VISA: An Ambiguous Subtitles Dataset
for Visual Scene-Aware Machine Translation

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, Sadao Kurohashi

Kyoto University, Kyoto, Japan

{liyh, sshimizu, gu, chu, kuro}@nlp.ist.i.kyoto-u.ac.jp

Abstract

Existing multimodal machine translation (MMT) datasets consist of images and video captions or general subtitles, which rarely contain linguistic ambiguity, making visual information not so effective to generate appropriate translations. We introduce VISA, a new dataset that consists of 40k Japanese–English parallel sentence pairs and corresponding video clips with the following key features: (1) the parallel sentences are subtitles from movies and TV episodes; (2) the source subtitles are ambiguous, which means they have multiple possible translations with different meanings; (3) we divide the dataset into *Polysemy* and *Omission* according to the cause of ambiguity. We show that VISA is challenging for the latest MMT system, and we hope that the dataset can facilitate MMT research. The VISA dataset is available at: https://github.com/ku-nlp/VISA.

Keywords: multimodality, machine translation, ambiguity, scene-awareness

1. Introduction

Neural machine translation (NMT) models relying on text data have achieved state-of-the-art performance for domains with little ambiguity in data. However, in several other domains, particularly real-time domains such as spoken language or sports commentary, the word sense ambiguity of verbs and nouns has a significant impact on translation quality (Gu et al., 2021). Multimodal machine translation (MMT) focuses on using visual data as auxiliary information to tackle the ambiguity problem. The contextual information in the visual data assists in reducing the ambiguity of nouns or verbs in the source text data.

Previous MMT studies mainly focus on the image-guided machine translation (IMT) task (Elliott et al., 2016; Zhao et al., 2020), where, given an image and a source sentence, the goal is to enhance the quality of translation by leveraging their semantic correspondence to the image. Resolving ambiguities through visual cues is one of the main motivations behind this task. However, videos are better information sources than images because one video contains an ordered sequence of frames and provides much more visual features such as motion features. Recently, some studies have started to focus on the video-guided machine translation (VMT) task (Wang et al., 2019; Hirasawa et al., 2020).

VMT datasets face the problem of data scarcity. How2 (Sanabria et al., 2018) and VATEX (Wang et al., 2019) datasets are recent efforts to alleviate the problem. In addition, previous datasets are limited to general subtitles or video captions which are actually descriptions of video clips. Besides the lack of practical use of caption MT, it has been shown that general subtitles and caption MT essentially do not require visual information due to the lack of language ambiguity in captions (Caglayan et al., 2019).

To this end, we construct a new large-scale parallel sub-titles dataset[1] for VMT research, VISA[2], which contains $39,880$ Japanese–English parallel subtitles pairs and corresponding video clips. Compared to previous large-scale VMT datasets, VISA is characterized by the following significant properties. (1) VISA uses subtitles from movies and TV series to construct parallel sentences for the translation task. Most of the subtitles are dialogues with short sentence lengths. (2) The source subtitles are all ambiguous subtitles, which means they have multiple possible translations with different meanings. In this case, it's difficult to determine which translation is correct without other information such as visual clues. (3) We focus on two causes of ambiguity, polysemy and omission, and divide the dataset accordingly. The ambiguity of *Polysemy* subtitles is mainly caused by one or more polysemy in the subtitles. In contrast, the ambiguity of *Omission* subtitles is mainly caused by the omission of the subject or object of subtitles. The examples are shown in Figure 1. The Polysemy part contains $20,666$ parallel subtitle pairs, while the *Omission* part contains $19,214$ parallel subtitle pairs.

We conduct experiments on VISA with one of the latest VMT architecture proposed in Gu et al. (2021) as our baseline. We tested the performance of four models on the dataset according to this architecture, using only the text feature, with the motion feature, with the spatial feature, and with all three features, respectively. The results show that VISA is challenging for the existing VMT models.

