Machine Translation (2020) 34:97–147
https://doi.org/10.1007/s10590-020-09250-0

Multimodal machine translation through visuals and speech

Umut Sulubacak[1] · Ozan Caglayan[3] · Stig-Arne Grönroos[2] · Aku Rouhe[2] · Desmond Elliott[4] · Lucia Specia[3] · Jörg Tiedemann[1]

Received: 5 December 2019 / Accepted: 22 July 2020 / Published online: 13 August 2020

Abstract
Multimodal machine translation involves drawing information from more than one modality, based on the assumption that the additional modalities will contain useful alternative views of the input data. The most prominent tasks in this area are spoken language translation, image-guided translation, and video-guided translation, which exploit audio and visual modalities, respectively. These tasks are distinguished from their monolingual counterparts of speech recognition, image captioning, and video captioning by the requirement of models to generate outputs in a different language. This survey reviews the major data resources for these tasks, the evaluation campaigns concentrated around them, the state of the art in end-to-end and pipeline approaches, and also the challenges in performance evaluation. The paper concludes with a discussion of directions for future research in these areas: the need for more expansive and challenging datasets, for targeted evaluations of model performance, and for multimodality in both the input and output space.

Keywords Natural language processing · Machine translation · Multimodal machine translation · Image-guided translation · Speech language translation

1 Introduction

Humans are able to make use of complex combinations of visual, auditory, tactile and other stimuli, and are capable of not only handling each sensory modality in isolation, but also simultaneously integrating them to improve the quality of perception and understanding (Stein et al. 2009). From a computational perspective, natural language processing (NLP) requires such abilities, too, in order to approach human-level grounding and understanding in various AI tasks.

✉ Umut Sulubacak
umut.sulubacak@helsinki.fi

Extended author information available on the last page of the article

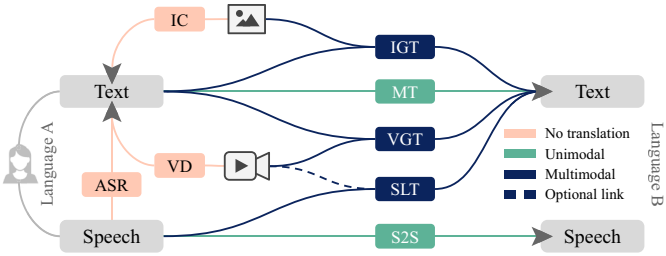

Fig. 1 Prominent examples of multimodal translation tasks, such as image-guided translation (IGT), video-guided translation (VGT), and spoken language translation (SLT), shown in contrast to unimodal translation tasks, such as text-based machine translation (MT) and speech-to-speech translation (S2S), and multimodal NLP tasks that do not involve translation, such as automatic speech recognition (ASR), image captioning (IC), and video description (VD)

While language covers written, spoken, and sign language in human communication; vision, speech, and language processing communities have worked largely apart in the past. As a consequence, NLP became more focused towards *textual* representations, which often disregard many other characteristics of communication such as non-verbal auditory cues, facial expressions, and hand gestures. Luckily, recent advances in multimodal machine learning have brought these different aspects of language together, through a plethora of multimodal NLP tasks. Specifically, these tasks involve more than one modality, either by (i) using one modality to aid the interpretation of language in another modality, or by (ii) converting one modality into another. Notable examples for the first category are extensions to initially unimodal problems, such as multimodal coreference resolution (Ramanathan et al. 2014), multimodal sentiment analysis (Zadeh et al. 2016), and visual question answering (Antol et al. 2015). For the second category that involves modality conversion, well-known examples are image captioning (IC) (Bernardi et al. 2016), where the task is to generate a textual description from an image, automatic speech recognition (ASR) (Yu and Deng 2016), where the task is to transcribe spoken language audio into text, and speech synthesis (Ling et al. 2015), which is the converse of ASR, with the goal of generating speech from written language.

Although more pointers exist in general surveys of multimodality in NLP (Bernardi et al. 2016; Baltrušaitis et al. 2017; Kafle and Kanan 2017; Mogadala et al. 2019), this article is concerned with tasks that involve both multiple modalities and different input and output languages, i.e. the tasks that fall under the umbrella of multimodal machine translation (MMT). The connection between modalities and translation tasks according to our definition is illustrated in Fig. 1, outlining the major tasks of spoken language translation (SLT) (Akiba et al. 2004), image-guided translation (IGT) (Elliott et al. 2015; Specia et al. 2016), and video-guided translation (VGT) (Sanabria et al. 2018; Wang et al. 2019b).

Today, the rising interest in MMT is largely driven by the state-of-the-art performance and the architectural flexibility of neural sequence-to-sequence models (Sutskever et al. 2014; Bahdanau et al. 2015; Vaswani et al. 2017). This flexibility, which is due to the end-to-end nature of these approaches, has the potential of bringing

the vision, speech and language processing communities back together. From a historical point of view however, there was already a great deal of interest in doing machine translation (MT) with non-text modalities, even before the arrival of successful statistical machine translation models. Among 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, making up a full speech-to-speech translation system. Further research has led to several other speech-to-speech translation systems (Lavie et al. 1997; Takezawa et al. 1998; Wahlster 2000).

In contrast, the use of visual modality in translation has not attracted comparable interest until recently. At present, there is a variety of multimodal task formulations including some form of machine translation, involving image captions, instructional text with photographs, video recordings of sign language, subtitles for videos (and especially movies), and descriptions of video scenes. As a consequence, modern multimodal MT studies dealing with visual (or audiovisual) information are becoming as prominent as those tackling audio. We believe that multimodal MT is a better reflection of how humans acquire and process language, with many theoretical advantages in language grounding over text-based MT as well as the potential for new practical applications like cross-modal cross-lingual information retrieval (Calixto and Liu 2017b; Gella et al. 2017; Kádár et al. 2018).

In the following, we will provide a detailed description of MMT tasks and approaches that have been proposed in the past. Sect. 2 contains an overview of the tasks of spoken language translation, image-guided translation and video-guided translation. Section 3 reviews the methods and caveats of evaluating MT performance, and discusses prominent evaluation campaigns, while Sect. 4 contains an overview of major datasets that can be used as training or test corpora. Section 5 discusses the state-of-the-art models and approaches in MMT, especially focusing on image-guided translation and spoken language translation. Section 6 outlines fruitful directions of future research in multimodal MT.

2 Tasks

While our definition of multimodal MT excludes both cross-modal conversion tasks with no cross-linguality (e.g. automatic speech recognition and video description), and machine translation tasks within a single modality (e.g. text-to-text and speech-to-speech translation), it is still general enough to accommodate a fair variety of tasks. Some of these tasks such as spoken language translation (SLT) and continuous sign language recognition (CSLR) meet the criteria because their source and target languages are, by definition, expressed through different modes. Other tasks like image-guided translation (IGT) and video-guided translation (VGT) are included on the grounds that they complement the source language with related visuals that constitute an extra modality. In some cases, a well-established multimodal machine translation task can be characterised by methodological constraints (e.g. simultaneous interpretation), or by domain and semantics (e.g. video description translation).

We observe that a shared modality composition is the foremost prerequisite that dictates the applicability of data, approaches and methodologies across multimodal translation tasks. For this reason, further in this article, we classify the studies we have surveyed according to the modality composition involved. We also restrict the scope of our discussions to the more well-recognised cases that involve audio and/or visual data in addition to text. It should be noted that, despite our grouping, there may be conceptual differences between the modalities involved in different multimodal MT tasks, where, for example, the audio in SLT corresponds to speech that is semantically equivalent to the associated text, while the visual modalities in IGT and VGT may merely serve to narrow down the context. In the following subsections, we explain our use of the terms *spoken language translation*, *image-guided translation*, and *video-guided translation*, and provide further discussions for each of these tasks.

2.1 Spoken language translation

Spoken language translation (SLT), also known as speech-to-text translation or automatic speech translation, comprises the translation of speech in a source language to text in a target language. As such, it differs from conventional MT in the source-side modality. The need to simultaneously perform both modality conversion and translation means that systems must learn a complex input–output mapping, which poses a significant challenge. The SLT task has been shaped by a number of influential early works (e.g. Vidal 1997; Ney 1999), and championed by the speech translation tasks of the IWSLT evaluation campaign since 2004 (see Sect. 3.2.2).

Traditionally, SLT was addressed by a pipeline approach (see Sect. 5 for more details), effectively separating multimodal MT into modality conversion followed by unimodal MT. More recently, end-to-end systems have been proposed, often based on NMT architectures, where the source language audio sequence is directly converted to the target language text sequence (Weiss et al. 2017; Bérard et al. 2018). Despite the short time during which end-to-end approaches have been developed, they have been rapidly closing the gap with the dominant paradigm of pipeline systems. The current state of end-to-end systems is discussed further in Sect. 5.2.3.

2.2 Image-guided translation

Image-guided translation can be defined as a contextual grounding task, where, given a set of images and associated documents, the aim is to enhance the translation of the documents by leveraging their semantic correspondence to the images. Resolving ambiguities through visual cues is one of the main motivating forces behind this task.

A well-known realisation of IGT is image caption translation, where the correspondence is related to sentences being the descriptions of the images. Initial attempts at image caption translation were mostly pipeline approaches: Elliott et al. (2015) proposed a pipeline of visually conditioned neural language models, while Hitschler et al. (2016) approached the problem from a multimodal retrieval

and reranking perspective. With the introduction of the WMT multimodal translation shared task (Specia et al. 2016, see Sect. 3.2.1), IGT attracted a lot more attention from the research community. Today, the prominent approaches rely on visually conditioning end-to-end neural MT systems with visual features extracted from state-of-the-art pretrained CNNs.

Although the utility of the visual modality has recently been disputed under specific dataset and task conditions (Elliott 2018; Caglayan et al. 2019), using images when translating captions is theoretically very advantageous to handle grammatical characteristics (e.g. noun genders) in translating between dissimilar languages, and resolving translational ambiguities. Also, Caglayan et al. (2019) shows how state-of-the-art models become capable of leveraging the visual signal when source captions are deliberately deteriorated in a simulated low-resource scenario. We discuss the current state of the art and the predominant approaches in IGT in Sect. 5.1.

2.3 Video-guided translation

We posit the task of video-guided translation (VGT) as a multimodal machine translation task similar to image-guided translation, but tackling video clips (and potentially audio clips as well) rather than static images associated with the textual input. Within video-guided translation, there can be variants depending on the textual content. The source text can be transcripts of speech from the video, which would be typically segmented as standard subtitles, or a textual description of the visual scene or an action demonstrated in the clip, often created for visually impaired people. As such, video-guided translation can be subject to particular challenges from both SLT (time-variant audiovisual input) and IGT (indirect correspondence between source modalities). On the other hand, these similarities could also indicate that it might be possible to adapt or reuse approaches from both of those areas to bootstrap VGT systems.

One major challenge hindering progress in video-guided translation is the relative scarcity of datasets. While a large collection such as the OpenSubtitles corpus[1] (Lison and Tiedemann 2016) can provide access to a considerable amount of parallel subtitles, there is no attached audiovisual content since the corresponding movies are not freely available. Recent efforts to compile freely accessible data for video-guided translation, like the How2 (Sanabria et al. 2018) and VaTeX (Wang et al. 2019b) datasets (both described in Sect. 4.3) have started to alleviate this bottleneck. Although there has been decidedly little time to observe the full impact of such initiatives, we hope that they will inspire further research in video-guided translation.

[1] Derived from https://www.opensubtitles.com/.

3 Evaluation

Evaluating the performance of a machine translation system is a difficult and controversial problem. Typically, there are numerous ways of translating even a single sentence which would be acceptably produced by human translators (or systems), and it is often unclear which one is (or which ones are) good or better, and in what respect, given that the pertinent evaluation criteria are multi-dimensional, context-dependent, and highly subjective (see for example Chesterman and Wagner 2002; Drugan 2013). Traditionally, human analysis of translation quality has often been divided into the evaluation of adequacy (semantic transfer from source language) and fluency (grammatical soundness of target language) (Doherty 2017). While this separation is considered somewhat artificial, it was created to make evaluation simpler and to allow comparison of translation systems in more specific terms. In practice, systems that are good at one criterion tend to be good at the other, and a lot of the more recent evaluation campaigns have focused on directly ranking systems for general quality rather than scoring individual systems on these criteria (relative ranking), or scoring systems for general quality instead (direct assessment).

Since human evaluation comes with considerable monetary and time costs (Castilho et al. 2018), evaluation efforts have converged to devising automatic metrics in recent years (Ma et al. 2018, 2019), which typically operate by comparing the output of a translation system against one or more human translations. While a number of metrics have been proposed over the last two decades, they are mostly based on statistics computed between the translation hypothesis and one or more references. Procuring reference translations in itself entails some costs, and any metrics and approaches that require multiple references to work well may therefore not be feasible for common use. Further in this section, we discuss the details of some of the dominant evaluation metrics as well as the most well-known shared tasks of multimodal MT that serve as standard evaluation settings to facilitate research.

3.1 Metrics

Among the various MT evaluation metrics in the literature, the most commonly used ones are BLEU (Papineni et al. 2001), METEOR (Lavie and Agarwal 2007; Denkowski and Lavie 2014) and TER (Snover et al. 2006). To summarise them briefly, BLEU is based on an aggregate precision measure of n-gram matches between the reference(s) and machine translation, and penalises translations that are too short. METEOR accounts for and gives partial credit to stem, synonyms, and paraphrase matches, and considers both precision and recall with configurable weights for both criteria. TER is a variant of word-level edit distance between the source and the target sentences, with an added operation for shifting one or more adjacent words. BLEU is by far the most commonly used automatic evaluation metric, despite its relative simplicity. Most quantitative comparisons of machine translation systems are reported using only BLEU scores. METEOR has been shown to correlate better with human judgements (especially for

adequacy) due to both its flexibility in string matching and its better balance between precision and recall, but its dependency on linguistic resources makes it less applicable in the general case. Both BLEU and METEOR, much like the majority of other evaluation metrics developed so far, are reference-based metrics. These metrics are inadvertently heavily biased on the translation styles that they see in the reference data, and end up penalising any alternative phrasing that might be equally correct (Fomicheva and Specia 2016).

Human evaluation is the optimal choice when a trustworthy measure of translation quality is needed and resources to perform it are available. The usual strategies for human evaluation are fluency and adequacy rankings, direct assessment (DA) (Graham et al. 2013), and post-editing evaluation (PE) (Snover et al. 2006). Fluency and adequacy rankings are conventionally between 1 and 5, while DA is a general scale between 0 and 100 indicating how "good" the translation is, either with respect the original sentence in the source language (DA-*src*), or the ground truth translation in the target language (DA-*ref*). 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, or other metrics such as post-editing time and keystrokes (Specia et al. 2017). For pragmatics reasons, these human evaluation methods are typically crowdsourced to non-expert annotators to reduce costs. While this may still result in consistent evaluation scores if multiple crowd annotators are considered, it is a well-accepted 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 are an active research subject in their own right (Specia et al. 2018; Ma et al. 2018, 2019). However, there is currently little research on developing evaluation approaches specifically tailored to multimodal translation. Fully-automatic evaluation is typically text-based, while methods that go beyond the text rely on manually annotated resources, and could rather be considered semi-automatic. One such method is multimodal lexical translation (MLT) (Lala and Specia 2018), which is a measure of translation accuracy for a set of ambiguous words given their textual context and an associated image that allows visual disambiguation. Even in human evaluation there are only a few examples where the evaluation is multimodal, such as the addition of images in the evaluation of image caption translations via direct assessment (Elliott et al. 2017; Barrault et al. 2018), or via qualitative comparisons of post-editing (Frank 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. One possible promising direction is the work of Madhyastha et al. (2019) for image captioning evaluation, where the content of the image is directly taken into account via the matching of detected objects in the image and concepts in the generated caption.

3.2 Shared tasks

A great deal of research into developing natural language processing systems 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-guided translation, both of which are represented in shared tasks organised by the machine translation community. The conference on machine translation (WMT) ran three shared tasks for image caption translation from 2016–2018, and the International Workshop on Spoken Language Translation (IWSLT) has led an annual evaluation campaign on speech translation since 2004.

3.2.1 Image-guided translation: 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. 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, and new test sets. The third shared task in 2018 added Czech as a third possible target language, and another new test set. This last[2] 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 ambiguous verbs (see Sect. 4.1). The translations generated by the submitted systems are scored by the METEOR, BLEU, and TER metrics. In addition, all participants are required to devote resources to manually scoring translations in a blind fashion. This scoring is done by direct assessment using the original source captions and the image as references. 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 submitted systems are analysed in relation to this.

[2] The multimodal translation task was not held in WMT 2019.

3.2.2 Spoken language translation: IWSLT evaluation campaign

The spoken language translation tasks have been held as part of the annual IWSLT evaluation campaign since Akiba et al. (2004). Following the earlier C-STAR evaluations, the aim of the campaign is to investigate newly-developing translation technologies as well as methodologies for evaluating them. The first years of the campaign were based on a basic travel expression corpus developed by C-STAR to facilitate standard evaluation, 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 spoken language 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 with Paul et al. (2010), the campaign adopted TED talks as their primary training data, and eventually shifted away from the tourism domain towards lecture transcripts.

