# MeMAD Deliverable

## *D4.1 Report on Multimodal Machine Translation*

| | |
|---|---|
| Grant agreement number | 780069 |
| Action acronym | MeMAD |
| Action title | Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy |
| Funding scheme | H2020–ICT–2016–2017/H2020–ICT–2017–1 |
| Version date of the Annex I against which the assessment will be made | 3.10.2017 |
| Start date of the project | 1.1.2018 |
| Due date of the deliverable | 31.12.2018 |
| Actual date of submission | 31.12.2018 |
| Lead beneficiary for the deliverable | University of Helsinki |
| Dissemination level of the deliverable | Public |

**Action coordinator's scientific representative**
Prof. Mikko Kurimo
AALTO–KORKEAKOULUSÄÄTIÖ, Aalto University School of Electrical Engineering, Department of Signal Processing and Acoustics
mikko.kurimo@aalto.fi

*MeMAD project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069. This document has been produced by theMeMAD project. The content in this document represents the views of the authors, and the European Commission has no liability in respect of the content.*

| Authors in alphabetical order | | |
|---|---|---|
| Name | Beneficiary | e-mail |
| Stig-Arne Grönroos | Aalto University | stig-arne.gronroos@aalto.fi |
| Umut Sulubacak | University of Helsinki | umut.sulubacak@helsinki.fi |
| Jörg Tiedemann | University of Helsinki | jorg.tiedemann@helsinki.fi |



**Abstract**

Multimodal machine translation involves drawing information from more than one modality (text, audio, and visuals), and is an emerging subject within the machine translation community. In MeMAD, multimodal translation is of particular interest in facilitating cross-lingual multimodal content retrieval, and is one of the main focuses of WP4. Though multimodal machine translation efforts have been emerging since the early 1990s, there has not been research on a large scale until the last decade. Especially prominent are the multimodal tasks of spoken language translation and image caption translation, exploiting audio and visual modalities respectively. Both of these tasks are championed by evaluation campaigns, acting as competitions to stimulate research and to serve as a regulated platform investigating evaluation methodologies. So far, one multimodal machine translation system has been developed within WP4 of the MeMAD project for either task, and especially the image caption translation system had great success. In this deliverable, we present a survey of the state of the art in machine translation with an emphasis on multimodal tasks and systems. Later, we describe our own multimodal machine translation efforts carried out in WP4 within the first year of MeMAD. Finally, to conclude our report, we discuss our plans of tackling video subtitle and audio description translations as the next steps in WP4.

# Contents

# 1 Introduction

From the viewpoint of linguistics, language/verbal mode covers written, spoken, and signed language, auditory mode includes non-verbal sound and music, and co-speech gestures such as facial expressions and hand gestures comprise modes of their own. In contrast, natural language processing (NLP) has a more practical division of modalities based on differences in data representation. The modalities typically associated with NLP are text, audio, and visuals. While text straightforwardly encodes written language, the other two modalities can correspond to different channels of communication. The obvious example for audio is spoken language, though it is possible to explore the modality in other sounds and even music. Visuals are diverse, and though the visual parallel to spoken language is perhaps signed language, visual co-speech gestures also play linguistic roles, and other visual media such as images and videos can be associated with language. Text is by far the most common modality among these three in NLP literature, likely owing to its ease of processing, and wide availability in a variety of different forms.

Multimodal NLP tasks are those involving more than one modality, either by using information from one modality to aid the interpretation of language in another modality, or by converting language between modalities. Many multimodal NLP tasks are multimodal extensions of language analysis tasks initially modelled as unimodal discrete classification tasks. Some examples of these tasks could be emotion detection, multimodal named entity recognition, multimodal sentiment analysis, and visual question answering. For other tasks that involve modality conversion, one well-known example is automatic speech recognition (ASR), the task of transcribing spoken language audio into text. Speech synthesis is the converse of ASR, with the goal of generating speech from written language. Media description tasks (*e.g.* image captioning, video summarisation) aim at processing visuals and/or audio (*e.g.* photos, video clips) to generate interpretive language in text (*e.g.* captions, summaries).

While transmodal conversion tasks necessarily exercise some language interpretation to restructure the input and re-express it in the output modality, this is still not regarded as translation, unless there would be more than one natural language involved. In contrast, the tasks that fall under the umbrella of multimodal machine translation (MT) both include multiple modalities and non-matching input and output languages. Some of the major multimodal MT tasks such as image caption translation, sign language translation, spoken language translation, and video subtitle translation are described in detail in Section 6.1. Dealing with intermodality and translation at the same time, multimodal MT tasks are fairly hard challenges, and while vision, speech, and language processing communities have worked largely apart in the past, the rising interest in tackling these tasks has brought them together.

Furthermore, conventional text-based MT has been recently enjoying widespread success with the adoption of deep learning architectures. One implication that this has had for

multimodal MT was the significantly increased need for large datasets in order to keep up with the heavily data-driven state-of-the-art methodologies. Considering that data availability varies greatly between different languages, introducing the additional requirement of another modality becomes very restrictive. For any given multimodal MT task, there is virtually no training data available for most language pairs, and only limited availability for others. Another matter is that the utility of multimodal MT is sometimes disputed in recognition of the success of text-based MT. Regardless, multimodal MT is a better reflection of how humans acquire and process language, with many theoretical advantages in language grounding over text-based MT (see Section 6) as well as the potential for new practical applications like cross-modal cross-lingual information retrieval.

General surveys of multimodality in NLP tasks exist (*e.g.* Baltrušaitis et al. (2017)), but so far, there is no similar study for multimodal MT in particular. For this reason, this deliverable was structured as a review of the multimodal MT literature with accompanying discussions especially as relating to the MeMAD project. Section 2 outlines the early efforts in multimodal MT predating the advance of the current state of the art in MT. Section 3 reviews the methods and caveats of evaluating MT performance, and discusses on multimodal MT evaluation campaigns. Section 4 contains an overview of the datasets suitable as training or test corpora for multimodal MT. Section 5 is a brief summary of the state of the art in unimodal MT, serving as a basis for Section 6 describing various multimodal MT tasks and the diverse set of approaches used to address them. Finally, Section 7 presents the current status and research agenda of WP4 in MeMAD, followed by Appendices A and B displaying system description papers that recently came out of WP4.

## 2 Early efforts

The current state of the art in text-based machine translation produces fairly satisfactory results in a restricted domain, and multimodality is often regarded as an additional challenge to implement in preparation for the next level in machine translation. However, there was a great deal of interest in doing machine translation with non-text modalities even before the subject was streamlined with the arrival of successful statistical machine translation models (*e.g.* the Candide system (Berger et al., 1994)). Among one of the earliest attempts is the Automatic Interpreting Telephony Research project (Morimoto, 1990), a 1986 proposal that aimed at implementing a pipeline of automatic speech recognition, rule-based machine translation, and speech synthesis, in order to have a full speech-to-speech translation pipeline. Unfortunately, the project completed its seven-year term in 1993 without delivering a finished product. The idea of *telephony interpretation* was quite well-received, and encouraged further research in speech-to-speech translation, making use of components that had been in development for the last two decades. Starting from this period until the early 2000s, several speech-to-speech translation systems were developed and re-

leased, using the same basic setup of components but with a wide array of methods to couple them and fine-tune translations (Zhang, 2003), such as the JANUS system (Lavie et al., 1997), MATRIX (Takezawa et al., 1998), VerbMobil (Wahlster, 2000), EUTRANS (Pastor et al., 2001), NESPOLE! (Lavie et al., 2001), the TONGUES system (Black et al., 2002), and IBM Mastor (Gao et al., 2006).

In contrast with audio data as in speech-to-speech translation, the use of visual data in translation has not attracted comparable interest until quite recently. This is perhaps owing to the fact that the equivalent of the Consortium for Speech Translation Advanced Processing (C-STAR) did not exist at the time to incentivise visual processing, or because visual cues such as facial expressions and gestures are not as salient as voice cues in human communication, resulting in a weaker analogy for machine translation. Regardless, at present, between image captions, instructional text with photographs, video recordings of sign language, and subtitles for videos (and especially movies), there are a multitude of new archetypes of multimodal composition for which machine translation has become more relevant than ever before. Due to this shift in priorities, modern multimodal MT subjects dealing with visual (or audiovisual) cues are just as prominent as those tackling audio.

## 3 Evaluation

Evaluating the performance of a machine translation system is a difficult and controversial problem in itself. Typically, there are numerous ways of translating even a single sentence for human translators, and it is often unclear which one is (or which ones are) better, and in what respect, given that the relevant evaluation criteria are multi-dimensional, context-dependent, and highly subjective. Human analysis of translation quality is often divided into the evaluation of adequacy (semantic transfer) and fluency (grammatical soundness). The reason for this is to manage the ambiguity about what is under evaluation as well as to be able to compare and contrast translation systems more clearly. Even then, this aspectual division is sometimes criticised for disregarding pragmatic and functional equivalence measures. Regardless, even with a fine-grained division, evaluation can still be quite arbitrary. In the midst of this controversy, certain well-defined human evaluation methods and automatic evaluation metrics stand out as good compromises between feasibility and completeness. While these metrics have changed over time and may yet change further, the accepted practice in the MT community is to agree upon the better metrics and use them consistently in evaluating translation systems, being conscious of the fact that these would be not absolute but relative measures.

### 3.1 Metrics

Among the various evaluation metrics in the literature, the most commonly used ones are BLEU (Papineni et al., 2001), METEOR (Lavie and Agarwal, 2007), and TER (Snover et al., 2006). To summarise briefly, BLEU is based on an aggregate precision measure of n-grams and penalises translations that are too short, METEOR accounts for and gives partial credit to stem matches, synonyms, and detected paraphrases, and TER is a variant of word-level edit distance between the source and the target. BLEU is typically associated with fluency, METEOR with adequacy, and TER with post-editing utility. BLEU is by far the most common automatic evaluation metric, and most non-detailed quantitative comparisons of machine translation models are content to use only BLEU scores. Both BLEU and METEOR, much like the majority of other evaluation metrics developed so far, are *example-based* metrics built on statistics drawn from the training data associated with the translation model. These metrics are inadvertently heavily biased on the translation styles that they see in the training data, and end up penalising any alternative phrasing that might be equally correct.

Human evaluation is the optimal choice when a trustworthy measure of translation quality is called for and resources to perform it are available. The usual strategies for human evaluation are fluency and adequacy rankings, direct assessment (DA), and post-editing evaluation (PE). Fluency and adequacy rankings are conventionally between 1–5, while DA is a general scale between 0–100 indicating how complete the translation is, either in reference to either the original sample in the source language (DA-*src*), or the ground truth sample in the target language (DA-*ref*). A common critique for these methods is that the assigned scores can be quite arbitrary. On the other hand, in PE, human annotators are asked to *correct* translations by changing the words and the ordering as little as possible, and the rest of the evaluation is based on an automatic edit distance measure between the original and post-edited translations. All of these human evaluation methods are typically crowdsourced to non-expert annotators to reduce expenses, due to limited research funding. While this may still result in consistent evaluation scores, it is a recognised fact that professional translators capture more details and are generally better judges than non-expert speakers (Bentivogli et al., 2018).

The problems recognised even in human evaluation methods substantiate the notion that no metric is perfect. In fact, evaluation methods have become an active research subject in their own right (Ma et al., 2018; Specia et al., 2018). However, there is currently little research on developing evaluation approaches specifically tailored to multimodal translation. Currently, all automatic evaluation is strictly text-based, while only indirect methods such as lexical translation accuracy (see Section 6.1.4) carry a focus on multimodality. In human evaluation, perhaps the only particular example is the addition of source images in the direct assessment of image caption translations (Elliott et al., 2017; Barrault et al., 2018). Having consistent methods to evaluate how well translation systems take

multimodal data into account would make it possible to identify bottlenecks and facilitate future development. For now, shared tasks are the flag-bearers in the multimodal machine translation community making the only organised effort to investigate the best practices for evaluation.

## 3.2 Shared tasks

A great deal of natural language processing system development research is made in preparation for shared tasks under academic conferences and workshops, and the relatively new subject of multimodal machine translation is not an exception. These shared tasks lay out a specific experimental setting, for which participants submit their own systems, often developed using the training data provided by the campaign. Currently, there are not many datasets encompassing both multiple languages and multiple modalities, that are also of sufficiently high quality and large size, and available for research purposes. However, multilingual datasets that augment text with only speech or only images are somewhat less rare than those with videos, given their utility for tasks such as automatic speech recognition and image captioning. Adding parallel text data in other languages enables such datasets to be used for spoken language translation and image caption translation, both of which are represented in shared tasks organised by the machine translation community. The International Workshop on Spoken Language Translation (IWSLT) has led an annual evaluation campaign on speech translation since 2004, and the Conference on Machine Translation (WMT) has been running shared tasks for image caption translation annually since 2016.

### 3.2.1 IWSLT evaluation campaign

The speech translation evaluation campaign started out in 2003 as an exclusive event for the members of the Consortium for Speech Translation Advanced Research (C-STAR), with the aim of investigating the application of evaluation methodologies to the newly-developing translation technologies at the time (see Section 2). After a closed first rendition, the campaign became an open shared task under IWSLT in 2004 (Akiba et al., 2004). The first years of the campaign were based on the internal Basic Travel Expression Corpus (BTEC), a dataset containing basic tourist utterances (*e.g. "Where is the restroom?"*) and their transcripts. The corpus was eventually extended with more samples (from a few thousand to tens of thousands) and more languages (from Japanese and English, to Arabic, Chinese, French, German, Italian, Korean, and Turkish). Each year also had a new challenge theme, such as robustness of speech translation, spontaneous (as opposed to scripted) speech, and dialogue translation, introducing corresponding data sections (*e.g.* running dialogues) as well as sub-tasks (*e.g.* translating from noisy ASR output) to facilitate the challenges. Starting from 2010, the campaign adopted TED talks as their primary

training data (Paul et al., 2010), and eventually shifted away from the tourism domain towards lecture transcripts.

Until IWSLT 2016 (Cettolo et al., 2016), the evaluation campaign has been handled under three main tracks: Automatic speech recognition, text-based machine translation, and spoken language translation. While these tasks involve different sources and diverging methodologies, they converge on plain text output. The organisers have made considerable effort to use several automatic metrics at once to evaluate participating systems, and to analyse the outputs from these metrics. Traditionally, there has also been human evaluation (fluency and adequacy assessment) only on the most successful systems for each track according to the automatic metrics. These assessments have been used to investigate which automatic metrics correlate with which human assessments to what extent, and to pick out and discuss drawbacks in evaluation methodologies.

Additional tasks such as multilingual translation and dialogue translation (Cettolo et al., 2017), and low-resource speech translation (Niehues et al., 2018) were reintroduced to the IWSLT evaluation campaign from 2017 on as both the TED data and machine translation literature grew richer. We submitted our own speech translation system to the 2018 lecture translation track representing the MeMAD project (Sulubacak et al., 2018).

#### 3.2.2 WMT multimodal translation task

The Conference on Machine Translation (WMT) has organised multimodal translation shared tasks annually since the first event (Specia et al., 2016) in 2016, and it is currently the only evaluation campaign of its kind. The first shared task was such that the participants were given images and an English caption for each image as input, and were required to generate a translated caption in German. The second shared task had a similar experimental setup, but added French to the list of target languages. The third shared task in 2018 added Czech as a third possible target language. This last task also had a secondary track which only had Czech on the target side, but allowed the use of English, French and German captions together along with the image in a multisource translation setting.