In summary, our contributions are mainly three-fold:

- We construct a new large-scale ambiguous parallel subtitles dataset to promote VMT research. The

---

[1]The dataset is for research use only and is not available to the public. If you need the dataset, please send an email to the author.

[2]VISA stands for VIsual Scene-Aware.

(a) *Polysemy* example. "放せ" (let...go) can be regarded as a polysemy because it can be translated in two possible ways.
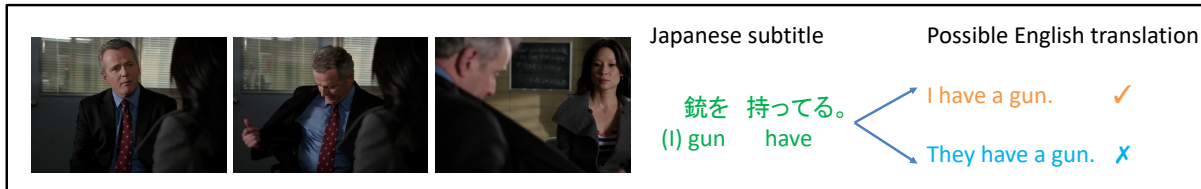
(b) *Omission* example. The subject in the Japanese subtitle is omitted.

Figure 1: Examples of *Polysemy* and *Omission* data. In both cases, the translation can be disambiguated using the visual clue.

dataset is divided into the *Polysemy* part and the *Omission* part according to the cause of ambiguity.

- We conduct experiments on the VISA dataset with the latest VMT architecture to set a baseline of the dataset.

- Based on the analysis of the dataset and experimental results, we discuss problems and future perspectives for VMT models.

2. Related Work

Multimodal Machine Translation. MMT involves drawing information from multiple modalities, assuming that the additional modalities will contain useful alternative views of the input data (Sulubacak et al., 2020). Previous studies mainly focus on IMT using images as the visual modality to help machine translation (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). The utility of the visual modality has recently been disputed under specific datasets, or task conditions (Elliott, 2018; Caglayan et al., 2019). However, using images in captions translation is theoretically helpful for handling grammatical characteristics and resolving translational ambiguities in translating between dissimilar languages (Sulubacak et al., 2020). VMT is a multimodal machine translation task similar to IMT but focuses on video clips rather than images associated with the textual input. Depending on the textual context, there may be variations in video-guided translation. The source text can be speech transcripts from the video, which are often split as standard subtitles, or a description of the visual scene or activity displayed in the movie, which is frequently made for visually impaired persons (Sulubacak et al., 2020).

VMT Datasets. The scarcity of datasets is one of the biggest roadblocks to the advancement of VMT. Recent efforts to compile freely accessible data for VMT, such as the How2 (Sanabria et al., 2018), and VATEX (Wang et al., 2019) datasets, have begun to alleviate this bottleneck.

The How2 dataset is a collection of $79,114$ instructional YouTube video clips (about $2,000$ hours) with associated English subtitles and summaries spanning a variety of topics. The average length of video clips is 90s. For multimodal translation, a 300-hour subset of How2 subtitles that covers 22 different topics is available with crowdsourced Portuguese translations. The Video and TeXt (VATEX) dataset (Wang et al. 2019b) is a bilingual collection of video descriptions containing over $41,250$ videos and $825,000$ captions in English and Chinese. With low-approval samples removed, the released version of the dataset includes $206,345$ translation pairs in total. However, both datasets do not require visual information due to the lack of language ambiguity. We follow in the footsteps of How2 and VATEX, further utilize ambiguous subtitles to construct the new dataset, VISA, containing $39,880$ Japanese–English parallel subtitles pairs and corresponding 10 seconds video clips, in total.

3. Dataset Construction

The main goal of our work is to construct an MMT dataset combined with parallel subtitles and corresponding video clips, in which the source subtitles must be ambiguous. Our work uses Japanese–English parallel subtitles and considers Japanese and English as the source language and target language of translation tasks, respectively.