Until Cettolo et al. (2016), the evaluation campaign had 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 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 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 dialogue translation (Cettolo et al. 2017) and low-resource spoken languagetranslation (Niehues et al. 2018) were reintroduced to the IWSLT evaluation campaign from 2017 on, as TED data and machine translation literature both grew richer. Niehues et al. (2019) introduced a new audiovisual spoken language translation task, leveraging the How2 corpus (Sanabria et al. 2018). In this task, video is included as an additional input modality, for the general case of subtitling audiovisual content.

4 Datasets

Text-based machine translation has recently enjoyed widespread success with the adoption of deep learning model architectures. The success of these data-driven systems rely heavily on the factor of data availability. An implication of this for multimodal MT is the need for large datasets in order to keep up with the data-driven state-of-the-art methodologies. Unfortunately, due to its simultaneous requirement of multimodality and multilinguality in data, multimodal MT is subject to an especially restrictive bottleneck. Datasets that are sufficiently large for training

Table 1 Summary statistics from most prominent multimodal machine translation datasets

Dataset	Media	Text	Languages	Tasks
IAPR TC-12 (Grubinger et al. 2006)	20k images	20k captions	DE, EN	IGT
Flickr8k (Rashtchian et al. 2010)	8k images	41k captions	EN, TR, ZH	IGT
Flickr30k (Young et al. 2014)	30k images	158k captions	DE, EN	IGT
Multi30k (Elliott et al. 2016)	30k images	30k captions	CS, DE, EN, FR	IGT
QED (Abdelali et al. 2014)	23.1k video clips	8k–335k segments	20 languages	SLT, VGT
How2 (Sanabria et al. 2018)	13k video clips	189k segments	EN, PT	SLT, VGT
VaTeX (Wang et al. 2019b)	41k video clips	206k segments	EN, ZH	SLT, VGT
WIT^3 (Cettolo et al. 2012)	2086 audio clips	3–575k segments	109 languages	SLT
Fisher & Callhome (Post et al. 2013)	38 h audio	171k segments	EN, ES	SLT
MSLT (Federmann and Lewis 2017)	4.5–10 h audio	7 k–18k segments	DE, EN, FR, JA, ZH	SLT
IWSLT '18 (Niehues et al. 2018)	1565 audio clips	171k segments	DE, EN	SLT
LibriSpeech (Kocabiyikoglu et al. 2018)	236 h audio	131k segments	EN, FR	SLT
MuST-C (Di Gangi et al. 2019b)	385–504 h audio	211k–280k segments	10 languages	SLT
MaSS (Boito et al. 2019)	18.5–23 h audio	8.2k segments	8 languages	SLT

We report image captions *per language*, and audio clips and segments *per language pair*

multimodal MT models are only available for a handful of languages and domain-specific tasks. The limitations imposed by this are increasingly well-recognised, as evidenced by the fact that most major datasets intended for multimodal MT were released relatively recently. Some of these datasets are outlined in Table 1, and explained in more detail in the subsections to follow.

4.1 Image-guided translation datasets

IAPR TC-12 The International Association of Pattern Recognition (IAPR) TC-12 benchmark dataset (Grubinger et al. 2006) was created for the cross-language image retrieval track of the CLEF evaluation campaign (ImageCLEF 2006) (Clough et al. 2006). The benchmark is structurally similar to the multilingual image caption datasets commonly used by contemporary image-guided translation systems. IAPR TC-12 contains 20,000 images from a collection of photos of landmarks taken in various countries, provided by a travel organisation. Each image was originally annotated with German descriptions, and later translated to English. These descriptions are composed of phrases that describe the visual contents of the photo following strict linguistic patterns, as shown in Fig. 2. The dataset al.so contains *light* annotations such as titles and locations in English, German, and Spanish.

EN: the courtyard of an orange, two-storey building with a footpath to a swimming pool in the shape of an eight and small palm trees to the left and right;

DE: der Innenhof eines zweistöckigen, orangen Gebäudes mit einem Weg zu einem achterförmigen Schwimmbecken und kleine Palmen rechts und links davon;

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.

Fig. 2 Examples from IAPR TC-12 image descriptions (top) and Multi30k image captions (bottom)

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 training corpus for automatic image captioning, the dataset comprises a set of 8,092 images extracted from the Flickr website, each with 5 crowdsourced captions in English that describe the image. Flickr8k has shorter captions compared to IAPR TC-12, focusing on the most salient objects or actions, rather than complete descriptions. As the dataset has been a popular and useful resource, it has been further extended with captions in other languages such as Chinese (Li et al. 2016) and Turkish (Unal et al. 2016). However, as these captions were independently crowdsourced, they are not translations of each other, which makes them less effective for MMT.

Flickr30k/Multi30k 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. Multi30k (Elliott et al. 2016) was initially released as a bilingual subset of Flickr30k captions, providing German translations for 1 out of the 5 English captions per image, with the aim of stimulating multimodal and multilingual research. In addition, the study collected 5 independent German captions for each image. The WMT multimodal translation tasks later introduced French (Elliott et al. 2017) and Czech (Barrault et al. 2018) extensions to Multi30k, making it a staple dataset for image-guided translation, and further expanding the set's utility to cutting-edge subtasks such as multisource training. An example from this dataset can be seen in Fig. 2.

WMT test sets The past 3 years of multimodal shared tasks at WMT each came with a designated test set for the task (Specia et al. 2016; Elliott et al. 2017; Barrault et al. 2018). 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 validation and internal evaluation. In parallel with the language expansion of Multi30k, the test set from 2016 contains only English and German captions, and the one from 2017 contains only English, German, and French. The 2018 test set contains English, German, French, and Czech captions that are not publicly available, though systems can be evaluated against it using an online server.[3]

MS COCO captions Introduced in 2015, the MS COCO Captions dataset (Chen et al. 2015) offers caption annotations for a subset of roughly 123,000 images from the large-scale object detection and segmentation training corpus MS COCO (Microsoft Common Objects in Context) (Lin et al. 2014b). Each image in this dataset is associated with up to 5 independently annotated English captions, with a total of 616,767 captions. Though originally a monolingual dataset, the dataset's large size makes it useful for data augmentation methods for image-guided translation, as demonstrated in Grönroos et al. (2018). There has also been some effort to add other languages to COCO. A small subset with only 461 captions containing ambiguous verbs was released as a test set for the WMT 2017 multimodal machine translation shared task, called Ambiguous COCO (Elliott et al. 2017), and is available in all target languages of the task. The YJ Captions dataset (Miyazaki and Shimizu 2016) and the STAIR Captions dataset (Yoshikawa et al. 2017) comprise, respectively, 132k and 820k crowdsourced Japanese captions for COCO images. However, these are not parallel to the original English captions, as they were independently annotated.

4.2 Spoken language translation datasets

The TED corpus TED is a nonprofit organisation that hosts talks in various topics, comprising a rich resource of spoken language produced by a variety of speakers in English. Video recordings of all TED talks are made available through the TED website,[4] as well as transcripts with translations in up to 116 languages. While the talks comprise a rich resource for language processing, the original transcripts are divided into arbitrary segments formatted like subtitles, which makes it difficult to get an accurate sentence-level parallel segmentation for use in translation systems. While resegmentation is possible with heuristic approaches, it comes with the additional challenge of aligning the new segments to the audiovisual content, and to each other in source and target languages.

The Web Inventory of Transcribed and Translated Talks (WIT^3) (Cettolo et al. 2012) is a resource with the aim of facilitating the use of the TED Corpus in MT. The initiative distributes transcripts organised in XML files through their website[5], as well as tools to process them in order to extract parallel sentences. Currently, WIT^3 covers 2086 talks in 109 languages containing anywhere between 3 and 575k segments in raw transcripts, and is continually growing.

[3] https://competitions.codalab.org/competitions/19917.
[4] http://www.ted.com/talks.
[5] http://wit3.fbk.eu.

Since 2011, the annual speech translation tracks of the IWSLT evaluation campaign (see Sect. 3.2.2) has used datasets compiled from WIT[3]. While each of these sets contain a high-quality selection of English transcripts aligned with the audio and the target languages featured each year, they are not useful for training SLT systems due to their small sizes. As part of the 2018 campaign, the organisers released a large-scale English–German corpus (Niehues et al. 2018) containing 1565 talks with 170,965 segments automatically aligned based on time overlap, which allows end-to-end training of SLT models.

The MuST-C dataset (Di Gangi et al. 2019b) is a more recent effort to compile a massively multilingual dataset from TED data, spanning 10 languages (English aligned with Czech, Dutch, French, German, Italian, Portuguese, Romanian, Russian, and Spanish translations), using more reliable timestamps for alignments than the IWSLT'18 dataset using a rigorous alignment process. The dataset contains a large amount of data for each target language, corresponding to a selection of English speech ranging from 385 h for Portuguese to 504 h for Spanish.

LibriSpeech The original LibriSpeech corpus (Panayotov et al. 2015) is a collection of 982 h of read English speech derived from audiobooks from the LibriVox project, automatically aligned to their text versions available from the Gutenberg project for the purpose of training ASR systems. Kocabiyikoglu et al. (2018) augments this dataset for use in training SLT systems by aligning chapters from LibriSpeech with their French equivalents through a multi-stage automatic alignment process. The result is a parallel corpus of spoken English to textual French, consisting of 1408 chapters from 247 books, totalling 236 h of English speech and approximately 131k text segments.

MSLT The Microsoft Speech Language Translation (MSLT) corpus (Federmann and Lewis 2016) consists of bilingual conversations on Skype, together with transcriptions and translations. For each bilingual speaker pair, there is one conversation where the first speaker uses their native language and the second speaker uses English, and another with the roles reversed. The first phase transcripts were annotated for disfluencies, noise and code switching. In a second phase, the transcripts were cleaned, punctuated and recased. The corpus contains 7 to 8 h of speech for each of English, German, and French. The English speech was translated to both German and French, while German and French speech was translated only to English. Federmann and Lewis (2017) repeat the process with Japanese and Chinese, expanding the dataset with 10 h of Japanese and 4.5 h of Chinese speech.

Fisher & Callhome Post et al. (2013) extends the Fisher[6] and Callhome[7] datasets of transcribed Spanish speech with English translations, developed by the Linguistic Data Consortium. The original Fisher dataset contains about 160 h of telephone conversations in various dialects of Spanish between strangers, while the Callhome dataset contains 20 h of telephone conversations between relatives and

[6] Speech: https://catalog.ldc.upenn.edu/LDC2010S01, Transcripts: https://catalog.ldc.upenn.edu/LDC2010T04.

[7] Speech: https://catalog.ldc.upenn.edu/LDC96S35, Transcripts: https://catalog.ldc.upenn.edu/LDC2010T04.

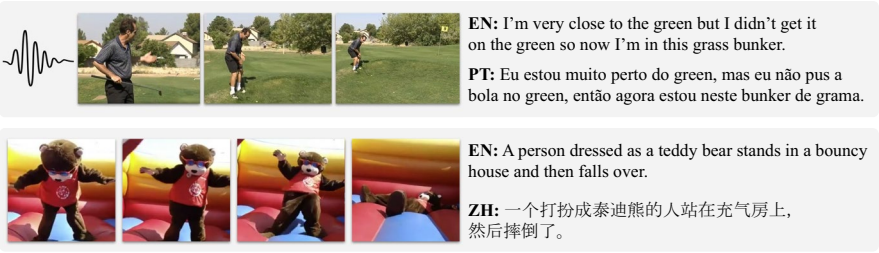

Fig. 3 Examples from How2 video subtitles (top) and VaTeX video descriptions (bottom), retrieved and adapted from Sanabria et al. (2018) and Wang et al. (2019b), respectively

friends. The translations were collected from non-professional translators on the crowdsourcing platform Mechanical Turk. Fisher & Callhome is distributed with predesignated development and test splits, a part of which contains four reference translations for each transcript segment. The data in the corpus also includes ground truth ASR lattices that facilitate the training of strong specialized ASR models, allowing pipeline SLT studies to focus on the MT component. As the largest SLT corpus available at the time of its release, the Fisher & Callhome corpus has been widely used, and remains relevant for SLT today.

MaSS The *M*ultilingu*a*l corpus of *S*entence-aligned **S**poken utterances (MaSS) (Boito et al. 2019) is a multilingual corpus of read bible verses and chapter names from the New Testament. It is fully multi-parallel across 8 languages (Basque, English, Finnish, French, Hungarian, Romanian, Russian, and Spanish), comprising 56 language pairs in total. The multi-parallel content makes this dataset suitable for training SLT systems for language pairs not including English, unlike other multilingual datasets such as MuST-C. The data is aligned on the level of verses, rather than sentences. In rare cases, the audio for some verses is missing for some languages. MaSS contains a total of 8,130 eight-way parallel text segments, corresponding to anywhere between 18.5 and 23 h of speech per language.

4.3 Video-guided translation datasets

The QED corpus The QCRI Educational Domain (QED) Corpus (Guzman et al. 2013; Abdelali et al. 2014), formerly known as the QCRI AMARA Corpus, is a large-scale collection of multilingual video subtitles. The corpus contains publicly available videos scraped from massive online open courses (MOOCs), spanning a wide range of subjects. The latest v1.4 release comprises a selection of 23.1k videos in 20 languages (Arabic, Bulgarian, Traditional and Simplified Chinese, Czech, Danish, Dutch, English, French, German, Hindi, Italian, Japanese, Korean, Polish, Portuguese, Russian, Spanish, Thai, and Turkish), subtitled in the collaborative Amara environment[8] (Jansen et al. 2014) by volunteers. A sizeable portion of the videos has parallel subtitles in multiple languages, varying in size from 8k segments

[8] https://amara.org/.

(for Hindi–Russian) to 335k segments (for English–Spanish). Of these, about 75% of the parallel segments align perfectly in the original data, while the rest were automatically aligned using heuristic algorithms. An alpha v2.0 of the QED corpus is currently underway, scheduled to appear in the OPUS repository (Tiedemann 2012), containing a large amount of (noisy) re-crawled subtitles.

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 around 2000 h 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-guided 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. For the purpose of multimodal translation, a 300-h subset of How2 that covers 22 different topics is available with crowdsourced Portuguese translations. This dataset has also recently been used for multimodal machine translation (Sanabria et al. 2018; Wu et al. 2019b). An example from this dataset can be seen in Fig. 3.

The VaTeX *dataset* The *V*ideo *a*nd *TeX*t (VaTeX) dataset (Wang et al. 2019b) is a bilingual collection of video descriptions, built on a subset of 41,250 video clips from the action classification benchmark DeepMind Kinetics-600 (Kay et al. 2017; Carreira et al. 2018). Each clip runs for about 10 seconds, showing one of 600 human activities. VaTeX adds 10 Chinese and 10 English crowdsourced captions describing each video, half of which are independent annotations, and the other half Chinese–English parallel sentences. With low-approval samples removed, the released version of the dataset contains 206,345 translation pairs in total. VaTeX is intended to facilitate research in multilingual video captioning and video-guided machine translation, and the authors keep a blind test set reserved for use in evaluation campaigns. The rest of the dataset is divided into training (26k videos), validation (3k videos), and public test splits (6k videos). The training and validation splits also have public action labels. An example from VaTeX is shown in Fig. 3.

5 Models and approaches

This section discusses some of the prominent models and approaches for multimodal MT tasks introduced in Sect. 2. In particular, we focus on IGT and SLT, and present our overview of the state-of-the-art models for either task.

For some multimodal MT tasks, the traditional approach is to put together a pipeline to divide the task into several sub-tasks, and cascade different modules to handle each of them. For instance, in the case of spoken language translation (SLT), this pipeline would first convert the input speech into text by an automatic speech recognition module (modality conversion), 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. Pipeline systems are less vulnerable to training data insufficiency

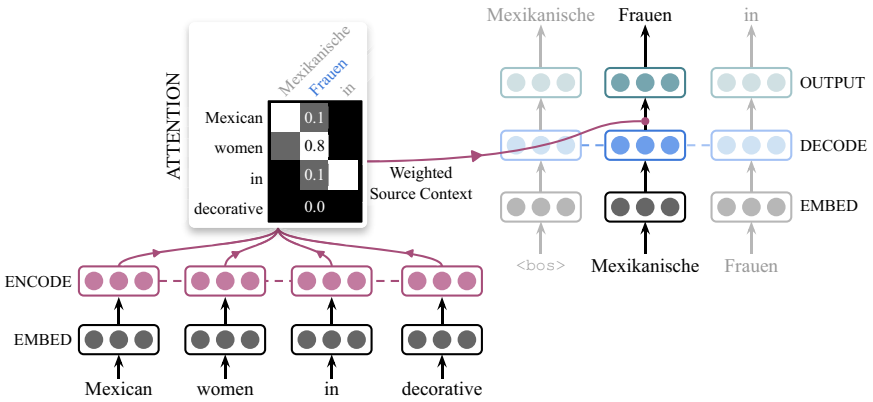

Fig. 4 A simplified view of encoder-decoder architecture with attention: an English sentence is first encoded into a latent space from which an attentive decoder sequentially generates the German sentence. The dashed *recurrent* connections are replaced by self-attention in fully-connected architectures such as transformers (Vaswani et al. 2017)

compared to data-driven end-to-end systems, since each component can be pre-trained in isolation on abundant sub-task resources. However, they carry the risk of error propagation between stages and ignore cross-modal transfer of implicit semantics. As an example for the latter, consider two languages which emphasise words via prosody and specific word order, respectively. Translating the transcript would make it impossible to reflect the word order in the target sentence as the semantic correspondence would be lost at transcription stage. Nevertheless, both pipeline and end-to-end approaches rely heavily on the sequence-to-sequence learning framework on account of its flexibility and good performance across tasks. In the following, we describe this framework in detail.

