Enhancing Video Translation Context with Object Labels

Jeremy Gwinnup[1,2], **Tim Anderson**[2], **Brian Ore**[2], **Eric Hansen**[2], **Kevin Duh**[1]
[1]Johns Hopkins University, [2]Air Force Research Laboratory
{jeremy.gwinnup.1, timothy,anderson.20, brian.ore.1, eric.hansen.5}@us.af.mil,
kevinduh@cs.jhu.edu

Abstract

We present a simple yet efficient method to enhance the quality of machine translation models trained on multimodal corpora by augmenting the training text with labels of detected objects in the corresponding video segments. We then test the effects of label augmentation in both baseline and two automatic speech recognition (ASR) conditions. In contrast with multimodal techniques that merge visual and textual features, our modular method is easy to implement and the results are more interpretable. Comparisons are made with Transformer translation architectures trained with baseline and augmented labels, showing improvements of up to +1.0 BLEU on the How2 dataset.

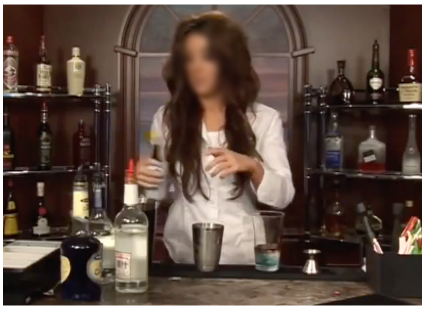
1 Introduction

Video streams are rich sources of content and the application of machine translation to videos present open research challenges. Specifically, we are interested in translating the speech content present in videos, using the visual modality as auxiliary input to improve translation quality. Intuitively, visual signals may help disambiguate under-specified words or correct speech recognition errors.

There has been much research in speech translation, which focuses on *speech* input, and multimodal machine translation, which focuses on *visual and textual* inputs; this work combines aspects of both areas. We assume a cascaded pipeline, where the speech in a video input is first passed to a speech recognition component, then the text transcripts together with the video frames are passed to a multimodal machine translation (MMT) system. Our contribution is a MMT system that augments text-based training data with labels obtained from a computer vision object detector (Fig. 1).

In contrast to more complex multimodal fusion techniques that combine vision and translation neural networks into end-to-end models, our modular approach is simple to implement, requiring no toolkit changes, and allows for easier interpretation of results.

On the How2 dataset (Sanabria et al., 2018), we experiment with using clean transcripts and automatic speech recognition transcripts of varying quality as input to our translation systems. This tests the effectiveness of our multimodal approach in noisy conditions, beneficial in real-world use cases. Results show gains of +0.4 to +1.0 BLEU on the How2 held-out test set.



src: And then you're going to stir it so have your stirrer available. PERSON CUP BOTTLE
tgt: E então você vai mexer, então tenha seu agitador disponível.

Figure 1: Demonstration of augmenting source data with detected object labels to provide additional context.

2 Object Class Label Augmentation

When considering the translation of instructional videos, the speaker's narration may use ambiguous language when describing the steps to the task as the viewer may be able to infer the intent through objects or actions in the scene. If MT systems are trained on the speaker's words and translations, these cues from the scene are not present. We proposed to address this omission by analyzing clips of the video and augmenting the text data with objects found in that clip.

Augmentation Process: To augment training data with object labels, an object recognition model

Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023), pages 130–137
July 13-14, 2023 ©2023 Association for Computational Linguistics