To construct the dataset, we select ambiguous parallel subtitles from the existing subtitle dataset and crop the corresponding video clips from movies or TV episodes. Subtitles selection is divided into two steps. The first step is pre-selection. Because ambiguity is often caused by polysemy, we select Japanese subtitles that contain polysemy, which is more likely to be ambiguous. The second step is crowdsourcing. In this

step, we use Yahoo Crowdsourcing[3] to further select Japanese sentences that are indeed ambiguous. After selecting subtitles, we first align subtitles with videos, then crop video clips according to the timestamp of the aligned subtitle. The whole procedure is shown in Figure 2.

3.1. Pre-selection

Through pre-selection, we obtain subtitles containing polysemy in four steps: (1) get Japanese–English parallel subtitles; (2) build vocabulary from Japanese subtitles; (3) utilize BabelNet, a multilingual dictionary, to select polysemy from the vocabulary; (4) select subtitles which contain polysemy.
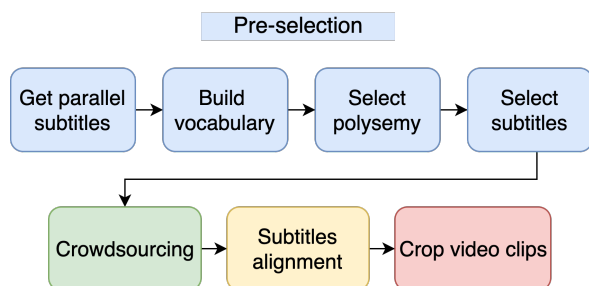

Figure 2: The procedure of dataset construction.

3.1.1. Get Parallel Subtitles

For the first part, we get Japanese–English parallel subtitles from the OpenSubtitles (Lison and Tiedemann, 2016) dataset. The main reason we chose Japanese as the source language is that it is a typical pro-drop language. OpenSubtitles is a subtitle dataset compiled from an extensive database of film and TV subtitles which includes a total of $1,689$ bitexts spanning 2.6 billion sentences across 60 languages. In OpenSubtitles, all parallel subtitles belonging to the same video are organized into a single subtitle file. In addition to subtitles, OpenSubtitles also contains the Internet Movie Database (IMDb) id of the video corresponding to each subtitle file and the timestamp of each subtitle in the video. From the OpenSubtitles dataset, we get the Japanese–English parallel subtitles, timestamps for each of them, and the corresponding video sources. In this way, we collected $1,213,468$ parallel subtitles pairs.

3.1.2. Build Vocabulary

After getting parallel subtitles, we can build a vocabulary for Japanese subtitles. Since different forms of the same word have the same polysemous property, we reduce words to their original form before recording them into the vocabulary. In addition, a word may have different polysemous properties under different parts of speech (POS). In order to determine the polysemous

---

property of words in a subtitle more precisely, we register each word into vocabulary with its POS. We use Juman++ (Tolmachev et al., 2018) to segment Japanese subtitles into words, convert words to the original form, and check the POS of words. If a word is a content word (i.e., POS is one of noun, adjective, verb, or adverb), then we combine the original form and POS of the word into an entry of the vocabulary. For example, for a noun "銃" (gun) in the subtitle, we register "銃｜noun" into the vocabulary.

3.1.3. Select Polysemy

BabelNet polysemy As the third step, we aim to select polysemy from the vocabulary. Polysemy is a word or phrase which can represent different but related concepts. Therefore, to determine whether a word is polysemous, we use BabelNet (Navigli and Ponzetto, 2010) to check whether the word represents more than one concept. BabelNet is an extensive, wide-coverage multilingual semantic network automatically constructed through a methodology that integrates lexicographic and encyclopedic knowledge from WordNet and Wikipedia. In BabelNet, the resulting set of multilingual lexicalizations of a given concept is called a synset. Therefore, if a word appears in more than one synset in BabelNet, we can regard it as a polysemy.

However, different synsets of BabelNet may represent very similar concepts. Therefore, we need to add restrictions to the process of determining polysemy. Due to the nature of BabelNet, a synset may contain lexicalizations from different sources. We can restrict the source to OmegaWiki and specify the POS of polysemy to reduce the number of synsets. In the sources provided by BabelNet, OmegaWiki has a clear division of words that represent different concepts, so it is more effective when used to determine polysemy.

