The IWSLT Evaluation Campaign: Challenges, Achievements, Future Directions

Luisa Bentivogli[1], Marcello Federico[1], Sebastian Stüker[2], Mauro Cettolo[1], Jan Niehues[2]
[1]FBK - Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy
[2] KIT - Karlsruhe Institut of Technology, Adenauerring 2, 76131 Karlsruhe, Germany

Abstract

Evaluation campaigns are the most successful modality for promoting the assessment of the state of the art of a field on a specific task. Within the field of Machine Translation (MT), the International Workshop on Spoken Language Translation (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. The IWSLT campaign, which is the only one addressing speech translation, started in 2004 and will feature its 13th installment in 2016. Since its beginning, the campaign attracted around 70 different participating teams from all over the world. In this paper we present the main characteristics of the tasks offered within IWSLT, as well as the evaluation framework adopted and the data made available to the research community. We also analyse and discuss the progress made by the systems along the years for the most addressed and long-standing tasks and we share ideas about new challenging data and interesting application scenarios to test the utility of MT systems in real tasks.

Keywords: Evaluation Campaign, Spoken Language Translation, Machine Translation

1. Introduction

Evaluation based on measurable and shared criteria has always been an essential component of scientific research, and constitutes the hallmark of any well established research field. Shared evaluation criteria and accepted evaluation practices help in promoting the most promising scientific approaches, and thus foster the quick production of technological advancements. They also contribute to strengthen the scientific relationships and the self-awareness within a research community, and they can encourage the involvement of newcomers in the field, by providing clearly defined scientific and technological objectives, and benchmarks for evaluating them. Evaluation campaigns are the most successful modality for promoting the assessment of the state of the art of a field on a specific task.

Within the field of Machine Translation (MT), the *International Workshop on Spoken Language Translation* (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation. The IWSLT campaign, which is the only one addressing speech translation, started in 2004 and will feature its 13th installment in 2016. IWSLT's evaluations are not competition-oriented, since their goal is to favor cooperative work and scientific exchange. In this respect, IWSLT proposes challenging research tasks and an open experimental infrastructure for the scientific community working on spoken language translation.

In the following, after introducing the evaluation campaign, we present the peculiarities and challenges of spoken language translation (Section 2). We then describe the main characteristics of the offered tasks, as well as the data sets and the evaluation infrastructure made available to the community (Section 3). We also present how human evaluation evolved from adequacy/fluency assessment to relative ranking, and finally to post-editing performed by professional translators, pursuing the objective of maximising the benefit to the research community, both in terms of information about MT systems and data and resources to be reused (Section 4). To complete the overview on the evolution of the evaluation campaign, we analyse the progresses made by the systems along the years for the most addressed and long-standing tasks (Section 5). Finally, we conclude presenting ideas about new challenging data and interesting application scenario to test the utility of MT systems in real tasks (Section 6).

2. The Evaluation Campaign

The IWSLT workshop was started in 2004 with the purpose of enabling the exchange of knowledge among researchers working on speech-to-speech translation and creating an opportunity to enhance the MT systems by comparing technologies on a common test bed. The campaign built on one of the outcomes of the C-STAR (Consortium for Speech Translation Advanced Research) project, namely the BTEC (Basic Travel Expression Corpus) multilingual spoken language corpus (Takezawa et al., 2002), which served as a primary source of evaluation. Since its beginning, increasingly challenging translation tasks were offered and new data sets covering a huge number of language pairs were shared with the research community.

In the twelve editions organized from 2004 to 2015, the campaign attracted around 70 different participating teams from all over the world. Figure 1 presents the number of different teams participating in each round of the campaign.

The task of speech translation is particularly challenging for a number of reasons. On one side, MT systems are required to deal with the specific features of spoken language. With respect to written language, speech is structurally less complex, formal and fluent. It is also characterized by shorter sentences with a lower amount of rephrasing but a higher pronoun density (Ruiz and Federico, 2014). On the other side, speech translation (Casacuberta et al., 2008) requires the integration of MT with automatic speech recognition, which brings with it the additional difficulty of translating content that may have been corrupted by speech recognition errors.