General purpose sequence-to-sequence learning is inspired by the pioneering works in unimodal neural machine translation (NMT). The state of the art in unimodal MT has been dominated by statistical machine translation (SMT) methodologies (Koehn 2009) for at least two decades, until the field drastically moved towards NMT techniques around 2015. Inspired by the successful use of deep neural networks in language modelling (Bengio et al. 2003; Mikolov et al. 2010) and automatic speech recognition (Graves et al. 2013), there has been a plethora of NMT studies featuring different neural architectures and learning methods. These architectures often rely on continuous word vector representations to encode various kinds of linguistic information in a common vector space, thereby eliminating the need for hand-crafted linguistic features. One of the first NMT studies by Kalchbrenner and Blunsom (2013) combined recurrent language modelling (Mikolov et al. 2010) and convolutional neural networks (CNN) to improve the performance of SMT systems through rescoring. Later on, the application of recurrent architectures, 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, eventually leading to the fundamental encoder-decoder architecture (Cho et al. 2014; Sutskever et al. 2014). Although the latter

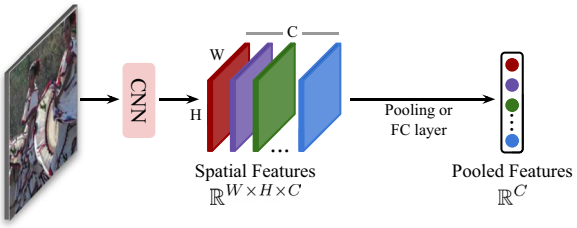

Fig. 5 An overview of two common types of visual featuers extracted from CNNs

RNN variants mitigated the problem of *vanishing gradients* through the use of gated computations, the idea of compressing a variable-length sequence into a fixed capacity vector turned out to be too restrictive for learning long-range dependencies such as grammatical agreement in very long sentences. The attention mechanism (Bahdanau et al. 2015) addressed this issue by simultaneously learning to align translation units and to translate, supplying a context window with the relevant input units at each decoding step, i.e. for each generated word in the target language (Fig. 4).

The performance of the NMT systems that followed came close to, and soon surpassed, that of the state-of-the-art SMT systems. Successful non-recurrent alternatives have also been proposed, such as convolutional encoders and decoders with attention (Gehring et al. 2017), and the fully-connected deep transformers which employ the idea of self-attention in addition to the default cross-attention mechanism (Vaswani et al. 2017). The main motivation behind these is to allow for efficient parallel training across multiple processing units, and to prevent learning difficulties such as vanishing gradients.

Lastly, we would like to mention some major open-source toolkits which contribute vastly to the state of the art in machine translation by allowing fast prototyping of new approaches as well as the extension of existing ones to new tasks and paradigms: Moses (Koehn et al. 2007) for SMT, and FairSeq (Ott et al. 2019), Joey-NMT (Kreutzer et al. 2019), Lingvo (Shen et al. 2019), Marian (Junczys-Dowmunt et al. 2018), Nematus (Sennrich et al. 2017), NeuralMonkey (Helcl et al. 2018a), nmtpytorch (Caglayan et al. 2017b), OpenNMT (Klein et al. 2017), Sockeye (Hieber et al. 2017) and Tensor2Tensor (Vaswani et al. 2018) for NMT.

5.1 Image-guided translation

In this section, we present the state-of-the-art models for the image-guided translation (IGT) task. We first discuss the visual feature extraction process, continue with reviews of the two main end-to-end neural approaches, and finally briefly cover retrieval and reranking methods.

5.1.1 Feature extraction

The practice of embedding translation units into continuous vector representations has become a standard in NMT. For compatibility with various NMT architectures,

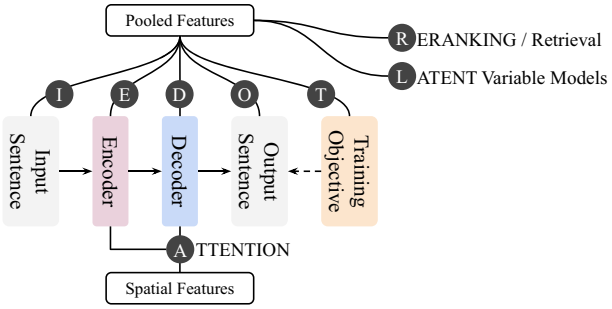

Fig. 6 A broad visualisation of the state of the art in image-guided translation

multimodal MT systems need to embed input data from other modalities, whether alongside or in place of the text, in a similar fashion. For visual information, the current best practice is to use a convolutional neural network (CNN) with multiple layers stacked on top of each other, train the system for a relevant computer vision task, and use the latent features extracted from the trained network as visual representations. Although these visual encoders are highly optimised for the underlying vision tasks such as large-scale image classification or object detection (Russakovsky et al. 2015), it has been shown that the learned representations transfer very well into vision-to-language tasks such as image captioning (Vinyals et al. 2015; Xu et al. 2015). Therefore, the majority of IGT approaches rely on features extracted from state-of-the-art CNNs (Simonyan and Zisserman 2015; Ioffe and Szegedy 2015; He et al. 2016) trained for the ImageNet (Deng et al. 2009) image classification task, where the output of the network is a distribution over 1000 object categories. These features usually come in two flavors (Fig. 5): (i) spatial features which are feature maps $\mathbf{V} \in \mathbb{R}^{W \times H \times C}$ extracted from specific convolutional layers, and (ii) a pooled feature vector $v \in \mathbb{R}^{C}$ which is the outcome of applying a projection or pooling layer on top of spatial features. The main difference between these features is that the former preserves spatial information, while the latter is a spatially unaware, compact vectorial representation. An even more compact representation is to use the posterior class probabilities ($v \in \mathbb{R}^{K}$) extracted from the output layer of a pretrained CNN, with K denoting the size of the task-specific label set (for ImageNet, K is 1000). Finally, it is also possible to obtain *a set of pooled feature vectors* (or *local* features) from salient regions of a given image, with regions predicted by object detection CNNs (Girshick et al. 2014).

5.1.2 Sequence-to-sequence grounding with pooled features

The simplest and the most intuitive way of visually conditioning a sequence-to-sequence model is to employ pooled features in a way that they will interact with various components of the architecture. These approaches are mostly inspired by the early works in neural image captioning (Kiros et al. 2014; Mao et al. 2015; Vinyals et al. 2015), and are categorised in Fig. 6 with respect to their entry points.

The very first attempt for neural image-guided translation comes from Elliott et al. (2015), where they formulate the problem as a semantic transfer from a source language model to a target language model, within an encoder-decoder framework without attention. They propose to initialise the hidden state(s) of the source language model (LM), the target LM, or both, using pretrained VGG features (Simonyan and Zisserman 2015). Later initialisation variants are applied to attentive NMTs: Calixto et al. (2016) and Libovický et al. (2016) experiment with recurrent decoder initialisation while Ma et al. (2017) initialise both the encoder and the decoder, with features from a state-of-the-art ResNet (He et al. 2016). Madhyastha et al. (2017) explore the expressiveness of the posterior probability vector as a visual representation, rather than the pooled features from the penultimate layer of a CNN.

Huang et al. (2016) take a different approach and enrich the source sentence representation with visual information by projecting the feature vector into the source language embedding space and then adding it to the beginning or the end of the embedding sequence. This allows the attention mechanism in the decoder to attend to a mixed-modality source representation instead of a purely textual one. Instead of the conventional ImageNet-extracted features, they make use of *local features* from RCNN (Girshick et al. 2014) to represent explicit visual semantics related to salient objects. In another model referred to as *Parallel-RCNN*, they build five different source embedding sequences, each being enriched with a visual feature vector extracted from a different salient region of the image. A shared LSTM encodes these five sequences and average pools them to end up with the final source representation.

Calixto and Liu (2017a) revisit the idea of source enrichment to extend it by simultaneously appending and prepending the projected visual features to the embedding sequence; and combining it with encoder and/or decoder initialisation. Caglayan et al. (2017a) explore different source and target interaction methods such as the element-wise multiplication between the visual features and the source/target word embeddings. Delbrouck and Dupont (2018) add another recurrent layer within the decoder in their *DeepGRU* model, conditioned on the visual features and the bottom layer hidden state. Both recurrent layers simultaneously decide on the output probability distribution by additively fusioning their respective unnormalised logits.

As for transformer-based architectures, Grönroos et al. (2018) revisit the source enrichment by adding the visual feature vector to the beginning of the embedding sequence (Huang et al. 2016). They also experiment with modulating the output probability distribution through a time-dependent visual decoder gate. More interestingly, they explore different pooled visual representations such as scene–type associations (Xiao et al. 2010), action–type associations (Yao et al. 2011), and object features from Mask R-CNN (He et al. 2017).

Multi-task learning Training an end-to-end neural model to perform multiple tasks at once can improve the model's task-specific performance by forcing it to exploit commonalities across the tasks involved (Caruana 1997; Dong et al. 2015; Luong et al. 2016). The *Imagination* architecture, initially proposed by Elliott and Kádár (2017) and later integrated into transformer-based NMTs by Helcl et al. (2018b), attempts to leverage the benefits of multi-tasking by proposing a one-to-many framework which shares the sentence encoder between the translation task and an auxiliary visual reconstruction task. Besides the usual cross-entropy translation

objective, the model weights are also optimised through a margin-based loss which minimises the distance between the ground-truth visual feature vector and the one predicted from the sentence encoding. The visual reconstruction task encourages the sentence encoding to be visually grounded, i.e. it needs to be suitable for both translation and predicting the visual features of the image associated with the sentence encoding. The visual features are only used at training time and are not needed when generating translations. Zhou et al. (2018) further extends the *Imagination* network by incorporating an attention[9] over source sentence encodings, with the query vector being the visual features. In this approach, the auxiliary margin-based loss is modified so that the output of the attention layer is considered a reconstruction of the pooled feature vector.

Other approaches All grounding approaches covered so far rely on the maximum-likelihood estimation (MLE) principle during training, which does not necessarily relate to maximising the translation performance as measured by proxy metrics such as BLEU. In order to alleviate this discrepancy, Zheng et al. (2018) apply a reinforcement learning based fine-tuning where model parameters are updated based on a reward signal measured by sentence BLEU scores. In terms of visual integration, they simply initialise the decoder with pooled features. Toyama et al. (2016), Calixto et al. (2019) and Delbrouck and Dupont (2019) cast the problem as a latent variable model and resort to techniques such as variational inference and generative adversarial networks (GANs). Finally, Nakayama and Nishida (2017) approach the problem from a *zero-resource* perspective: they encode {source caption, image} pairs into a multimodal vectorial space using a max-margin loss. In a second step, they train the decoder using {target caption, image} pairs. Specifically, they do a forward-pass with the image as input and obtain the multimodal embedding, from which the recurrent decoder is trained to generate the target caption as usual. The image encoder is a pretrained VGG CNN. The *zero-resource* aspect comes from the fact that the sets of pairs do not overlap i.e. the approach does not require parallel IGT corpus. Chen et al. (2018) tackle the same problem from a multi-agent communication game perspective where a translator and a captioner agent cooperatively engage with each other to maximise task-specific rewards.

5.1.3 Visual attention

Inspired by the previous success of visual attention in image captioning (Xu et al. 2015), attentive approaches explore how to efficiently integrate a visual attention (approach A in Fig. 6) over the spatial features, alongside the language attention in NMTs. The most interesting research questions about visual attention are as follows: where to apply the visual attention, what kind of parameter sharing should be preferred and, how to fuse the output of language and visual attention layers. Caglayan et al. (2016a) and Calixto et al. (2016) are the first works to tackle these questions, through a visual attention which uses the hidden state of the decoder as *query* into

[9] It should be noted that the attention here is over the source language encodings, and hence not a visual/spatial attention.

the set of $W \times H$ spatial features. Their implementation is quite similar to the language attention, which results in two modality-specific contexts that should be fused before the output layer of the network. One notable difference is that Caglayan et al. (2016a) experiment with a single multimodal attention layer *shared* across modalities while Calixto et al. (2016) keep the attention layers separate. Later on, Caglayan et al. (2016b) evaluate both *shared* and *separate* attentions with *additive* and *concatenative* fusion, and discover that proper feature normalisation is crucial for their recurrent approaches (Caglayan et al. 2018). Delbrouck and Dupont (2017b) propose a different fusion operation based on *compact bilinear pooling* (Fukui et al. 2016), to efficiently realise the computationally expensive outer product. Unlike *additive* and *concatenative* fusions, outer product ensures that each dimension of the language context vector interacts with each dimension of the visual context vector and vice-versa. Follow-up studies extend the *decoder-based visual attention* approach in different ways: Calixto et al. (2017) reimplement the *gating mechanism* (Xu et al. 2015) to rescale the magnitude of the visual information before the fusion, while Libovický and Helcl (2017) introduce the *hierarchical attention* which replaces the *concatenative* fusion with a new attention layer that dynamically weighs the modality-specific context vectors. Finally, Arslan et al. (2018) and Libovický et al. (2018) introduce the same idea into the Transformer-based (Vaswani et al. 2017) architectures. Besides revisiting the *hierarchical* attention, Libovický et al. (2018) also introduce *parallel* and *serial* variants. The former is quite similar to Arslan et al. (2018) and simply performs *additive* fusion while the latter first applies the language attention, which produces the *query* vector for the subsequent visual attention. Ive et al. (2019) extend Libovický et al. (2018) to add a 2-stage decoding process where visual features are only used in the second stage, through a visual cross-modal attention. They also experiment with another model where the attention is applied over the embeddings of object labels detected from the images.

In contrast to the *decoder-based visual attention*, encoder-based approaches are relatively less explored. To that end, Delbrouck and Dupont (2017a) propose conditional batch normalisation, a technique to modulate the batch normalisation layer (Ioffe and Szegedy 2015) of ResNet. Specifically, they condition the mean and the variance of the batch normalisation layer on the source sentence representation for informed feature extraction. In the same work, Delbrouck and Dupont (2017a) also propose to apply an *early visual attention* inside the encoder, to yield inherently multimodal source encodings, on top of which the usual language attention would be applied by the decoder.

5.1.4 Reranking and retrieval based approaches

The most typical pipeline for MT is to obtain an *n-best* list of translation candidates from an arbitrary MT system and select the best candidate amongst them after *reranking* with respect to an aggregated score. This score is often a combination of several models that are able to quantitatively assess translation-related qualities of a candidate sentence, such as the adequacy or the fluency, for example. Each model is assigned a coefficient and an optimisation step is executed to find the best set of coefficients that maximise the translation performance on an held-out development

set (Och 2003). The challenge for the IGT task is how to incorporate the visual modality into this pipeline in order to assign a better rank to visually plausible translations. To this end, Caglayan et al. (2016a) combine a feed-forward language model (Bengio et al. 2003; Schwenk et al. 2006) and a recurrent NMT to rerank the translation candidates obtained from an SMT system. The language model is special in the sense that it is not only conditioned on *n-gram* contexts but also on the pooled visual feature vector. In contrast, Shah et al. (2016) conjecture that the posterior class probabilities may be more expressive than a pooled representation for reranking, and treat each probability v_i as an independent score for which a coefficient is learned. In a recent work, Lala et al. (2018) demonstrate that for the Multi30k dataset, better translations are available inside an *n-best* list obtained from a text-only NMT model, which allow up to 10 points absolute improvement in METEOR score. They propose the multimodal lexical translation (MLT) model where they rerank the n-best list with scores assigned by a multimodal word sense disambiguation system based on pooled features.

Another line of work considers the task as a joint retrieval and reranking problem, which can be useful in overcoming data sparsity issues with small multilingual multimodal datasets. Hitschler et al. (2016) construct a multimodal/cross-lingual retrieval pipeline to *rerank* SMT translation candidates. Specifically, they use a large corpus of target {caption, image} pairs, and retrieve a set of pairs similar to the translation candidates and the associated image. The visual similarity is computed using the Euclidean distance in the pooled CNN feature space. The initial translation candidates are then reranked with respect to their—inverse document frequency based—relevance to the retrieved captions. Zhang et al. (2017) also employ a combined framework of retrieval and reranking. For a given {caption, image} pair, they first retrieve a set of similar training images. The target captions associated with these images are considered as candidate translations. They learn a multimodal word alignment between source and candidate words and select the most probable target word for each source word. An *n-best* list from their SMT is reranked using a bi-directional NMT trained on the aforementioned source/target word sequences. Finally, Duselis et al. (2017) and Gwinnup et al. (2018) propose a pure retrieval system without any reranking involved. For a given image, they first obtain a set of candidate captions from a pretrained image captioning system. Two distinct neural encoders are used to encode the source and the candidate captions, respectively. A mapping is then learned from the hidden space of the source encoder to the target one, allowing the retrieval of the candidate caption which minimises the distance with respect to the source caption representation.