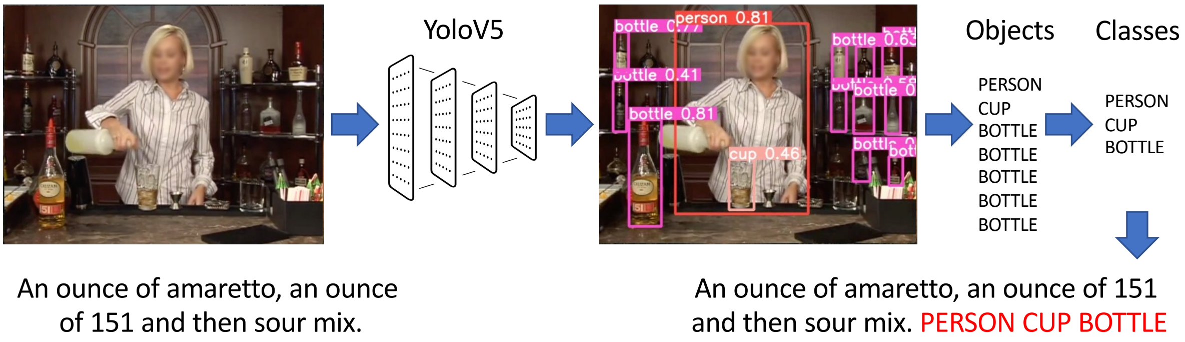

YoloV5
person 0.81
bottle 0.81
bottle 0.41
bottle 0.63
cup 0.46
Objects
Classes
PERSON
CUP
BOTTLE
BOTTLE
BOTTLE
BOTTLE
BOTTLE
PERSON
CUP
BOTTLE
An ounce of amaretto, an ounce of 151 and then sour mix.
An ounce of amaretto, an ounce of 151 and then sour mix. PERSON CUP BOTTLE

Figure 2: Illustration of the object label augmentation processing pipeline.

was applied to each of the videos in the training set in order to generate lists of objects present. To that end, we apply the YOLOv5[1] (Jocher et al., 2021) model (specifically `yolov5s`) to the 189k video clips corresponding to the utterances from the How2 training data. The object detection model can detect 80 types of objects as outlined in the COCO (Lin et al., 2015) dataset.

The detected labels for the time-slices in the video clip are collated and collapsed in order to keep final sentence length to a manageable size - we are interested in the presence of an object class versus how many times that class has occurred in the scene or the time slices in the video clip.

Once processed, the per-clip labels are appended to the source side of the training, dev and test sets as "context-markers". We do not apply these labels to the target side as we wish to generate coherent sentences in the target language. This processing pipeline is illustrated in Figure 2.

In particular, we note in the example in Figure 1 that the transcription discusses a stirrer but does not give context to what kind of stirrer: A laboratory sample stirrer, a paint stirrer, or in this case a stirrer to mix a drink. Using the object labels from the example, we see that the stirrer in this case refers to a drink - adding valuable context.

The augmented How2 corpus will be available for download at a future date.

Distribution of Augmentation Labels: When examining the counts of per-segment object class annotations in the training set (shown in Figure 3), we note that over 64% of the segments have between one and three object classes present, 13% have no detected object classes, and the remaining 23% have four or greater classes present with higher class counts forming a long tail. Full class object counts are shown in Table 1.

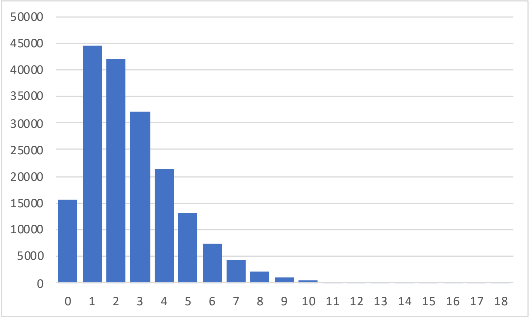


Figure 3: Training segments with N object classes detected.

Observing the most-detected class labels in training segments (shown in Figure 4), we see that PERSON is by far the most common object class with over 164k occurrences, while CUP and BOTTLE are the next most common with around 23.8k occurrences each. As How2 is comprised of instructional videos in which the authors are demonstrating how to perform a task, PERSON's high occurrence rate seems reasonable. The figure shows the top 15 object classes detected, the full list of detection counts is shown in Table 2.

While the above analyses focus on the training portion of the dataset, similar distributions are present in both the validation and test sets.