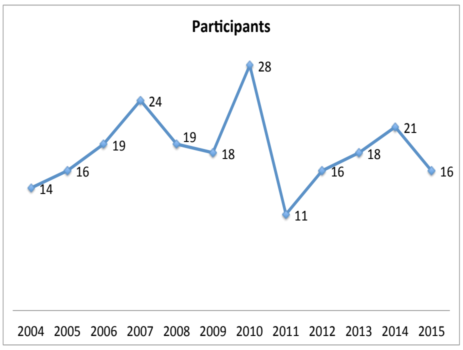

Figure 1: Number of teams that participated in the IWSLT evaluation campaigns.

Along the years, three main evaluation tracks were progressively introduced, addressing all the core technologies involved in the spoken language translation task, namely:

- Automatic speech recognition (ASR), *i.e.* the conversion of a speech signal into a transcript
- Machine translation (MT), *i.e.* the translation of a polished transcript into another language
- Spoken language translation (SLT), *i.e.* the conversion and translation of a speech signal into a transcript in another language

In the first IWSLT campaign in 2004 only the MT track was offered. Since correct human transcriptions were given as input to the MT systems, the task allowed to focus on the specific challenges related to the translation of spoken language.

Starting from 2005, also the SLT track was proposed, in order to include the additional challenge of dealing with automatic transcriptions of the audio signal, and thus investigating the impact of recognition errors on the MT performance. Participants in the SLT track could either use their own ASR systems or the ASR outputs provided by the organizers to facilitate participation. Depending on the year, different types of ASR outputs were released, such as first best output, n-best lists, lattices, ROVER combination of the outputs submitted to the ASR tracks.

The ASR track, which was offered starting from IWSLT 2011, is out of the scope of this paper since it is specifically devoted to the evaluation of speech recognition systems and does not address MT evaluation.

3. Tasks and Challenges

The first IWSLT task (Akiba et al., 2004) addressed the translation of *read-speech* transcripts in the travel domain. It was based on the BTEC corpus, which is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country and cover utterances for every potential subject in travel situations. The BTEC task was replicated in the second round of IWSLT (Eck and Hori, 2005) and was offered as "Classical" task until 2010 so to give continuity with previous editions and allow new and old participants to test their systems against a standard setting.

Starting from 2006, new and progressively more challenging tasks were added to the BTEC task, aiming at keeping the interest of the research community high by introducing more realistic scenarios. The new focus was the translation of *spontaneous speech*, while the tourism domain was maintained.

For these so-called "Challenge" tasks, different types of speech data – recorded in realistic settings – were collected, namely *answers* to travel-related questions (Paul, 2006), *monolingual dialogues* from travel agent and client interactions via telephone (Fordyce, 2007), *machine-mediated dialogues* where foreign travelers were asked to use a state-of-the-art speech-to-speech translation device to communicate with local staff (Paul, 2008), *cross-lingual human-mediated dialogues* in travel situations, where the uttered sentences were simultaneously translated by interpreters (Paul, 2009; Paul et al., 2010), and finally human dialogs in travel situations closely related to the Beijing 2008 Olympic Games (Federico et al., 2012b).

In 2010, the seventh round of IWSLT presented a mixture of innovation and continuity with the previous campaigns. Besides the Classical BTEC and Challenge Dialog tasks, a completely new task was piloted, which marked a major change with respect to previous tasks.

The new pilot task was based on TED Talks,[1] a collection of recordings of public speeches covering a wide variety of topics. Each talk is delivered in a brilliant and original style by a very skilled speaker and, while addressing a wide audience, it pursues the goal of both entertaining and persuading the listeners on a specific idea. For each talk, transcriptions and translations into several languages are provided by volunteers worldwide.