5.1.5 Comparison of approaches

Table 2 presents BLEU and METEOR scores on the English→German `test2016` set of Multi30k dataset, as this is the test set that most studies report against. When possible, we annotate each score with the associated gain or loss with respect to the underlying unimodal MT baseline reported in the respective papers. The results concentrate around *constrained* systems, which only allow the use of parallel Multi30k corpus during training. A few studies experiment with

Table 2 Automatic scores of state-of-the-art IGT methods on Multi30k English→German `test2016`: the table is clustered (and sorted by METEOR) across years for constrained systems, followed by unconstrained ones

	BLEU ↑	METEOR ↑	Type	Description	Arch.
Elliott et al. (2015)†	9.7 (N/A)	24.7 (N/A)	E,D	Conditional LMs	RNN
Caglayan et al. (2016a)†	29.3 (↓4.6)	48.5 (↓4.3)	A	Shared attention	RNN
Calixto et al. (2016)†	28.8 (N/A)	49.6 (N/A)	A	Separate attention	RNN
Huang et al. (2016)	36.8 (↑2.0)	54.4 (↑2.3)	I*	Parallel RCNN-LSTMs	RNN
Hitschler et al. (2016)†	34.3 (N/A)	56.0 (N/A)	R	Retrieval + reranking	SMT
Toyama et al. (2016)	36.5 (↑1.6)	56.0 (↑0.7)	L	Variational	RNN
Shah et al. (2016)†	34.8 (↑0.2)	56.7 (↑0.1)	R	Visual reranking	SMT
Caglayan et al. (2016a)†	36.2 (– 0.0)	57.5 (↑0.1)	R	Visual reranking	SMT
Helcl and Libovický (2017)	31.9 (↓2.7)	49.4 (↓2.3)	A	Hierarchical attention	RNN
Calixto and Liu (2017a)	36.9 (↑3.2)	54.3 (↑2.0)	I	Input prepend & append	RNN
Calixto et al. (2017)	36.5 (↑2.8)	55.0 (↑2.7)	A	Gated attention	RNN
Calixto and Liu (2017a)	37.3 (↑3.6)	55.1 (↑2.8)	D	Decoder init.	RNN
Elliott and Kádár (2017)	36.8 (↑1.3)	55.8 (↑1.8)	T	Imagination	RNN
Caglayan et al. (2017a)	38.2 (↑0.1)	57.6 (↑0.3)	E,D	Encoder decoder init.	RNN
	37.8 (↓0.3)	57.7 (↑0.4)	O	Multiplicative interaction	RNN
Delbrouck and Dupont (2017a)	40.5 (N/A)	57.9 (N/A)	A	Encoder attention + CBN	RNN
Arslan et al. (2018)	41.0 (↑2.4)	53.5 (↓1.5)	A	Parallel attention	Transformer
Calixto et al. (2019)	37.7 (↑2.7)	56.0 (↑1.1)	L	Variational	RNN
Helcl et al. (2018b)	38.8 (↑0.7)	56.4 (↑0.2)	T	Imagination	Transformer
Libovický et al. (2018)	38.5 (↑0.2)	56.5 (↓0.2)	A	Hierarchical attention	Transformer
	38.6 (↑0.3)	57.4 (↑0.7)	A	Parallel attention	Transformer
Ive et al. (2019)	38.0 (↑0.1)	55.6 (↓0.3)	D*	2-stage decoder + label embs.	Transformer
Libovický (2019)	37.6 (↑0.9)	56.0 (↑0.9)	A	Hierarchical attention	RNN
Caglayan (2019)	39.0 (↑0.1)	58.5 (↑0.1)	E,D	Encoder decoder init.	RNN
	39.4 (↑0.5)	58.7 (↑0.3)	A	Separate attention + L_2 Norm.	RNN
Unconstrained ensembles					
Helcl et al. (2018b)	42.6 (↑2.2)	59.4 (↑0.4)	T	Imagination	Transformer
Grönroos et al. (2018)	45.5 (– 0.0)	(N/A)	I*	Input prepend	Transformer

Systems marked with † are re-evaluated with tokenised sentences, * denotes the use of visual features other than ImageNet CNNs. The gains and losses are with respect to the MT baselines reported in the papers. The types refer to Fig. 6

using external resources (Calixto et al. 2017; Helcl and Libovický 2017; Elliott and Kádár 2017; Grönroos et al. 2018) for pretraining the MT system and then fine-tuning it on Multi30k, or directly training the system on the combination of Multi30k and the external resource. Two such *unconstrained* systems are also reported.

At a first glance, the automatic results reveal that (i) initially, neural systems were not able to surpass the SMT systems, (ii) the use of external resources is beneficial to boost the underlying baseline performance, which further manifests itself as a boost in the multimodal scores and (iii) careful tuning allows RNN-based models to reach and even surpass Transformer-based models. From a multimodal perspective, the results are not very conclusive as there does not seem to be a single architecture, feature type or integration type that brings consistent improvements. Elliott (2018) attempted to answer the question of how efficiently state-of-the-art models were integrating information from the visual modality and concluded that when models were adversarially challenged with wrong images at test time, the quality of the produced translations was not that much affected as one would expect. Later on, Caglayan et al. (2019) showed how these seemingly insensitive architectures start to significantly rely on the visual modality, once words were systematically removed from source sentences during training and test. We believe that this latter finding may also be connected to the fact that better baselines benefit less from the visual modality (Table 2) i.e. sub-optimal architectures may leverage more from the visual information when compared to well trained NMT models. In fact, even the choice of vocabulary size may simulate *systematic word removal*, if a significant portion of the source vocabulary are mapped to unknown tokens. The same experimental pipeline of Caglayan et al. (2019) also paved the way for assessing the particular strengths of some of the covered IGT approaches and showed that, the use of spatial features through visual attention is superior than initialising the encoders and the decoders using pooled features.

Lastly, if we take a look at the human evaluation rankings conducted throughout the WMT shared tasks, we see that the top three ranks for English→German and English→French are occupied by two unconstrained ensembles (Grönroos et al. 2018; Helcl et al. 2018b), the MLT Reranking (Lala et al. 2018) and the DeepGRU (Delbrouck and Dupont 2018) systems in 2018. In 2017, the multiplicative interaction (Caglayan et al. 2017a), unimodal NMT reranking (Zhang et al. 2017), unconstrained *Imagination* (Elliott and Kádár 2017), encoder enrichment (Calixto and Liu 2017a) and hierarchical attention (Helcl and Libovický 2017) were ranked as top three, again for both language pairs.

5.2 Spoken language translation

In spoken language translation, the non-text modality is the source language audio, which is translated into target language text. While source language transcripts may be available for training, at translation time the speech is typically the only input modality. We begin this section with a brief introduction to speech-specific feature extraction (Sect. 5.2.1). Section 5.2.2 reviews the current state of the art for the traditional pipeline methods and finally, Sect. 5.2.3 covers the end-to-end methods which saw a rapid development in recent years.

5.2.1 Feature extraction

Even though many deep learning applications use raw input data, it is still common to use somewhat engineered features in speech applications. The relevant information in the speech signal is in the temporal variation of frequency content, and therefore a *spectrogram* representation is computed. It discards the phase information of the signal and captures signal activity at different frequencies in short, consecutive, and typically overlapping frames. The frame length trades off time and frequency precision: longer frames capture finer *spectral* (i.e. frequency) detail, but also describe a longer segment of time, which can be problematic as certain speech events (e.g. the stop consonants p, t) can have a very short duration.

Next, a *Mel-scale filterbank* is applied to each frame, and the logarithm of each filter's output is computed. This leads to *log Mel-filterbank* features. The filterbank operation reduces the number of dimensions. However, these operations are also perceptually motivated: the filterbank by the masking of frequencies close to each other in the ear, the Mel-scale as it relates frequency to perceived pitch, and the logarithm by the relation of perceived loudness to signal activity (Pulkki and Karjalainen 2015).

Continued efforts in learning deep representations from raw samples exist, with some success (Sainath et al. 2015). However, log Mel-filterbank vectors as input to deep neural network models (Mohamed et al. 2012) are the standard choice. Additional, more complex features may be used to aid robustness to speaker variability (Saon et al. 2013) or recognition in tonal languages (Ghahremani et al. 2014).

5.2.2 State of the art in pipeline methods

Pipeline approaches in SLT chain together separate ASR and MT modules, and these naturally follow progress in their respective fields. A popular ASR system architecture is an HMM-DNN hybrid acoustic model (Yu and Li 2017), followed by an n-gram language model in the first decoding pass, and a neural language model for rescoring. This type of HMM-based ASR is essentially *pipeline ASR*. In addition to pipeline ASR, *end-to-end ASR* methods have recently gained popularity. Particularly, encoder-decoder architectures with attention have been successful, although on standard publicly available datasets HMM-based models still narrowly outperform end-to-end ones (Lüscher et al. 2019). Chiu et al. (2018) show that encoder-decoder with attention ASR can outperform HMM-based models on an very large (12,500 h) proprietary dataset. Another common end-to-end ASR method is Connectionist Temporal Classification (CTC) (e.g. Li et al. (2019)).

Wang et al. (2018c) and Liu et al. (2018) place first and second, respectively, in the IWSLT 2018 evaluation campaign. Both apply similar pipeline architectures: a system combination of multiple different HMM-DNN acoustic models and LSTM rescoring for ASR, followed by a system combination of multiple Transformer NMT models for translation. Liu et al. (2018) additionally use an encoder-decoder with attention ASR to improve the system combination ASR results, although individually the end-to-end model is clearly outperformed by the HMM-DNN models. Wang et al. (2018c) use an additional target-to-source

NMT system for rescoring to improve adequacy. The systems also differ in interfacing strategies between ASR and MT.

In the latest IWSLT evaluation campaign in 2019, end-to-end SLT models were encouraged. However, the best performance was still achieved with a pipeline SLT approach, where Pham et al. (2019) use end-to-end ASR and a Transformer NMT model. In the ASR module, an LSTM-based approach outperforms a Transformer model, though combining both in an ensemble proved beneficial. Weiss et al. (2017) and Pino et al. (2019) also report competitive results using end-to-end ASR, with Pino et al. (2019) surpassing the state-of-the-art in SLT. End-to-end ASR has attracted attention in SLT, because it allows for parameter transfer in end-to-end SLT (e.g. Bérard et al. (2018), and Fig. 8).

Challenges in pipeline SLT Research in pipeline SLT has specifically focused on the interface between ASR and MT. There is a clear mismatch between MT training data and ASR output, caused by the ASR noise characteristics (i.e. transcription errors), and the ASR output dissimilarity with respect to the written text due to lack of capitalisation and punctuation, and the disfluencies (e.g. repetitions and hesitations), which naturally occur in speech. Ruiz and Federico (2014, 2015), Ruiz et al. (2017) quantify the effect of ASR errors on MT. In a linear mixed-effects model, the amount of WER added on top of gold standard transcripts has a direct effect on TER increase. The results do not vary over different ASR systems. Minor localised ASR errors can result in longer distance errors or duplication of content words in NMT. Homophonic substitution error spans (e.g. *anatomy* → *and that to me*) are shown to account for a significant portion of ASR errors and to have a large impact on translation quality. With regards to noise robustness, it is noted that the utterances which were best translated by phrase-based MT, had higher average WER than utterances which were best translated by NMT. In general, NMT has been established as particularly sensitive to noisy inputs (Belinkov and Bisk 2018; Cheng et al. 2018).

One approach to address the mismatch is training the MT system on noisy, ASR-like input. Peitz et al. (2012) use an additional phrase-table trained on ASR-outputs on the SLT corpus. Tsvetkov et al. (2014) augment a phrase-table with plausible ASR misrecognitions. These errors are synthesised by mapping each phrase to phones via a pronunciation dictionary, and randomly applying heuristic phone-level edit operations.

Sperber et al. (2017b) first train an NMT system on reference transcripts, and then fine-tune on noisy transcripts. The noise is sampled from a uniform distribution over insertions, deletions or substitutions, with optional unigram weighting for the substitutions and insertions. Additionally, a deletion-only noise is used. Smaller amounts of noise are shown to improve SLT results, but increasing noise levels to actual test-time ASR levels (rather high, at 40%) only degrades performance. Increased noise is noted to produce shorter outputs, which in turn are punished by the BLEU brevity penalty. A precision-recall tradeoff is observed: the system could either drop uncertain inputs (better precision) or try to guess translations (better recall). Fine-tuning with deletion-only noise biases the system to produce longer outputs, which is shown to counteract the effect of noisy inputs producing shorter outputs. Pham et al. (2019) use the data augmentation method SwitchOut (Wang et al. 2018b), to make

Table 3 SLT search in mathematical formulation, for translation y, source language transcript z, source language speech x, and set of all possible transcripts Z

End-to-end search	$\underset{y}{\text{argmax}}\, P(y\|x)$
General pipeline search	$\underset{y}{\text{argmax}} \sum_{z \in Z'(x)} P(y\|z)P(z\|x)$
Pure serial pipeline	$Z'(x) = \left\{ \underset{z}{\text{argmax}}\, P(z\|x) \right\}$
Loosely coupled pipeline	$Z'(x) \subset Z$
Tightly coupled pipeline	$Z'(x) = Z$

their NMT models more robust to ASR errors. During training, SwitchOut randomly replaces words in both the source and the target sentences.

Another approach to cope with the mismatch is to transform the ASR-output into written text. Wang et al. (2018c) apply a Transformer-based punctuation restoration and heuristic rules which remove disfluencies and transform written out numbers and quantities into numerals. Liu et al. (2018) experiment with NMT-based transformations in both directions: producing ASR-like text from written text for training the translation system, or producing written text from ASR-like text as a test-time bridge between ASR and translation. Transforming the MT training data into an ASR-like format consistently outperforms inverse normalization of ASR-output, though both are beneficial in the final system combination.

Long audio streams typically need to be segmented into manageable length pieces using voice activity detection (Ramirez et al. 2007), or more elaborate speaker diarisation methods (Anguera et al. 2012). These methods may not produce clean sentence boundaries. This is a clear problem in MT, as the boundaries can cut between actual sentences. Liu et al. (2018) alleviate the problem by applying an LSTM-based resegmenter after the ASR system. Pham et al. (2019) combine resegmentation, and casing and punctuation restoration into a single ASR post-processing task, and apply an NMT model.

Coupling between ASR and MT The SLT search problem is often described mathematically as shown in Table 3. Generally, pipeline search is based on the assumption that $P(y|z, x) = P(y|z)$, i.e. given the source language transcript, the translation does not depend on the speech. It is still possible to take the uncertainty of the transcription into account under this conditional independence assumption, but it rules out the use of paralinguistic cues, e.g. prosody. In *pure serial* pipeline search, first the 1-best ASR result is decoded, then only this 1-best result is translated. The hard choice in 1-best decoding is especially susceptible to error propagation. Early work in SLT found consistent improvements with *loosely coupled* search, where a rich representation carrying the ASR uncertainty, such as an N-best list or word lattice, is used in translation. *Tightly coupled* search, i.e. *joint decoding*, is also possible, although the application is limited by excessive computational demands. In tightly coupled search, the translation model would also influence which ASR hypotheses were searched further. This was done by representing both the ASR and the phrase-based MT search spaces as Weighted Finite State Transducers (WFST). (Matusov et al. 2006; Zhou 2013)

Osamura et al. (2018) implement a type of loose coupling by using the softmax posterior distribution from the ASR module as the input for NMT. Loose coupling

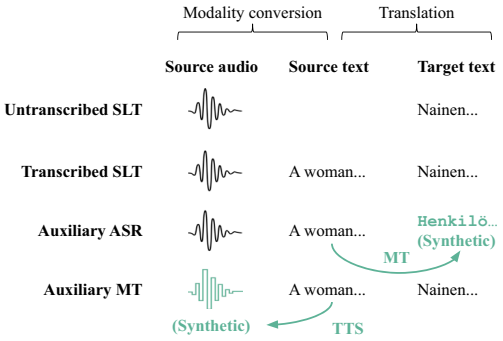

Fig. 7 Four types of data that can be used to train SLT systems. Untranscribed SLT is the minimal type of data for end-to-end systems. Adding source text transcripts completes the triple. The source text is an intermediate representation which divides the SLT mapping into a modality conversion and a translation. Two types of auxiliary data, ASR and MT data, form adjacent pairs in the triple, leaving one of the ends empty. The auxiliary data can be used as is for pretraining or multi-task learning, or it can be completed into synthetic triples using external TTS or MT systems

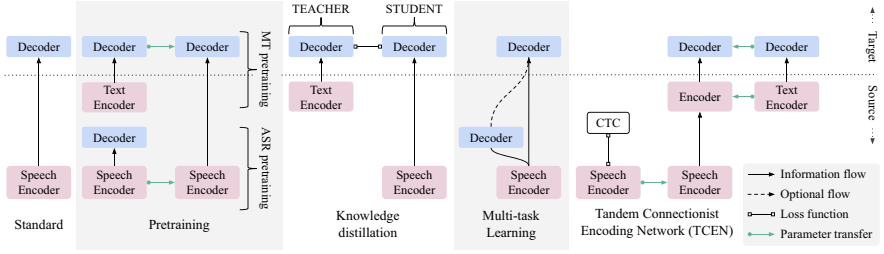

Fig. 8 Learning setups for end-to-end SLT: The *standard* framework uses untranscribed SLT data. Auxiliary data can be exploited in different ways such as by *pretraining* the encoder through ASR, *pretraining* the decoder through MT, *knowledge distillation*, or *multi-task learning*. The optional link in *multi-task learning* results in 2-step decoding. TCEN combines multiple types of pretraining

via using lattices as input in NMT is not straightforward. Sperber et al. (2017a) implement LatticeLSTM for lattice inputs in RNN-based NMT, and find that preserving the uncertainty in the ASR output is beneficial for SLT. Zhang et al. (2019) further propose a Transformer model which can use lattice inputs, and find that it outperforms both a standard Transformer and a LatticeLSTM baseline in an SLT task. However, tight coupling of NMT and ASR has not been proposed in pipeline SLT.