Symilarity between synsets Furthermore, we compute the similarity between synsets and only preserve those words with small similarities. The algorithm we used to compute the similarity is Jiang and Conrath's similarity algorithm (Jiang and Conrath, 1997), which uses WordNet as its central resource to quantify lexical semantic relatedness. It is found to be superior to other WordNet-based measures (Budanitsky and Hirst, 2006). We use BabelNet to export words' synsets in WordNet and then use this algorithm to calculate the synsets similarity. We consider a word as polysemous only if its minimum synsets similarity is less than a threshold of 0.06, which we set based on experiments.

Removing specific words Specifically, we do not regard "の" (of) and "来る" (come) as polysemy. They have a high frequency in Japanese sentences but do not cause ambiguity in most situations.

3.1.4. Select Subtitles

Select subtitles containing polysemy At last, we select subtitles that contain polysemy. Again, we use Juman++ to segment the subtitles and determine the original form and the POS of each word in the sentence.

Then we check whether the word-POS pair is in the polysemy vocabulary. We only keep subtitles that contain at least one polysemy.

Length ratio   Sometimes, the subtitles pairs are not parallel subtitles. For example, the source sentence is a complete Japanese sentence, while the target sentence is a phrase. In order to ensure that the subtitles pairs are parallel, we further select the subtitles with length ratio. Given a source sentence s and a target sentence t, we define a length ratio r as:

$$r(s,t) = \frac{\max(|s|, |t|)}{\min(|s|, |t|)} \qquad (1)$$

in which $|x|$ is the number of words in sentence x. We only preserve subtitles where $r \leq 2$.

3.2. Crowdsourcing

After selecting subtitles containing polysemy, we further selected the subtitles that were indeed ambiguous on sentence level through crowdsourcing. Because the purpose of crowdsourcing is to determine whether Japanese subtitles are ambiguous, for each group of duplicate Japanese subtitles, we only preserve one of them.

For each subtitle, we present it to workers together with annotations about the polysemy of the subtitle. Each annotation consists of polysemy and the main lemma of all synsets corresponding to the polysemy in both Japanese and English (e.g., 放す: 逃す/let_go, 切り放す/let_loose, 逃す/free, 見捨てる/abandon).

Furthermore, we ask workers whether the Japanese subtitle is ambiguous when translated into English and provide five choices to them: (A) yes, and caused by polysemy; (B) yes, and caused by omission; (C) yes, and caused by other reasons; (D) no, it is not ambiguous; (E) the sentence is unnatural. Because sentence ambiguity may be caused by more than one reason, workers can choose multiple choices. The interface is shown in Figure 3. In these choices, we only use the ambiguous subtitles caused by polysemy and omission to construct the dataset and give up the ambiguous subtitles caused by other reasons. The main reason we provide this choice to workers is to ensure that the options cover all possibilities.

Three different workers answered each question. Each worker may choose multiple choices, and if an option is chosen by at least two workers of the three workers, we consider it a valid answer and consider the subtitle as a valid subtitle. Because the crowdsourcing was conducted on a large scale, we split it into multiple crowdsourcing tasks and posted one of them each day.

To ensure the quality of workers' answers, we set up hidden check questions in crowdsourcing questions. We accepted a worker's answers only if s/he has correctly answered the check question. We select check questions using valid subtitles of completed tasks and manually selected questions as shown in Figure 4. Because determining the ambiguity of sentences requires

Is this sentence ambiguous when translated into English?

| sentence | 彼女のことはハンから聞きました |

☐ Yes, and caused by polysemy

☐ Yes, and caused by omission

☐ Yes, and caused by other reason

☐ No, it is not ambiguous

☐ The sentence is unnatural

・聞き: かぐ/smell, 聞く/hear, 聞く/heed, 質問/ask, 聞く/listen, 下問/ask, 与える/allow

Figure 3: The interface for crowdsourcing.