The proposed new challenge departed from and completed the application scenarios proposed till then in the IWSLT evaluations. On one side, the communication modality changed from dialogue to monologue and the language style passed from spontaneous to planned. On the other side, TED Talks data are recordings of really occurring open-domain speeches vs. speeches recorded in realistic situations within a restricted domain. Furthermore, from an application perspective, the TED Talks task is a captioning scenario, which suggests translation tasks ranging from off-line translation of written captions, up to on-line speech translation, requiring a tight integration of MT with ASR possibly handling stream-based processing.

The TED Talks task embeds interesting research challenges which are unique among the available speech recognition and machine translation benchmarks, such as coping with *(i)* background noise—e.g., applause and laughter from the audience—, *(ii)* different speakers—e.g., accents including non native speakers, varying speaking rates, prosodic aspects—, and *(iii)* limited in-domain training data and variability of topics and styles.

The TED Talks task became the main IWSLT task in 2011, and was offered to participants up to the last IWSLT edition in 2015 (Federico et al., 2011; Federico et al., 2012b; Cet-

[1] www.ted.com

tolo et al., 2013; Cettolo et al., 2014; Cettolo et al., 2015).

A major benefit to the community with respect to previous tasks lies in the public availability of TED Talks. While the BTEC corpus and all the other datasets used in the "Challenge" tasks were licenced only to IWSLT participants, TED talks video recordings, transcripts, and translations are distributed from the TED website under a Creative Commons license. Aiming at maximizing the sharing of resources, starting from 2012, the TED datasets used in the IWSLT evaluations were distributed through the WIT^3 web repository (Cettolo et al., 2012).[2] The purpose of this repository is to make the collection of TED talks effectively usable by the NLP community. Besides offering ready-to-use parallel corpora, the WIT^3 repository also offers MT benchmarks and text-processing tools designed for the TED talks collection.

The various IWLST tasks described above were offered for a remarkable number of language pairs which changed along the years. Both distant language pairs—typically involving English and Japanese, Chinese, Korean, Arabic, Turkish—and languages belonging to the same family, such as German, French, Italian, English, and many others, were addressed. All details can be found in the IWSLT overview papers.

Finally, official evaluation specifications were defined for the IWSLT tasks and require MT output to be *(i)* case sensitive and *(ii)* with punctuation marks. These specifications were chosen to serve the double purpose of delivering usable translations and making IWSLT evaluation results comparable to outcomes of other MT evaluation initiatives. In addition, automatic evaluation scores have always been calculated also for the case-insensitive (lower-case only) and no-punctuation setting.

In line with other major evaluation campaigns in the MT field, both automatic metrics and human assessments are used to evaluate submissions to IWLST. As for automatic metrics, BLEU has always been the primary metric used to rank the participating systems; furthermore, along the years additional standard metrics have been calculated, such as METEOR, WER, PER, TER, GTM, and NIST.

An important novelty introduced in IWSLT 2015 is the availability of an evaluation server, developed with the purpose of allowing participants to assess their progresses automatically and in identical conditions. Participants could submit the translation of any development set to the evaluation server, receiving scores calculated with BLEU, NIST, and TER. The evaluation server was used by the organizers for the automatic evaluation of the official submissions and, after the evaluation period, the evaluation on test sets was enabled to all participants as well. The evaluation server is maintained active and new datasets will be added for evaluations in the next campaigns.

4. Human Evaluation

Although automatic evaluation plays a very important role in fostering MT research, human evaluation is crucial aspect for an evaluation campaign. On the one hand, it provides the most direct and reliable assessment of translation quality; on the other, it is used to validate and improve automatic metrics by measuring their correlation with human judgments.

A distinguishing characteristic of IWSLT is the attention paid to the quality of human evaluation. For this reason, human evaluation was not done on a voluntary basis but was typically carried out by paid evaluators. However, it is well-known that collecting human judgments of MT outputs is time consuming and expensive, especially on the scale of an evaluation campaign. In order to find a trade off between human evaluation quality and costs, evaluation was limited to a subset of submitted runs and test data.

