EAMT-funded Project
"Sponsorship of the IWSLT 2012 Evaluation Campaign"

Final Report

Luisa Bentivogli
CELCT -Centre for the Evaluation of Language and Communication Technologies
Via alla Cascata, 56/c - 38123 Povo (TN) - Italy

1. The IWSLT 2012 Evaluation Campaign

The International Workshop on Spoken Language Translation (IWSLT) is a yearly scientific workshop, associated with an open evaluation campaign on spoken language translation, where both scientific papers and system descriptions are presented. The 9th International Workshop on Spoken Language Translation took place in Hong Kong on December 6-7, 2012.

The focus of the 2012 IWSLT Evaluation Campaign (http://iwslt2012.org/) was translation of lectures and dialogs.

The task of translating lectures was built around the TED talks[1], a collection of public lectures covering a variety of topics, and for which high quality transcriptions and translations into several languages are available. The ***TED Task*** offered three distinct tracks addressing:

- automatic speech recognition (ASR) in English
- spoken language translation (SLT) from English to French
- machine translation (MT) from English to French and from Arabic to English.

In addition to the official MT language pairs, ten other unofficial translation directions were offered, with English as the target language.

Moreover, the so-called ***OLYMPICS Task*** was launched, which addressed the MT of transcribed dialogs from Chinese to English in a limited domain (the Olympic Games).

A total of 16 teams from 11 countries took part in the IWSLT 2012 evaluation campaign. More precisely: 7 teams participated in the TED-ASR task, 4 in the TED-SLT task, 7 in the TED-MT task from English to French, 5 in the TED-MT task from Arabic to English, and 4 in the OLYMPICS-MT task.

[1] http://www.ted.com

2. IWSLT 2012 HUMAN EVALUATION

Human evaluation for IWSLT 2012 focused on the collection of *Relative Ranking* judgments through crowdsourcing, and was carried out on all primary runs submitted by participants to the official SLT and MT tracks, namely:

- OLYMPICS task (Chinese-English)
- TED task: SLT track (English-French) and MT track (English-French and Arabic-English).

For each task, systems were evaluated on an evaluation set composed of 400 sentences randomly taken from the test set used for automatic evaluation.

The goal of the Ranking evaluation is to produce a complete ordering of the systems participating in a given task. The ranking task requires human judges to decide whether one system output is better than another for a given input sentence. Judges are also given the possibility to assign a tie in case both translations are equally good or bad. The judgments collected through these comparisons are then used to produce the final overall ranking.

The Ranking evaluation was carried out relying on crowdsourced data. All the pairwise comparisons to be evaluated were posted to Amazon's Mechanical Turk[2] through the CrowdFlower[3] interface.

All details about how human evaluation was carried out can be found in (Federico et al., 2012). In the following, we present the main characteristics of each task and we give information about the collection of human judgments through crowdsourcing, which was sponsored by the EAMT grant.

2.1 TED TASK

For the TED task, as a novelty for 2012, we introduced a double-elimination tournament - Double Seeded Knockout with Consolation (DSKOC) - that previous experiments showed to provide rankings very similar to the more exhaustive but more costly round-robin scheme. Moreover, system ranking was performed on a progress test - i.e. on the 2011 evaluation set. This choice was due to the fact that our aim for IWSLT 2012 was not only to evaluate all the primary runs submitted for IWSLT 2012, but also to assess their progress with respect to the best 2011 systems. Given that an 8-player tournament was adopted, different system selection criteria were applied, depending on the participants in each track:

[2] http://www.mturk.com

[3] http://www.crowdflower.com

- SLT *En-Fr*: all four 2012 runs were evaluated together with the four best runs of 2011
- MT *En-Fr*: as eight primary runs had to be evaluated for 2012 (seven submitted runs plus a baseline created by the organizers), two subsequent tournaments were carried out. In the first tournament, only the bottom four runs of 2012 were ranked. The top four runs of 2012 were ranked jointly with the top four 2011 runs in the second tournament.
- MT *Ar-En*: all five 2012 runs were evaluated together with the baseline created by the organizers and the top two runs of 2011. In this track, only the top two 2011 systems were selected. This is due to the fact that the evaluation of last year showed a large gap between the two top-ranked systems and the last two systems, which obtained poor results both in terms of automatic metrics and subjective ranking.

The following table presents all the details regarding crowdsourcing. For each pairwise comparison we requested three redundant judgments from different MTurk contributors. This means that for each task we collected three times the number of the necessary judgments (to compute label aggregation in order to ensure the quality of data, and to calculate inter-annotator agreement). As far as crowdsourcing costs are concerned, for the English tasks we paid 3 cents for 5 pairwise comparisons (0.6 cents per judgment), whereas for the French tasks the cost was slightly higher as we paid 5 cents for 5 pairwise comparisons (1 cent per judgment).

Task	# Ranked systems	# Different HtH matches to be evaluated	# Collected judgment	Total cost ($)
SLT *En-Fr*	8	12	14,400	289.68
MT *En-Fr*	8 (partial) +8	11 + 8	13,200 + 9,600	265.54+241.40
MT *Ar-En*	8	14	16,800	202.79
TOTAL		45	54,000	999.41

2.2 OLYMPICS TASK

For the OLYMPICS Task, system ranking was based on a round-robin tournament structure, following the evaluation scheme adopted in IWSLT 2011. In this type of tournament, each of the four participating systems was compared against every other system for all the 400 evaluation sentences.

The following table presents all the details regarding crowdsourcing for the OLY task. For each pairwise comparison we requested three redundant judgments from different MTurk contributors, and we paid 3 cents for 5 pairwise comparisons (0.6 cents per judgment).

Task	# Ranked systems	# Different HtH matches to be evaluated	# Collected judgment	Total cost ($)
OLY *Ch-En*	4	6	7,200	86.92

Conclusions

The EAMT grant supported the human evaluation activities carried out for the Evaluation Campaign associated to IWSLT 2012, with a specific focus on the collection of *Relative Ranking* judgments through crowdsourcing.

Crowdsourced data were collected for 4 for different tasks, namely 2 for English language and 2 for French language.

The 61,200 judgments to be collected were divided in 21 different jobs posted to MTurk through the CrowdFlower interface. Data collection lasted for around 5 weeks, as the first job was launched on 18 October 2012, and the last finished on 23 November 2012.

The total cost for crowdsourcing was 1,086.33$, corresponding to around 840€ (exchange rate for November 2012).

Following the spirit of the IWSLT evaluation campaigns, all the human evaluation data crowdsourced for IWSLT 2012 are freely distributed to the community at:

https://wit3.fbk.eu/

References

Marcello Federico, Mauro Cettolo, Luisa Bentivogli, Michael Paul, Sebastian Stuker. "Overview of the IWSLT 2012 Evaluation Campaign". In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, Hong Kong, 6-7 December 2012.