The WMT multimodal translation shared tasks evaluate the performances of submitted systems on several test sets at once, including the Ambiguous COCO test set (Elliott et al., 2017), which incorporates image captions that contain visually-resolvable ambiguity (see Section 4.2). The translations generated by the submitted systems are scored initially by the BLEU, METEOR, and TER metrics. In addition, all participants are required to devote resources to manually scoring anonymised translations. This scoring is done by direct assessment, variously using the original source captions and the ground truth translations as reference. The images are also shown to the annotators as an additional reference for scoring, observing the multimodality aspect. During the assessment, ground truth translations are shuffled into the outputs from the submissions, and scored just like them. This establishes an approximate reference score for the ground truth, and the aggregate scores

| Dataset | Samples | Languages | Audio | Visuals |
|---|---|---|---|---|
| Flickr8k | 8k images, 41k captions | EN, TR, ZH | | ✓ |
| Flickr30k | 30k images, 158k captions | DE, EN | | ✓ |
| Multi30k | 30k images, 30k captions | CS, DE, EN, FR | | ✓ |
| MS COCO (captions) | 123k images, 617k–820k captions | EN, JA | | ✓ |
| MS COCO (ambiguous) | 461 images and captions | CS, DE, EN, FR | | ✓ |
| How2 | 13,493 clips, 189k segments | EN, PT | ✓ | ✓ |
| TED-LIUM 3 | 2,351 talks, 268k segments | EN | ✓ | |
| TED-TRANS | 1,565 talks, 171k segments | DE, EN | ✓ | |
| YLE - May 2014 | 70 clips, 2026 segments | FI, SV | ✓ | ✓ |
| YLE - Strömsö | 35 clips, 7090 segments | FI, SV | ✓ | ✓ |
| YLE - Spotlight | 15 clips, 2456 segments | FI, SV | ✓ | ✓ |

**Table 1:** A summary of statistics from some datasets relevant to multimodal machine translation.

for the submissions are analysed in proportion to this.

While the first two WMT multimodal translation shared tasks predate the MeMAD project, we participated in the 2018 event. Following the languages the project focuses on, we only participated in the single-source multimodal translation track, and only for the target languages French and German. Our system (Grönroos et al., 2018) had great success (see Appendix A) according to both automatic and human evaluation results, not only coming first place in the evaluation campaign, but also surpassing the second-best systems by a large margin (Barrault et al., 2018).

## 4 Datasets

Data availability is often cited as the most important factor in successfully training the data-driven NLP architectures commonly used today. Unfortunately, all multimodal NLP tasks, and especially multimodal machine translation (due to its simultaneous requirement of multimodality and multilinguality in training data) is subject to a data bottleneck. Thankfully, this issue is eventually getting better recognition, and there are more and more datasets released, that are useful for multimodal machine translation. Some of these datasets are outlined in Table 1, and explained in more detail in the subsections to follow.

### 4.1 Flickr image captions

**Flickr8k** Released in 2010, the Flickr8k dataset (Rashtchian et al., 2010) has been one of the most widely-used multimodal corpora. Originally intended as a high-quality


**Figure 1:** An example from the Flickr8k dataset.

A man in street racer armor is examining the tire of another racers motor bike.
The two racers drove the white bike down the road.
Two motorists are riding along on their vehicle that is oddly designed and colored.
Two people are in a small race car driving by a green hill.
Two people in racing uniforms in a street car.

training corpus for automatic image captioning, Flickr8k comprises a diverse set of 8,092 images extracted from Flickr, each with 5 crowdsourced captions in English that describe the image. Unlike some of the earlier image caption datasets such as Grubinger et al. (2006), Flickr8k incorporates shorter captions that focus on describing only the most salient objects and actions. As the dataset has been a popular and useful resource, it has been extended over the years with other languages such as Chinese (Li et al., 2016) and Turkish (Ünal et al., 2016) with independently crowdsourced captions. Despite its relatively small size, Flickr8k retains a strong relevance to multimodal machine translation with these multilingual extensions. An example from this dataset can be seen in Figure 1.

**Flickr30k** The Flickr30k dataset (Young et al., 2014) was released in 2014 as a larger dataset following in the footsteps of Flickr8k. Collected using the same crowdsourcing approach for independent captions as its predecessor, Flickr30k contains 31,783 photos depicting common scenes, events, and actions, each annotated with 5 independent English captions. Unlike Flickr8k, this dataset has not been extended with independently annotated image captions in other languages. The collective effort that led to the creation of Multi30k (Elliott et al., 2016), a multilingual subset of Flickr30k, also produced professional human translations of all Flickr30k captions into German, and collected 5 independent German captions per image. This version of the dataset has been used as training and development data for the first (Specia et al., 2016) and second (Elliott et al.,

**Figure 2:** An example from the Flickr30k dataset.

Ballerinas performing a dance routine wearing various colored dresses holding hands with there right legs held up.
Several women are performing ballet on stage with colorful leotards and tutus.
A harmonized moment in a colorful ballerina show.
Dancers on a stage in different color dresses
ballerinas dancing on a stage

2017) multimodal translation shared tasks at WMT, with an overt disclaimer that the German captions are not independent annotations. An example from this dataset can be seen in Figure 2.

**Multi30k** Multi30k (Elliott et al., 2016) was initially released in 2016 as a bilingual dataset of English and German image captions following the Flickr sets, with the aim of stimulating multimodal and multilingual research from the beginning. Although the dataset uses the same selection of images as Flickr30k, it includes only 1 out of 5 captions for each image. The WMT multimodal translation shared tasks of the last two years introduced French (Elliott et al., 2017) and Czech (Barrault et al., 2018) extensions to Multi30k, making it a staple dataset for the task, and further expanding the set's utility to cutting-edge applications such as multisource training. An example from this dataset can be seen in Figure 3.

**WMT datasets** The past three years of multimodal shared tasks at WMT each came with a designated test set for the task. Totalling 3,017 images in the same domain as the Flickr sets (including Multi30k), these sets are too small to be used for training purposes, but could smoothly blend in with the other Flickr sets to expand their size. So far, test sets from the previous shared tasks (each containing roughly 1,000 images with captions) have been allowed for augmenting the validation set for the current year's task. In parallel with the language expansion of Multi30k, the test from 2016 contains only English and German captions, and the one from 2017 contains only English, German, and French.

**Figure 3:** An example from the Multi30k dataset.

(EN) Mexican women in decorative white dresses perform a dance as part of a parade.
(DE) Mexikanische Frauen in hübschen weißen Kleidern führen im Rahmen eines Umzugs einen Tanz auf.
(FR) Les femmes mexicaines en robes blanches décorées dansent dans le cadre d'un défilé.
(CS) Součástí průvodu jsou mexičanky tančící v bílých ozdobných šatech.

## 4.2 MS COCO

The Microsoft Common Objects in Context (MS COCO) dataset (Lin et al., 2014) was first released in 2014 as a large-scale training corpus for object detection and segmentation. The corpus features around 330,000 images, out of which over 200,000 have been annotated with the bounds and labels of 1.5 million concrete objects from 80 categories in a variety of visual contexts. The initial object detection and segmentation layers of COCO are intended for image processing tasks, but a large subset of these images were also subsequently annotated with captions, enabling the set to be used in multimodal NLP tasks such as automatic image captioning (see MeMAD Deliverable 2.1 for further discussions).

**COCO Captions** Introduced in 2015, the COCO Captions set (Chen et al., 2015) forms an additional caption annotation layer for a subset of roughly 123,000 images from MS COCO. Each image in this dataset is associated with up to 5 independently annotated captions in English, with a total of 616,767 captions (414,113 in the official training set split, with 202,654 reserved for validation). Although this is a monolingual dataset, its large size alleviates a well-recognised bottleneck of insufficient multimodal data. For this reason, it is still useful for multimodal machine translation in combination with data augmentation methods such as synthetic data generation, as demonstrated in Grönroos et al. (2018). Crowdsourcing methods have been recently used to generate Japanese captions parallel to the original English, resulting in the STAIR Captions dataset (Yoshikawa et al., 2017). So far, these are the only two languages in which captions are available for the COCO images. An example from this dataset can be seen in Figure 4.

**Ambiguous COCO** Another dataset introduced to the literature by the WMT multimodal

**Figure 4:** An example from the COCO Captions and Ambiguous COCO datasets.

(EN) A wedding being set up with one of the umbrellas blowing away.
(DE) Vorbereitungen für eine Hochzeit, wobei einer der Sonnenschirme weggeweht wird.
(FR) Un mariage est en train d'être installé et un des parasols qui s'envole.

translation shared task organisation is Ambiguous COCO (Elliott et al., 2017). Released in 2017 as an additional test set for the year's shared task, Ambiguous COCO constitutes a small subset of COCO with English captions containing potentially ambiguous words (note the translation of the word 'umbrella' to *'Sonnenschirm'* in German rather than *'Regenschirm'*, and to *'parasol'* in French rather than *'parapluie'* in Figure 4). These captions were then translated to German and French, and further filtered so that there is a balanced representation of each word sense for each verb. Both the 2017 and 2018 shared tasks attested to translation systems that consistently fared worse on Ambiguous COCO compared to the Flickr test sets. Regardless, it is under question whether Ambiguous COCO is a difficult test set because of its ambiguous samples, and whether these ambiguities are visually resolvable after all (Elliott et al., 2017).

### 4.3 TED talk transcripts

TED talks comprise a rich resource of spoken language produced by a large variety of English speakers. Since both video recordings and transcripts of the talks are available for research purposes, they have been useful as training data for speech processing systems. Furthermore, many TED talks have professional human translations of the transcripts to provide access to speakers of different languages. Currently, the TED Corpus Search Engine (TCSE) (Hasebe, 2015) has indexed a total of 2,857 talks with parallel transcripts available in up to 29 languages. However, these transcripts are segmented into utterances and arbitrarily divided for subtitling, and it is difficult to get accurate sentence-level segmentations of the transcripts. Even then, aligning these segments to the audiovisual content or to each other in source and target languages are two further challenges. While

a number of corpora have been made available that address either challenge, there haven't been many that address both.

**WIT**$^3$ The Web Inventory of Transcribed and Translated Talks (WIT$^3$) (?) is a resource released in 2012, with the aim of utilising the freely available parallel data published by TED and distributing aligned transcripts for use in machine translation. At the time of release, the corpus was compiled from around 1,000 english talks translated into 80 languages. WIT$^3$ is still being maintained and continually growing, currently incorporating thousands of talks in 109 languages. While it is a very large resource for parallel texts, the distributed transcripts are not aligned with the audiovisual content, and so they are not in their raw form suitable for multimodal language processing systems.

**TED-LIUM** Introduced first in 2012 and later augmented with more data in 2014 and 2018, the TED-LIUM corpus (Rousseau et al., 2014; Hernandez et al., 2018) currently contains 452 hours of transcribed TED talks in English automatically segmented and aligned to time spans on audio. The entire set contains 2,351 talks that amount to 268,231 segments and roughly 4.9 million words, making the corpus one of the largest datasets available as training data for ASR systems. In their latest release, the authors also provide a subset filtered from the full data that is balanced in certain characteristics (duration, speaker gender, number of speakers) to adapt the set for speaker adaptation experiments. However, this data set is currently monolingual, and it is not suitable for use in speech translation experiments without preprocessing.

**IWSLT datasets** The IWSLT workshop organises a speech translation evaluation campaign each year and publishes a new test set for each of the campaigns. Since 2011, IWSLT test sets have been compiled by WIT$^3$ from TED talks and their transcripts. While their sizes are too small to allow training, these sets feature English transcripts aligned both with the audio and with the translations. For the first time as part of the 2018 evaluation campaign, the organisers released the IWSLT Speech Translation TED Corpus (Niehues et al., 2018). This corpus is a large speech translation dataset composed of 1,565 English TED talks with 170,965 segments aligned with both the audio and German translations. Collectively, these IWSLT datasets are more suitable for multimodal machine translation systems to exploit compared to the other TED corpora.

### 4.4 The How2 dataset

The How2 dataset (Sanabria et al., 2018) is a collection of 79,114 clips with an average length of 90 seconds, containing a total of roughly 2,000 hours of instructional YouTube videos in English, spanning a variety of topics. The dataset is intended as a resource for several multimodal tasks, such as multimodal ASR, multimodal summarisation, spoken

language translation, and video subtitle translation. To establish cross-modal associations, the videos in the dataset were annotated with word-level alignments to ground truth English subtitles. There are also English descriptions of each video written by the users who uploaded the videos, added to the dataset as metadata corresponding to video-level summaries. Currently, for multimodal translation purposes, a 300-hour subset covering 22 different topics is available with crowdsourced Portuguese translations, but translations are ongoing, and a 480-hour expanded subset is under preparation and scheduled to be released soon. While the dataset is quite promising for the MeMAD project and for multimodal MT in general, it is also quite recent, and it has not yet been used as a translation benchmark. An example from this dataset can be seen in Figure 5.

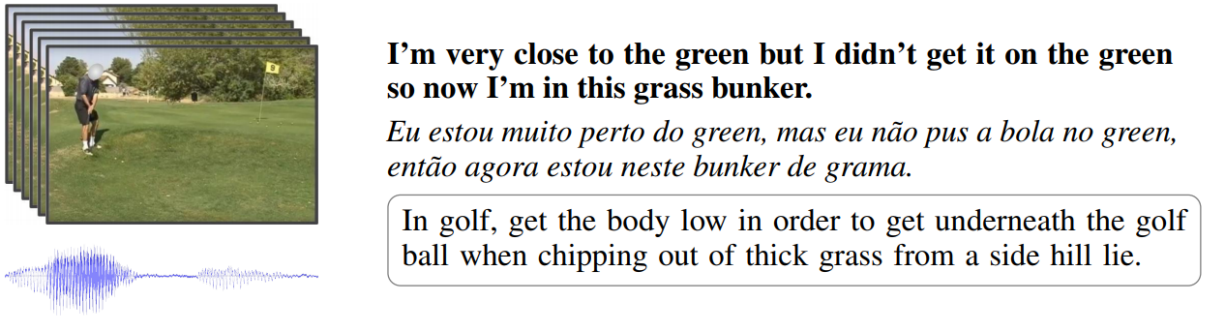


**Figure 5:** An example from the How2 dataset, retrieved from Sanabria et al. (2018).

### 4.5 YLE show transcripts

In addition to the freely available datasets outlined in this section, partners of the MeMAD project will also have access to a substantial amount of proprietary audiovisual data provided by the Finnish broadcasting company Yle. Throughout the course of the project, Yle has agreed to provide 500 hours of TV programs from its archives, which will be built and released iteratively to cater to the different lines of research going on in the project (*e.g.* multilingual subtitling, content description, multimodal machine translation). Currently, there are approximately 66 hours of released content, and a large part of it is already suitable for benchmarking the multimodal translation models to be developed within WP4. The three datasets corresponding to this part (also outlined previously in deliverable 1.2) are summarised again below.

**May 2014 archives** The May 2014 archive data (`Yle_MeMAD_006_may2014AV2_1`) contains 85 items of Yle TV news, current affairs, and factual programming of in-house production transmitted during the period 19 to 26 May 2014. As relevant to WP4, the dataset contains

approximately 23.5 hours of video with time-aligned content descriptions, and bilingual subtitles in Finnish and Swedish.

**Strömsö** The Strömsö set (`Yle_MeMAD_001_Stromso01_1`) contains 35 episodes of the lifestyle program "Strömsö", from the year 2017. As relevant to WP4, the dataset contains approximately 16.5 hours of video with time-aligned content descriptions, and bilingual subtitles in Finnish and Swedish.

**Spotlight** The Spotlight set (`Yle_MeMAD_003_Spotlight01_1`) contains 15 episodes of the current affairs documentary series "Spotlight", from the year 2017. As relevant to WP4, the dataset contains approximately 7 hours of video with time-aligned content descriptions, and bilingual subtitles in Finnish and Swedish.

## 5 Unimodal machine translation

While the state of the art in machine translation was defined by statistical machine translation (SMT) methodologies for at least two decades, the field made a shift towards neural machine translation (NMT) techniques in the early 2010s. Inspired by the successful use of deep neural machine learning architectures in NLP systems such as automatic speech recognition (Graves et al., 2013), the pioneering study on NMT by Kalchbrenner and Blunsom (2013) used recurrent and convolutional neural networks to tackle machine translation. The continuous vector representations used in NMT encode various kinds of linguistic information in a shared space, fully automating the learning task and eliminating the need for hand-crafted linguistic features.

After this outbreak of interest in NMT research, there has been a plethora of studies featuring different deep neural architectures and learning methods. The application of RNNs (Elman, 1990) and other recurrent architectures (see Figure 6), such as bidirectional RNNs (Schuster and Paliwal, 1997), LSTMs (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) and GRUs (Chung et al., 2014), introduced further diversity into the field. These more advanced neural units were not as susceptible to the problems initially perceived in NMT: They were naturally dealing with variable-length sequences, and had clear computational advantages as well as superior performance. While these architectures were also used for training models for phrase-based SMT as in Cho et al. (2014b), the early encoder-decoder (sequence-to-sequence) NMT applications by Sutskever et al. (2014) and Cho et al. (2014a) that used them had an overall much larger impact. The performance of the NMT systems that followed came close to, and eventually surpassed, that of the state-of-the-art SMT systems.