In the first IWSLT campaign, the standard methodology followed in other MT evaluations was adopted, where systems were judged on the basis of *fluency* and *adequacy* (White et al., 1994). Fluency refers to the degree to which the translation is well-formed according to the grammar of the target language, while adequacy refers to the degree to which the translation contains the information present in the source. This methodology was used for the first three evaluation campaigns, while in IWSLT 2007 a new methodology was introduced. In fact, studies on the reliability of human evaluation demonstrated that ranking judgments, in which annotators rank MT systems with respect to each other, are shown to have higher inter-annotator and intra-annotator agreement than adequacy and fluency judgments (Callison-Burch et al., 2007). For this reason, in IWSLT 2007 the *Ranking* task was introduced. In this task, for each source sentence five MT outputs (randomly sampled from those submitted) are presented to the evaluator, who must rank them from best to worst using a five-point scale. The collected judgments are used to obtain the ranking scores, which are calculated as the average number of times that a system was judged better than any other system. In addition to the ranking task, the evaluation based on fluency and adequacy was also carried out until IWSLT 2010 for comparison purposes.

IWSLT 2011 represents a major change in the evolution of human evaluation, since it focused solely on the ranking task and introduced a number of novelties with respect to the traditional ranking evaluation carried out in previous campaigns.

The major change was that the evaluation was not carried out by hired expert graders but relying on crowdsourced data. This choice was motivated by the results of a previous experiment on IWSLT data (Bentivogli et al., 2011), which demonstrated the feasibility of using crowdsourcing methodologies as an effective way to reduce the costs of MT evaluation without sacrificing quality.

The cost reduction obtained by using crowdsourcing allowed the modification of the ranking methodology in various respects, with the aim of maximizing the overall evaluation reliability. First, the traditional five-fold ranking task involving the evaluation of five translated sentences at a time was abandoned in favor of a direct comparison between two translated sentences only, which is a more reliable task due to the lower cognitive load required to perform it. Second—and differently from previous campaigns—to ensure system ranking reliability, full coverage of pairwise comparisons was achieved following

[2] http://wit3.fbk.eu

a *round-robin* tournament, in which each system competes against every other system.

Following the practice consolidated in the previous campaign, the IWSLT 2012 evaluation was also carried out with ranking judgments collected through crowdsourcing. However, the goal for 2012 was to find a tournament structure comparable with round robin in terms of reliability, but requiring less comparisons in favor of cost effectiveness. The most suitable structure, given its ability of ranking all competitors and the relatively few comparisons required, turned out to be the *Double Seeded Knockout with Consolation* tournament, which was thus adopted for the evaluation.

IWSLT 2013 saw the introduction of the last major novelty in human evaluation. The Ranking task was substituted by a *Post-Editing* task and, accordingly, HTER (Human-mediated Translation Edit Rate) was adopted as the official evaluation metric to rank the systems.

Post-Editing, i.e. the manual correction of machine translation output, has long been investigated by the translation industry as a form of machine assistance to reduce the costs of human translation. Nowadays, Computer-aided translation (CAT) tools incorporate post-editing functionalities, and a number of studies (Federico et al., 2012a; Green et al., 2013) demonstrate the usefulness of MT to increase professional translators' productivity. The MT TED task offered in IWSLT can be seen as an interesting application scenario to test the utility of MT systems in a real subtitling task.

From the point of view of the evaluation campaign, the goal was to adopt a human evaluation framework able to maximize the benefit to the research community, both in terms of information about MT systems and data and resources to be reused. With respect to previously adopted evaluation methodologies (i.e. adequacy/fluency and ranking tasks), the post-editing task has the double advantage of producing *(i)* a set of edits pointing to specific translation errors, and *(ii)* a set of additional reference translations. Both these byproducts are very useful for MT system development and evaluation. Human evaluation based on post-editing was adopted also in IWSLT 2014 and 2015.