some associative ability, we mainly select unambiguous or unnatural sentences from the completed tasks to construct the check questions. However, to ensure the balance of the check questions, we also included two manually selected simple questions, a polysemy question and an omission question, in each day's check questions. By doing so, we ensured that these inspection questions were not too difficult but effective so that workers who answered carefully could be paid.
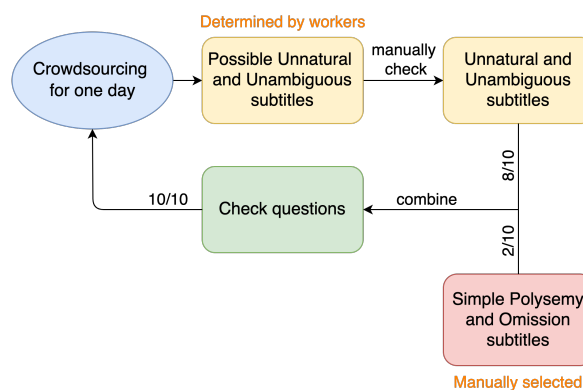

Figure 4: The procedure of check questions selection.

3.3. Subtitles Alignment

We collected $2,337$ videos corresponding to the subtitles. The videos are either movies or TV episodes. Different from video captions in which we do not need to care about timestamps, for subtitles, we must ensure that the timestamps of the subtitles correspond exactly to the time when the subtitles appear. However, many timestamps of the subtitles are not well aligned to the videos. For example, there may be a constant shift between the subtitles and the videos. In order to find the exact video clip corresponding to the subtitles, we need to align the original subtitles from the OpenSubtitles dataset with the videos.

We utilize alass[4] to align the original subtitles with videos. Alass can perform subtitles alignment in two

---

[4] https://github.com/kaegi/alass

ways. One is to align subtitle files with incorrect timestamps to subtitle files with correct timestamps, such as those that come with videos. The other is to align the incorrect subtitle file with the corresponding video using voice activity detection (VAD). The former performs significantly better and almost always aligns perfectly. After alignment, alass outputs a subtitle file with correct timestamps.
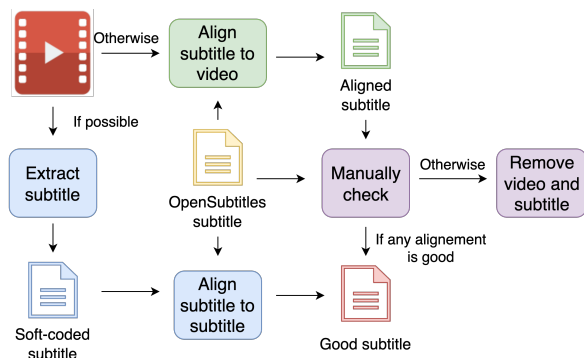

Figure 5: The procedure for subtitles alignment. The blue path is the first path, while the green and the purple path is the second path.

Based on the two functions of alass, we designed two subtitle alignment paths. The procedure is shown in Figure 5. The first one is based on aligning original subtitles to extracted subtitles with alass. Many videos contain soft-coded subtitles, which can be extracted from videos and are correct. Therefore, we utilize FFmpeg to extract soft-coded subtitles from videos and align original subtitles from the OpenSubtitles dataset with the extracted subtitles. As a result, we get well-aligned subtitles.

If the video does not contain a soft-coded subtitle file, we need to perform the second path, which aligns the subtitles with the video. However, the aligned subtitles may still be incorrect. Therefore, we manually check if there is a well-aligned subtitle file in the aligned subtitle file and the original subtitle file. If both subtitle files are incorrect, we remove the video and the corresponding subtitles.

In this way, we can align the subtitles and the videos to determine precisely where each subtitle appears in the video. This lays the groundwork for cropping the video clips corresponding to the subtitles.