Until the introduction of the attention mechanism (see Figure 7) by Bahdanau et al. (2015), one of the problems persisting in NMT was the difficulty of learning long-range de-

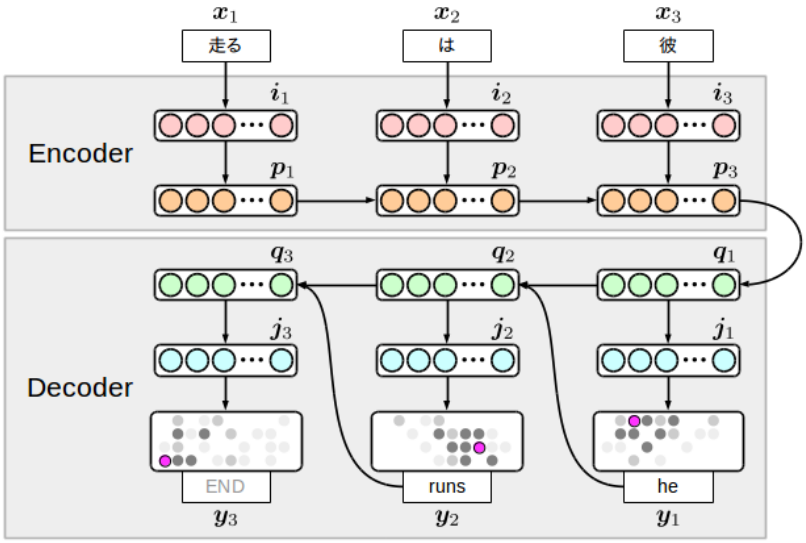


**Figure 6:** A visualisation of the encoder-decoder architecture, showing the intermediate states of recurrent units in a Japanese-to-English translation example. (Retrieved from `https://goo.gl/XDf66h`.)

pendencies in translation sequences (*e.g.* grammatical agreement in very long sentences). The attention mechanism addressed this issue by simultaneously learning to align translation units (see Figure 8), and providing a window into the relevant input units for each decoding step. Providing partial human readability for translation processes as well as a way for translation systems to avoid having to cram too much information in a fixed-size vector, the attention mechanism became a staple in sequence-to-sequence NMT. Successful alternative approaches are still brewing, such as Gehring et al. (2017) passing up RNNs in favor of convolutional learning with attentional layers, and the fully self-attentional transformer by Vaswani et al. (2017) which is causing a great deal of hype as evident from the MT submissions to all major NLP conferences this year.

So far, virtually all state-of-the-art NMT systems have used supervised deep learning methods that rely on large amounts of parallel data. However, parallel datasets are a resource with varied availability, and can be very challenging to find for some low-resource languages. In contrast, monolingual datasets are typically easier to obtain, and reasonably-sized monolingual data may be available even for some under-resourced languages. Nonetheless, utilising monolingual data short of synthetic data generation methods such as back-translation is not possible with the bilingual-data-driven architectures in use. Considering this fact, the pioneering work by Ravi and Knight (2011) used monolingual corpora to train a translation system to *decipher* the source language to produce maximally fluent output in the target language. Recently, combining the idea of unsupervised MT with

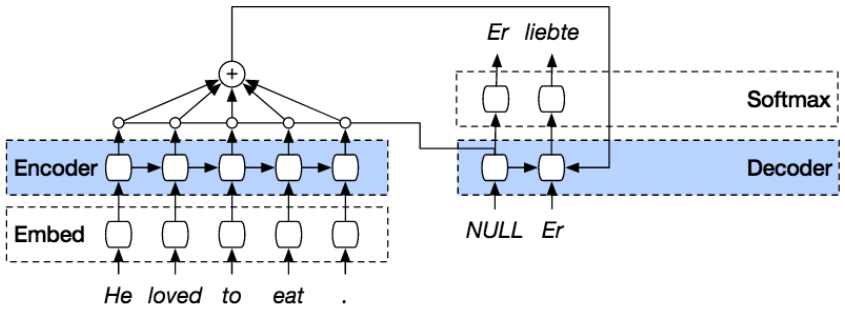


**Figure 7:** A visualisation of the attention layer over the entire input sequence, conditioned on the previously decoded unit for a given decoding step. (Retrieved from `https://goo.gl/c8KF6V`.)

state-of-the-art translation architectures, Artetxe et al. (2018) and Lample et al. (2018a,b) learn unsupervised word embeddings from monolingual corpora and iteratively map them to a shared latent space to bootstrap unsupervised MT systems. While these systems do not perform as well as supervised models yet, they still seem to be able to produce fairly fluent translations given the constraints on training.

Currently, open-source implementations of the aforementioned state-of-the-art machine translation architectures can be accessed through several freely available MT toolkits, such as Moses (Koehn et al., 2007) for SMT, and Sockeye (Hieber et al., 2017), OpenNMT (Klein et al., 2017, 2018), Marian (Junczys-Dowmunt et al., 2018), and Tensor2Tensor (Vaswani et al., 2018) for NMT.

Unless specified otherwise, *machine translation* is typically interpreted as a task of text-based translation between a single pair of languages. However, recently there has been growing interest in other translation configurations, fuelled in part by the flexibility of neural methods in incorporating different data sources. One example of this development is multilingual translation, in which the constraint of a single language pair is lifted. Different parameter sharing strategies for multilingual translation have been explored, including separate encoders/decoders (Luong et al., 2015), full sharing combined with target language tag (Johnson et al., 2016), and more recently contextual parameter generators (Platanios et al., 2018). Within the MeMAD project, we have explored how translation into a morphologically complex target language can be improved using resources from a related higher resource language. Improving the consistency of segmentation between the related target languages improves the cross-lingual transfer (Grönroos et al., 2018). The success of neural machine translation methods in learning to model multiple languages simultaneously, raises hopes that similar architectures will have success incorporating inputs of different modalities.

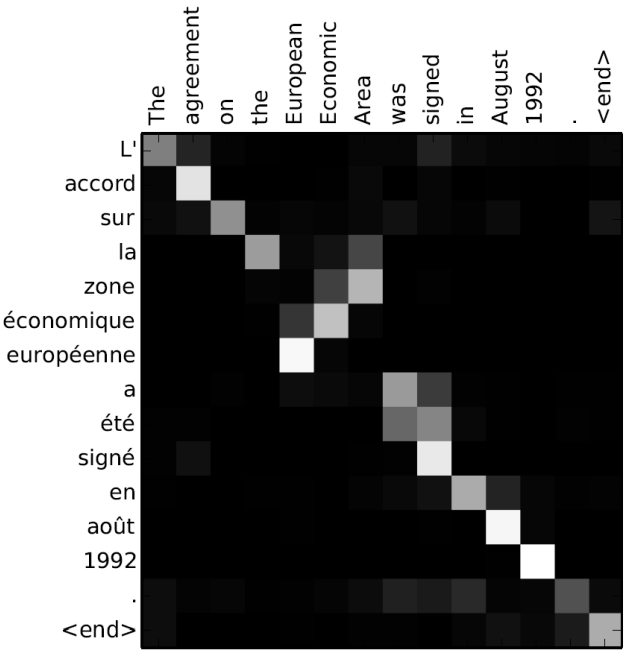


**Figure 8:** A visualisation (Bahdanau et al., 2015) of attention values for an English-to-French translation example, demonstrating the learned alignment model.

# 6 Multimodal machine translation

An example in which the default machine translation assumptions are widened is multimodal machine translation, where the requirement is to have at least two different modalities (*e.g.* text, audio, visuals) collectively pertaining to the source and target side. While this still allows many possible configurations, the typical configuration has text output as the target side, and the multimodality comes from a single non-text modality either augmenting or replacing the source text as the input.

## 6.1 Tasks

### 6.1.1 Image caption translation

The task of translating image captions has become well-known in the multimodal MT community, owing to the WMT multimodal translation shared tasks (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018) which have used it as the basis for their evaluation benchmark since 2016. The caption translation task provides a set of images with a text captions for each, while the goal is to translate the captions from the source language to a target language, considering also the visual cues in the image that might be relevant.

For this task, there have been 25 dedicated systems (*e.g.* Caglayan et al. (2016), Caglayan et al. (2017), Grönroos et al. (2018)) submitted to the WMT shared tasks alone over the last three years, and the state of the art is currently held by MeMAD with our submission to the WMT 2018 shared task (Grönroos et al., 2018).

### 6.1.2 Spoken language translation

Spoken language translation (SLT), also known as speech translation, undertakes the translation of spoken language audio in a source language to text in a target language. As such, it differs from conventional MT only in the source side modality, though this already introduces a multimodality challenge for MT. The SLT task has been first defined by the Consortium for Speech Translation Advanced Research (C-STAR) in 2003, and championed by the IWSLT speech translation shared tasks (*e.g.* Niehues et al. (2018)) since 2004. Thanks to this long-running campaign, nearly a hundred participants developed and submitted translation systems for this task (*e.g.* Cho et al. (2016), Nguyen et al. (2017), and Wang et al. (2018) are some of the recent successful systems). MeMAD is also represented in this task by our 2018 shared task submission (Sulubacak et al., 2018).

Traditionally, SLT is addressed by a pipeline approach (see Section 6.2), effectively separating multimodal MT into modality conversion followed by unimodal MT. More recently, end-to-end systems have started to be implemented, often based on NMT architectures, where the source language audio sequence is directly converted to the target language text sequence. Although end-to-end systems currently appear to be both less common and less successful than pipeline systems, their research is relatively fresh, and may still yield good results in the near future.

### 6.1.3 Sign language translation

As the primary languages of the deaf community, sign languages are major languages that warrant interest from the viewpoint of MT research. Sign language translation typically addresses the translation between a sign language and the corresponding standard written language. This does not mean that the task is trivial, because often there would be limited word-to-word correspondence, word sense divergences, and a significant difference in word order and grammatical structures within these language pairs. From a technical viewpoint, sign language translation is largely analogous to speech translation in how it incorporates multimodality. Like speech translation, the source and target sides each contain a single, different modality, and both pipeline and end-to-end approaches are plausible. In contrast, sign language translation has video on either the source or the target side, while speech translation has audio exclusively on the source side. These differences further limit the availability of parallel training data for sign language translation, as well as multimodal MT architectures to serve as precedent.

So far, studies on sign language translation have been relatively uncommon, and the subject seems to have fallen behind recent advances in MT. Text-to-sign translation is commonly implemented as a rule-based system that makes use of syntactic transfer to generate enriched representations for the signs, which are then typically sent to an avatar generation system to visualise the signing. Conversely, sign-to-text translation is structured as a continuous sign language recognition (CSLR) problem, where the pipeline is even further divided into video segmentation, isolated word recognition (i.e. isolated sign language recognition, or SLR), and finally target sentence synthesis. A very recently published study by Camgoz et al. (2018) describes the first end-to-end NMT approach to translating sign language directly from videos, and also introduces the RWTH-PHOENIX-Weather 2014T Continuous SLT Dataset containing segmented videos, gloss annotations, and translations into written language.

#### 6.1.4 Multimodal lexical translation

The task of multimodal lexical translation (MLT) was introduced very recently (Lala and Specia, 2018), with the principal goal of evaluating the multimodal disambiguation capabilities of translation systems. Unlike the other multimodal MT tasks, the basic translation sample for MLT is a word rather than a sentence. More specifically, the task involves the translation of ambiguous words in a source language (in such a way as to be visually disambiguable) to the correct words corresponding to them in a target language. The MLT dataset released with the introductory study demonstrates the structure of the data needed for the task: A set of 4-tuples containing an ambiguous source word, its textual context (the sentence in which the word occurs), its visual context (an image of which the sentence is a caption), and the corresponding target word. In this study, the authors also report that MLT accuracy is correlated with both automatic and human evaluation scores for MT, proposing it as an alternative targeted evaluation metric for multimodal MT.

#### 6.1.5 Video subtitle translation

The OpenSubtitles corpus (Tiedemann, 2016a; Lison and Tiedemann, 2016) from the OPUS collection (Tiedemann, 2016b) is a freely-available resource containing large amounts of parallel movie subtitles in 62 languages. Although there is a substantial amount of noise in the OpenSubtitles data, the corpus has become a common training resource to use for various flavours of MT. However, these subtitle samples do not have attached audio or video, therefore training an MT system on OpenSubtitles does not suffice in making it multimodal.

We construe the subtitle translation task as a multimodal MT task similar to caption translation, but tackling movies rather than images, and video and audio rather than still visuals attached to the text input. With these specifications, subtitle translation is likely

to be a much more advanced and challenging task than caption translation, since it calls for multiple non-text modalities in the source that are time-variant, and not directly corresponding to the subtitles for each sample. Subtitle translation is intended to be the main task for evaluating multimodal MT performance within the MeMAD project, and multimedia datasets that would enable experimentation toward this task are still being prepared by project partners Yle and INA.

## 6.2 Approaches

For some multimodal language processing tasks, the traditional way is to put together a pipeline as a divide-and-conquer method. The pipeline would divide the task into several sub-tasks, and cascade different modules to handle each of them. Typically, one of the tasks would involve modality conversion. For instance, in the case of spoken language translation, this pipeline would first convert the input speech into text by an automatic speech recognition module, and then redirect the output to a text-based MT module. This is in contrast to end-to-end models, where the source language would be encoded into an intermediate representation, and decoded directly into the target language. While pipeline systems are less vulnerable to training data insufficiency compared to data-driven end-to-end systems, they also eliminate intermodal transfer of implicit semantics, bear a risk of error propagation between stages, and may not be easily applicable to some multimodal NLP tasks. For their theoretical advantages and relevance to the state-of-the-art learning architectures, end-to-end systems are of greater interest to the MeMAD project.

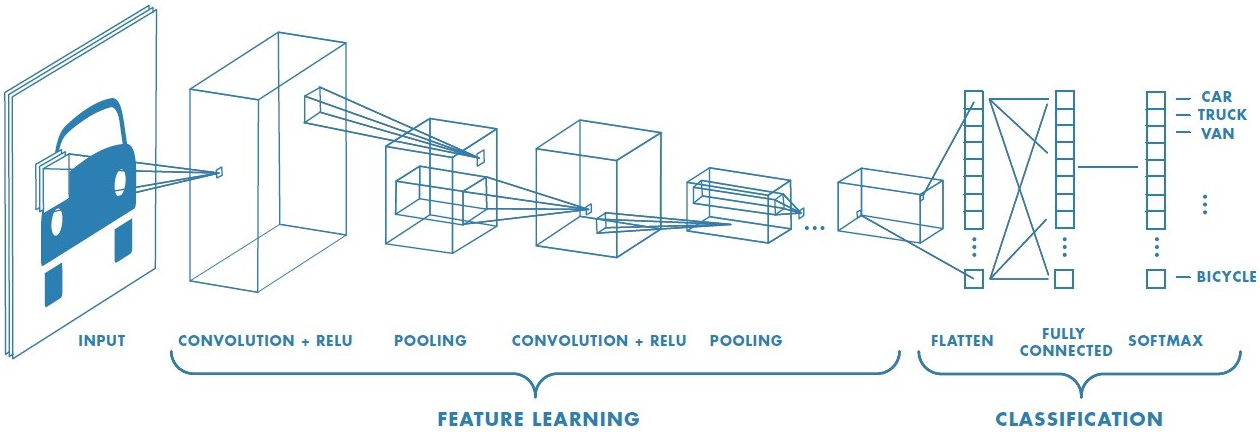


**Figure 9:** A visualisation of the convolutional neural network architecture used to encode images. (Retrieved from `https://goo.gl/9C8w6v`.)