3 How2 Dataset

The How2 (Sanabria et al., 2018) dataset is a collection of instructional videos hosted on YouTube that are paired with spoken utterances, English subtitles and a set of crowdsourced Portuguese translations. Additional metadata such as video descriptions and summaries are also available. The dataset contains upwards of 2,000 hours of videos, but only a 300

[1] You Only Look Once

Classes	Segments	Classes	Segments	Classes	Segments
0	15,544	6	7,508	12	143
1	44,496	7	4,300	13	79
2	41,950	8	2,259	14	42
3	32,077	9	1,166	15	14
4	21,428	10	626	16	7
5	13,011	11	293	17	3

Table 1: Video segments with n object classes present.

Class	Count	Class	Count	Class	Count
PERSON	164,605	MICROWAVE	4,298	TOILET	1,333
CUP	23,870	REFRIGERATOR	4,014	BROCCOLI	1,327
BOTTLE	23,809	CAKE	3,911	SURFBOARD	1,281
CHAIR	17,806	DONUT	3,729	HORSE	1,222
CELL_PHONE	17,016	DOG	3,496	BED	1,141
REMOTE	16,127	TOOTHBRUSH	2,839	BOAT	1,056
BOWL	13,524	SUITCASE	2,730	BACKPACK	1,034
POTTED_PLANT	13,045	APPLE	2,714	TRUCK	924
TV	11,455	BASEBALL_GLOVE	2,682	TRAFFIC_LIGHT	919
SPORTS_BALL	10,290	SPOON	2,636	ORANGE	841
TIE	9,971	HANDBAG	2,352	COW	794
LAPTOP	9,066	COUCH	2,316	SANDWICH	763
VASE	9,033	BASEBALL_BAT	2,293	FIRE_HYDRANT	722
BOOK	7,612	BIRD	2,292	TEDDY_BEAR	713
WINE_GLASS	7,229	BANANA	2,145	AIRPLANE	576
DINING_TABLE	6,315	PIZZA	2,103	BUS	516
TENNIS_RACKET	5,922	CAT	2,054	SKIS	456
KNIFE	5,355	CARROT	1,986	SNOWBOARD	387
CAR	5,198	BENCH	1,899	TRAIN	338
MOUSE	5,107	MOTORCYCLE	1,872	ELEPHANT	265
SINK	4,688	BICYCLE	1,856	STOP_SIGN	246
FRISBEE	4,675	HOT_DOG	1,652	PARKING_METER	218
OVEN	4,450	SCISSORS	1,529	SHEEP	215
CLOCK	4,382	FORK	1,480	BEAR	198
KEYBOARD	4,353	UMBRELLA	1,408	GIRAFFE	177
SKATEBOARD	4,304	KITE	1,384	ZEBRA	158

Table 2: Detected class counts for training segments.

hour subset contains the full set of annotations. This work focuses on that subset.

	Videos	Hours	Sentences
train	13,168	298.2	184,949
validation	150	3.2	2,022
test	175	3.7	2,305

Table 3: How2 300h subset statistics

This portion consists of 13,493 videos consisting of a total run-time of 305.1 hours from which 189,276 utterances are extracted. These videos and segments are then segregated into training, validation and test sets as shown in Table 3. These segments are then used to train systems in downstream tasks such as MT.

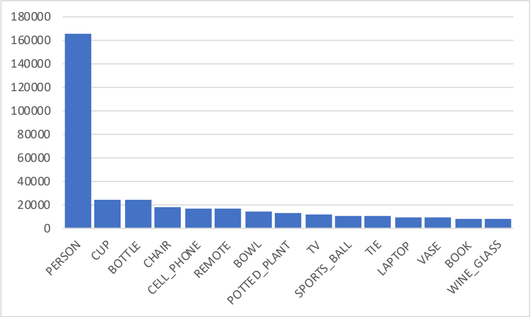


Figure 4: Top 15 classes present in training video snippets.

4 Experiments

To gauge the effectiveness of the label augmentation approach, we train baseline and object-label augmented systems in Marian (Junczys-Dowmunt et al., 2018) with a transformer-base (Vaswani et al., 2017) architecture. We also replicate the baseline and image feature augmented shallow recurrent neural network (RNN systems) described in (Sanabria et al., 2018) for comparison.