3.4. Video Cropping

With well-aligned subtitle files and videos, we can crop video clips for the ambiguous parallel subtitles. We standardized the format of the video clips as 10-second 25-fps mp4 files. From the midpoint of each subtitle's period, each video clip takes 5 seconds before and after, respectively. We only keep the video content of video clips and remove the audio content. At last, we combine ambiguous parallel subtitles and corresponding video clips to construct the dataset.

4. Dataset Statistics through Construction

4.1. Pre-selection

The core task of the pre-selection is to build the polysemy vocabulary. As shown in Table 1, the number of subtitles corresponding to the polysemy vocabulary is reduced step by step during the selection process. As a result, we select $4,875$ polysemy from the word-POS vocabulary of size $55,414$ and get $329,211$ parallel subtitles.

Select setting	Number of subtitles
Original	1,213,468
+ BabelNet polysemy	815,705
+ Synsets similarity	352,344
+ Remove specific words	329,211
+ Length ratio	309,895

Table 1: The number of subtitles during pre-selection.

4.2. Crowdsourcing

In crowdsourcing, we present five choices to workers. We classify each subtitle according to how many (none, two, or three) of the three workers agree with its category. Because the questions are multiple-choice, a subtitle may belong to more than one category. Only if a subtitle is not a valid subtitle, we classify it as no agreement. For example, if the three workers' choices are $\{A, B\}$, $\{A, B, C\}$, and $\{A\}$, the subtitle will belong to both A of three agreements category and B of two agreements category. The statistics of the subtitles are shown in Table 2. After keeping only those answers agreed by at least two workers, we get the valid subtitles as shown in Table 3. We can see that most of the subtitles are not ambiguous, while the number of polysemy part and omission part are similar.

After crowdsourcing, we only preserve the ambiguous subtitles caused by polysemy or omission and sample 50 polysemy subtitles and 50 omission subtitles from them. We asked two native Japanese speaker to help us manually check whether these subtitles are really ambiguous and what caused the ambiguity. As a result, 82% *Polysemy* subtitles and 88% *Omission* subtitles have correct categories after crowdsourcing.

4.3. Subtitles Alignment

There are $2,751$ videos corresponding to the subtitles. We had $2,240$ videos of them at the beginning, and $1,978$ videos of them became well-aligned subtitles after subtitles alignment. The specific number of different alignments is shown in Table 4. The result shows that most subtitles can be well aligned to videos, and subtitles alignment works better for movies rather than TV episodes. We only preserve well-aligned subtitles to construct the dataset.

	Agreement		
Choice	Three	Two	None
Polysemy	3,099	22,065	28,464
Omission	2,818	20,888	26,843
Other	13	966	8,915
Not ambiguous	40,005	91,769	32,787
Unnatural	1,851	8,851	14,495

Table 2: The number of subtitles during crowdsourcing.

Type	**Number**	**Percentage**
Polysemy	25,164	13.08%
Omission	23,706	12.33%
Other	979	0.51%
Not ambiguous	131,774	68.52%
Unnatural	10,702	5.56%

Table 3: The number of valid subtitles after crowdsourcing.

	Movies	**TV episodes**	**Total**
Good alignment	757	1,221	1,978
Bad alignment	57	205	262
Good ratio	93.00%	85.62%	88.30%

Table 4: The number of videos during subtitles alignment.

4.4. Final Result

As a result, we constructed a dataset that consists of $39,880$ parallel subtitles and corresponding video clips. This dataset is divided into two categories, *Polysemy* and *Omission*, based on what causes the subtitle ambiguity. The *Polysemy* part has $20,666$ parallel subtitles, while the *Omission* part has $19,214$ parallel subtitles. There is a slight overlap between these two parts containing $1,971$ parallel subtitle pairs. The final number of subtitles is more than the number of subtitles in the crowdsourcing result because we removed the duplicate Japanese subtitles before crowdsourcing, and we restored the duplicate subtitles after crowdsourcing.

5. Experimental Settings

5.1. Datasets

For the dataset, in addition to experimenting on the whole VISA dataset, we also conducted experiments on the *Polysemy* and *Omission* datasets separately. The splits of the three datasets are shown in Table 5.

