 770–778, Las Vegas, NV, USA, June. IEEE.

Hewitt, J., Ippolito, D., Callahan, B., Kriz, R., Wijaya, D. T., and Callison-Burch, C. (2018). Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2566–2576, Melbourne, Australia, July. Association for Computational Linguistics.

Hollink, L., Bedjeti, A., van Harmelen, M., and Elliott, D. (2016). A corpus of images and text in online news. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

IMDb. (2019). IMDb. `https://www.imdb.com/interfaces/`. Accessed: 2018-12-17.

Kiela, D., Vulić, I., and Clark, S. (2015). Visual bilingual lexicon induction with transferred ConvNet features. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Lisbon, Portugal, September. Association for Computational Linguistics.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In David Fleet, et al., editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 740–755, Zurich, Switzerland, September. Springer International Publishing.

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Marín, J., Biswas, A., Ofli, F., Hynes, N., Salvador, A., Aytar, Y., Weber, I., and Torralba, A. (2018). Recipe1m: A dataset for learning cross-modal embeddings for cooking recipes and food images. *CoRR*, abs/1810.06553.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, et al., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

Miyazaki, T. and Shimizu, N. (2016). Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790, Berlin, Germany, August. Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

OpenSubtitles. (2019). Subtitles - download movie and TV Series subtitles. `http://www.opensubtitles.org/`. Accessed: 2018-12-17.

Paetzold, G. and Specia, L. (2016). Inferring psycholinguistic properties of words. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 435–440, San Diego, California, June. Association for Computational Linguistics.

Ramisa, A., Yan, F., Moreno-Noguer, F., and Mikolajczyk, K. (2018). BreakingNews: Article annotation by image and text processing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1072–1085, May.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, December.

Schamoni, S., Hitschler, J., and Riezler, S. (2018). A dataset and reranking method for multimodal MT of user-generated image captions. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Papers)*, pages 40–153, Boston, MA, USA, March.

Snoek, C. G. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, January.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption gen-

erator. In *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition (CVPR)*, pages 3156–3164, Boston, MA, USA, June. IEEE.

Wang, J., Madhyastha, P., Figueiredo, J., Lala, C., and Specia, L. (2021). Multisubs: A large-scale multimodal and multilingual dataset. *CoRR*, abs/2103.01910.

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, February.