4.1 Training Hyperparameters

The Marian (Junczys-Dowmunt et al., 2018) systems trained for our experiments use transformer-base settings as described in Vaswani et al. (2017): 6-layer encoder, 6-layer decoder, 8 transformer heads, 2048 hidden units. These training sessions were performed on 2 NVidia Titan-X Pascal devices each with 12Gb GPU RAM, taking 6.5-7.5 hours per model.

4.2 Data preprocessing

In order to prepare the augmented data for use in training MT systems, we employ SentencePiece (Kudo and Richardson, 2018) unigram-model subword processing with a disjoint[2] vocabulary size of 32k. One important change we introduce is to preserve each of the COCO class labels as atomic tokens that are not broken apart. These labels are additionally in all caps to both disambiguate from natural occurrences of the label words and provide a convenient marker for diagnosis.

4.3 Pruning Over-represented Object Labels

As noted in Section 2, PERSON is by far the most represented object class label. We posit this prevalence may have a negative effect on performance. To investigate this hypothesis, we examine three methods to prune over-prevalent or under-represented object class labels: naïve dropping of the N most-represented labels, inverse document frequency (IDF) thresholding and normalized term frequency-inverse document frequency (TF-IDF) thresholding. For the first method, object labels are simply removed in the most common order - e.g. drop-3 removes the three most common classes: PERSON, CUP, and BOTTLE.

$$IDF_T = log_2 \frac{\text{Total Corpus Lines}}{\text{\# Lines with T present}} \qquad (1)$$

Inverse document frequency thresholding (as calculated by Equation 1) removes labels that fall below a specified threshold compared to a precomputed table of IDF scores for each class, effectively removing the most represented labels.

Lastly, normalized TF-IDF thresholding does the same using the product of TF (calculated by the number of times an object label occurs in video time-slices[3]) and IDF scores normalized from 0 to 1 - this tries to bring a balance between most represented labels and more unique labels that may add a distinct contribution to a translation.

4.4 ASR-Degraded experiments

The How2 dataset is provided with reference speech transcription, but in realistic settings one may need to derive these automatically. Automatic speech recognition (ASR) errors may lead to additional ambiguity in the MT input, but hopefully can be recovered partially with image context. We build Kaldi (Povey et al., 2011) ASR systems to recognize the speech of the speakers in the How2 videos, then match the ASR output timings to those of the gold-standard utterances. These new utterances are used as the source side of the training corpus for both the baseline and object label augmented condition.

In a second experiment, we add 5 dB of background noise to the audio in the How2 videos using noise samples from the MUSAN corpus (Snyder et al., 2015). The same ASR system described above is then evaluated on the noisy audio to produce a second set of ASR hypotheses.

The English speech recognition system was trained using the Kaldi ASR toolkit. The acoustic models utilized 2400 hours of audio from Fisher

[2] Separate vocabularies for English and Portuguese.

[3] This is different than our use of object class occurrences in augmentation; the larger video-timeslice object count is needed for the TF-IDF calculation to work properly.

(Cieri et al., 2004–2005), TEDLIUM-v3 (Hernandez et al., 2018), and ATC (Godfrey, 1994); the language models (LM) were estimated on 1 billion words from Fisher, News-Crawl 2007-2017 (Kocmi et al., 2022), News-Discuss 2014-2017 (Kocmi et al., 2022), and TED. This system used Mel frequency cepstral coefficient (MFCC) features as input to a factorized time delay neural network (TDNN) with residual network style skip connections. Initial decoding was performed using a finite state transducer (FST) built from a bigram LM, and the resulting lattices were rescored with a RNN LM. The vocabulary included 100k words.

4.5 Results

Armed with an array of label pruning strategies, we run a series of experiments to determine the effectiveness of each method.