5.2. Models

To test the performance of MMT systems on our dataset, we implement the VMT translation (VMT) architecture described in Gu et al. (2021) and consider

Split	**Train**	**Validation**	**Test**
Polysemy	18,666	1,000	1,000
Omission	17,214	1,000	1,000
Combined	35,880	2,000	2,000

Table 5: The splits of *Polysemy*, *Omission*, and combined dataset.

four models based on this architecture as our baseline models.

The main idea of the VMT architecture is to use both spatial and motion representations in videos as auxiliary information to translate sentences. The architecture consists of the following four modules: (1) Text Encoder: A Bi-GRU (Schuster and Paliwal, 1997) encoder transforms source sentences into text representations; (2) Motion Encoder: After extracting motion features with pretrained I3D (Carreira and Zisserman, 2017) model for action recognition, it uses a Bi-GRU motion encoder and a positional encoding (PE) layer to encode the motion features into ordered motion representations; (3) Spatial HAN: The architecture uses a hierarchical attention network to model the spatial information from object-level to video-level and extract contextual spatial representations; (4) Target Decoder: All representations from respective encoders are processed with attention mechanisms (Luong et al., 2015) and help decode the target word embedding with a GRU decoding layer.

Based on this VMT architecture, we performed experiments with four models: (1) NMT model: only use the text encoder; (2) VMT (motion) model: use both the text and motion encoders; (3) VMT (spatial) model: use both the text encoder and spatial HAN; (4) VMT (both): use the whole architecture using text, motion, and spatial representations.

We trained the models on the three datasets following most of the model settings described in Gu et al. (2021). As a special case, for the experiment on the *Polysemy* dataset using VMT (both) model, we set drop out as 0.4 instead of 0.5 because it's difficult for the model to converge.

5.3. Evaluation Metrics

We adopt three automatic evaluation metrics: BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and RIBES (Isozaki et al., 2010). The reason we introduced METEOR and RIBES scores as measures in addition to the BLEU score is that subtitles are similar to dialogue sentences and most subtitles are very short, with an average length of only around 7 words.

Because the calculation of the BLEU score is based on n-grams, the BLEU score cannot accurately reflect the quality of translation when the sentence length is very short. On the other hand, the METEOR score is based on the harmonic mean of unigram precision and recall, and the RIBES score is based on rank correlation coef-

Model	Polysemy			Omission			Combined		
	BLEU	METEOR	RIBES	BLEU	METEOR	RIBES	BLEU	METEOR	RIBES
NMT	**7.71**	**24.67**	**14.37**	5.58	18.54	9.86	13.14	**29.00**	**20.66**
VMT (motion)	7.09	22.81	13.31	5.97	20.09	**11.11**	**13.19**	28.84	19.70
VMT (spatial)	6.99	22.33	13.32	**5.99**	19.76	10.94	12.99	28.26	19.65
VMT	7.40	22.33	12.64	5.63	**20.10**	10.98	12.89	28.45	19.45

Table 6: VMT results reported on BLEU, METEOR, and RIBES. Models are described in section 5.2.

ficients modified with precision and mainly applied to distant language pairs such as Japanese and English. In practice, we utilize the Smooth 4 method described in Chen and Cherry (2014) to correct the BLEU score.

6. Results and Discussion

Table 6 shows the results of the four baseline models evaluated using the three metrics. Which of the NMT model or VMT models performs better depends on the dataset. On the *Omission* dataset, all three VMT models perform significantly better than the NMT model. The VMT model achieves 0.41 BLEU score, 1.56 METEOR score, and 1.25 RIBES score over the NMT model, respectively. However, on the whole dataset and the *Polysemy* dataset, the NMT model performs better than VMT models in most situations. Therefore, it is still challenging for the current VMT model to handle the polysemy problem.

Why doesn't the current VMT model work well on VISA? By analyzing the translation results and the dataset, we found the following reasons.