The practice of embedding words (or more accurately, translation units) into fixed-size, dense, continuous vector representations has become a unanimous practice in NMT. For compatibility with various NMT architectures as the state of the art further develops, mul-

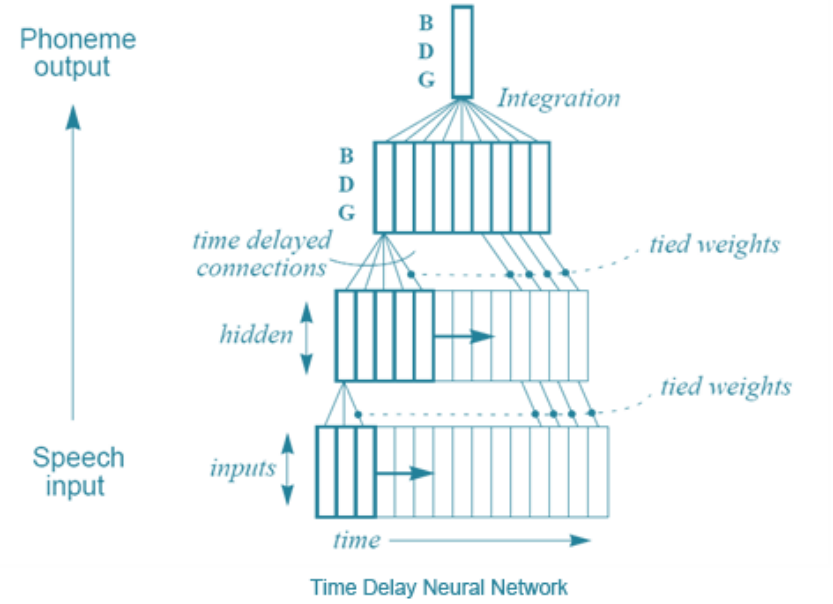


**Figure 10:** A visualisation of the time-delay neural network architecture used to encode audio. (Retrieved from `https://goo.gl/WdycBd`.)

timodal MT systems are required to embed input data from other modalities, whether alongside the text or instead of it, in a similar fashion. For visuals alone, the current best practice is to use many convolutional neural network (CNN) layers stacked on top of each other (see Figure 9), train the system for a relevant image processing task (*e.g.* object detection), and use the output of the final hidden layer from the trained network as the embedding for the image. The state of the art in encoding audio involves time-delay neural networks (TDNN), which are modular feed-forward neural networks (see Figure 10) efficient at modeling temporal context, to classify spans of audio into phones, which are then decoded into sequences of words.

It should be noted that the best practices outlined above are for non-text modalities in isolation, while the best practice of encoding the entire multimodal input is still debated. In the multimodal MT literature, one group of studies seems to prefer employing multiple encoders for different modalities and letting the attention mechanism handle the rest (*e.g.* Libovický et al. (2016)), whereas another group is investigating ways to encode multimodal input jointly into the same latent space (*e.g.* Elliott and Kádár (2017) and Zhou et al. (2018)), as shown in Figure 12. For the former case, there are further questions of how to plug the embeddings into the current MT architectures (see Figure 11 for visual examples and Appendix A (Grönroos et al., 2018) for a summary), and how to adapt attention to multivariate input (*e.g.* by concatenating the input (Huang et al., 2016), employing

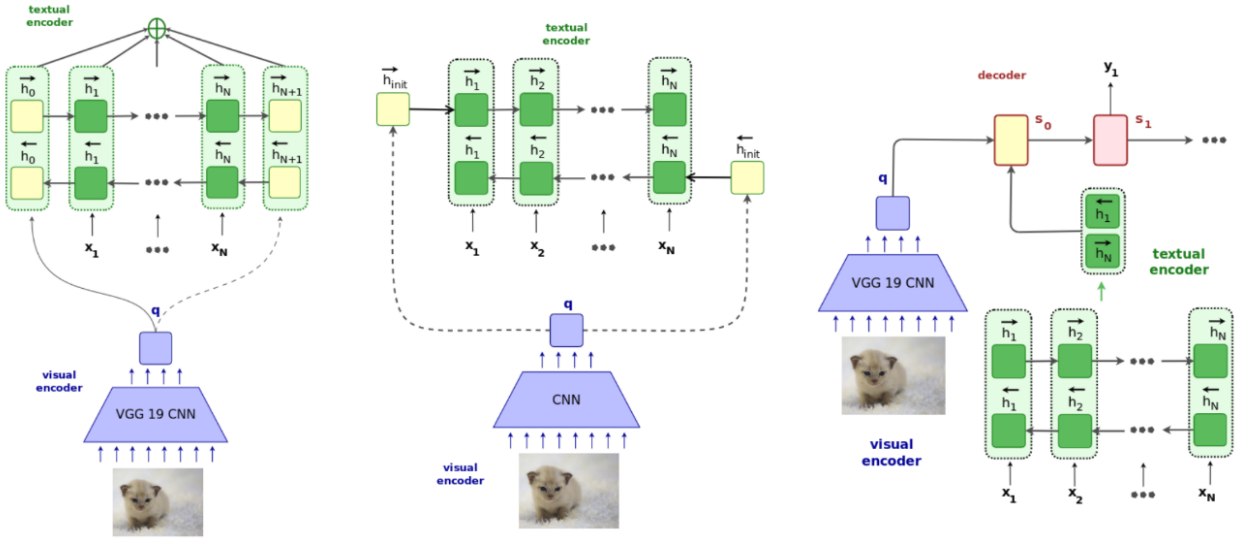


**Figure 11:** Three different ways of plugging embeddings of non-text modalities (Calixto et al., 2017) in a multimodal translation architecture.

separate attention heads (Libovický et al., 2016), or using hierarchical multi-layered attention (Moon et al., 2018)). For the latter case, the translation architecture learns to ground knowledge between the modalities, which is only meaningful when the different modalities describe different aspects of one notion (*e.g.* an image and its description, but not a video clip and subtitles). Investigation of the trade-offs between these practices is of particularly interest to MeMAD.

Furthermore, training a deep learning model to perform multiple tasks at once can improve the model's general performance at those by forcing the model to represent shared knowledge in a broader sense, as long as the objectives of the tasks are relevant for each other. While the idea has been implemented for multimodal MT in *e.g.* Elliott and Kádár's Imagination architecture (Elliott and Kádár, 2017) (see Figure 13), it is not nearly as well-studied as in text-based multilingual MT (*e.g.* Johnson et al. (2016)). Furthermore, recent studies of compound attention (Cífka and Bojar, 2018; Vázquez et al., 2018) described attempts to reinstate the aggregate sequence embedding point between the encoder and the decoder, and as a good side effect, they may have provided a mechanism for efficiently learning to attend to multiple aspects of the same input data. This could be used to encode *e.g.* both text and images into the same shared intermediate representation and enable a multi-task training schedule that simultaneously learns to caption images and (unimodally) translate captions, possibly bootstrapping an effective multimodal caption translation model.

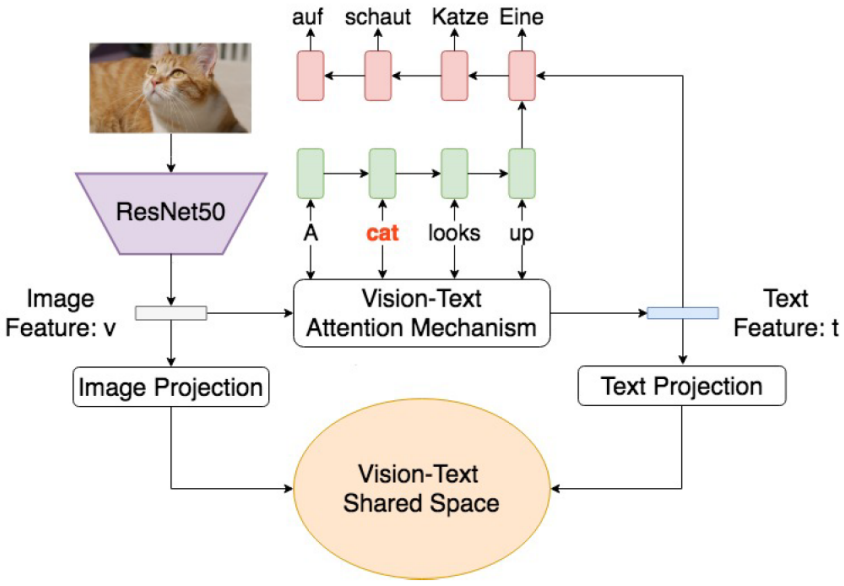


**Figure 12:** An example architecture (Zhou et al., 2018) showing how to encode multimodal input jointly into a shared space.

# 7 MeMAD status and plans

This section gives a summary of the current status and plans of multimodal translation in the MeMAD project.

## 7.1 Image caption translation

The system description paper introducing our image caption translation system is included as Appendix A (Grönroos et al., 2018). In this work, we performed a large number of preliminary experiments, to determine the type of text-only translation system to extend, and the optimal visual features to use. In the final system, we adapted the Transformer architecture (Vaswani et al., 2017) to a multimodal setting by incorporating global image features based on Mask R-CNN (He et al., 2017) object localisation outputs. For extracting the visual features, we used the Detectron software[1] using ResNeXt-152 (Xie et al., 2017) as the basic image features. The final feature is an 80-dimensional vector expressing the image surface area covered by each of the MS-COCO classes, based on the Mask R-CNN masks. These visual features are then projected into a pseudo-word embedding which is concatenated to the word embeddings of the source sentence.

We use two additional training corpora: COCO Captions (Chen et al., 2015) and Open-

[1] https://github.com/facebookresearch/Detectron

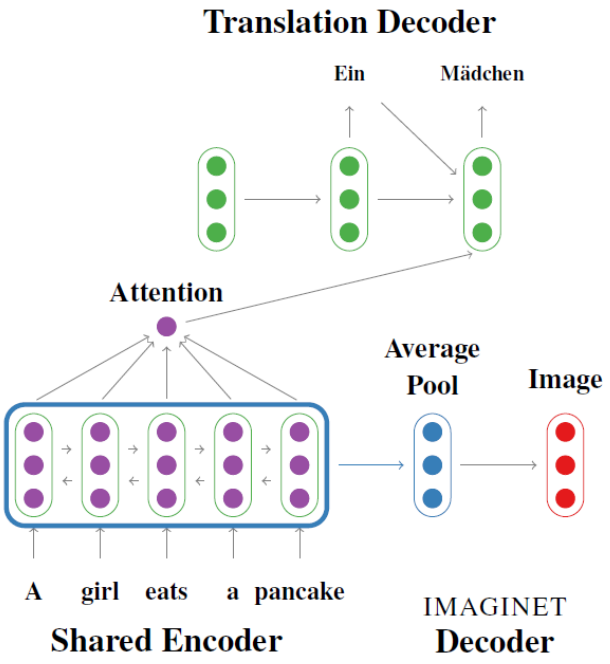


**Figure 13:** A representation (Elliott and Kádár, 2017) of the Imagination architecture, representing multimodal translation in a multi-task learning setting.

Subtitles2018 (Tiedemann, 2016a; Lison and Tiedemann, 2016). We extended COCO Captions to a synthetic multimodal and multilingual training set by forward translation from English into French and German using a text-only translation system. To enable training with the text-only OpenSubtitles data, we feed in a dummy feature consisting of the average vector of the visual features in the training data.

We apply a language model filtering technique to simultaneously remove noise from the OpenSubtitles data and adapt it towards the image caption domain.

In the WMT multimodal machine translation shared task (Elliott et al., 2017), we have the top scoring system for both English-to-German and English-to-French, according to both the automatic metrics on the Flickr18 test set, and the human evaluations.

Our ablation experiments show that the effect of the visual features in our system is small. Our largest gains come from the quality of the underlying text-only NMT system. We find that appropriate use of additional data is effective for improving fluency and overall translation quality. However, the large synthetic data does not contain examples where visual disambiguation is possible, which also biases the model away from using the visual information.

We consider the work on translation of still image captions to be completed for now, and our aim is to use this work as a step towards translation of video captions with the ultimate

goal of automatic generation of multilingual video captions.

### 7.2 Spoken language translation

We developed a pipeline SLT system for the English-to-German IWSLT speech translation evaluation campaign. Our system description is included here as Appendix B (Sulubacak et al., 2018). Our pipeline consists of a conventional ASR system, which converts English speech into text. The ASR output text is then translated into German using an NMT system.

We did not apply any SLT-specific considerations in the ASR module. The module was trained on the TED-LIUM corpus (release 2) (Rousseau et al., 2014), although we filtered out some data from the training set to comply with the restrictions of the evaluation campaign. We use the Kaldi toolkit (Povey et al., 2011) and a standard recipe included with it. The recipe trains a TDNN acoustic model using the lattice-free maximum mutual information criterion (Povey et al., 2016). For language modelling, in addition to the audio transcripts, the TED-LIUM corpus also includes a large amount of news domain text data. Using these, we trained 4-gram language models as required by our module.

The ASR system achieves a word error rate (WER) of 8.83 on the standard test set of the TED-LIUM corpus. This quantifies the amount of errors already present in the input text of the NMT system. The ASR system output is also missing case and punctuation information.

The NMT system incorporates some well-established good practices, such as normalisation, byte-pair encoding, and the Transformer architecture. The competition provides a set of TED talks transcribed into English, and translated by humans into German. These English transcripts and German translations, as well as English-German data from the OpenSubtitles corpus, were used for training the NMT module.

We experimented with some special measures in the NMT training to help the model cope with ASR-produced text as input. First, we extracted lists of the 50-best ASR decoding hypotheses for the TED talks. This gave us actual ASR output, for which we also had the corresponding German translations. Using 50-best lists rather than a single top-scoring hypothesis, we hoped to capture some additional errors characteristic to ASR. However, as the training data for the NMT module, we eventually used only the top 10 hypotheses.

We also briefly experimented with using a separate case and punctuation restoration phase after translation. This was contrasted with a system that directly output full-cased and punctuated text. The results from this experiment were inconclusive.

The OpenSubtitles data does not contain corresponding audio, so we could not extract actual ASR output for it. Instead, we trained a separate NMT system to convert the subtitle data into an ASR-like form using the 50-best lists and the corresponding ground truth English transcripts. Visually inspecting the results, the system seemed to successfully remove case and punctuation, and make some modifications to the text resembling ASR errors. In our development experiments, training on this data provided a small performance improvement, but the improvement was not reflected on the competition test set, and this

method slightly degraded our performance instead.

On the test set, our best model used the ASR-decoded training data, and additional data from the OpenSubtitles corpus. On the development set, the model had a BLEU score of 20.58, but on the test set, the BLEU score was only 16.45.

We also started developing an end-to-end system for the same task, as an extension of the OpenNMT-py neural translation system. The work was not finished in time for the shared task deadline. The architecture consisted of multiple encoders and decoders, trained in a multi-task learning setting. We are exploring continuations for this work, implemented on either OpenNMT-py or the ESPnet ASR system (Kim et al., 2017).

The USTC entry in the IWSLT competition included an end-to-end speech translation system, which had a test set BLEU score of 19.4 (Liu et al., 2018). This was about nine points lower than the best pipeline systems in the competition, which means there remains much work to be done in end-to-end SLT research. However, the system was still somewhat competitive (*e.g.* outperforming our pipeline system), which we take as an encouraging sign for the end-to-end approach.

### 7.3 Video subtitle translation

Video subtitle translation will be our main focus in the near future. Our project-internal data resources place us in a favourable position to address this task. We plan to explore discourse-aware systems for subtitle translation. Subtitles can be augmented with metadata that could be helpful in translation. Speaker information and dialog structures make it possible to adapt translation models, and the dynamic structure of the narrative could be explored to improve coherence of the translations. We also plan to involve audio and video signals to improve the integration of contextual features in this process. For example, stress patterns and talking speed, pause duration, and other features could help in generating textual representations that better match the scenes. Visual information may help to disambiguate certain expressions, even though this turns out to be difficult from our experience on image caption translation. The narrative structure might help to accumulate useful information from the video. For this, we will collaborate closely with WP2 on video content description.

Another focus in subtitle translation is the combination of multimodal input with multilingual models. Ultimately, we aim for the generation of translations in multiple languages from video without further intervention. But even for training alone, it is useful to learn from translations into several languages to pick up more information about the semantics of the input. The multi-task learning setup that we envision for the end-to-end speech translation model (see Section 7.2) will be employed for the integration of multiple languages enabling effective transfer learning. We recently developed a multilingual MT model that combines language-specific encoders and decoders via a shared intermediate layer (which we call the *attention bridge* (Vázquez et al., 2018). That system will be extended with

multimodal support to serve our needs in MeMAD.