4.5.1 Marian Label Augmented Systems

Marian label augmentation and pruning results are shown in Table 4 reporting scores for BLEU (Papineni et al., 2002), chrF2 (Popović, 2015) and TER (Snover et al., 2006) as calculated by SacreBLEU (Post, 2018) and COMET (Rei et al., 2020) with the default `wmt20-comet-da` model.

We note that drop-3, tfidf at 0.20, and idf at 4.0 each yield a +0.9-1.0 gain in BLEU over baseline. We also report the number of labels pruned at each experimental threshold noting that drop and tfidf remove approximately 42-43% of object class labels at maximum performance, while idf removes a much larger 74.73%.

As we see from the results, each of the three label pruning methods yields improvements over both the text-only and non-pruned augmented systems. Using the `compare-mt` (Neubig et al., 2019) tool, we take a closer look at various characteristics of the translation hypotheses of each of these five systems to see if any trends emerge. Table 5 shows averaged sentence BLEU scores for hypotheses with outputs of varying lengths. The intuition is that these average scores will help determine if a given system or pruning strategy is better at certain output lengths.

From these averaged scores, we note that plain label augmentation tends to improve over baseline with hypothesis lengths between 30 and 60 tokens but performs worse when outside of those ranges. Of the three pruning strategies, drop 3 tends to bring the most improvement, especially with shorter hypotheses and idf 4.0 tends to help the longer sequences.

4.5.2 Nmtpytorch Baseline Experiments

For nmtpytorch baseline comparison systems, we note that maximum training sequence has an effect on system performance, most likely due to the shallow RNN architecture. Table 6 shows that using the default 120 max token limit from Sanabria et al. (2018) yields better performance (+0.9-1.1 BLEU) with both the visual perturbation and our label augmentation approach. These results show our approach yields a similar performance gain.

4.5.3 ASR Noise Experiments

For the ASR-based experiments shown in Table 7, we see improvements of +0.7 BLEU with both the clean and noisy Kaldi systems. We expect that the speech-recognition based systems would not perform as well as the gold-standard systems, but the use of object labels can help mitigate this loss in performance.

4.6 Analyzing Attention Outputs

We use Marian's ability to output soft attention weights to compare an augmented system against its baseline counterpart, as shown in Figure 5. For this example, line 221 of the test set, the baseline system scores a sentence-BLEU of 30.66 versus the augmented system's 61.32. We note the attention contributions of the object labels on the output tokens. Utilizing this feature as part of an unaltered MT toolkit allows for quick and easy analysis of the benefits of object label augmentation.

5 Related Work

Perhaps most closely related to our approach is ViTA (Gupta et al., 2021), which adds object labels extracted from images in an image captioning translation task. While the motivation of adding object labels are similar, there are important differences with our setup: 1) We work on video narration of an author's task demonstration where objects appear at different points in the clip, which differs significantly from static image captions. 2) Our work focuses on training MT systems from scratch as opposed to fine-tuning existing models.

For a broad survey of multimodal translation, refer to Sulubacak et al. (2020). Specifically for video translation on How2, Sanabria et al. (2018) investigates a MT system that adds a 2048-dimensional feature vector averaging features for every 16 frames to create a global feature vector for

System	BLEU	chrF2	TER	COMET	Dropped Labels
Marian baseline	57.9	75.0	29.6	0.6819	–
nmtpy baseline	56.2	74.2	30.7	0.6234	–
nmtpy visual	55.9	74.0	31.1	0.6090	–
drop 0	57.6	74.9	29.9	0.6732	0 (0%)
drop 1	58.6	75.4	28.9	0.6785	164,605 (33.55%)
drop 2	58.7	75.5	28.9	0.6840	188,475 (38.41%)
drop 3	**58.9**	**75.7**	**28.7**	**0.6907**	**212,284 (43.26%)**
drop 4	58.5	75.3	29.1	0.6766	230,090 (46.89%)
drop 5	58.5	75.2	29.3	0.6687	247,106 (50.36%)
tfidf 0.10	58.3	75.1	29.5	0.6778	162,762 (33.17%)
tfidf 0.20	**58.8**	75.4	**28.8**	**0.6817**	**205,938 (41.97%)**
tfidf 0.30	58.8	**75.5**	29.0	0.6812	398,643 (81.24%)
idf 3.0	58.4	75.2	29.2	0.6832	212,284 (43.26%)
idf 4.0	**58.9**	**75.5**	**29.0**	**0.6887**	**366,695 (74.73%)**
idf 5.0	58.5	75.4	29.0	0.6857	428,655 (87.36%)