In addition to coupled decoding, end-to-end SLT leverages coupled training. This can avoid suboptimization; for phrase-based MT and HMM-GMM ASR, He et al. (2011) show how optimizing the ASR component purely for WER can produce worse results in SLT. He and Deng (2013) foreshadow end-to-end neural SLT systems, proposing a joint, end-to-end optimization procedure for a pipeline of HMM-GMM ASR and phrase-based MT. In the proposed approach, the ASR and MT components are first trained separately, and then the whole pipeline is jointly optimized

for sentence-level BLEU, by iteratively sampling sets of competing hypotheses from the pipeline and updating the parameters of the submodels discriminatively.

5.2.3 End-to-end spoken language translation

The first attempts to use end-to-end methods for SLT were published in 2016. This period saw experimentation with a wide variety of approaches, before research focus converged on sequence-to-sequence architectures. These early methods (Duong et al. 2016; Anastasopoulos et al. 2016; Bansal et al. 2017) were able to align source language audio to target language text, but they were not able to perform translation. The first true end-to-end SLT system is presented by Bérard et al. (2016). Still a proof-of-concept, it was trained on BTEC French→English with synthetic audio containing a small number of speakers.

Figure 7 shows the different types of training data applicable for SLT. The standard learning setup for end-to-end SLT is only able to train from untranscribed SLT data. The task is very challenging, as data of this type is scarce, and the representation gap between source audio and target text is large. The source transcript is useful as an intermediary representation, a stepping stone to divide the gap into two smaller ones: modality conversion and translation. Many learning setups (see Fig. 8), e.g. pretraining, multi-task learning, and knowledge distillation, have been applied for exploiting the source transcripts. In early experiments, no new examples are introduced for the auxiliary task(s); Only source transcript labels for the SLT examples were added. Later the same learning setups have been applied to exploit more abundant auxiliary ASR and MT data.

An important milestone towards parity with pipeline approaches was to achieve better translation quality when both the end-to-end system and the pipeline system are trained on the same SLT data. This milestone was reached by Weiss et al. (2017), training on the 163h Fisher&Callhome Spanish→English data set. As pipeline methods are naturally capable of exploiting the more abundant paired ASR and MT data, but in this case this condition was unrealistically constrained. When the constraint is lifted, pipeline methods improve to a level that is difficult or impossible to reach on small amounts of source audio-translated text data. The effective use of auxiliary data was a key insight going forward towards achieving parity with pipeline approaches.

Figure 8 shows learning setups that have been applied for exploiting source transcripts and auxiliary data. Weiss et al. (2017) use a multi-task learning procedure with ASR as the auxiliary task, training only on transcribed SLT data. In multi-task learning (Caruana 1997), multiple tasks are trained in parallel, with some network components shared between the tasks. Bérard et al. (2018) compare *pretraining* (sequential transfer) with *multi-task learning* (parallel transfer), finding very little difference between the two. In pretraining, some of the parameters from a network trained to perform an auxiliary task are used to initialise parameters in the network for the main task. The system is trained only on transcribed SLT data, with two auxiliary tasks: pretraining the encoder and decoder with ASR and textual MT respectively. Stoian et al. (2019) compare the effects of pretraining on auxiliary ASR

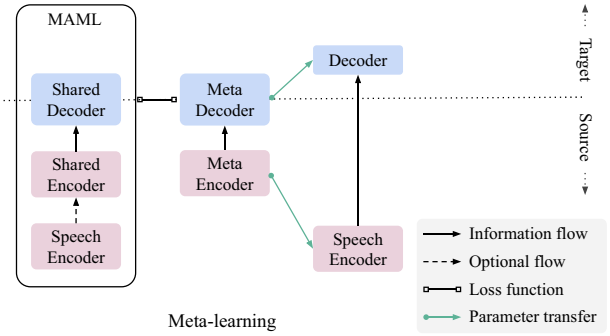

Fig. 9 *Meta-learning* expressed in the visual terms of the learning setups in Fig. 8. Meta-learning can be used to find a good initialization for standard training. The MAML algorithm explicitly learns to learn fast on the ASR and MT tasks

datasets of different languages and sizes, concluding that the WER of the ASR system is more predictive of the final translation quality than language relatedness.

Anastasopoulos and Chiang (2018) make the line between pipeline and end-to-end approaches more blurred by using a multi-task learning setup with two-step decoding. First the source transcript is decoded using the ASR decoder. A second SLT decoder attends to both the speech input and the hidden states of the ASR decoder. While the system is trained end-to-end, the two-step decoding is still necessary at translation time. The system is trained only on transcribed SLT data. Liu et al. (2019) focus on exploiting source transcripts by means of knowledge distillation. They train the student SLT model to match the output probabilities of a text-only MT teacher model, finding that knowledge distillation is better than pretraining. Inaguma et al. (2019b) also see substantial improvements from knowledge distillation when adding auxiliary textual parallel data. Wang et al. (2019a) introduce the *Tandem Connectionist Encoding Network* (TCEN), which allows neural network components to be pretrained while minimising both the number of parameters not transferred from the pretraining phase, and the mismatch of components between pretraining and finetuning. The final network consists of four components: ASR encoder, MT encoder, MT attention and MT decoder. The ASR encoder is pretrained with a Connectionist Temporal Classification objective function, which does not require a separate ASR decoder which would go to waste after pretraining. The last three parts can be pretrained with a textual MT task.

Jia et al. (2019) show that *augmenting auxiliary data* is more effective than multi-task learning. MT data is augmented with synthesised speech, while ASR data is augmented with synthetic target text by forward translation using a text-only MT system (see Fig. 7). These kinds of synthetic data augmentation are conceptually similar to the highly successful practice of using backtranslation (Sennrich et al. 2016a) to exploit monolingual data in textual MT. With both pretraining and multi-task learning, the end-to-end system slightly outperforms the pipeline. Adding synthetic data substantially outperforms the pipeline. The systems are both trained on exceptionally large proprietary corpora: ca 1300h translated

speech and 49000h transcribed speech. Controversially the system is also evaluated on a proprietary test set. The speech encoder is divided into two parts, of which only the first is pretrained on an ASR auxiliary task. The entire decoder is pretrained on the text MT task. Pino et al. (2019) evaluate several pretraining and data augmentation approaches. They use TTS to synthesise source audio for parallel text data, finding that the effect depends on the quality and quantity of the synthetic data. Using textual MT to synthesise target text from ASR data is clearly beneficial. Pretraining the speech encoder on an ASR task is useful for the lower resourced English→Romanian, but not for English→French. Pretraining on ASR is not a good substitute for using textual MT for augmenting the ASR data, but does speed up convergence of the SLT model. Using a combination of a VGG Transformer speech encoder and decoder, they very nearly reach parity with a strong pipeline system.

One more indirect way to exploit auxiliary data is meta-learning. In general, the goal of meta-learning is to use multiple related tasks to learn how to learn fast. A widely applicable meta-learning approach is the Model-agnostic Meta Learning (MAML) (Finn et al. 2017) algorithm. In normal MAML, a model is explicitly optimized to require few updates in learning a new task by backpropagating through training. Indurthi et al. (2020) adapt MAML to SLT by first meta-training the model on both ASR and MT. To deal with the multiple modalities and languages inherent, a shared vocabulary is used for the target and source sides, and CNN layers with stride larger than one are used to compress the input when in the speech modality. After meta-learning, the model is finetuned in the SLT task. Figure 9 relates this approach to the learning setups of Fig. 8 visually. Improvements are found over transfer- and multi-task learning on the MuST-C English→German and English→French tasks. When both subword units and synthetically augmented SLT data are leveraged, even a pipeline baseline is outperformed.

Bansal et al. (2019) apply *crosslingual pretraining*, by pretraining on high-resource ASR to improve low-resource SLT. They use a small Mboshi→French SLT corpus without source transcripts. As Mboshi has no official orthography, transcripts may be difficult to collect. Pretraining the speech encoder using a completely unrelated high-resource language, English, effectively allows to account for acoustic variability, such as speaker and channel differences. Di Gangi et al. (2019d) train a one-to-many multilingual system to translate from English to all 8 target languages of the MuST-C corpus, with an additional task pair for English ASR. Prepending a target language tag to the input (Johnson et al. 2017), is not effective in multilingual SLT, resulting in many acceptable translations into the wrong language. Better results are achieved with a stronger language signal using *merge*, a language-dependent shifting operation. Inaguma et al. (2019a) train multilingual models for {EN, ES} → {EN, FR, DE} SLT. They achieve better results with the multilingual models than with bilingual ones, including pipeline methods for some test sets.

Noise-based data augmentation methods have also been applied to the speech audio. Bahar et al. (2019) and Di Gangi et al. (2019a) apply spectral augmentation (SpecAugment), which randomly masks blocks of features that are consecutive in time and/or frequency.

5.2.4 End-to-end SLT architectures

There is a large variety of architectures that have been applied to end-to-end SLT, with no clear favourite having emerged. However, recent architectures all follow some type of sequence-to-sequence architectures that makes use of attention mechanisms.

Two varieties of LSTM layers have been used: standard bi-LSTM (e.g. Jia et al. 2019) and pyramidal bi-LSTM (e.g. Duong et al. 2016; Bérard et al. 2016; Bahar et al. 2019). The pyramidal construction of the encoder downsamples the long speech input sequence, making subsequent bi-LSTM layers and the attention mechanism faster and alignment easier. Following Weiss et al. (2017), Bérard et al. (2018) move away from the pyramidal bi-LSTM encoder architecture to convolution followed by bi-LSTM. The prepended convolutional layers perform the downsampling of the audio signal, making the pyramidal construction unnecessary.

Transformers have also been used in many SLT systems. Liu et al. (2019) propose an architecture in which all encoders and decoders are standard Transformer encoders and decoders respectively. Pino et al. (2019) further prepend VGG-style convolutional blocks to Transformer encoders and decoders, in order to replace the positional embedding layer of the standard Transformer architecture and to downsample the signal. A convolution is able to encode local word order information, even though a subsequent pooling layer will remove global order information. Di Gangi et al. (2019d) use a speech encoder which begins with stacks of convolutional layers interleaved with 2D self-attention (Dong et al. 2018), followed by a stack of Transformer layers. Salesky et al. (2019) revisit the network-in-network (Lin et al. 2014a) architecture to achieve downsampling: parameters are shared spatially in a similar way to CNN, but a full multi-layer perceptron network is applied to each window.

Convolutional Neural Networks are used in many SLT architectures, but only in combination with LSTM or Transformer, not in isolation. The combined CNN-LSTM architecture is popular in end-to-end ASR (Watanabe et al. 2018). The CNN is well suited for reduction of the time scale to something manageable, and modeling short range dependencies. The appended LSTM or Transformer is useful for encoding the semantic information for translation. The CNNs used in SLT are typically 2D convolutions (parameter sharing across both time and frequency). Time Delay Neural Networks (TDNN) are still popular in ASR, but have not to the best of our knowledge been used in end-to-end SLT. TDNNs can be seen as a 1D convolution, only sharing parameters across time. The VGG (Simonyan and Zisserman 2015) architecture of CNNs is used in SLT, but not ResNet (He et al. 2016).

Comparison of architectures In SLT, the choice between LSTM and Transformer architectures doesn't seem to be a settled matter: recent papers use both. Both architectures are powerful enough, when stacked into sufficiently deep networks. Pino et al. (2019) present a result in favour of the Transformer, as they only reach parity with their pipeline using Transformers, but not LSTMs. Inaguma et al. (2019b) find that Transformers consistently outperform LSTMs in their experiments. A downside of LSTM is slow training on the very long sequences encountered in speech translation. While the Transformer parallelises

to a larger extent, making training fast, it is not immune to long sequences, as the self-attention is quadratic in memory w.r.t. the length. The Transformer also lacks explicit modelling of short range dependencies, due to the self-attention learning dependencies of any range with equal difficulty. Di Gangi et al. (2019c) attempt to augment the Transformer to alleviate some of its shortcomings.

Decoding units In textual NMT, subword-level decoders have become the standard choice (Sennrich et al. 2016b). Most end-to-end SLT systems use character-level decoders. Although word level decoding is rare, Bansal et al. (2018) focus on a low-computation setting, deciding to use word-level decoding to shorten the sequence length. Some well-performing recent systems use subword units (Liu et al. 2019; Jia et al. 2019; Pino et al. 2019; Bansal et al. 2019; Indurthi et al. 2020). Wang et al. (2019a) find characters to work better than subwords in their system.

Has parity with pipeline approaches been reached? Recent results (Jia et al. 2019; Pino et al. 2019; Indurthi et al. 2020) show that on certain tasks with large enough datasets of high-quality, end-to-end systems can reach the same or even better performance than pipeline systems. On the other hand, Di Gangi et al. (2019c) show that when both pipeline systems and end-to-end models are restricted to pure SLT-data only, end-to-end methods do not lag far behind in performance. In low-resource settings, end-to-end systems do not perform as well. Furthermore, in the IWSLT 2019 evaluation campaign (Niehues et al. 2019), the pipeline system of Schneider and Waibel (2019) clearly outperforms all end-to-end submissions. Sperber et al. (2019) find that current methods do not use auxiliary data effectively enough. The amount of transcribed SLT data is critical: When the size of the data containing all three of source audio, source text and target text is sufficient, end-to-end methods outperform pipeline methods. In lower resource settings where the amount of SLT data is insufficient, pipeline methods are better.

Table 4 shows results on various SLT tasks. The English→French Augmented LibriSpeech test set is one of the most competed test sets for SLT, particularly end-to-end SLT. It shows the rapid increase in performance during the last two years, and the importance of maximally exploiting available training data. MuST-C has gained popularity since its recent release and shows a similar pattern

6 Future directions

The previous sections provide a detailed overview of resources, definitions of various kinds of multimodal MT, and the extensive work that has been devoted to develop models for the different tasks. However, multimodal MT is still in its infancy. This is especially the case for truly end-to-end models, which have only appeared in recent years. Future work should explore more realistic settings that go beyond restricted domains and rather artificial problems such as visually-guided image caption translation.

Table 4 BLEU scores for SLT methods on various tasks. All results reported on respective tasks' test sets

Approach	BLEU ↑	Training data			Description
		SLT (h)	ASR (h)	MT (sent)	
English → French Augmented LibriSpeech					
Bérard et al. (2018)	13.4	100			CNN+LSTM. Multi-task.
Di Gangi et al. (2019c)	13.8	236			CNN+Transformer.
Bahar et al. (2019)	17.0	100	130	95k	Pyramidal LSTM. Pretraining, augmentation.
Liu et al. (2019)	17.0	100			Transformer. Knowledge distillation.
Inaguma et al. (2019a)	17.3	472			CNN+LSTM. Multilingual.
Pino et al. (2019)	21.7	100	902	29M	CNN+Transformer. Pretraining, augmentation.
Pino et al. (2019)†	21.8	100	902	29M	End-to-end ASR. CNN+LSTM.
English → German Must-C					
Indurthi et al. (2020)	17.2	408	395	4.2M	CNN+Transformer. Meta-learning.
Di Gangi et al. (2019c)	17.3	408			CNN+Transformer.
Di Gangi et al. (2019d)	17.6	850			CNN+Transformer. Multilingual.
Di Gangi et al. (2019c)†	18.5	408			SLT-data-only pipeline. End-to-end ASR.
Indurthi et al. (2020)†	20.9	408	395	4.2M	End-to-end ASR. Augmentation.
Indurthi et al. (2020)	22.1	408	395	4.2M	CNN+Transformer. Meta-learning, augmentation.
English → French Must-C					
Di Gangi et al. (2019c)	26.9	492			CNN+Transformer.
Di Gangi et al. (2019c)†	27.9	492			SLT-data-only pipeline. End-to-end ASR.
Indurthi et al. (2020)	29.2	492	395	29M	CNN+Transformer. Meta-learning.
Indurthi et al. (2020)†	33.7	492	395	29M	End-to-end ASR. Augmentation.
Indurthi et al. (2020)	34.1	492	395	29M	CNN+Transformer. Meta-learning, augmentation.
English → Romanian Must-C					
Di Gangi et al. (2019c)	16.5	432			CNN+Transformer.
Di Gangi et al. (2019c)x†	16.8	432			SLT-data-only pipeline. End-to-end ASR.