### 7.4 Audio description translation

We intend to explore data availability and feasibility of translating audio descriptions, as a case study of domain adaptation. Translated audio description basically does not exist in large quantities for training our models. Instead, we plan to adapt subtitle translation models (see Section 7.3) to the scenario of audio descriptions by fine-tuning models with limited data resources that we will have access to. Monolingual audio descriptions will be useful via back-translation—a standard technique that improves domain adaptation in NMT. We will also explore movie scripts and other manuscripts that can be linked to video material. As datasets are scarce, audio description translation will remain a limited case study and will need further attention in future work. Audio description translation will also heavily depend on the success of WP2 and the automatic generation of such descriptions. An end-to-end audio description generation model with translation capabilities will be discussed with the developers of the video content description system. Overall, we will focus on the development of end-to-models in close collaboration with work package 2 and our partners in MeMAD.

## 8 References

Yasuhiro Akiba, Marcello Federico, Noriko Kando, Hiromi Nakaiwa, Michael Paul, and Jun'ichi Tsujii. 2004. Overview of the IWSLT 2004 evaluation campaign. In *Proceedings of the 2004 International Workshop on Spoken Language Translation*, Kyoto, Japan.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised Neural Machine Translation.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proceedings of the 3rd International Conference on Learning Representations*, pages San Diego, CA, USA. ArXiv: 1409.0473.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2017. Multimodal Machine Learning: A Survey and Taxonomy. *arXiv:1705.09406 [cs]*. ArXiv: 1705.09406.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the Third Shared Task on Multimodal Machine Translation.

Luisa Bentivogli, Mauro Cettolo, Marcello Federico, and Christian Federmann. 2018. Machine Translation Human Evaluation: An investigation of evaluation based on Post-Editing and its relation with Direct Assessment. In *Proceedings of the 2018 International Workshop on Spoken Language Translation*, pages 62–69, Bruges, Belgium.

Adam L. Berger, Peter F. Brown, Stephen A. della Pietra, Vincent J. della Pietra, John R. Gillett, John D. Lafferty, Robert L. Mercer, Harry Printz, and Luboš Ureš. 1994. The Candide system for machine translation. In *Proceedings of the workshop on Human Language Technology - HLT '94*, Plainsboro, NJ. Association for Computational Linguistics.

Alan W Black, Ralf Brown, Robert Frederking, Rita Singh, John Moody, and Eric Steinbrecher. 2002. TONGUES: Rapid development of a speech-to-speech translation system. In *Proceedings of the second international conference on Human Language Technology Research -*, pages 183–186, San Diego, California. Association for Computational Linguistics.

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

Ozan Caglayan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, and Joost van de Weijer. 2016. Does Multimodality Help Human and Machine for Translation and Image Captioning? *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 627–633. ArXiv: 1605.09186.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Incorporating Global Visual Features into Attention-Based Neural Machine Translation. *arXiv:1701.06521 [cs]*. ArXiv: 1701.06521.

Necati Cihan Camgoz, Simon Hadfield, Oscar Koller, Hermann Ney, and Richard Bowden. 2018. Neural Sign Language Translation.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 2017 International Workshop on Spoken Language Translation*, pages 2–14, Tokyo, Japan.

Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, Roldano Cattoni, and Marcello Federico. 2016. The IWSLT 2016 evaluation campaign. In *Proceedings of the 2016 International Workshop on Spoken Language Translation*.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv:1504.00325 [cs]*. ArXiv: 1504.00325.

Eunah Cho, Jan Niehues, Thanh-Le Ha, Matthias Sperber, Mohammed Mediani, and Alex Waibel. 2016. Adaptation and Combination of NMT Systems: The KIT Translation Systems for IWSLT 2016.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the Properties of Neural Machine Translation: Encoder-Decoder Approaches. *arXiv:1409.1259 [cs, stat]*. ArXiv: 1409.1259.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. *arXiv:1412.3555 [cs]*. ArXiv: 1412.3555.

Ondřej Cífka and Ondřej Bojar. 2018. Are BLEU and Meaning Representation in Opposition? *arXiv:1805.06536 [cs]*. ArXiv: 1805.06536.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the Second Shared Task on Multimodal Machine Translation and Multilingual Image Description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual English-German Image Descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves Multimodal Translation. *arXiv:1705.04350 [cs]*. ArXiv: 1705.04350.

Jeffrey L. Elman. 1990. Finding Structure in Time. *Cognitive Science*, 14(2):179–211.

Yuqing Gao, Wei Zhang, Laurent Besacier, Liang Gu, Bowen Zhou, Ruhi Sarikaya, Mohamed Afify, Hong-Kwang Kuo, Wei-zhong Zhu, Yonggang Deng, and Charles Prosser. 2006. IBM MASTOR system: multilingual automatic speech-to-speech translator. In *Proceedings of the Workshop on Medical Speech Translation - MST '06*, pages 53–56, New York, New York. Association for Computational Linguistics.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. Convolutional Sequence to Sequence Learning. *arXiv:1705.03122 [cs]*. ArXiv: 1705.03122.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, Vancouver, BC, Canada. IEEE.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 4, pages 2047–2052, Montreal, Que., Canada. IEEE.

Stig-Arne Grönroos, Sami Virpioja, and Mikko Kurimo. 2018. Cognate-aware morphological segmentation for multilingual neural translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 386–393.

Michael Grubinger, Paul Clough, Henning Müller, and Thomas Deselaers. 2006. The IAPR TC-12 Benchmark: A New Evaluation Resource for Visual Information Systems. In *Proceedings of the OntoImage Workshop on Language Resources for Content-based Image Retrieval*, pages 13–23, Genoa, Italy.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphaël Troncy, and Raúl Vázquez. 2018. The MeMAD Submission to the WMT18 Multimodal Translation Task. *arXiv:1808.10802 [cs]*. ArXiv: 1808.10802.

Yoichiro Hasebe. 2015. Design and Implementation of an Online Corpus of Presentation Transcripts of TED Talks. *Procedia - Social and Behavioral Sciences*, 198:174–182.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988. IEEE.

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. *arXiv:1805.04699 [cs]*. ArXiv: 1805.04699.

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

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Po-Yao Huang, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer. 2016. Attention-based Multimodal Neural Machine Translation. In *Proceedings of the 1st Conference on Machine Translation*, volume 2, pages 639–645, Berlin, Germany. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *arXiv:1611.04558 [cs]*. ArXiv: 1611.04558.

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 Neural Machine Translation in C++. *arXiv:1804.00344 [cs]*. ArXiv: 1804.00344.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 4835–4839. IEEE.

Guillaume Klein, Yoon Kim, Yuntian Deng, Vincent Nguyen, Jean Senellart, and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. *arXiv:1805.11462 [cs]*. ArXiv: 1805.11462.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-Source Toolkit for Neural Machine Translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Philipp Koehn, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, Evan Herbst, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, and Christine Moran. 2007. Moses: open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions - ACL '07*, Prague, Czech Republic. Association for Computational Linguistics.

Chiraag Lala and Lucia Specia. 2018. Multimodal Lexical Translation. In *Proceedings of the 11th Conference on Language Resources and Evaluation*, Miyazaki, Japan.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. *arXiv:1804.07755 [cs]*. ArXiv: 1804.07755.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation - StatMT '07*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.

Alon Lavie, Chad Langley, Alex Waibel, Fabio Pianesi, Gianni Lazzari, Paolo Coletti, Loredana Taddei, and Franco Balducci. 2001. Architecture and Design Considerations

in NESPOLE!: a Speech Translation System for E-commerce Applications. In *Proceedings of the 1st International Conference on Human Language Technology Research*, pages 1–4.

Alon Lavie, Alex Waibel, Lori Levin, Michael Finke, Donna Gates, Marsal Gavalda, Torsten Zeppenfeld, and Puming Zhan. 1997. JANUS-III: Speech-to-speech translation in multiple languages. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 99–102, Munich, Germany. IEEE Comput. Soc. Press.

Xirong Li, Weiyu Lan, Jianfeng Dong, and Hailong Liu. 2016. Adding Chinese Captions to Images. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval - ICMR '16*, pages 271–275, New York, New York, USA. ACM Press.

Jindřich Libovický, Jindřich Helcl, Marek Tlustý, Pavel Pecina, and Ondřej Bojar. 2016. CUNI System for WMT16 Automatic Post-Editing and Multimodal Translation Tasks. *arXiv:1606.07481 [cs]*. ArXiv: 1606.07481.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Proceedings of the 13th European Conference on Computer Vision*, volume 8693, pages 740–755, Zurich, Switzerland. Springer International Publishing.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles.

Dan Liu, Junhua Liu, Wu Guo, Shifu Xiong, Zhiqiang Ma, Rui Song, Chongliang Wu, and Quan Liu. 2018. The USTC-NEL Speech Translation system at IWSLT 2018. *arXiv:1812.02455 [cs, eess]*. ArXiv: 1812.02455.

Minh-Thang Luong, Quoc V Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2015. Multi-task sequence to sequence learning. *arXiv preprint arXiv:1511.06114*.

Qingsong Ma, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 Metrics Shared Task: Both characters and embeddings achieve good performance.

Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal Named Entity Recognition for Short Social Media Posts. In *Proceedings of the 16th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 852–860, New Orleans, Louisiana. Association for Computational Linguistics.

Tsuyoshi Morimoto. 1990. Automatic interpreting telephony research at ATR. In *Proceedings of a Workshop on Machine Translation*, UMIST.

Thai-Son Nguyen, Markus Muller, Matthias Sperber, Thomas Zenkel, Sebastian Stuker, and Alex Waibel. 2017. The 2017 KIT IWSLT Speech-to-Text Systems for English and German.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Mauro Cettolo, Marco Turchi, and Marcello Federico. 2018. The IWSLT 2018 Evaluation Campaign. In *Proceedings of the 2018 International Workshop on Spoken Language Translation*, Bruges, Belgium.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Moisés Pastor, Alberto Sanchis, Francisco Casacuberta, and Enrque Vidal. 2001. EUTRANS: a Speech-to-Speech Translator Prototype.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the IWSLT 2010 evaluation campaign. In *Proceedings of the 2010 International Workshop on Spoken Language Translation*.

Emmanouil Antonios Platanios, Mrinmaya Sachan, Graham Neubig, and Tom Mitchell. 2018. Contextual parameter generation for universal neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 425–435.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukaš Burget, Ondřej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlíček, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The Kaldi Speech Recognition Toolkit.

Daniel Povey, Vijayaditya Peddinti, Daniel Galvez, Pegah Ghahremani, Vimal Manohar, Xingyu Na, Yiming Wang, and Sanjeev Khudanpur. 2016. Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI. pages 2751–2755.

Cyrus Rashtchian, Peter Young, Micah Hodosh, and Julia Hockenmaier. 2010. Collecting image annotations using Amazon's Mechanical Turk. In *Proceedings of the Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 139–147. Association for Computational Linguistics.

Sujith Ravi and Kevin Knight. 2011. Deciphering Foreign Language. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*.

Anthony Rousseau, Paul Deléglise, and Yannick Estève. 2014. Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks. In *Proceedings of 9th International Conference on Language Resources and Evaluation*, pages 3935–3939, Reykjavík, Iceland.

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. *arXiv:1811.00347 [cs]*. ArXiv: 1811.00347.

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

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation.

Lucia Specia, Frédéric Blain, Varvara Logacheva, Ramón F. Astudillo, and André Martins. 2018. Findings of the WMT 2018 Shared Task on Quality Estimation. page 21.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A Shared Task on Multimodal Machine Translation and Crosslingual Image Description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Umut Sulubacak, Aku Rouhe, Jörg Tiedemann, Stig-Arne Grönroos, and Mikko Kurimo. 2018. The MeMAD Submission to the IWSLT 2018 Speech Translation Task.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks.

Toshiyuki Takezawa, Tsuyoshi Morimoto, Yoshinori Sagisaka, Nick Campbell, Hitoshi Iida, Fumiaki Sugaya, Akio Yokoo, and Seiichi Yamamoto. 1998. A Japanese-to-English Speech Translation System: ATR-MATRIX.

Jörg Tiedemann. 2016a. Finding Alternative Translations in a Large Corpus of Movie Subtitles.

Jörg Tiedemann. 2016b. Parallel Data, Tools and Interfaces in OPUS.

Ashish Vaswani, Samy Bengio, Eugene Brevdo, Francois Chollet, Aidan N. Gomez, Stephan Gouws, Llion Jones, Łukasz Kaiser, Nal Kalchbrenner, Niki Parmar, Ryan Sepassi, Noam Shazeer, and Jakob Uszkoreit. 2018. Tensor2tensor for Neural Machine Translation. *arXiv:1803.07416 [cs, stat]*. ArXiv: 1803.07416.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need.

Raúl Vázquez, Alessandro Raganato, Jörg Tiedemann, and Mathias Creutz. 2018. Multilingual NMT with a language-independent attention bridge. *arXiv:1811.00498 [cs]*. ArXiv: 1811.00498.

Wolfgang Wahlster. 2000. Mobile Speech-to-Speech Translation of Spontaneous Dialogs: An Overview of the Final Verbmobil System. In Wolfgang Wahlster, editor, *Verbmobil: Foundations of Speech-to-Speech Translation*, pages 3–21. Springer Berlin Heidelberg, Berlin, Heidelberg.

Yuguang Wang, Liangliang Shi, Linyu Wei, Weifeng Zhu, Jinkun Chen, Zhichao Wang, Shixue Wen, Wei Chen, Yanfeng Wang, and Jia Jia. 2018. The Sogou-TIIC Speech Translation System for IWSLT 2018. In *Proceedings of the 2018 International Workshop on Spoken Language Translation*, pages 112–117.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5987–5995. IEEE.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. 2017. STAIR Captions: Constructing a Large-Scale Japanese Image Caption Dataset. *arXiv:1705.00823 [cs]*. ArXiv: 1705.00823.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions.

Ying Zhang. 2003. Survey of Current Speech Translation Research. Technical report.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018. A Visual Attention Grounding Neural Model for Multimodal Machine Translation. *arXiv:1808.08266 [cs]*. ArXiv: 1808.08266.

Mesut Erhan Ünal, Begüm Çıtamak, Semih Yağcıoğlu, Aykut Erdem, Erkut Erdem, Nazlı İkizler Cinbiş, and Ruket Çakıcı. 2016. TasvirEt: Görüntülerden Otomatik Türkçe Açıklama Oluşturma İçin Bir Denektaşı Veri Kümesi (TasvirEt: A Benchmark Dataset for Automatic Turkish Description Generation from Images).

# A Appendix: WMT MeMAD system description paper

# The MeMAD Submission to the WMT18 Multimodal Translation Task

**Stig-Arne Grönroos**
Aalto University

**Benoit Huet**
EURECOM

**Mikko Kurimo**
Aalto University

**Jorma Laaksonen**
Aalto University

**Bernard Merialdo**
EURECOM

**Phu Pham**
Aalto University

**Mats Sjöberg**
Aalto University

**Umut Sulubacak**
University of Helsinki

**Jörg Tiedemann**
University of Helsinki

**Raphael Troncy**
EURECOM

**Raúl Vázquez**
University of Helsinki

## Abstract

This paper describes the MeMAD project entry to the WMT Multimodal Machine Translation Shared Task.

We propose adapting the Transformer neural machine translation (NMT) architecture to a multi-modal setting. In this paper, we also describe the preliminary experiments with text-only translation systems leading us up to this choice.

We have the top scoring system for both English-to-German and English-to-French, according to the automatic metrics for *flickr18*.

Our experiments show that the effect of the visual features in our system is small. Our largest gains come from the quality of the underlying text-only NMT system. We find that appropriate use of additional data is effective.