Table 4: Marian system scores for How2 en–pt test set, measured in BLEU, chrF2, TER and COMET. There are 490,697 object class labels present in the entire augmented training corpus.

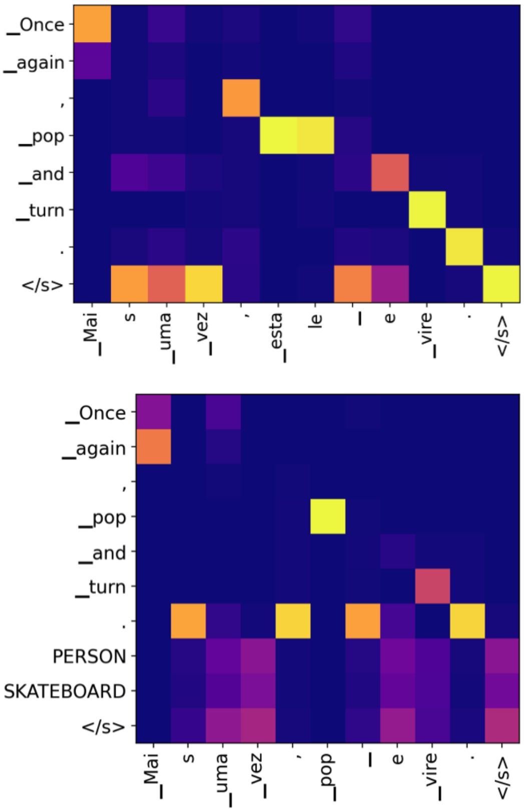


Figure 5: Attention grid for the same output sentence for Baseline (top, 30.66 sentence-BLEU) and Augmented (bottom, 61.32 sentence-BLEU) systems. We note the attention contributions of the augmented object labels.

length	base	aug	drop3	tfidf0.2	idf4.0
<10	52.7	51.8	**53.4**	52.8	53.1
[10,20)	57.6	57.1	**58.7**	58.3	57.8
[20,30)	53.7	53.6	54.8	**55.1**	55.2
[30,40)	53.1	54.1	55.4	54.9	**55.8**
[40,50)	52.4	52.0	52.9	52.6	**53.1**
[50,60)	48.3	49.3	**52.1**	49.8	48.8
>=60	46.6	44.6	45.5	47.3	**48.8**

Table 5: Averaged sentence BLEU scores for hypotheses in incremental length bins.

that entire video. This differs from our approach of creating labels solely for the objects in a clip directly corresponding to that text segment. Madhyastha et al. (2017) uses a similar approach as How2 on static imagery.

The Vatex (Wang et al., 2020) video description dataset includes a Video-guided Machine Translation (VMT) approach that utilizes an action detection model feeding a video encoder with temporal attention and a text source encoder with attention that both inform the target decoder, producing translated output from a unified network. The authors perform experiments in an video captioning setting, as opposed How2's task narration setting.

As part of the work in Calixto and Liu (2017), the authors project static image features into the

System	Max Tok	BLEU
nmtpy base	120	55.0
nmtpy vis	120	56.1
nmtpy aug	120	55.9
nmtpy base	250	56.2
nmtpy vis	250	55.9
nmtpy aug	250	55.7

Table 6: Max token length effect on BLEU for nmtpytorch baseline, visual perturbation and our label augmented systems.

System	BLEU	COMET
Kaldi clean base	52.0	0.556
Kaldi clean aug	52.7	0.583
Kaldi 5 dB noise base	50.8	0.459
Kaldi 5 dB noise aug	51.5	0.459

Table 7: Results for clean and noisy Kaldi systems for both baseline and augmented conditions.

word embedding space to produce image-based first and last words to influence word choice in their bidirectional RNN systems.