Table 4 (continued)

Approach	BLEU ↑	Training data			Description
		SLT (h)	ASR (h)	MT (sent)	
Pino et al. (2019)	17.3	432		849k	CNN+LSTM, MT-pretraining.
Pino et al. (2019)†	21.0	432	902	849k	End-to-end ASR. CNN+LSTM.

More language pairs are available in MuST-C, but these three pairs were the most competitive. Pipeline systems are marked with a dagger (†), otherwise system is end-to-end

6.1 Datasets and resources

Image-guided translation has, thus far, been studied with small-scale datasets (Elliott et al. 2016), and there is a need for larger-scale datasets that bring the resources for this task closer to the size of image captioning (Chen et al. 2015) and machine translation datasets (Tiedemann 2012). Larger-scale datasets have started to appear for video-guided translation (Sanabria et al. 2018; Wang et al. 2019b). Spoken-language translation datasets (Kocabiyikoglu et al. 2018; Niehues et al. 2018) are smaller than standard automatic speech recognition datasets. A common challenge in multimodal translation is the need for crosslingually aligned resources, which are expensive to collect (Elliott et al. 2016), or can result in a small dataset of *clean* examples (Kocabiyikoglu et al. 2018). Future work will obviously benefit from larger datasets, however, researchers should further explore the role of data augmentation strategies (Jia et al. 2019) in both spoken language translation and visually-guided translation.

6.2 Evaluation and "verification"

A significant challenge in image-guided translation has been to demonstrate that a model definitively improves translation with image guidance. This has resulted in more focused evaluation datasets that test noun sense disambiguation (Elliott et al. 2017; Lala and Specia 2018) and verb sense disambiguation (Gella et al. 2019). In addition to new evaluations, researchers are focusing their efforts on determining whether image-guided translation models are sensitive to perturbations in the inputs. Elliott (2018) showed that the translations of some trained models are not affected when guided by incongruent images (i.e. the translation models were not guided by the image that the source language sentence describes, instead they are guided by a randomly selected image; see Sect. 5.1.5 for more details); Caglayan et al. (2019) demonstrated that training models with masked tokens increases the sensitivity of models to incongruent image guidance; and, more recently, Dutta Chowdhury and Elliott (2019) showed that trained models are more sensitive to textual perturbations than incongruent image guidance. Overall, there is a need for more focused evaluations, especially in a wider variety of language pairs, and for models to be explicitly evaluated in these more challenging conditions. Future research on visually-guided translation should also ensure that new models are actually using the visual guidance in the translation process.

In spoken language translation, this line of research into focused evaluations might involve digging into the cases where a good transcript is not enough to disambiguate the translation. One possible case is translating into a language where the speaker's gender matters, such as French or Arabic (Elaraby et al. 2018). End-to-end SLT systems have the potential to use non-linguistic information from the speech signal to tackle these challenges, but it is currently unknown to which extent they are able to do so.

6.3 Shared tasks

In addition to stimulating research interest, shared task evaluation campaigns enable easier comparison of results by encouraging the use of standardised data conditions. The choice of data condition can be made with many aims in mind. To set up a race for state-of-the-art results using any and all available resources, it is enough to define a common test set. For this goal, any additional restrictions are unnecessary or even detrimental. For example the GLUE natural language understanding task (Wang et al. 2018a) takes this approach.

On the other hand, if the goal is to achieve as fair as possible comparison between architectures, then strict limitations on the training data are required as well. Most evaluation campaigns choose this approach. However, it is far from trivial to select an appropriate set of data types to include in the condition. In many tasks, the use of auxiliary or synthetic data has proved vitally useful, e.g. exploiting monolingual data in textual MT using backtranslation (Sennrich et al. 2016a). In spoken language translation, the use of auxiliary data has prompted some discussion of when end-to-end systems are considered to have reached parity with pipeline systems. To answer this question in a fair comparison, both types of systems should be evaluated under standardised data conditions.

6.4 *Multi*modality and new tasks

Most previous work on multimodal translation emphasises multimodal *inputs* and unimodal outputs, mainly text. The integration of speech synthesis, and also a better integration of visual signals in generated communication is required for improved intelligent systems and interactive artificial agents. In addition to multimodal outputs, there should be a stronger emphasis on real-time language processing and translation. This new emphasis would also result in a closer integration of models for spoken language translation models and visually-guided translation.

In SLT, the visual modality could contribute both complementary and disambiguating information. In addition, visual speech recognition, automatic lip reading in particular (e.g. Chung et al. 2017), could aid SLT for example in audio noise robustness. The How2 dataset should allow a flurry of research in the nascent field of audio-visual SLT. Wu et al. (2019a) present exploratory first results. BLEU improvements over the best non-visual baseline are not found, although the visual modality improves results when comparing between model using cascaded deliberation.

In zero-shot translation, a multilingual model is used for translating between a language pair that was not included in the parallel training data (Firat et al. 2016; Johnson et al. 2017). For example, if a model does zero-shot French→Chinese translation, the training data contains language pairs with French as the source language and Chinese as the target language but no parallel French→Chinese data. Considering ongoing research into multilingual translation models also in

multimodal translation (e.g. Inaguma et al. 2019a), and the fact that multimodal translation training data of sufficient size is available for a very limited number of language pairs, we expect an interest in zero-shot multimodal language translation in the future.

7 Conclusions

Multimodal machine translation provides an exciting framework for further development in grounded cross-lingual natural language understanding combining work in NLP, computer vision and speech processing. This paper provides a thorough survey of the current state of the art in the field focusing on specific tasks and benchmarks that drive the research. This survey details the essential language, vision, and speech resources that are available to researchers, and discusses the models and learning approaches in the extensive literature on various multimodal translation paradigms. Combining these different paradigms into truly multimodal end-to-end models of natural cross-lingual communication will be the goal of future developments, given the foundations laid out in this survey.

Acknowledgements Open access funding provided by University of Helsinki including Helsinki University Central Hospital. This study has been supported by the MeMAD project, funded by the European Union's Horizon 2020 research and innovation programme (grant agreement No 780069), the FoTran and MultiMT projects, funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No 771113 and No 678017 respectively), and the MMVC project, funded by the Newton Fund Institutional Links grant programme (grant ID 352343575). We would also like to thank Maarit Koponen for her valuable feedback and her help in establishing our discussions of machine translation evaluation.

References

Abdelali A, Guzman F, Sajjad H, Vogel S (2014) The AMARA corpus: building parallel language resources for the educational domain. In: Proceedings of the 9th international conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Reykjavík, Iceland, pp 1856–1862

Akiba Y, Federico M, Kando N, Nakaiwa H, Paul M, Tsujii J (2004) Overview of the IWSLT 2004 evaluation campaign. In: Proceedings of the 2004 international workshop on spoken language translation (IWSLT), Kyoto, pp 1–12

Anastasopoulos A, Chiang D (2018) Tied multitask learning for neural speech translation. In: Proceedings of the 2018 conference of the North American chapter of the association for computational

linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), New Orleans, Louisiana, pp 82–91
Anastasopoulos A, Chiang D, Duong L (2016) An unsupervised probability model for speech-to-translation alignment of low-resource languages. In: Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Austin, pp 1255–1263
Anguera X, Bozonnet S, Evans N, Fredouille C, Friedland G, Vinyals O (2012) Speaker diarization: a review of recent research. IEEE Trans Audio Speech Lang Process 20(2):356–370
Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D (2015) VQA: visual question answering. In: Proceedings of the 2015 IEEE international conference on computer vision (ICCV), Santiago, pp 2425–2433
Arslan HS, Fishel M, Anbarjafari G (2018) Doubly attentive transformer machine translation. Computing research repository. arXiv:1807.11605
Bahar P, Zeyer A, Schlüter R, Ney H (2019) On using SpecAugment for end-to-end speech translation. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong
Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd international conference on learning representations (ICLR), San Diego
Baltrušaitis T, Ahuja C, Morency LP (2017) Multimodal machine learning: a survey and taxonomy. Computing research repository arXiv:1705.09406
Bansal S, Kamper H, Lopez A, Goldwater S (2017) Towards speech-to-text translation without speech recognition. In: Proceedings of the 15th conference of the european chapter of the association for computational linguistics (EACL), Association for computational linguistics (ACL), Valencia, pp 474–479
Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S (2018) Low-resource speech-to-text translation. In: Proceedings of interspeech, Hyderabad, pp 1298–1302
Bansal S, Kamper H, Livescu K, Lopez A, Goldwater S (2019) Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 58–68
Barrault L, Bougares F, Specia L, Lala C, Elliott D, Frank S (2018) Findings of the third shared task on multimodal machine translation. In: Proceedings of the 3rd conference on machine translation (WMT), association for computational linguistics (ACL), Belgium, pp 308–327
Belinkov Y, Bisk Y (2018) Synthetic and natural noise both break neural machine translation. In: Proceedings of the 6th international conference on learning representations (ICLR), Vancouver
Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3(Feb):1137–1155
Bentivogli L, Cettolo M, Federico M, Federmann C (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 (IWSLT), Bruges, pp 62–69
Bérard A, Pietquin O, Servan C, Besacier L (2016) Listen and translate: A proof of concept for end-to-end speech-to-text translation. In: Proceedings of the 29th neural information processing systems conference (NeurIPS) end-to-end learning for speech and audio processing workshop, Barcelona
Bérard A, Besacier L, Kocabiyikoglu AC, Pietquin O (2018) End-to-end automatic speech translation of audiobooks. 2018 international conference on acoustics, speech and signal processing (ICASSP). IEEE, Calgary, pp 6224–6228
Bernardi R, Cakici R, Elliott D, Erdem A, Erdem E, Ikizler-Cinbis N, Keller F, Muscat A, Plank B (2016) Automatic description generation from images: a survey of models, datasets, and evaluation measures. J Artif Intell Res 55:409–442
Boito MZ, Havard WN, Garnerin M, Ferrand ÉL, Besacier L (2019) MaSS: a large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. Computing research repository arXiv:1907.12895
Caglayan O (2019) Multimodal machine translation. PhD thesis, Université du Maine
Caglayan O, Aransa W, Wang Y, Masana M, García-Martínez M, Bougares F, Barrault L, van de Weijer J (2016a) Does multimodality help human and machine for translation and image captioning? In:

Proceedings of the 1st conference on machine translation (WMT), association for computational linguistics (ACL), Berlin, pp 627–633
Caglayan O, Barrault L, Bougares F (2016b) Multimodal attention for neural machine translation. Computing research repository arXiv:1609.03976
Caglayan O, Aransa W, Bardet A, García-Martínez M, Bougares F, Barrault L, Masana M, Herranz L, van de Weijer J (2017a) LIUM-CVC submissions for WMT17 multimodal translation task. In: Proceedings of the 2nd conference on machine translation, association for computational linguistics (ACL), Copenhagen, pp 432–439
Caglayan O, García-Martínez M, Bardet A, Aransa W, Bougares F, Barrault L (2017b) NMTPY: a flexible toolkit for advanced neural machine translation systems. Prague Bull Math Linguistics 109:15–28
Caglayan O, Bardet A, Bougares F, Barrault L, Wang K, Masana M, Herranz L, van de Weijer J (2018) LIUM-CVC submissions for WMT18 multimodal translation task. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 603–608
Caglayan O, Madhyastha P, Specia L, Barrault L (2019) Probing the need for visual context in multimodal machine translation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 4159–4170
Calixto I, Liu Q (2017a) Incorporating global visual features into attention-based neural machine translation. In: Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Copenhagen, pp 992–1003
Calixto I, Liu Q (2017b) Sentence-level multilingual multi-modal embedding for natural language processing. In: Proceedings of the international conference recent advances in natural language processing (RANLP), INCOMA Ltd., Varna, Bulgaria, pp 139–148
Calixto I, Elliott D, Frank S (2016) Dcu-uva multimodal mt system report. In: Proceedings of the 1st conference on machine translation (WMT), Association for Computational Linguistics (ACL), Berlin, pp 634–638
Calixto I, Liu Q, Campbell N (2017) Doubly-attentive decoder for multi-modal neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Vancouver, pp 1913–1924
Calixto I, Rios M, Aziz W (2019) Latent variable model for multi-modal translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Florence, pp 6392–6405
Carreira J, Noland E, Banki-Horvath A, Hillier C, Zisserman A (2018) A short note about Kinetics-600. Computing research repository arXiv:1808.01340
Caruana R (1997) Multitask learning. Mach Learn 28(1):41–75
Castilho S, Doherty S, Gaspari F, Moorkens J (2018) Approaches to human and machine translation quality assessment. In: Translation quality assessment: from principles to practice, machine translation: technologies and applications, Springer, Berlin, pp 9–38
Cettolo M, Girardi C, Federico M (2012) WIT3: web inventory of transcribed and translated talks. In: Proceedings of the 16th conference of the European Association for Machine Translation (EAMT), European Association for Machine Translation (EAMT), Trento, pp 261–268
Cettolo M, Niehues J, Stüker S, Bentivogli L, Cattoni R, Federico M (2016) The IWSLT 2016 evaluation campaign. In: Proceedings of the (2016) International workshop on spoken language translation (IWSLT), Tokyo
Cettolo M, Federico M, Bentivogli L, Niehues J, Stüker S, Sudoh K, Yoshino K, Federmann C (2017) Overview of the IWSLT 2017 evaluation campaign. In: Proceedings of the 2017 international workshop on spoken language translation (IWSLT), Tokyo, pp 2–14
Chen X, Fang H, Lin TY, Vedantam R, Gupta S, Dollar P, Zitnick CL (2015) Microsoft COCO captions: data collection and evaluation server. Computing research repository arXiv:1504.00325
Chen Y, Liu Y, Li V (2018) Zero-resource neural machine translation with multi-agent communication game. In: 32nd AAAI conference on artificial intelligence, association for the advancement of artificial intelligence (AAAI)
Cheng Y, Tu Z, Meng F, Zhai J, Liu Y (2018) Towards robust neural machine translation. In: Proceedings of the 56th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Melbourne, pp 1756–1766

Chesterman A, Wagner E (2002) Can theory help translators?. Routledge, a dialogue between the Ivory Tower and the Wordface

Chiu C, Sainath TN, Wu Y, Prabhavalkar R, Nguyen P, Chen Z, Kannan A, Weiss RJ, Rao K, Gonina E, Jaitly N, Li B, Chorowski J, Bacchiani M (2018) State-of-the-art speech recognition with sequence-to-sequence models. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). Calgary, pp 4774–4778

Cho K, van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder–decoder approaches. In: Proceedings of SSST-8, 8th workshop on syntax, semantics and structure in statistical translation, Association for Computational Linguistics (ACL), Doha, pp 103–111

Chung J, Gulcehre C, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. In: Proceedings of the 27th neural information processing systems conference (NeurIPS) workshop on deep learning, Montreal

Chung JS, Senior A, Vinyals O, Zisserman A (2017) Lip reading sentences in the wild. 2017 IEEE conference on computer vision and pattern recognition (CVPR). Honolulu, Hawaii, pp 3444–3453

Clough P, Grubinger M, Deselaers T, Hanbury A, Müller H (2006) Overview of the ImageCLEF 2006 photographic retrieval and object annotation tasks. In: Proceedings of the 7th international conference on cross-language evaluation forum (CLEF), Springer, Alicante, pp 579–594

Delbrouck JB, Dupont S (2017a) Modulating and attending the source image during encoding improves multimodal translation. In: NIPS Workshop on Visually Grounded Interaction and Language (ViGIL). Long Beach, California

Delbrouck JB, Dupont S (2017b) Multimodal compact bilinear pooling for multimodal neural machine translation. Computing research repository arXiv:1703.08084

Delbrouck JB, Dupont S (2018) UMONS submission for WMT18 multimodal translation task. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 643–647

Delbrouck JB, Dupont S (2019) Adversarial reconstruction for multi-modal machine translation. Computing research repository arXiv:1910.02766

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009) Imagenet: a large-scale hierarchical image database. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, pp 248–255

Denkowski M, Lavie A (2014) Meteor universal: language specific translation evaluation for any target language. In: Proceedings of the 9th workshop on statistical machine translation, Association for Computational Linguistics (ACL), Baltimore, pp 376–380

Di Gangi M, Negri M, Nguyen VN, Tebbifakhr A, Turchi M (2019a) Data augmentation for end-to-end speech translation: FBK @ IWSLT'19. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong

Di Gangi MA, Cattoni R, Bentivogli L, Negri M, Turchi M (2019b) MuST-C: a multilingual speech translation corpus. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 2012–2017

Di Gangi MA, Negri M, Turchi M (2019c) Adapting transformer to end-to-end spoken language translation. In: Proceedings of interspeech, international speech communication association (ISCA), Graz, pp 1133–1137

Di Gangi MA, Negri M, Turchi M (2019d) One-to-many multilingual end-to-end speech translation. In: Proceedings of the (2019) IEEE workshop on automatic speech recognition and understanding (ASRU). Sentosa

Doherty S (2017) Issues in human and automatic translation quality assessment. In: Kenny D (ed) Human issues in translation technology: the IATIS yearbook. Routledge, pp 131–148

Dong D, Wu H, He W, Yu D, Wang H (2015) Multi-task learning for multiple language translation. In: Proceedings of the 53rd annual meeting of the association for computational linguistics (ACL) and the 7th international joint conference on natural language processing (IJCNLP), Association for Computational Linguistics (ACL), Beijing, pp 1723–1732

Dong L, Xu S, Xu B (2018) Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Calgary, pp 5884–5888

Drugan J (2013) Quality in professional translation: assessment and improvement. Continuum Advances in Translation, Bloomsbury Academic

Duong L, Anastasopoulos A, Chiang D, Bird S, Cohn T (2016) An attentional model for speech translation without transcription. In: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), San Diego, pp 949–959

Duselis J, Hutt M, Gwinnup J, Davis J, Sandvick J (2017) The AFRL-OSU WMT17 multimodal translation system: an image processing approach. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, pp 445–449

Dutta Chowdhury K, Elliott D (2019) Understanding the effect of textual adversaries in multimodal machine translation. In: Proceedings of the Beyond Vision and LANguage: inTEgrating Real-world kNowledge (LANTERN), Association for Computational Linguistics (ACL), Hong Kong, pp 35–40

Elaraby M, Tawfik AY, Khaled M, Hassan H, Osama A (2018) Gender aware spoken language translation applied to English–Arabic. In: Proceedings of the 2nd international conference on natural language and speech processing (ICNLSP), IEEE, Algiers, pp 1–6

Elliott D (2018) Adversarial evaluation of multimodal machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), pp 2974–2978

Elliott D, Kádár Á (2017) Imagination improves multimodal translation. In: Proceedings of the 8th international joint conference on natural language processing (IJCNLP), Asian Federation of Natural Language Processing, Taipei, pp 130–141

Elliott D, Frank S, Hasler E (2015) Multi-language image description with neural sequence models. Computing research repository arXiv:1510.04709

Elliott D, Frank S, Sima'an K, Specia L (2016) Multi30k: multilingual English-German image descriptions. In: Proceedings of the 5th workshop on vision and language, Association for Computational Linguistics (ACL), Berlin, pp 70–74

Elliott D, Frank S, Barrault L, Bougares F, Specia L (2017) Findings of the second shared task on multimodal machine translation and multilingual image description. In: Proceedings of the 2nd conference on machine translation, Association for Computational Linguistics (ACL), Copenhagen, pp 215–233

Federmann C, Lewis WD (2016) Microsoft speech language translation (MSLT) corpus: the IWSLT 2016 release for English, French and German. In: Proceedings of the 13th international workshop on spoken language translation (IWSLT), Seattle

Federmann C, Lewis WD (2017) The Microsoft speech language translation (MSLT) corpus for Chinese and Japanese: conversational test data for machine translation and speech recognition. In: Proceedings of the machine translation summit XVI (MT Summit), Nagoya, pp 72–85

Finn C, Abbeel P, Levine S (2017) Model-agnostic meta-learning for fast adaptation of deep networks. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning (ICML), PMLR, Sydney, proceedings of machine learning research, vol 70, pp 1126–1135

Firat O, Sankaran B, Al-onaizan Y, Yarman Vural FT, Cho K (2016) Zero-resource translation with multi-lingual neural machine translation. In: Proceedings of the 2016 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Austin, pp 268–277

Fomicheva M, Specia L (2016) Reference bias in monolingual machine translation evaluation. In: Proceedings of the 54th annual meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (ACL), Berlin, ACL, pp 77–82

Frank S, Elliott D, Specia L (2018) Assessing multilingual multimodal image description: studies of native speaker preferences and translator choices. Nat Lang Eng 24(03):393–413

Fukui A, Park DH, Yang D, Rohrbach A, Darrell T, Rohrbach M (2016) Multimodal compact bilinear pooling for visual question answering and visual grounding. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics (ACL), Austin, pp 457–468

Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Proceedings of the 34th international conference on machine learning (ICML), JMLR.org, Sydney, ICML'17, pp 1243–1252

Gella S, Sennrich R, Keller F, Lapata M (2017) Image pivoting for learning multilingual multimodal representations. In: Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Copenhagen, pp 2839–2845

Gella S, Elliott D, Keller F (2019) Cross-lingual visual verb sense disambiguation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 1998–2004

Ghahremani P, BabaAli B, Povey D, Riedhammer K, Trmal J, Khudanpur S (2014) A pitch extraction algorithm tuned for automatic speech recognition. 2014 IEEE international conference on acoustics, speech and signal processing (ICASSP). Florence, pp 2494–2498

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: The IEEE conference on computer vision and pattern recognition (CVPR), Columbus

Graham Y, Baldwin T, Moffat A, Zobel J (2013) Continuous measurement scales in human evaluation of machine translation. In: Proceedings of the 7th linguistic annotation workshop and interoperability with discourse, association for computational linguistics (ACL), Sofia, pp 33–41

Graves A, Schmidhuber J (2005) Framewise phoneme classification with bidirectional LSTM networks. In: Proceedings. 2005 IEEE international joint conference on neural networks, 2005., IEEE, Montreal, vol 4, pp 2047–2052

Graves A, Ar Mohamed, Hinton G (2013) Speech recognition with deep recurrent neural networks. 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Vancouver, pp 6645–6649

Grönroos SA, Huet B, Kurimo M, Laaksonen J, Merialdo B, Pham P, Sjöberg M, Sulubacak U, Tiedemann J, Troncy R, Vázquez R (2018) The MeMAD submission to the WMT18 multimodal translation task. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 609–617

Grubinger M, Clough P, Müller H, Deselaers T (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, Genoa, pp 13–23

Guzman F, Sajjad H, Vogel S, Abdelali A (2013) The AMARA corpus: building resources for translating the web's educational content. In: Proceedings of the 10th international workshop on spoken language translation (IWSLT), Heidelberg

Gwinnup J, Sandvick J, Hutt M, Erdmann G, Duselis J, Davis J (2018) The AFRL-Ohio State WMT18 multimodal system: combining visual with traditional. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 618–621

He X, Deng L (2013) Speech-centric information processing: an optimization-oriented approach. Proc IEEE 101(5):1116–1135

He X, Deng L, Acero A (2011) Why word error rate is not a good metric for speech recognizer training for the speech translation task? 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). Prague, pp 5632–5635

He K, Xiangyu Z, Shaoqing R, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), Las Vegas, pp 770–778

He K, Gkioxari G, Dollár P, Girshick R (2017) Mask R-CNN. In: Proceedings of the 2017 IEEE international conference on computer vision (ICCV), Venice, pp 2980–2988

Helcl J, Libovický J (2017) CUNI system for the WMT17 multimodal translation task. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, pp 450–457

Helcl J, Libovický J, Kocmi T, Musil T, Cífka O, Variš D, Bojar O (2018a) Neural Monkey: the current state and beyond. In: Proceedings of the 13th conference of the association for machine translation in the Americas (AMTA), Association for Machine Translation in the Americas, Boston, pp 168–176

Helcl J, Libovický J, Varis D (2018b) CUNI system for the WMT18 multimodal translation task. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 622–629

Hieber F, Domhan T, Denkowski M, Vilar D, Sokolov A, Clifton A, Post M (2017) Sockeye: a toolkit for neural machine translation. Computing Research Repository arXiv:1712.05690

Hitschler J, Schamoni S, Riezler S (2016) Multimodal pivots for image caption translation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Berlin, pp 2399–2409

Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780
Huang PY, Liu F, Shiang SR, Oh J, Dyer C (2016) Attention-based multimodal neural machine translation. In: Proceedings of the 1st conference on machine translation, Association for Computational Linguistics (ACL), Berlin, vol 2, pp 639–645
Inaguma H, Duh K, Kawahara T, Watanabe S (2019a) Multilingual end-to-end speech translation. In: Proceedings of the (2019) IEEE workshop on automatic speech recognition and understanding (ASRU). Sentosa, Singapore
Inaguma H, Kiyono S, Soplin NEY, Suzuki J, Duh K, Watanabe S (2019b) ESPnet How2 speech translation system for IWSLT 2019: Pre-training, knowledge distillation, and going deeper. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong
Indurthi S, Han H, Lakumarapu NK, Lee B, Chung I, Kim S, Kim C (2020) End-end speech-to-text translation with modality agnostic meta-learning. In: 2020 ieee international conference on acoustics, speech and signal processing (ICASSP), Barcelona, pp 7904–7908
Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the 32nd international conference on machine learning (ICML), Lille, pp 448–456
Ive J, Madhyastha P, Specia L (2019) Distilling translations with visual awareness. In: Proceedings of the 57th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Florence, pp 6525–6538
Jansen D, Alcala A, Guzman F (2014) AMARA: a sustainable, global solution for accessibility, powered by communities of volunteers. Universal access in human-computer interaction. Springer, Design for all and accessibility practice, pp 401–411
Jia Y, Johnson M, Macherey W, Weiss RJ, Cao Y, Chiu CC, Ari N, Laurenzo S, Wu Y (2019) Leveraging weakly supervised data to improve end-to-end speech-to-text translation. 2019 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, Brighton, pp 7180–7184
Johnson M, Schuster M, Le QV, Krikun M, Wu Y, Chen Z, Thorat N, Viégas F, Wattenberg M, Corrado G, Hughes M, Dean J (2017) Google's multilingual neural machine translation system: enabling zero-shot translation. Trans Assoc Comput Linguistics 5:339–351
Junczys-Dowmunt M, Grundkiewicz R, Dwojak T, Hoang H, Heafield K, Neckermann T, Seide F, Germann U, Aji AF, Bogoychev N, Martins AFT, Birch A (2018) Marian: fast neural machine translation in C++. In: Proceedings of the 56th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Melbourne, pp 116–121
Kádár Á, Elliott D, Côté MA, Chrupała G, Alishahi A (2018) Lessons learned in multilingual grounded language learning. In: Proceedings of the 22nd conference on computational natural language learning (CoNLL), Association for Computational Linguistics (ACL), Brussels, pp 402–412
Kafle K, Kanan C (2017) Visual question answering: datasets, algorithms, and future challenges. Comput Vis Image Underst 163:3–20
Kalchbrenner N, Blunsom P (2013) Recurrent continuous translation models. In: Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Seattle, pp 1700–1709
Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, Suleyman M, Zisserman A (2017) The Kinetics human action video dataset. Computing Research Repository arXiv:1705.06950
Kiros R, Salakhutdinov R, Zemel R (2014) Multimodal neural language models. In: Proceedings of the 31st international conference on machine learning (ICML), Beijing
Klein G, Kim Y, Deng Y, Senellart J, Rush A (2017) OpenNMT: open-source toolkit for neural machine translation. In: Proceedings of the 55th annual meeting of the association for computational linguistics, Association for Computational Linguistics (ACL), Vancouver, pp 67–72
Kocabiyikoglu AC, Besacier L, Kraif O (2018) Augmenting Librispeech with French translations: a multimodal corpus for direct speech translation evaluation. In: Proceedings of the 11th conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Miyazaki
Koehn P (2009) Statistical machine translation. Cambridge University Press, Cambridge
Koehn P, Zens R, Dyer C, Bojar O, Constantin A, Herbst E, Hoang H, Birch A, Callison-Burch C, Federico M, Bertoldi N, Cowan B, Shen W, Moran C (2007) Moses: open source toolkit for statistical

machine translation. In: Proceedings of the 45th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Prague

Kreutzer J, Bastings J, Riezler S (2019) Joey NMT: a minimalist NMT toolkit for novices. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP–IJCNLP): system demonstrations, Association for Computational Linguistics (ACL), Hong Kong, pp 109–114

Lala C, Specia L (2018) Multimodal lexical translation. In: Proceedings of the 11th international conference on language resources and evaluation (LREC), Miyazaki

Lala C, Madhyastha PS, Scarton C, Specia L (2018) Sheffield submissions for WMT18 multimodal translation shared task. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 630–637

Lavie A, Agarwal A (2007) METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In: Proceedings of the 2nd workshop on statistical machine translation - StatMT '07, Association for Computational Linguistics (ACL), Prague, pp 228–231

Lavie A, Waibel A, Levin L, Finke M, Gates D, Gavalda M, Zeppenfeld T, Zhan P (1997) JANUS-III: Speech-to-speech translation in multiple languages. In: 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP), IEEE Comput. Soc. Press, Munich, vol 1, pp 99–102

Li X, Lan W, Dong J, Liu H (2016) Adding Chinese captions to images. In: Proceedings of the 2016 ACM on international conference on multimedia retrieval-ICMR'16, ACM Press, New York, pp 271–275

Li J, Lavrukhin V, Ginsburg B, Leary R, Kuchaiev O, Cohen JM, Nguyen H, Gadde RT (2019) Jasper: an end-to-end convolutional neural acoustic model. In: Proceedings of Interspeech, Graz, pp 71–75

Libovický J (2019) Multimodality in machine translation. PhD thesis, Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague

Libovický J, Helcl J (2017) Attention strategies for multi-source sequence-to-sequence learning. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL), Association for Computational Linguistics (ACL), Vancouver, pp 196–202

Libovický J, Helcl J, Tlustý M, Bojar O, Pecina P (2016) CUNI system for WMT16 automatic post-editing and multimodal translation tasks. In: Proceedings of the 1st Conference on Machine Translation (WMT), Association for Computational Linguistics (ACL), Berlin, pp 646–654

Libovický J, Helcl J, Mareček D (2018) Input combination strategies for multi-source transformer decoder. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 253–260

Lin M, Chen Q, Yan S (2014a) Network in network. In: Proceedings of the 2nd international conference on learning representations (ICLR), Scottsdale

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL (2014b) Microsoft COCO: Common Objects in Context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T (eds) Proceedings of the 13th European Conference on Computer Vision (ECCV), Springer International Publishing, Zurich, vol 8693, pp 740–755

Ling ZH, Kang SY, Zen H, Senior A, Schuster M, Qian XJ, Meng HM, Deng L (2015) Deep learning for acoustic modeling in parametric speech generation: a systematic review of existing techniques and future trends. IEEE Signal Process Mag 3(32):35–52

Lison P, Tiedemann J (2016) OpenSubtitles2016: extracting large parallel corpora from movie and TV subtitles. In: Chair) NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the 10th international conference on language resources and evaluation (LREC), European Language Resources Association (ELRA), Portorož

Liu D, Liu J, Guo W, Xiong S, Ma Z, Song R, Wu C, Liu Q (2018) The USTC-NEL speech translation system at IWSLT 2018. In: Proceedings of the 15th international workshop on spoken language translation (IWSLT), pp 70–75

Liu Y, Xiong H, He Z, Zhang J, Wu H, Wang H, Zong C (2019) End-to-end speech translation with knowledge distillation. In: Proceedings of Interspeech, Graz

Luong T, Le QV, Sutskever I, Vinyals O, Kaiser L (2016) Multi-task sequence to sequence learning. In: Proceedings of the 4th international conference on learning representations (ICLR), San Juan

Lüscher C, Beck E, Irie K, Kitza M, Michel W, Zeyer A, Schlüter R, Ney H (2019) RWTH ASR systems for LibriSpeech: hybrid vs attention. In: Proceedings of interspeech, Graz, pp 231–235

Ma M, Li D, Zhao K, Huang L (2017) OSU multimodal machine translation system report. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, pp 465–469

Ma Q, Bojar O, Graham Y (2018) Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 682–701

Ma Q, Wei J, Bojar O, Graham Y (2019) Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In: Proceedings of the 4th conference on machine translation (WMT), Association for Computational Linguistics (ACL), Florence, pp 62–90

Madhyastha PS, Wang J, Specia L (2017) Sheffield MultiMT: using object posterior predictions for multimodal machine translation. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, pp 470–476

Madhyastha P, Wang J, Specia L (2019) VIFIDEL: evaluating the visual fidelity of image descriptions. In: Proceedings of the 57th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Florence, pp 6539–6550

Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A (2015) Deep captioning with multimodal recurrent neural networks (m-rnn). In: Proceedings of the 3rd international conference on learning representations (ICLR), Banff

Matusov E, Kanthak S, Ney H (2006) Integrating speech recognition and machine translation: Where do we stand? In: 2006 IEEE international conference on acoustics speech and signal processing proceedings, Toulouse, vol 5, pp 1217–1220