5. Trends in System Performance

Our analysis focuses on the MT tracks organised over the period 2012-2015, which considered the translation of TED talks from English into language X, as well as the translation of TEDx talks given in language X into English. Tracking the progress on this task is not straightforward, as every year new evaluation sets and new training data were released. In fact, machine translation performance varies from evaluation set to evaluation set, independently from the relative improvements of the systems over the years. These random variations can be so large that they may hide the progress of the systems. Another factor that influences the absolute performance of a system is the amount of available training data. Exploiting more data, especially in-domain data, generally leads to better performance.

In order to neutralize the random effects introduced by the different test sets and the different in-domain training sets, we do compare performance of systems relative to standardised baseline systems. In particular, each baseline system is trained in exactly the same way, over the years, with the in-domain training data for each year and tested on the corresponding evaluation set.

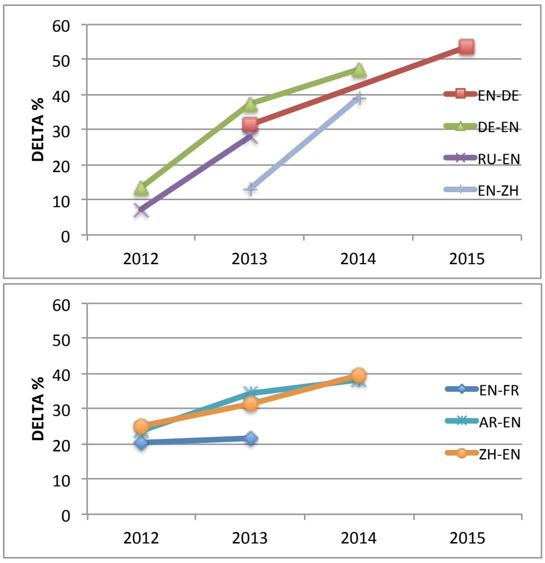

Figure 2: Performance trends over popular language pairs in terms of relative (%) improvement over the standardised baseline system.

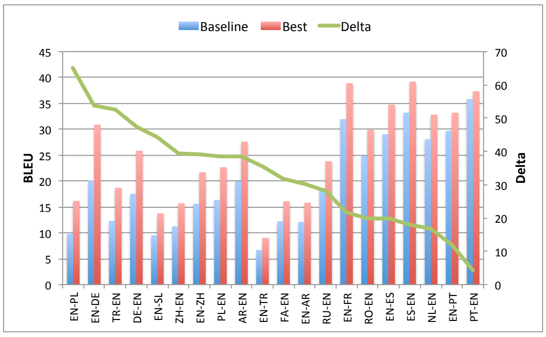

Figure 3: Best relative improvements (Delta) measured for each language pairs covered by at least two IWSLT editions, during the period 2012-2015. BLEU scores of the corresponding systems and baselines are reported, too.

Figure 2 plots the relative differences between the BLEU scores of the best system and its corresponding baseline, for a range of language pair and years. The figures are provided for the most popular translation directions and to the years where a positive trend was observed. More precisely, this excludes the cases in which the best system in one year overtakes the baseline by less than the systems of the previous years did. Our underlying assumption is that performance of MT systems developed for this task should not get worse over time. The fact that this *monotonic progress* behavior is not observed in the IWSLT evaluation is mainly due to the participant turnover, i.e. the top system of one year does not show up in the following years.

Figure 3 shows instead the overall best results in terms of BLEU score improvement - i.e. best system vs. base-

Table 1: Example of a sentence pair from the QED data

Language	Transcript
English	So in this video I'm just going to do a ton of examples.
German	Daher werde ich in diesem Video viele Beispiele durchrechnen.

line - for all translation directions that were proposed at least twice during the period 2012-2015. Language pairs are sorted by the observed improvement (Delta) with respect to the reference baseline system. This plot also shows the actual BLEU scores of the systems and baselines. Although BLEU scores on different language pairs are not directly comparable among each other, they can give a rough idea of the level of performance achieved by the baseline systems and consequently of the level of difficulty of each translation direction.