## 1 Introduction

In multi-modal translation, the task is to translate from a source sentence and the image that it describes, into a target sentence in another language. As both automatic image captioning systems and crowd captioning efforts tend to mainly yield descriptions in English, multi-modal translation can be useful for generating descriptions of images for languages other than English. In the MeMAD project[1], multi-modal translation is of interest for creating textual versions or descriptions of audio-visual content. Conversion to text enables both indexing for multi-lingual image and video search, and increased access to the audio-visual materials for visually impaired users.

| Data set | images | en | de | fr | sentences |
|---|---|---|---|---|---|
| Multi30k | ✓ | ✓ | ✓ | ✓ | 29k |
| MS-COCO | ✓ | ✓ | + | + | 616k |
| OpenSubtitles | | ✓ | ✓ | ✓ | 23M/42M |
| | | 1M, 3M, and 6M subsets used. | | | |

Table 1: Summary of data set sizes. ✓means attribute is present in original data. + means data set augmented in this work.

We adapt[2] the Transformer (Vaswani et al., 2017) architecture to use global image features extracted from Detectron, a pre-trained object detection and localization neural network. We use two additional training corpora: MS-COCO (Lin et al., 2014) and OpenSubtitles2018 (Tiedemann, 2009). MS-COCO is multi-modal, but not multi-lingual. We extended it to a synthetic multi-modal and multilingual training set. OpenSubtitles is multilingual, but does not include associated images, and was used as text-only training data. This places our entry in the unconstrained category of the WMT shared task. Details on the architecture used in this work can be found in Section 4.1. Further details on the synthetic data are presented in Section 2. Data sets are summarized in Table 1.

## 2 Experiment 1: Optimizing Text-Based Machine Translation

Our first aim was to select the text-based MT system to base our multi-modal extensions on.

[1] https://www.memad.eu/

[2] Our fork available from https://github.com/Waino/OpenNMT-py/tree/develop_mmod

| EN-FR | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|
| multi30k | 61.4 | 54.0 | 43.1 |
| +SUBS$_{\text{full}}$ | 53.7 | 48.9 | 47.0 |
| +domain-tuned | 66.1 | 59.7 | **51.7** |
| +ensemble-of-3 | **66.5** | **60.2** | 51.6 |

| EN-DE | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|
| multi30k | 38.9 | 32.0 | 27.7 |
| +SUBS$_{\text{full}}$ | 41.3 | 34.1 | 31.3 |
| +domain-tuned | 43.3 | 38.4 | 35.0 |
| +ensemble-of-3 | **43.9** | **39.6** | **37.0** |

Table 2: Adding subtitle data and domain tuning for image caption translation (BLEU% scores). All results with Marian Amun.

| | EN-FR | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|---|
| A | SUBS1M$_H$+MS-COCO | 66.3 | 60.5 | 52.1 |
| A | +domain-tuned | 66.8 | 60.6 | 52.0 |
| A | +labels | **67.2** | 60.4 | 51.7 |
| T | SUBS1M$_{LM}$+MS-COCO | 66.9 | 60.3 | **52.8** |
| T | +labels | **67.2** | **60.9** | 52.7 |

| | EN-DE | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|---|
| A | SUBS1M$_H$+MS-COCO | 43.1 | 39.0 | 35.1 |
| A | +domain-tuned | 43.9 | 39.4 | 35.8 |
| A | +labels | 43.2 | 39.3 | 34.3 |
| T | SUBS1M$_{LM}$+MS-COCO | **44.4** | 39.4 | 35.0 |
| T | +labels | 44.1 | **39.8** | **36.5** |

Table 3: Using automatically translated image captions and domain labels (BLEU% scores). A is short for Amun, T for Transformer.

We tried a wide range of models, but only include results with the two strongest systems: Marian NMT with the *amun* model (Junczys-Dowmunt et al., 2018), and OpenNMT (Klein et al., 2017) with the *Transformer* model.

We also studied the effect of additional training data. Our initial experiments showed that movie subtitles and their translations work rather well to augment the given training data. Therefore, we included parallel subtitles from the OpenSubtitles2018 corpus to train better text-only MT models. For these experiments, we apply the Marian amun model, an attentional encoder-decoder model with bidirectional LSTM's on the encoder side. In our first series of experiments, we observed that domain-tuning is very important when using Marian. The domain-tuning was accomplished by a second training step on in-domain data after training the model on the entire data set. Table 2 shows the scores on development data. We also tried decoding with an ensemble of three independent runs, which also pushed the performance a bit.

Furthermore, we tried to artificially increase the amount of in-domain data by translating existing English image captions to German and French. For this purpose, we used the large MS-COCO data set with its 100,000 images that have five image captions each. We used our best multidomain model (see Table 2) to translate all of those captions and used them as additional training data. This procedure also transfers the knowledge learned by the multidomain model into the caption translations, which helps us to improve the coverage of the system with less out-of-domain data. Hence, we filtered the large collection of translated movie subtitles to a smaller portion of reliable sentence pairs (one million in the experiment we report) and could train on a smaller data set with better results.

We experimented with two filtering methods. Initially, we implemented a basic heuristic filter (SUBS$_H$), and later we improved on this with a language model filter (SUBS$_{LM}$). Both procedures consider each sentence pair, assign it a quality score, and then select the highest scoring 1, 3, or 6 million pairs, discarding the rest. The SUBS$_H$ method counts terminal punctuation (*'.', '...', '?', '!'*) in the source and target sentences, initializing the score as the negative of the absolute value of the difference between these counts. Afterwards, it further decrements the score by 1 for each occurrence of terminal punctuation beyond the first in each of the sentences. The SUBS$_{LM}$ method first preprocesses the data by filtering samples by length and ratio of lengths, applying a rule-based noise filter, removing all characters not present in the Multi30k set, and deduplicating samples. Afterwards, target sentences in the remaining pairs are scored using a character-based deep LSTM language model trained on the Multi30k data. Both selection procedures are intended for noise filtering, and SUBS$_{LM}$ additionally acts as domain adaptation. Table 3 lists the scores we obtained on development data.

To make a distinction between automatically translated captions, subtitle translations and human-translated image captions, we also

introduced domain labels that we added as special tokens to the beginning of the input sequence. In this way, the model can use explicit information about the domain when deciding how to translate given input. However, the effect of such labels is not consistent between systems. For Marian amun, the effect is negligible as we can see in Table 3. For the Transformer, domain labels had little effect on BLEU but were clearly beneficial according to chrF-1.0.

### 2.1 Preprocessing of textual data

The final preprocessing pipeline for the textual data consisted of lowercasing, tokenizing using Moses, fixing double-encoded entities and other encoding problems, and normalizing punctuation. For the OpenSubtitles data we additionally used the SUBS$_{LM}$ subset selection.

Subword decoding has become popular in NMT. Careful choice of translation units is especially important as one of the target languages of our system is German, a morphologically rich language. We trained a shared 50k subword vocabulary using Byte Pair Encoding (BPE) (Sennrich et al., 2015). To produce a balanced multi-lingual segmentation, the following procedure was used: First, word counts were calculated individually for English and each of the 3 target languages Czech[3], French and German. The counts were normalized to equalize the sum of the counts for each language. This avoided imbalance in the amount of data skewing the segmentation in favor of some language. Segmentation boundaries around hyphens were forced, overriding the BPE.

Multi-lingual translation with target-language tag was done following Johnson et al. (2016). A special token, e.g. <TO_DE> to mark German as the target language, was prefixed to each paired English source sentence.

## 3 Experiment 2: Adding Automatic Image Captions

Our first attempt to add multi-modal information to the translation model includes the incorporation of automatically created image captions in a purely text-based translation engine. For this, we generated five English captions for each of the images in the provided training and test data. This was done by using our in-house captioning system (Shetty et al., 2018). The image captioning system uses a 2-layer LSTM with residual connections to generate captions based on scene context and object location descriptors, in addition to standard CNN-based features. The model was trained with the MS-COCO training data and used to be state of the art in the COCO leaderboard[4] in Spring 2016. The beam search size was set to five.

| EN-FR | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|
| multi30k | 61.4 | 54.0 | 43.1 |
| +autocap (dual attn.) | 60.9 | 52.9 | 43.3 |
| +autocap 1 (concat) | 61.7 | 53.7 | 43.9 |
| +autocap 1-5 (concat) | **62.2** | **54.4** | **44.1** |
| EN-DE | flickr16 | flickr17 | mscoco17 |
| multi30k | 38.9 | 32.0 | 27.7 |
| +autocap (dual attn.) | 37.8 | 30.2 | 27.0 |
| +autocap 1 (concat) | 39.7 | **32.2** | **28.8** |
| +autocap 1-5 (concat) | **39.9** | 32.0 | 28.7 |

Table 4: Adding automatic image captions (only the best one or all 5). The table shows BLEU scores in %. All results with Marian Amun.

We tried two models for the integration of those captions: (1) a dual attention multi-source model that adds another input sequence with its own decoder attention and (2) a concatenation model that adds auto captions at the end of the original input string separated by a special token. In the second model, attention takes care of learning how to use the additional information and previous work has shown that this, indeed, is possible (Niehues et al., 2016; Östling et al., 2017). For both models, we applied Marian NMT that already includes a working implementation of dual attention translations. Table 4 summarizes the scores on the three development test sets for English-French and English-German.

We can see that the dual attention model does not work at all and the scores slightly drop. The concatenation approach works better probably because the common attention

[3]Czech was later dropped as a target language due to time constraints.

[4]https://competitions.codalab.org/competitions/3221

model learns interactions between the different types of input. However, the improvements are small if any and the model basically learns to ignore the auto captions, which are often very different from the original input. The attention pattern in the example of Figure 1 shows one of the very rare cases where we observe at least some attention to the automatic captions.

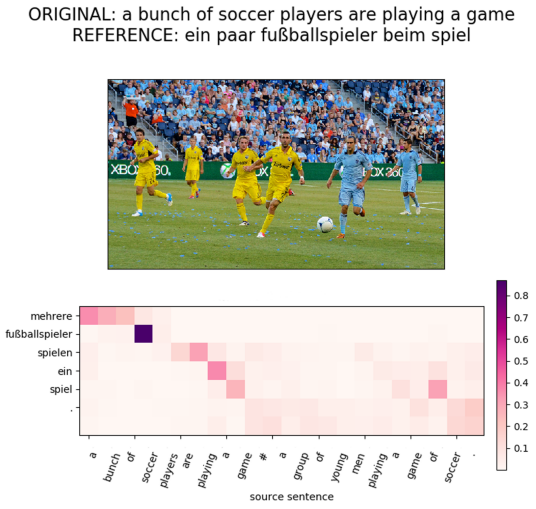


Figure 1: Attention layer visualization for an example where at least one of the attention weights for the last part of the sentence, which corresponds to the automatically generated captions, obtains a value above 0.3

## 4 Experiment 3: Multi-modal Transformer

One benefit of NMT, in addition to its strong performance, is its flexibility in enabling different information sources to be merged. Different strategies to include image features both on the encoder and decoder side have been explored. We are inspired by the recent success of the Transformer architecture to adapt some of these strategies for use with the Transformer.

Recurrent neural networks start their processing from some **initial hidden state**. Normally, a zero vector or a learned parameter vector is used, but the initial hidden state is also a natural location to introduce additional context e.g. from other modalities. Initializing can be applied in either the encoder ($\text{IMG}_E$) or decoder ($\text{IMG}_D$) (Calixto et al., 2017). These approaches are not directly applicable to the Transformer, as it is not a recurrent model, and lacks a comparable initial hidden state.

**Double attention** is another popular choice, used by e.g. Caglayan et al. (2017). In this approach, two attention mechanisms are used, one for each modality. The attentions can be separate or hierarchical. While it would be possible to use double attention with the Transformer, we did not explore it in this work. The multiple multi-head attention mechanisms in the Transformer leave open many challenges in how this integration would be done.

**Multi-task learning** has also been used, e.g. in the Imagination model (Elliott and Kádár, 2017), where the auxiliary task consists of reconstructing the visual features from the source encoding. Imagination could also have been used with the Transformer, but we did not explore it in this work.

The **source sequence** itself is also a possible location for including the visual information. In the $\text{IMG}_W$ approach, the visual features are encoded as a pseudo-word embedding concatenated to the word embeddings of the source sentence. When the encoder is a bidirectional recurrent network, as in Calixto et al. (2017), it is beneficial to add the pseudo-word both at the beginning and the end to make it available for both encoder directions. This is unnecessary in the Transformer, as it has equal access to all parts of the source in the deeper layers of the encoder. Therefore, we add the pseudo-word only to the beginning of the sequence. We use an affine projection of the image features $V \in \mathbb{R}^{80}$ into a pseudo-word embedding $x_I \in \mathbb{R}^{512}$

$$x_I = W_{src} \cdot V + b_I.$$

In the LIUM *trg-mul* (Caglayan et al., 2017), the **target embeddings** and visual features are interacted through elementwise multiplication.

$$y'_j = y_j \odot \tanh(W^{dec}_{mul} \cdot V)$$

Our initial gating approach resembles *trg-mul*.

### 4.1 Architecture

The baseline NMT for this experiment is the OpenNMT implementation of the Transformer. It is an encoder-decoder NMT system

using the Transformer architecture (Vaswani et al., 2017) for both the encoder and decoder side. The Transformer is a deep, non-recurrent network for processing variable-length sequences. A Transformer is a stack of layers, consisting of two types of sub-layer: multi-head (MH) attention (Att) sub-layers and feed-forward (FF) sub-layers:

$$\begin{aligned} \text{Att}(Q,K,V) &= \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \\ a_i &= \text{Att}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{MH}(Q,K,V) &= [a_1;\ldots;a_h]W^O \\ \text{FF}(x) &= \max(0, xW_1 + b_1)W_2 + b_2 \end{aligned} \tag{1}$$

where $Q$ is the input query, $K$ is the key, and $V$ the attended values. Each sub-layer is individually wrapped in a residual connection and layer normalization.

When used in translation, Transformer layers are stacked into an encoder-decoder structure. In the encoder, the layer consists of a self-attention sub-layer followed by a FF sub-layer. In self-attention, the output of the previous layer is used as queries, keys and values $Q = K = V$. In the decoder, a third context attention sub-layer is inserted between the self-attention and the FF. In context attention, $Q$ is again the output of the previous layer, but $K = V$ is the output of the encoder stack. The decoder self-attention is also masked to prevent access to future information. Sinusoidal position encoding makes word order information available.

**Decoder gate**. Our first approach is inspired by *trg-mul*. A gating layer is introduced to modify the pre-softmax prediction distribution. This allows visual features to directly suppress a part of the output vocabulary. The probability of correctly translating a source word with visually resolvable ambiguity can be increased by suppressing the unwanted choices.

At each timestep the decoder output $s_j$ is projected to an unnormalized distribution over the target vocabulary.

$$y_j = W \cdot s_j + b$$

Before normalizing the distribution using a softmax layer, a gating layer can be added.

$$\begin{aligned} g &= \sigma(W_{gate}^{dec} \cdot V + b_{gate}^{dec}) \\ y'_j &= y_j \odot g \end{aligned} \tag{2}$$

Preliminary experiments showed that gating based on only the visual features did not work. Suppressing the same subword units during the entire decoding of the sentence was too disruptive. We addressed this by using the decoder hidden state as additional input to control the gate. This causes the vocabulary suppression to be time dependent.

$$g_j = \sigma(U_{gate}^{dec} \cdot s_j + W_{gate}^{dec} \cdot V + b_{gate}^{dec}) \tag{3}$$

**Encoder gate**. The same gating procedure can also be applied to the output of the encoder. When using the encoder gate, the encoded source sentence is disambiguated, instead of suppressing part of the output vocabulary.

$$\begin{aligned} g_i &= \sigma(U_{gate}^{enc} \cdot h_i + W_{gate}^{enc} \cdot V + b_{gate}^{enc}) \\ h'_i &= h_i \odot g_i \end{aligned} \tag{4}$$

The gate biases $b_{gate}^{dec}$ and $b_{gate}^{enc}$ should be initialized to positive values, to start training with the gates opened. We also tried combining both forms of gating.

| EN-FR | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|
| $\text{IMG}_W$ | *68.30* | **62.45** | 52.86 |
| enc-gate | 68.01 | 61.38 | **53.40** |
| dec-gate | 67.99 | 61.53 | 52.38 |
| enc-gate + dec-gate | **68.58** | *62.14* | *52.98* |
| **EN-DE** | **flickr16** | **flickr17** | **mscoco17** |
| $\text{IMG}_W$ | *45.09* | 40.81 | 36.94 |
| enc-gate | 44.75 | **41.44** | **37.76** |
| dec-gate | **45.21** | 40.79 | 36.47 |
| enc-gate + dec-gate | 44.91 | *41.06* | *37.40* |

Table 5: Comparison of strategies for integrating visual information (BLEU% scores). All results using Transformer, Multi30k+MS-COCO+SUBS3M$_{LM}$, Detectron mask surface, and domain labeling.