Mikolov T, Karafiát M, Burget L, Cernocký J, Khudanpur S (2010) Recurrent neural network based language model. In: Kobayashi T, Hirose K, Nakamura S (eds) Proceedings of interspeech, ISCA, Makuhari, Chiba, pp 1045–1048

Miyazaki T, Shimizu N (2016) Cross-lingual image caption generation. In: Proceedings of the 54th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Berlin, pp 1780–1790

Mogadala A, Kalimuthu M, Klakow D (2019) Trends in integration of vision and language research: A survey of tasks, datasets, and methods. Computing research repository arXiv:1907.09358

Mohamed A, Hinton G, Penn G (2012) Understanding how deep belief networks perform acoustic modelling. 2012 IEEE international conference on acoustics, speech and signal processing (ICASSP). Kyoto, pp 4273–4276

Morimoto T (1990) Automatic interpreting telephony research at ATR. In: Proceedings of a workshop on machine translation, UMIST, Manchester

Nakayama H, Nishida N (2017) Zero-resource machine translation by multimodal encoder–decoder network with multimedia pivot. Mach Transl 31(1–2):49–64

Ney H (1999) Speech translation: coupling of recognition and translation. In: 1999 IEEE international conference on acoustics, speech, and signal processing (ICASSP), IEEE, Phoenix, Arizona, vol 1, pp 517–520

Niehues J, Cattoni R, Stüker S, Cettolo M, Turchi M, Federico M (2018) The IWSLT 2018 evaluation campaign. In: Proceedings of the (2018) International workshop on spoken language translation (IWSLT). Bruges

Niehues J, Cattoni R, Stüker S, Negri M, Turchi M, Ha TL, Salesky E, Sanabria R, Barrault L, Specia L, Federico M (2019) The IWSLT 2019 evaluation campaign. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT)

Och FJ (2003) Minimum error rate training in statistical machine translation. In: Proceedings of the 41st annual meeting on association for computational linguistics-volume 1, Association for Computational Linguistics (ACL), Stroudsburg, ACL'03, pp 160–167

Osamura K, Kano T, Sakti S, Sudoh K, Nakamura S (2018) Using spoken word posterior features in neural machine translation. In: Proceedings of the 15th international workshop on spoken language translation (IWSLT), Bruges, pp 189–195

Ott M, Edunov S, Baevski A, Fan A, Gross S, Ng N, Grangier D, Auli M (2019) fairseq: A fast, extensible toolkit for sequence modeling. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 48–53

Panayotov V, Chen G, Povey D, Khudanpur S (2015) Librispeech: an ASR corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, South Brisbane, pp 5206–5210

Papineni K, Roukos S, Ward T, Zhu WJ (2001) BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Philadelphia, pp 311–318

Paul M, Federico M, Stüker S (2010) Overview of the IWSLT 2010 evaluation campaign. In: Proceedings of the (2010) international workshop on spoken language translation (IWSLT). France

Peitz S, Wiesler S, Nussbaum-Thom M, Ney H (2012) Spoken language translation using automatically transcribed text in training. In: Proceedings of the 9th international workshop on spoken language translation (IWSLT), Hong Kong, pp 276–283

Pham NQ, Nguyen TS, Ha TL, Hussain J, Schneider F, Niehues J, Stüker S, Waibel A (2019) The iwslt 2019 kit speech translation system. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong

Pino J, Puzon L, Gu J, Ma X, McCarthy AD, Gopinath D (2019) Harnessing indirect training data for end-to-end automatic speech translation: tricks of the trade. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT)

Post M, Kumar G, Lopez A, Karakos D, Callison-Burch C, Khudanpur S (2013) Improved speech-to-text translation with the Fisher and Callhome Spanish-English speech translation corpus. In: Proceedings of the 10th international workshop on spoken language translation (IWSLT), Heidelberg

Pulkki V, Karjalainen M (2015) Communication acoustics: an introduction to speech, audio and psychoacoustics. Wiley, Chichester

Ramanathan V, Joulin A, Liang P, Fei-Fei L (2014) Linking people in videos with "their" names using coreference resolution. In: Proceedings of the 13th European conference on computer vision (ECCV), Springer, pp 95–110

Ramirez J, Gorriz JM, Segura JC (2007) Voice activity detection. Fundamentals and speech recognition system robustness. In: Grimm M, Kroschel K (eds) Robust speech, IntechOpen, Rijeka, chap 1

Rashtchian C, Young P, Hodosh M, Hockenmaier J (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, Association for Computational Linguistics (ACL), pp 139–147

Ruiz N, Federico M (2014) Assessing the impact of speech recognition errors on machine translation quality. In: Proceedings of the 11th conference of the association for machine translation in the Americas (AMTA), Vancouver, pp 261–274

Ruiz N, Federico M (2015) Phonetically-oriented word error alignment for speech recognition error analysis in speech translation. In: Proceedings of the 2015 IEEE workshop on automatic speech recognition and understanding (ASRU), Scottsdale, Arizona, pp 296–302

Ruiz N, Gangi MAD, Bertoldi N, Federico M (2017) Assessing the tolerance of neural machine translation systems against speech recognition errors. In: Proceedings of Interspeech, Stockholm, pp 2635–2639

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. IJCV 115(3):211–252

Sainath TN, Weiss RJ, Senior A, Wilson KW, Vinyals O (2015) Learning the speech front-end with raw waveform cldnns. In: 16th annual conference of the international speech communication association (ISCA), Dresden

Salesky E, Sperber M, Waibel A (2019) Fluent translations from disfluent speech in end-to-end speech translation. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies (NAACL-HLT), Association for Computational Linguistics (ACL), Minneapolis, pp 2786–2792

Sanabria R, Caglayan O, Palaskar S, Elliott D, Barrault L, Specia L, Metze F (2018) How2: a large-scale dataset for multimodal language understanding. In: NeurIPS, workshop on visually grounded interaction and language (ViGIL). Montreal

Saon G, Soltau H, Nahamoo D, Picheny M (2013) Speaker adaptation of neural network acoustic models using i-vectors. In: Proceedings of the 2013 IEEE workshop on automatic speech recognition and understanding (ASRU), Olomouc, pp 55–59

Schneider F, Waibel A (2019) KIT's submission to the IWSLT 2019 shared task on text translation. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong

Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681

Schwenk H, Dechelotte D, Gauvain JL (2006) Continuous space language models for statistical machine translation. In: Proceedings of the 2006 joint conference on computational linguistics (COLING) and annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Sydney, pp 723–730

Sennrich R, Haddow B, Birch A (2016a) Improving neural machine translation models with monolingual data. In: Proceedings of the 54th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), pp 86–96

Sennrich R, Haddow B, Birch A (2016b) Neural machine translation of rare words with subword units. In: Proceedings of the 54th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Berlin, pp 1715–1725

Sennrich R, Firat O, Cho K, Birch-Mayne A, Haddow B, Hitschler J, Junczys-Dowmunt M, Läubli S, Miceli Barone A, Mokry J, Nadejde M (2017) Nematus: a toolkit for neural machine translation. In: Proceedings of the conference of the European chapter of the association for computational linguistics (EACL): software demonstrations, Association for Computational Linguistics (ACL), Valencia, pp 65–68

Shah K, Wang J, Specia L (2016) Shef-multimodal: Grounding machine translation on images. In: Proceedings of the 1st conference on machine translation (WMT), Association for Computational Linguistics (ACL), Berlin, pp 660–665

Shen J, Nguyen P, Wu Y, Chen Z et al (2019) Lingvo: a modular and scalable framework for sequence-to-sequence modeling. Computing research repository arXiv:1902.08295

Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the 3rd international conference on learning representations (ICLR), Banff

Snover M, Dorr B, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th conference of the association for machine translation in the Americas (AMTA), Cambridge, pp 223–231

Specia L, Frank S, Sima'an K, Elliott D (2016) A shared task on multimodal machine translation and crosslingual image description. In: Proceedings of the 1st conference on machine translation: volume 2, shared task papers, Association for Computational Linguistics (ACL), Berlin, pp 543–553

Specia L, Harris K, Blain F, Burchardt A, Macketanz V, Skadina I, Negri M, Turchi M (2017) Translation quality and productivity: a study on rich morphology languages. Machine Translation Summit XVI. Nagoya, Japan, pp 55–71

Specia L, Blain F, Logacheva V, Astudillo RF, Martins A (2018) Findings of the WMT 2018 shared task on quality estimation. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 702–722

Sperber M, Neubig G, Niehues J, Waibel A (2017a) Neural lattice-to-sequence models for uncertain inputs. In: Proceedings of the 2017 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Copenhagen, pp 1380–1389

Sperber M, Niehues J, Waibel A (2017b) Toward robust neural machine translation for noisy input sequences. In: Proceedings of the 14th international workshop on spoken language translation (IWSLT), Tokyo, pp 90–96

Sperber M, Neubig G, Niehues J, Waibel A (2019) Attention-passing models for robust and data-efficient end-to-end speech translation. Trans Assoc Comput Linguistics 7:313–325

Stein BE, Stanford TR, Rowland BA (2009) The neural basis of multisensory integration in the midbrain: its organization and maturation. Hear Res 258(1):4–15

Stoian MC, Bansal S, Goldwater S (2019) Analyzing ASR pretraining for low-resource speech-to-text translation. Computing research repository arXiv:1910.10762

Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: Proceedings of the 27th international conference on neural information processing systems (NeurIPS), MIT Press, Montreal, NIPS'14, pp 3104–3112

Takezawa T, Morimoto T, Sagisaka Y, Campbell N, Iida H, Sugaya F, Yokoo A, Yamamoto S (1998) A Japanese-to-English speech translation system: ATR-MATRIX. In: Proceedings of the 5th international conference on spoken language processing (ICSLP), Sydney

Tiedemann J (2012) Parallel data, tools and interfaces in OPUS. In: Calzolari N, Choukri K, Declerck T, Doğan MU, Maegaard B, Mariani J, Moreno A, Odijk J, Piperidis S (eds) Proceedings of the 8th international conference on language resources and evaluation (LREC), European language resources association (ELRA), Istanbul

Toyama J, Misono M, Suzuki M, Nakayama K, Matsuo Y (2016) Neural machine translation with latent semantic of image and text. Computing research repository arXiv:1611.08459

Tsvetkov Y, Metze F, Dyer C (2014) Augmenting translation models with simulated acoustic confusions for improved spoken language translation. In: Proceedings of the 14th conference of the European chapter of the association for computational linguistics (EACL), Association for Computational Linguistics (ACL), Gothenburg, pp 616–625

Unal ME, Citamak B, Yagcioglu S, Erdem A, Erdem E, Cinbis NI, Cakici R (2016) Tasviret: a benchmark dataset for automatic Turkish description generation from images. In: 2016 24th signal processing and communication application conference (SIU), pp 1977–1980

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R (eds) Advances in neural information processing systems 30, Curran Associates, Inc., pp 5998–6008

Vaswani A, Bengio S, Brevdo E, Chollet F, Gomez A, Gouws S, Jones L, Kaiser Ł, Kalchbrenner N, Parmar N, Sepassi R, Shazeer N, Uszkoreit J (2018) Tensor2Tensor for neural machine translation. In: Proceedings of the 13th conference of the association for machine translation in the Americas (AMTA), Association for Machine Translation in the Americas, Boston, pp 193–199

Vidal E (1997) Finite-state speech-to-speech translation. In: 1997 IEEE international conference on acoustics, speech, and signal processing (ICASSP), IEEE, Munich, vol 1, pp 111–114

Vinyals O, Toshev A, Bengio S, Erhan D (2015) Show and tell: a neural image caption generator. In: Proceedings of the IEEE conference on computer vision and pattern recognition, IEEE, pp 3156–3164

Wahlster W (2000) Mobile speech-to-speech translation of spontaneous dialogs: an overview of the final Verbmobil system. In: Wahlster W (ed) Verbmobil: foundations of speech-to-speech translation. Springer, Heidelberg, pp 3–21

Wang A, Singh A, Michael J, Hill F, Levy O, Bowman S (2018a) GLUE: A multi-task benchmark and analysis platform for natural language understanding. In: Proceedings of the 2018 EMNLP workshop BlackboxNLP: analyzing and interpreting neural networks for NLP, association for computational linguistics (ACL), Brussels, pp 353–355

Wang X, Pham H, Dai Z, Neubig G (2018b) SwitchOut: an efficient data augmentation algorithm for neural machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Brussels, pp 856–861

Wang Y, Shi L, Wei L, Zhu W, Chen J, Wang Z, Wen S, Chen W, Wang Y, Jia J (2018c) The Sogou-TIIC speech translation system for IWSLT 2018. In: Proceedings of the 2018 international workshop on spoken language translation (IWSLT), Bruges, pp 112–117

Wang C, Wu Y, Liu S, Yang Z, Zhou M (2019a) Bridging the gap between pre-training and fine-tuning for end-to-end speech translation. Computing research repository arXiv:1909.07575

Wang X, Wu J, Chen J, Li L, Wang Y, Wang WY (2019b) VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. Computing research repository arXiv:1904.03493

Watanabe S, Hori T, Karita S, Hayashi T, Nishitoba J, Unno Y, Soplin NEY, Heymann J, Wiesner M, Chen N, et al (2018) ESPnet: end-to-end speech processing toolkit. In: Proceedings of Interspeech, Hyderabad, pp 2207–2211

Weiss RJ, Chorowski J, Jaitly N, Wu Y, Chen Z (2017) Sequence-to-sequence models can directly translate foreign speech. In: Proceedings of Interspeech, Stockholm

Wu Z, Caglayan O, Ive J, Wang J, Specia L (2019a) Transformer-based cascaded multimodal speech translation. In: Proceedings of the 16th international workshop on spoken language translation (IWSLT), Hong Kong

Wu Z, Ive J, Wang J, Madhyastha P, Specia L (2019b) Predicting actions to help predict translations. In: Proceedings of the how2 challenge: new tasks for vision and language, Long Beach

Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) SUN database: Large-scale scene recognition from abbey to zoo. The 23rd IEEE conference on computer vision and pattern recognition. CVPR, San Francisco, pp 3485–3492

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: neural image caption generation with visual attention. In: Proceedings of the 32nd international conference on machine learning (ICML), JMLR workshop and conference proceedings, Lille, pp 2048–2057

Yao B, Jiang X, Khosla A, Lin AL, Guibas L, Fei-Fei L (2011) Human action recognition by learning bases of action attributes and parts. In: Proceedings of the 2011 IEEE international conference on computer vision (ICCV), Barcelona, pp 1331–1338

Yoshikawa Y, Shigeto Y, Takeuchi A (2017) STAIR captions: Constructing a large-scale Japanese image caption dataset. In: Proceedings of the 55th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Vancouver, pp 417–421

Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. Trans Assoc Comput Linguistics 2:67–78

Yu D, Deng L (2016) Automatic speech recognition: a deep learning approach. Springer, Berlin

Yu D, Li J (2017) Recent progresses in deep learning based acoustic models. IEEE/CAA J Autom Sin 4(3):396–409

Zadeh A, Zellers R, Pincus E, Morency LP (2016) MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. Computing research repository arXiv:1606.06259

Zhang J, Utiyama M, Sumita E, Neubig G, Nakamura S (2017) NICT-NAIST system for WMT17 multimodal translation task. In: Proceedings of the 2nd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Copenhagen, pp 477–482

Zhang P, Ge N, Chen B, Fan K (2019) Lattice transformer for speech translation. In: Proceedings of the 57th annual meeting of the association for computational linguistics (ACL), Association for Computational Linguistics (ACL), Florence, pp 6475–6484

Zheng R, Yang Y, Ma M, Huang L (2018) Ensemble sequence level training for multimodal MT: OSU-Baidu WMT18 multimodal machine translation system report. In: Proceedings of the 3rd conference on machine translation (WMT), Association for Computational Linguistics (ACL), Belgium, pp 638–642

Zhou B (2013) Statistical machine translation for speech: a perspective on structures, learning, and decoding. Proc IEEE 101(5):1180–1202

Zhou M, Cheng R, Lee YJ, Yu Z (2018) A visual attention grounding neural model for multimodal machine translation. In: Proceedings of the 2018 conference on empirical methods in natural language processing (EMNLP), Association for Computational Linguistics (ACL), Brussels, pp 3643–3653

Affiliations

Umut Sulubacak[1] · Ozan Caglayan[3] · Stig-Arne Grönroos[2] · Aku Rouhe[2] · Desmond Elliott[4] · Lucia Specia[3] · Jörg Tiedemann[1]

Ozan Caglayan
o.caglayan@imperial.ac.uk

Stig-Arne Grönroos
stig-arne.gronroos@aalto.fi

Aku Rouhe
aku.rouhe@aalto.fi

Desmond Elliott
de@di.ku.dk

Lucia Specia
l.specia@imperial.ac.uk

Jörg Tiedemann
jorg.tiedemann@helsinki.fi

[1] University of Helsinki, Helsinki, Finland

[2] Aalto University, Espoo, Finland

[3] Imperial College London, London, UK

[4] University of Copenhagen, Copenhagen, Denmark