By considering the language pairs in Figure 3, the average relative BLEU score improvement over the considered period is about 33%. In particular, remarkable performance gains were achieved for English-German (53.64%), German-English (47.26%), Chinese-English (39.48%), English-Chinese (38.57%) and Arabic-English (38.43%). In fact, these improvements are the result of significant progressions in performance over time (see Figure 2). On the other hand, less progress (21.67%) has been observed on a very popular language pair such as English-French (Figure 3). A probable explanation could be that this translation pair is hard to improve because its performance is already high (BLEU score is over 35). In fact, Figure 2 confirms that lower improvements (Delta values) are in general observed for languages having better performing baselines.

6. Future Directions for Spoken Language Translation Evaluation

The TED translation task of IWSLT has become a seasoned task by now. Its introduction was motivated by its higher complexity with respect to the previous travel tasks, and by the availability of high quality data. In order to keep the tasks interesting and to follow current trends in research and industry, we are going to expand and develop the IWSLT tasks further, starting with the evaluation campaign of 2016. We will augment the TED Talk task by including more challenging data from the QCRI Educational Domain (QED) Corpus[3] (Abdelali et al., 2014). Further, we will introduce a new task on Skype conversations. Unlike in previous years we will limit the scope of the evaluation to few languages: English, German, French, and one low resourced European language. The main reason for this is to avoid dispersion of participants in too many tasks.

6.1. Extended Lecture Task

TED talks are challenging due to their variety in topics, which can be considered unlimited for all practical purposes. With respect to the type of language, TED talks are, however, very well behaved. Before being delivered, TED talks are rehearsed rigorously. Therefore, the talks tend not to show spontaneous speech phenomena, but are rather well formed. However, the majority of talks held in the world are not that well formed and well rehearsed, but rather more spontaneous and of lower quality. A prominent example of such type of talk is given by academic lectures. In order to address more lifelike talks, we are going to include data from the QED corpus (Abdelali et al., 2014) into our lecture task. This data is obtained from subtitles created on the Amara platform of videos from Khan Academy, Coursera, Udacity, etc. Table 1 gives an example of a transcription and translation from the corpus.

Table 2: Example of a sentence pair from the Skype data

Language	Transcript
German	ähm wir haben grade über Platten geredet, und über, über Musik, Musik Stream, was mich halt irgendwie nervt ist das bei so vielen Platten vorn so krass viel Werbung dazwischen geschaltet wird, und das find ich äh sehr störend, ja.
English	We just talked about albums and about streaming music, which just bugs me somehow, that for so many albums, so much advertising is placed before and in between them. And I find that very disruptive, yes.

6.2. Skype Translation Task

Recently Microsoft has introduced its Skype Translator.[4] Translating Skype or video conference conversations is a challenging task due to the nature of the language used in conversations, which is often not planned, informal in nature, ungrammatical, using special idioms etc. Therefore, while maybe not as broad in domain as talks and lectures, this task represents a challenge that goes beyond the translation of TED talks.

The test data that will be made available from Microsoft Research consist of bilingual conversations, where each speaker was speaking in his own language but was able to understand the other dialog partner's language. In this way natural conversations could be recorded. Audio was then manually processed to produce transcripts, transformed transcripts (cleaned of disfluencies), and translations (in or out of English). Table 2 shows an example from such a dialogue in English and German.

6.3. Evaluation

We expect to evaluate the extended lecture task under the post-editing perspective, exactly as we have done for the TED talk task. For the Skype Translator task, instead, we plan to opt for an adequacy-oriented evaluation, given that the focus of this communication scenario is the exchange of content between two parties. For the incoming campaign, we plan to apply human evaluation only for the extended

[3] http://alt.qcri.org/resources/qedcorpus/

[4] (http://research.microsoft.com/en-us/about/speech-to-speech-milestones.aspx

lecture task and to ground it again on the post-edition of MT outputs by professional translators. For the Skype Translator task, on the basis of the performance and output variability that we will observe, we will decide if to apply in the future (starting from 2017) human evaluations based on ranking or Likert scales.