## 4.2 Visual feature selection

Image feature selection was performed using the LIUM-CVC translation system (Caglayan et al., 2017) training on the WMT18 training

| EN-FR | flickr16 | flickr17 | mscoco17 |
|---|---|---|---|
| SUBS3M$_{LM}$ detectron | 68.30 | 62.45 | 52.86 |
| +ensemble-of-3 | 68.72 | 62.70 | 53.06 |
| −visual features | **68.74** | **62.71** | 53.14 |
| −MS-COCO | 67.13 | 61.17 | **53.34** |
| −multi-lingual | 68.21 | 61.99 | 52.40 |
| SUBS6M$_{LM}$ detectron | 68.29 | 61.73 | 53.05 |
| SUBS3M$_{LM}$ gn2048 | 67.74 | 61.78 | 52.76 |
| SUBS3M$_{LM}$ text-only | 67.72 | 61.75 | 53.02 |
| EN-DE | flickr16 | flickr17 | mscoco17 |
| SUBS3M$_{LM}$ detectron | 45.09 | 40.81 | 36.94 |
| +ensemble-of-3 | 45.52 | **41.84** | **37.49** |
| −visual features | **45.59** | 41.75 | 37.43 |
| −MS-COCO | 45.11 | 40.52 | 36.47 |
| −multi-lingual | 44.95 | 40.09 | 35.28 |
| SUBS6M$_{LM}$ detectron | 45.50 | 41.01 | 36.81 |
| SUBS3M$_{LM}$ gn2048 | 45.38 | 40.07 | 36.82 |
| SUBS3M$_{LM}$ text-only | 44.87 | 41.27 | 36.59 |
| +multi-modal finetune | 44.56 | 41.61 | 36.93 |

Table 6: Ablation experiments (BLEU% scores). The row SUBS3M$_{LM}$ *detectron* shows our best single model. Individual components or data choices are varied one by one. + stands for adding a component, and − for removing a component or data set. Multiple modifications are indicated by increasing the indentation.

data, and evaluating on the *flickr16, flickr17* and *mscoco17* data sets. This setup is different from our final NMT architecture as the visual feature selection stage was performed at an earlier phase of our experiments. However, the LIUM-CVC setup without training set expansion was also faster to train which enabled a more extensive feature selection process.

We experimented with a set of state-of-the-art visual features, described below.

**CNN-based features** are 2048-dimensional feature vectors produced by applying reverse spatial pyramid pooling on features extracted from the $5^{th}$ Inception module of the pre-trained GoogLeNet (Szegedy et al., 2015). For a more detailed description, see (Shetty et al., 2018). These features are referred to as gn2048 in Table 6.

**Scene-type features** are 397-dimensional feature vectors representing the association score of an image to each of the scene types in SUN397 (Xiao et al., 2010). Each association score is determined by a separate Radial Basis Function Support Vector Machine (RBF-SVM) classifier trained from pre-trained GoogLeNet CNN features (Shetty et al., 2018).

**Action-type features** are 40-dimensional feature vectors created with RBF-SVM classifiers similarly to the scene-type features, but using the Stanford 40 Actions dataset (Yao et al., 2011) for training the classifiers. Pre-trained GoogLeNet CNN features (Szegedy et al., 2015) were again used as the first-stage visual descriptors.

**Object-type and location features** are generated using the Detectron software[5] which implements Mask R-CNN (He et al., 2017) with ResNeXt-152 (Xie et al., 2017) features. Mask R-CNN is an extension of Faster R-CNN object detection and localization (Ren et al., 2015) that also generates a segmentation mask for each of the detected objects. We generated an 80-dimensional *mask surface* feature vector by expressing the image surface area covered by each of the MS-COCO classes based on the detected masks.

We found that the Detectron mask surface resulted in the best BLEU scores in all evaluation data sets for improving the German translations. Only for *mscoco17* the results could be slightly improved with a fusion of mask surface and the SUN 397 scene-type feature. For French, the results were more varied, but we focused on improving the German translation results as those were poorer overall. We experimented with different ways of introducing the image features into the translation model implemented in LIUM-CVC, and found as in (Caglayan et al., 2017), that *trg-mul* worked best overall.

Later we learned that the *mscoco17* test set has some overlap with the COCO 2017 training set, which was used to train the Detectron models. Thus, the results on that test set may not be entirely reliable. However, we still feel confident in our conclusions as they are also confirmed by the *flickr16* and *flickr17* test sets.

### 4.3 Training

We use the following parameters for the network:[6] 6 Transformer layers in both encoder and decoder, 512-dimensional word embeddings and hidden states, dropout 0.1, batch

[5] https://github.com/facebookresearch/Detectron

[6] Parameters were chosen following the OpenNMT FAQ http://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model

Figure 2: Image 117 was translated correctly as feminine "eine besitzerin steht still und ihr brauner hund rennt auf sie zu ." when not using the image features, but as masculine "ein besitzer …" when using them. The English text contains the word "her". The person in the image has short hair and is wearing pants.

size 4096 tokens, label smoothing 0.1, Adam with initial learning rate 2 and $\beta_2$ 0.998.

For decoding, we use an ensemble procedure, in which the predictions of 3 independently trained models are combined by averaging after the softmax layer to compute combined prediction.

We evaluate the systems using uncased BLEU using multibleu. During tuning, we also used characterF (Popovic, 2015) with $\beta$ set to 1.0.

There are no images paired with the sentences in OpenSubtitles. When using OpenSubtitles in training multi-modal models, we feed in the mean vector of all visual features in the training data as a dummy visual feature.

### 4.4 Results

Based on the previous experiments, we chose the Transformer architecture, Multi30k+MS-COCO+SUBS3M$_{LM}$ data sets, Detectron mask surface visual features, and domain labeling.

Table 5 shows the BLEU scores for this configuration with different ways of integrating the visual features. The results are inconclusive. The ranking according to chrF-1.0 was not any clearer. Considering the results as a whole and the simplicity of the method, we chose IMG$_W$ going forward.

Table 6 shows results of ablation experiments removing or modifying one component or data choice at a time, and results when using ensemble decoding. Using ensemble decoding gave a consistent but small improvement. Multi-lingual models were clearly better than mono-lingual models. For French, 6M sentences of subtitle data gave worse results than 3M.

We experimented with adding multi-modality to a pre-trained text-only system using a fine tuning approach. In the fine tuning phase, a *dec-gate* gating layer was added to the network. The parameters of the main network were frozen, allowing only the added gating layer to be trained. Despite the freezing, the network was still able to unlearn most of the benefits of the additional text-only data. It appears that the output vocabulary was reduced back towards the vocabulary seen in the multi-modal training set. When the experiment was repeated so that the fine-tuning phase included the text-only data, the performance returned to approximately the same level as without tuning (+multi-modal finetune row in Table 6).

To explore the effect of the visual features on the translation of our final model, we performed an experiment where we retranslated using the ensemble while "blinding" the model. Instead of feeding in the actual visual features for the sentence, we used the mean vector of all visual features in the training data. The results are marked *-visual features* in Table 6. The resulting differences in the translated sentences were small, and mostly consisted of minor variations in word order. BLEU scores for French were surprisingly slightly improved by this procedure. We did not find clear examples of successful disambiguation. Figure 2 shows one example of a detrimental use of visual features.

It is possible that adding to the training data forward translations of MS-COCO captions from a text-only translation system introduced a biasing effect. If there is translational ambiguity that should be resolved using the image, the text-only system will not be able to resolve it correctly, instead likely yielding the word that is most frequent in that textual context. Using such data for training a multi-modal system might bias it towards ignoring the image.

On this year's *flickr18* test set, our system scores 38.54 BLEU for English-to-German and 44.11 BLEU for English-to-French.

## 5 Conclusions

Although we saw an improvement from incorporating multi-modal information, the improvement is modest. The largest differences in quality between the systems we experimented with can be attributed to the quality of the underlying text-only NMT system.

We found the amount of in-domain training data and multi-modal training data to be of great importance. The synthetic MS-COCO data was still beneficial, despite being forward translated, and the visual features being overconfident due to being extracted from a part of the image classifier training data.

Even after expansion with synthetic data, the available multi-modal data is dwarfed by the amount of text-only data. We found that movie subtitles worked well for this purpose. When adding text-only data, domain adaptation was important, and increasing the size of the selection met with diminishing returns.

Current methods do not fully address the problem of how to efficiently learn from both large text-only data and small multi-modal data simultaneously. We experimented with a fine tuning approach to this problem, without success.

Although the effect of the multi-modal information was modest, our system still had the highest performance of the task participants for the English-to-German and English-to-French language pairs, with absolute differences of +6.0 and +3.5 BLEU%, respectively.

## Acknowledgments

This work has been supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780069, and by the Academy of Finland in the project 313988. In addition the Finnish IT Center for Science (CSC) provided computational resources. We would also like to acknowledge the support by NVIDIA and their GPU grant.

## References

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost Van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*.

Iacer Calixto, Koel Dutta Chowdhury, and Qun Liu. 2017. DCU system report on the WMT 2017 multi-modal machine translation task. In *Proceedings of the Second Conference on Machine Translation*. pages 440–444.

Desmond Elliott and Àkos Kádár. 2017. Imagination improves multimodal translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. volume 1, pages 130–141.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pages 2980–2988.

Melvin Johnson, Mike Schuster, Quoc V Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, et al. 2016. Google's multilingual neural machine translation system: enabling zero-shot translation. *arXiv preprint arXiv:1611.04558* .

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 neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*. Association for Computational Linguistics, Melbourne, Australia, pages 116–121. http://www.aclweb.org/anthology/P18-4020.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*. https://doi.org/10.18653/v1/P17-4012.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR* abs/1405.0312. http://arxiv.org/abs/1405.0312.

Jan Niehues, Eunah Cho, Thanh-Le Ha, and Alex Waibel. 2016. Pre-translation for neural machine translation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. pages 1828–1836.

Robert Östling, Yves Scherrer, Jörg Tiedemann, Gongbo Tang, and Tommi Nieminen. 2017. The helsinki neural machine translation system. In *Proceedings of the Second Conference on Machine Translation*. pages 338–347.

Maja Popovic. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *WMT15*. pages 392–395.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems (NIPS)*. pages 91–99.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *ACL16*.

Rakshith Shetty, Hamed Rezazadegan Tavakoli, and Jorma Laaksonen. 2018. Image and video captioning with augmented neural architectures. *IEEE MultiMedia* To appear.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 1–9.

Jörg Tiedemann. 2009. News from OPUS - A collection of multilingual parallel corpora with tools and interfaces. In N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, editors, *Recent Advances in Natural Language Processing*, John Benjamins, Amsterdam/Philadelphia, Borovets, Bulgaria, volume V, pages 237–248.

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.

Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. 2010. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE conference on Computer vision and pattern recognition (CVPR)*. IEEE, pages 3485–3492.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pages 5987–5995.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas J. Guibas, and Fei-Fei Li. 2011. Human action recognition by learning bases of action attributes and parts. In *International Conference on Computer Vision (ICCV)*. Barcelona, Spain, pages 1331–1338.

# B Appendix: IWSLT MeMAD system description paper

# The MeMAD Submission to the IWSLT 2018 Speech Translation Task

*Umut Sulubacak* * *Jörg Tiedemann* * *Aku Rouhe* † *Stig-Arne Grönroos* † *Mikko Kurimo* †

∗ Department of Digital Humanities / HELDIG
University of Helsinki, Finland
{umut.sulubacak | jorg.tiedemann}@helsinki.fi

† Department of Signal Processing and Acoustics
Aalto University, Finland
{aku.rouhe | stig-arne.gronroos | mikko.kurimo}@aalto.fi

## Abstract

This paper describes the MeMAD project entry to the IWSLT Speech Translation Shared Task, addressing the translation of English audio into German text. Between the pipeline and end-to-end model tracks, we participated only in the former, with three contrastive systems. We tried also the latter, but were not able to finish our end-to-end model in time.

All of our systems start by transcribing the audio into text through an automatic speech recognition (ASR) model trained on the TED-LIUM English Speech Recognition Corpus (TED-LIUM). Afterwards, we feed the transcripts into English-German text-based neural machine translation (NMT) models. Our systems employ three different translation models trained on separate training sets compiled from the English-German part of the TED Speech Translation Corpus (TED-TRANS) and the OPENSUBTITLES2018 section of the OPUS collection.

In this paper, we also describe the experiments leading up to our final systems. Our experiments indicate that using OPENSUBTITLES2018 in training significantly improves translation performance. We also experimented with various pre- and postprocessing routines for the NMT module, but we did not have much success with these.

Our best-scoring system attains a BLEU score of 16.45 on the test set for this year's task.

## 1. Introduction

The evident challenge of speech translation is the transfer of implicit semantics between two different modalities. An end-to-end solution to this task must deal with the challenge posed by intermodality simultaneously with that of interlingual transfer. In a traditional pipeline approach, while speech-to-text transcription is abstracted from translation, there is then the additional risk of error transfer between the two stages. The MeMAD project[1] aims at multilingual description and search in audiovisual data. For this reason, multimodal translation is of great interest to the project.

Our pipeline submission to this year's speech translation task incorporates one ASR model and three contrastive NMT models. For the ASR module, we trained a time-delay neural network (TDNN) acoustic model using the Kaldi toolkit [1] on the provided TED-LIUM speech recognition corpus [2]. We used the transformer implementation of MarianNMT [3] to train our NMT models. For these models, we used contrastive splits of data compiled from two different sources: The $n$-best decoding hypotheses of the TED-TRANS [4] in-domain speech data, and a version of the OPENSUBTITLES2018 [5] out-of-domain text data (SUBS), further "translated" to an ASR-like format (SUBS-ASR) using a sequence-to-sequence NMT model. The primary system in our submission uses the NMT model trained on the whole data including SUBS-ASR, whereas one of the two contrastive systems uses the original SUBS before the conversion to an ASR-like format, and the other omits OPENSUBTITLES2018 altogether.

We provide further details about the ASR module in Section 2. Later, we provide a review of our experiments on the NMT module in Section 3. The first experiment we describe involves a pre-processing step where we convert our out-of-domain training data to an ASR-like format to avoid mismatch between source-side training samples. Afterwards, we report a postprocessing experiment where we retrain our NMT models with lowercased data, and defer case restoration to a subsequent procedure, and another where we translate several ASR hypotheses at once for each source sample, re-rank their output translations by a language model, and then choose the best-scoring translation for that sample. We present our results in Section 4 along with the relevant discussions.

## 2. Speech Recognition

The first step in our pipeline is automatic speech recognition. The organizers provide a baseline ASR implementation,

[1] https://www.memad.eu/

which consists of a single, end-to-end trained neural network using a Listen, Attend and Spell (LAS) architecture [6]. The baseline uses the XNMT toolkit [7]. However, we were not able to compile the baseline system, so we trained our own conventional, hybrid TDNN-HMM ASR system using the Kaldi toolkit.

### 2.1. Architecture

Our ASR system uses the standard Kaldi recipe for the TED-LIUM dataset (release 2), although we filter out some data from the training set to comply with the IWSLT restrictions. The recipe trains a TDNN acoustic model using the lattice-free maximum mutual information criterion [8]. The audio transcripts and large amount of out-of-domain text data included with the TED-LIUM dataset are used to train a heavily pruned 4-gram language model for first-pass decoding and less pruned 4-gram model for rescoring.

### 2.2. Word Error Rates

The LAS architecture has achieved state-of-the-art word error rates (WER) on a task with two orders of magnitude more training data than here [9], but on smaller datasets hybrid TDNN-HMM ASR approaches are still considerably better. Table 1 shows the results of our ASR model contrasted with those reported by XNMT in [7], on the TED-LIUM development and test sets.

| Model | Dev WER | Test WER |
|---|---|---|
| TDNN + large 4-gram | 8.24 | 8.83 |
| LAS | 15.83 | 16.16 |

Table 1: Word error rates on the TED-LIUM dataset.