First, the videos do not necessarily contribute to the disambiguation. For example, many video clips only focus on the speaker's face when a subtitle appears. In movies and TV series, when a subtitle appears, the video tends to show the speaker's face. Although we can use this to get emotional information, it is more difficult to get other information such as object information from the video. Therefore, it is more challenging to translate the subtitles than to translate the descriptions of video clips.

The second reason is the lack of speaker recognition, which means we do not know who said the sentence. If we cannot determine who the speaker is, we cannot use the visual information of the speaker in the video. For example, in the omission example in Figure 1, if the speaker is the man with a gun, the omitted subject would be "I", while if the speaker is the woman, the omitted subject would be "you". In VISA, because the subtitles and the videos are well-aligned, we can try to determine the speaker based on the actions of the character's mouth.

Thirdly, in current VMT models, emotional information is not well captured. Object and spatial information are often not sufficient for disambiguation, and VMT models need to be able to obtain more information from the video. We observe that emotional infor-

mation is necessary for disambiguation in many cases. For example, "何言ってんだよ" would be translated into "What are you talking about?" if the speaker is confused, while it would be translated into "It does not even make sense." if the speaker is angry or has no particular emotion.

7. Conclusion and Future Work

In this paper, we introduced a new large-scale ambiguous parallel subtitles dataset for VMT research. The dataset is divided into two parts according to the cause of ambiguity. Experiments conducted on the dataset showed that the dataset is challenging for the latest VMT model. Based on the analysis of the dataset and experimental results, we pointed out possible directions to improve the VMT model.

However, there are still some problems with the dataset. The main problem is that although the subtitles are ambiguous, the videos do not necessarily contribute to the disambiguation. In the future, we plan to construct a test set for the VMT task and ensure that the videos in it can help disambiguate the ambiguous subtitles. And based on the test set, we plan to design a new VMT architecture that can obtain more information from video clips to help translate subtitles.

Acknowledgement

8. Bibliographical References

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (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, October. Association for Computational Linguistics.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1):13–47.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (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, June. Association for Computational Linguistics.

Carreira, J. and Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733. IEEE Computer Society.

Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA, 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, September. Association for Computational Linguistics.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978, Brussels, Belgium, October-November. Association for Computational Linguistics.

Gu, W., Song, H., Chu, C., and Kurohashi, S. (2021). Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92, Online, August. Association for Computational Linguistics.

Hirasawa, T., Yang, Z., Komachi, M., and Okazaki, N. (2020). Keyframe segmentation and positional encoding for video-guided machine translation challenge 2020. *arXiv preprint arXiv:2006.12799*.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. (2010). Automatic evaluation of translation quality for distant language pairs. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 944–952, Cambridge, MA, October. Association for Computational Linguistics.

Jiang, J. J. and Conrath, D. W. (1997). Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of the 10th Research on Computational Linguistics International Conference*, pages 19–33, Taipei, Taiwan, August. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Lison, P. and Tiedemann, J. (2016). OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In Lluís Màrquez, et al., editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421. The Association for Computational Linguistics.

Navigli, R. and Ponzetto, S. P. (2010). BabelNet: Building a very large multilingual semantic network. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 216–225, Uppsala, Sweden, July. Association for Computational Linguistics.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Sanabria, R., Caglayan, O., Palaskar, S., Elliott, D., Barrault, L., Specia, L., and Metze, F. (2018). How2: A large-scale dataset for multimodal language understanding. In *NeurIPS*.

Schuster, M. and Paliwal, K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

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

Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., and Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, 34(2):97–147.

Tolmachev, A., Kawahara, D., and Kurohashi, S. (2018). Juman++: A morphological analysis toolkit for scriptio continua. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Lan-*

guage Processing: System Demonstrations, pages 54–59, Brussels, Belgium, November. Association for Computational Linguistics.

Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., and Wang, W. Y. (2019). Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591.

Zhao, Y., Komachi, M., Kajiwara, T., and Chu, C. (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, November. European Association for Machine Translation.