7. Acknwoledgments

Work by FBK's authors was partially funded by the EU H2020 project CRACKER, grant agreement no. 645357.

8. Bibliographical References

Abdelali, A., Guzman, F., Sajjad, H., and Vogel, S. (2014). The amara corpus: Building parallel language resources for the educational domain. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Akiba, Y., Federico, M., Kando, N., Nakaiwa, H., Paul, M., and Tsujii, J. (2004). Overview of the IWSLT04 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Kyoto, Japan.

Bentivogli, L., Federico, M., Moretti, G., and Paul, M. (2011). Getting Expert Quality from the Crowd for Machine Translation Evaluation. In *Proceedings of the MT Summmit XIII*, pages 521–528, Xiamen, China.

Callison-Burch, C., Fordyce, C., Koehn, P., Monz, C., and Schroeder, J. (2007). (Meta-) Evaluation of Machine Translation. In *Proceedings of the 2nd Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic.

Casacuberta, F., Federico, M., Ney, H., and Vidal, E. (2008). Recent efforts in spoken language processing. *IEEE Signal Processing Magazine*, 25(3):80–88, May.

Cettolo, M., Girardi, C., and Federico, M. (2012). WIT^3: Web Inventory of Transcribed and Translated Talks. In *Proceedings of the Annual Conference of the European Association for Machine Translation (EAMT)*, Trento, Italy, May.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2013). Report on the 10th IWSLT Evaluation Campaign. In *Proceedings of the Tenth International Workshop on Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT Evaluation Campaign, IWSLT 2014. In *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, USA.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., Cattoni, R., and Federico, M. (2015). The IWSLT 2015 Evaluation Campaign. In *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam.

Eck, M. and Hori, C. (2005). Overview of the IWSLT 2005 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–22, Pittsburgh, PA.

Federico, M., Bentivogli, L., Paul, M., and Stüker, S. (2011). Overview of the IWSLT 2011 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, San Francisco, USA.

Federico, M., Cattelan, A., and Trombetti, M. (2012a). Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the Tenth Conference of the Association for Machine Translation in the Americas (AMTA)*.

Federico, M., Cettolo, M., Bentivogli, L., Paul, M., and Stüker, S. (2012b). Overview of the IWSLT 2012 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 11–27, Hong Kong, HK.

Fordyce, C. S. (2007). Overview of the IWSLT 2007 evaluation campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–12, Trento, Italy.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 439–448. ACM.

Paul, M., Federico, M., and Stüker, S. (2010). Overview of the IWSLT 2010 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 3–27, Paris, France.

Paul, M. (2006). Overview of the IWSLT 2006 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–15, Kyoto, Japan.

Paul, M. (2008). Overview of the IWSLT 2008 Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Translation*, pages 1–17, Waikiki, Hawaii.

Paul, M. (2009). Overview of the IWSLT 2009 Evaluation Campaign. In *Proceedings of the sixth International Workshop on Spoken Language Translation*, pages 1–18, Tokyo, Japan.

Ruiz, N. and Federico, M. (2014). Complexity of spoken versus written language for machine translation. In *Proceedings of the 17th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 173–180, Dubrovnik, Croatia.

Takezawa, T., Sumita, E., Sugaya, F., Yamamoto, H., and Yamamoto, S. (2002). Toward a broad-coverage bilingual corpus for speech translation of travel conversations in the real world. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pages 147–152.

White, J. S., OConnell, T., and OMara, F. (1994). The arpa mt evaluation methodologies: evolution, lessons, and future approaches. In *Proceedings of AMTA*, pages 193–205.