## 3. Text-Based Translation

The ASR stage of our pipeline effectively converts the task of speech translation to text-based machine translation. For this stage, we build a variety of NMT setups and assess their performances. We experiment variously with the training architecture, different compositions of the training data, and several pre- and postprocessing methods. We present these experiments in detail in the subsections to follow, and then discuss their results in Section 4.

### 3.1. Data Preparation

We used the development and test sets from 2010's shared task for validation during training, and the test sets from the tasks between 2013 and 2015 for testing performance during development. In all of our NMT models, we preprocessed our data using the punctuation normalization and tokenization utilities from Moses [10], and applied byte-pair encoding [11] through full-cased and lowercased models as relevant, trained on the combined English and German texts in TED-TRANS and SUBS using 37,000 merge operations to create the vocabulary.

We experiment with attentional sequence-to-sequence models using the Nematus architecture [12] with tied embeddings, layer normalization, RNN dropout of 0.2 and source/target dropout of 0.1. Token embeddings have a dimensionality of 512 and the RNN layer units a size of 1024. The RNNs make use of GRUs in both, encoder and decoder. We use validation data and early stopping after five cycles (1,000 updates each) of decreasing cross-entropy scores. During training we apply dynamic mini-batch fitting with a workspace of 3GB. We also enable length normalization.

For the experiments with the transformer architecture we apply the standard setup with six layers in encoder and decoder, eight attention heads and a dynamic mini-batch fit to 8GB of work space. We also add recommended options such as transformer dropout of 0.1, label smoothing of 0.1, a learning rate of 0.0003, a learning-rate warmup with a linearly increasing rate during the first 16,000 steps, a decreasing learning rate starting at 16,000 steps, a gradient clip norm of 5 and exponential smoothing of parameters.

All translations are created with a beam decoder of size 12.

#### *3.1.1. ASR Output for TED Talks*

Translation models trained on standard language are not a good fit for a pipeline architecture that needs to handle noisy output from the ASR component discussed previously in Section 2. Therefore, we ran speech recognition on the entire TED-TRANS corpus in order to replace the original, human-produced English transcriptions with ASR output, which has realistic recognition errors.

To generate additional speech recognition errors to the training transcripts, we selected the top-50 decoding hypotheses. We did the same also for the development data to test our approach. We can now sample from those ASR hypotheses to create training data for our translation models that use the output of English ASR as its input. We experimented with various strategies varying from a selection of the top $n$ ASR candidates to different mixtures of hypotheses of different ranks of confidence. Some of these are shown in Table 2. In the end, there was not a lot of variance between the scores resulting from this selection, and we decided to use the top-10 ASR outputs in the remaining experiments to encourage some tolerance for speech recognition errors in the system.

#### *3.1.2. Translating Written English to ASR-English*

The training data that includes audio is very limited and much larger resources are available for text-only systems. Especially useful for the translation of TED talks is the collection of movie subtitles in OPENSUBTITLES2018. For English-German, there is a huge amount of movie subtitles (roughly 22 million aligned sentences with over 170 million

| Training data | Model | BLEU |
|---|---|---|
| TED-ASR-TOP-1 | AMUN | 16.65 |
| TED-ASR-TOP-10 | AMUN | 16.28 |
| TED-ASR-TOP-50 | AMUN | 15.88 |
| TED-ASR-TOP-1 | TRANSFORMER | 18.25 |
| TED-ASR-TOP-10 | TRANSFORMER | 17.90 |
| TED-ASR-TOP-50 | TRANSFORMER | 18.14 |

Table 2: Translating the development test set with different models and different selections of ASR output and German translations from the parallel TED-TRANS training corpus.

tokens per language) that can be used to boost the performance of the NMT module.

The problem is, of course, that the subtitles come in regular language, and, again, we would see a mismatch between the training data and the ASR output in the speech translation pipeline. In contrast to approaches that try to normalize ASR output to reflect standard text-based MT input such as [13], we had the idea to transform regular English into ASR-like English using a translation model trained on a parallel corpus of regular TED talk transcriptions and the ASR output generated for the TED talks that we described in the previous section. We ran a number of experiments to test the performance of such a model. Some of the results are listed in Table 3.

| Training data | Model | BLEU |
|---|---|---|
| TED-ASR-TOP-10 | AMUN | 61.87 |
| TED-ASR-TOP-10 | TRANSFORMER | 61.91 |
| TED-ASR-TOP-50 | AMUN | 61.82 |

Table 3: Translating English into ASR-like English using a model trained on TED-TRANS and tested on the development test set with original ASR output as reference.

As expected, the BLEU scores are rather high as the target language is the same as the source language, and we only mutate certain parts of the incoming sentences. The results show that there is not such a dramatic difference between the different setups (with respect to the model architecture and the data selection) and that a plain attentional sequence-to-sequence model with recurrent layers (AMUN) performs as well as a transformer model (TRANSFORMER) in this case. This makes sense, as we do not expect many complex long-distance dependencies that influence translation quality in this task. Therefore, we opted for the AMUN model trained on the top-10 ASR outputs, which we can decode efficiently in a distributed way on the CPU nodes of our computer cluster. With this we managed to successfully translate 99% of the entire SUBS collection from standard English into ASR-English. We refer to this set as SUBS-ASR.

We did a manual inspection on the result as well to see what the system actually learns to do. Most of the transformations are quite straightforward. The model learns to lowercase and to remove punctuation as our ASR output does not include it. However, it also does some other modifications that are more interesting from the viewpoint of an ASR module. While we do not have systematic evidence, Table 4 shows a few selected examples that show interesting patterns. First of all, it learns to spell out numbers (see *"2006"* in the first example). This is done consistently and quite accurately from what we have seen. Secondly, it replaces certain tokens with variants that resemble possible confusions that could come from a speech recognition system. The replacement of *"E.U."* with *"you"* and *"Stasi"* with *"stars he"* in these examples are quite plausible and rather surprising for a model that is trained on textual examples only. However, to conclude that the model learns some kind of implicit acoustic model would be a bit far-fetched, even though we would like to investigate the capacity of such an approach further in the future.

| | |
|---|---|
| **Original** | **Because in the summer of 2006, the E.U. Commission tabled a directive.** |
| ASR-REF | because in the summer of two thousand and six the e u commission tabled directive |
| ASR-OUT | because in the summer of two thousand and six you commission tabled a directive |
| **Original** | **Stasi was the secret police in East Germany.** |
| ASR-REF | what is the secret police in east germany |
| ASR-OUT | stars he was the secret police in east germany |

Table 4: Examples from the translations to ASR-like English. In the first column, ASR-REF refers to the top decoding hypothesis from the ASR model, while ASR-OUT is the output of the model translating the output to an ASR-like format.

In Section 4, we report on the effect of using synthetic ASR-like data on the translation pipeline.

### 3.2. Recasing Experiments

Our first attempt at a post-processing experiment involved using case-insensitive translation models, and deferring case restoration to a separate process unconditioned by the source side that we would apply after translation. We used the Moses toolkit [10] to train a recaser model on TED-TRANS. Afterwards, we re-trained a translation model on TED-ASR-TOP-10 and SUBS-ASR after lowercasing the training and validation sets, re-translated the development test set with this model, and then used the recaser to restore cases in the lowercase translations that we obtained. As shown in Table 5, evaluating the translations produced through these additional steps yielded scores that were very similar to those

obtained by the original case-sensitive translation models, and the result of this experiment was inconclusive.

| Training data | BLEU | BLEU-LC |
|---|---|---|
| TED-ASR-TOP-10+SUBS-ASR | 19.79 | 20.43 |
| TED-ASR-TOP-10+SUBS-ASR-LC | 19.73 | 20.91 |

Table 5: Case-sensitive models (TRANSFORMER) versus lowercased models with subsequent recasing. Recasing causes a larger drop than the model gains from training on lowercased training data. BLEU-LC refers to case-insensitive BLEU scores.

### 3.3. Reranking Experiments

In addition to using different subsets of the $n$-best lists output by the ASR model as additional training samples for the translation module, we also tried reranking alternatives using KenLM [14]. We initially generated a tokenized and lowercased version of TED-TRANS with all punctuation stripped, and then trained a language model on this set. We used this model to score and rerank samples in the 50-best lists, and then generated a new top-10 subset from this reranked version. However, when we re-trained translation models from these alternative sets, we observed that the model trained on the top-10 subsets before reranking exhibited a significantly better translation performance. We suspect that this is because, while the language model is useful for assessing the surface similarity of the ASR outputs to the source-side references, it was not uncommon for it to assign higher scores to ASR outputs that are semantically inconsistent with the target-side references, causing the NMT module to produce erroneous translations.

Similarly, we experimented with another language model trained on the target side of TED-TRANS, without the preprocessing. We intended this model to score and rerank outputs of the translation models, rather than the ASR module. To measure the effect of this language model, we fed the audio of our internal test set split through the ASR module, and produced 50-best lists for each sample. Afterwards, we used the language model to score and rerank the alternative transcripts for each sample produced by translating this set, and then selected the highest-scoring output for each sample. As in the previous language model experiment, employing this additional procedure significantly crippled the performance of our translation models.

## 4. Results

The results on development data reveal expected tendencies that we report below. First of all, as consistent with a lot of related literature, we can see a boost in performance when switching from a recurrent network model to the transformer model with multiple self-attention mechanisms. Table 6 shows a clear pattern of the superior performance of the transformer model that is also visible in additional runs that we do not list here. Secondly, we can see the importance of additional training data even if they come from slightly different domains. The vast amount of movie subtitles in OPENSUBTITLES2018 boosts the performance by about 3 absolute BLEU points. Note that the scores in Table 6 refer to models that do not use subtitles transformed into ASR-like English (SUBS-ASR) and which are not fine-tuned to TED talk translations.

| Training data | Model | BLEU |
|---|---|---|
| TED-ASR-TOP-10 | AMUN | 16.28 |
| TED-ASR-TOP-10+SUBS | AMUN | 19.93 |
| TED-ASR-TOP-10 | TRANSFORMER | 17.90 |
| TED-ASR-TOP-10+SUBS | TRANSFORMER | 20.44 |

Table 6: Model performance on the development test set when adding movie subtitles to the training data.

The effect of pre-processing by producing ASR-like English in the subtitle corpus is surprisingly negative. If we look at the scores in Table 7, we can see that the performance actually drops in all cases when considering only the untuned systems. We did not really expect that with the rather positive impression that we got from the manual inspection of the English-to-ASR translation discussed earlier. However, it is interesting to see the effect of fine-tuning. Fine-tuning here refers to a second training procedure that continues training with pure in-domain data (TED talks) after training the general model on the entire data set until convergence on validation data. Table 7 shows an interesting effect that may explain the difficulties of the integration of the synthetic ASR data. The fine-tuned model actually outperforms the model trained on standard data, which is due to a substantial jump from untuned models to the tuned version. The difference between those models with standard data is, on the other hand, only minor.

| | BLEU | |
|---|---|---|
| Training data | Untuned | Tuned |
| TED-ASR-TOP-10+SUBS | 20.44 | 20.58 |
| TED-ASR-TOP-10+SUBS-ASR | 19.79 | 20.80 |

Table 7: Training with original movie subtitles versus subtitles with English transformed into ASR-like English, before and after fine-tuning on TED-ASR-TOP-10 as pure in-domain training data (TRANSFORMER).

The synthetic ASR data look more similar to the TED-ASR data and, therefore, the model might get more confused between in-domain and out-of-domain data than it does for the model trained on the original subtitle data in connection with TED-ASR. Fine-tuning to TED-ASR brings the model

back on track again and synthetic ASR data becomes modestly beneficial.

Also of note is the contrast between the evaluation scores we obtained in development and those from the official test set. The translations we submitted obtain the BLEU scores shown in Table 8 on this year's test set.

| Training data | BLEU |
|---|---|
| TED-ASR-TOP-10 | 14.34 |
| TED-ASR-TOP-10+SUBS | 16.45 |
| TED-ASR-TOP-10+SUBS-ASR | 15.80 |

Table 8: BLEU scores from our final models (TRANSFORMER)—respectively, the 2nd contrastive, 1st contrastive, and primary submission—on this year's test set. The scores from the two models with SUBS in their training data were obtained after fine-tuning on TED-ASR-TOP-10.

## 5. Conclusions

Apart from employing well-established practices such as normalization and byte-pair encoding as well as the benefits of using the transformer architecture, the only substantial boost to translation performance came from our data selection for the NMT module. The NMT module of our best-performing system on this year's test set was trained on TED-ASR-TOP-10 and the raw SUBS, and later fine-tuned on TED-ASR-TOP-10.

Although we ran many experiments to improve various steps of our speech translation pipeline, their influence on translation performance has been marginal at best. The effects of training with different TED-ASR subsets were hard to distinguish. While using SUBS-ASR in training seemed to provide a modest improvement in development, this effect was not carried over to the final results on the test set. The later experiments with lowercasing and recasing had an ambiguous effect, and those with reranking had a noticeably negative outcome.

In future work, our aim is to further investigate what factors in a good speech translation model, and continue experimenting in relation to these on the NMT module. We will also try to improve our TDNN-HMM ASR module by replacing the n-grams with an RNNLM, and try see how our complete end-to-end speech-to-text translation model performs after having sufficient training time.

## 6. Acknowledgements

This work has been supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 780069, and by the Academy of Finland in the project 313988. In addition the Finnish IT Center for Science (CSC) provided computational resources.

## 7. References

[1] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society, Dec. 2011, iEEE Catalog No.: CFP11SRW-USB.

[2] A. Rousseau, P. Deléglise, and Y. Esteve, "TED-LIUM: an automatic speech recognition dedicated corpus," in *LREC*, 2012, pp. 125–129.

[3] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, and A. Birch, "Marian: Fast neural machine translation in C++," in *Proceedings of ACL 2018, System Demonstrations*. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 116–121. [Online]. Available: http://www.aclweb.org/anthology/P18-4020

[4] M. Cettolo, C. Girardi, and M. Federico, "WIT$^3$: Web inventory of transcribed and translated talks," in *Proceedings of the 16$^{th}$ Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May 2012, pp. 261–268.

[5] J. Tiedemann, "News from OPUS - A collection of multilingual parallel corpora with tools and interfaces," in *Recent Advances in Natural Language Processing*, N. Nicolov, K. Bontcheva, G. Angelova, and R. Mitkov, Eds. Borovets, Bulgaria: John Benjamins, Amsterdam/Philadelphia, 2009, vol. V, pp. 237–248.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4960–4964.

[7] G. Neubig, M. Sperber, X. Wang, M. Felix, A. Matthews, S. Padmanabhan, Y. Qi, D. S. Sachan, P. Arthur, P. Godard, *et al.*, "XNMT: The extensible neural machine translation toolkit," *arXiv preprint arXiv:1803.00188*, 2018.

[8] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely sequence-trained neural networks for ASR based on lattice-free MMI," in *Interspeech*, 2016, pp. 2751–2755.

[9] C.-C. Chiu, T. N. Sainath, Y. Wu, R. Prabhavalkar, P. Nguyen, Z. Chen, A. Kannan, R. J. Weiss, K. Rao, K. Gonina, *et al.*, "State-of-the-art speech recognition

with sequence-to-sequence models," *arXiv preprint arXiv:1712.01769*, 2017.

[10] P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, *et al.*, "Moses: Open source toolkit for statistical machine translation," in *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*. Association for Computational Linguistics, 2007, pp. 177–180.

[11] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL16*, 2015.

[12] R. Sennrich, O. Firat, K. Cho, A. Birch, B. Haddow, J. Hitschler, M. Junczys-Dowmunt, S. Läubli, A. V. M. Barone, J. Mokry, and M. Nadejde, "Nematus: a toolkit for neural machine translation," *CoRR*, vol. abs/1703.04357, 2017. [Online]. Available: http://arxiv.org/abs/1703.04357

[13] E. Matusov, A. Mauser, and H. Ney, "Automatic sentence segmentation and punctuation prediction for spoken language translation," in *International Workshop on Spoken Language Translation*, Kyoto, Japan, Nov 2006, pp. 158–165.

[14] K. Heafield, "KenLM: Faster and smaller language model queries," in *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2011, pp. 187–197.