While there are a few examples of object detection as a separate task (including our work), Baltrusaitis et al. (2019) notes the rapid jump to joint representations as neural networks became popular tools for a variety of multimodal tasks, explaining the prevalence of work following that approach.

6 Future Work

Having proven our object label augmentation technique on How2, future work includes applying label augmentation to other datasets such as the VATEX (Wang et al., 2020) video description and VISA (Li et al., 2022) ambiguous subtitles datasets. Further research into the effects of ASR degraded speech and examining task-agnostic image-language models such as CLIP (Radford et al., 2021) for label augmentation may also be useful.

7 Conclusion

We present a straight-forward method to improve MT context quality by augmenting training data with objects detected in corresponding video clips. Using these augmented corpora, we realize gains of up to +1.0 BLEU over baselines without changes to the underlying MT toolkits used to build models. We additionally show improvements of up to +0.7 BLEU with object label augmentation when substituting ASR speech for gold standard inputs.

References

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

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

Christopher Cieri, David Graff, Owen Kimball, David Miller, and Kevin Walker. 2004–2005. Fisher English Training Part 1 and 2 Speech and Transcripts. Linguistic Data Consortium, Philadelphia.

John Godfrey. 1994. Air Traffic Control Complete. Linguistic Data Consortium, Philadelphia.

Kshitij Gupta, Devansh Gautam, and Radhika Mamidi. 2021. ViTA: Visual-linguistic translation by aligning object tags. In *Proceedings of the 8th Workshop on Asian Translation (WAT2021)*, pages 166–173, Online. Association for Computational Linguistics.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer*, pages 198–208, Cham. Springer International Publishing.

Glenn Jocher, Alex Stoken, Ayush Chaurasia, Jirka Borovec, NanoCode012, TaoXie, Yonghye Kwon, Kalen Michael, Liu Changyu, Jiacong Fang, Abhiram V, Laughing, tkianai, yxNONG, Piotr Skalski, Adam Hogan, Jebastin Nadar, imyhxy, Lorenzo Mammana, AlexWang1900, Cristi Fati, Diego Montes, Jan Hajek, Laurentiu Diaconu, Mai Thanh Minh, Marc, albinxavi, fatih, oleg, and wanghaoyang0106. 2021. ultralytics/yolov5: v6.0 - YOLOv5n 'Nano' models, Roboflow integration, TensorFlow export, OpenCV DNN support.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast

The views expressed are those of the authors and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government. Distribution Statement A. Approved for public release: distribution is unlimited. Originator reference number RH-22-123269. Case number AFRL-2022-3098.

neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Yihang Li, Shuichiro Shimizu, Weiqi Gu, Chenhui Chu, and Sadao Kurohashi. 2022. VISA: an ambiguous subtitles dataset for visual scene-aware machine translation. *CoRR*, abs/2201.08054.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context.

Pranava Swaroop Madhyastha, Josiah Wang, and Lucia Specia. 2017. Sheffield MultiMT: Using object posterior predictions for multimodal machine translation. In *Proceedings of the Second Conference on Machine Translation*, pages 470–476, Copenhagen, Denmark. Association for Computational Linguistics.

Graham Neubig, Zi-Yi Dou, Junjie Hu, Paul Michel, Danish Pruthi, Xinyi Wang, and John Wieting. 2019. compare-mt: A tool for holistic comparison of language generation systems. *CoRR*, abs/1903.07926.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

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

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Matthew Snover, Bonnie Dorr, Rich Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA. Association for Machine Translation in the Americas.

David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. ArXiv:1510.08484v1.

Umut Sulubacak, Ozan Çağlayan, Stig-Arne Grönroos, Aku Rouhe, Desmond Elliott, Lucia Specia, and Jörg Tiedemann. 2020. Multimodal machine translation through visuals and speech. *Machine Translation*, 34.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2020. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research.