The IWSLT 2011 Evaluation Campaign on Automatic Talk Translation

Marcello Federico[1], Sebastian Stüker[2], Luisa Bentivogli[3], Michael Paul[4], Mauro Cettolo[1], Teresa Herrmann[2], Jan Niehues[2], Giovanni Moretti[3]

[1]Fondazione Bruno Kessler, via Sommarive 18, 38123 Povo (Trento), Italy
[2]Karlsruhe Institute of Technology, Adenauerring 2, 76131 Karlsruhe, Germany
[3]CELCT, Via alla Cascata 56/c, 38123 Povo (Trento), Italy
[4]NICT, Hikaridai 3-5, 619-0289 Kyoto, Japan
{federico|bentivo|cettolo}@fbk.eu, {sebastian.stueker|teresa.herrmann|jan.niehues}@kit.edu,
michael.paul@nict.go.jp, moretti@celct.it

Abstract

We report here on the eighth evaluation campaign organized in 2011 by the IWSLT workshop series. That IWSLT 2011 evaluation focused on the automatic translation of public talks and included tracks for speech recognition, speech translation, text translation, and system combination. Unlike in previous years, all data supplied for the evaluation has been publicly released on the workshop website, and is at the disposal of researchers interested in working on our benchmarks and in comparing their results with those published at the workshop. This paper provides an overview of the IWSLT 2011 evaluation campaign, and describes the data supplied, the evaluation infrastructure made available to participants , and the subjective evaluation carried out.

Keywords: evaluation campaign, speech translation, talk translation

1. Introduction

The *International Workshop on Spoken Language Translation* (IWSLT) is an annual scientific workshop, associated with an open evaluation campaign on spoken language translation. IWSLT's evaluations are not competition-oriented, but their goal is to foster cooperative work and scientific exchange. In this respect, IWSLT proposes every year challenging research tasks and an open experimental infrastructure for the scientific community working on spoken and written language translation. Openness is an important feature of the IWSLT workshop and is reflected by two major aspects: the language resources required for each evaluation are made available to the participants for free, and all results of the evaluations and papers presented at the workshop are published on the workshop's website.

Since its first edition in 2004, IWSLT's evaluations have mainly addressed application scenarios around the travel domain, featuring both dialogue and single utterance translation. For most of the tasks, parallel data covering several translation directions were extracted from the *Basic Travelling Expressions Corpus* (BTEC) (Takezawa et al., 2007), a multilingual corpus containing tourism-related sentences similar to those that are usually found in phrase books for tourists going abroad. Since IWSLT 2010 (Paul et al., 2010), a new challenge was introduced: the translation of public talks. This task is based on the TED talks collection,[1] a Web repository of recordings of public speeches, mostly held in English, covering a variety of topics, and for which high quality transcriptions and translations into several languages are available.

We call this new challenge the *TALK* task. It clearly departs from the application scenarios proposed in the previous IWSLT evaluations and completes them. Macroscopic differences between the TALK, the BTEC and the dialogue translation scenarios are in the assumed communication modality (monologue vs. dialogue), the spoken language style (planned vs. spontaneous), and the semantic context (open vs. limited).

From a translation point of view, the TALK task is basically a subtitling translation task, in which the ideal translation unit is a single caption as defined by the original transcript. In fact, some word re-ordering across consecutive captions is also permitted in order to accommodate syntactic differences between the source and target languages. The wide variety of topics covered by the TED talks has determined the type and volume of training data that has been prepared and released for the 2011 edition of this challenge (Federico et al., 2011).[2] For each language pair in question, it contains a roughly 2-million word parallel corpus of talks. The training data further contains several large out-of-domain parallel corpora, including texts from the United Nations, European Parliament, news commentaries, and the Web.

This paper gives an overview of the evaluation tracks around the TALK task that were organized at IWSLT 2011, and describes both, the language resources that have been packaged and published on the workshop website, as well as the evaluation infrastructure that has been set-up to automatically and manually evaluate the submitted runs.

2. Evaluation Tracks

The 2011 TALK evaluation campaign consisted of four tracks: a) an *automatic speech recognition* (ASR) track that targeted the automatic transcription of talks from the English audio of the TED talks to text, b) a *spoken language translation* (SLT) track for the translation of the automatic English transcriptions of the talks (ASR output) to French, c) a *machine translation* (MT) track for the translation of the manual transcriptions and translations of the talks in the directions of English to French, Arabic to English, and Chinese to English, d) a *system combination* (SC) track for

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

[2]http://iwslt2011.org

combining the English ASR outputs and/or the MT outputs in English and French.

2.1. Submission Formats

For all translation tasks the output was evaluated taking casing and punctuation into account. Especially for the SLT task this was a challenge, since the participants had to generate cased and punctuated translations from ASR output with non-reliable case and punctuation information.

3. Language Resources

As mentioned in the introduction, TED talks, mostly held in English, are available through the TED web site with manually created subtitles in English and many other languages. Translations are provided by volunteers worldwide, in tens of languages (82 at the end of 2011, to be extended to 90 in the near future). Most of the available talks have been translated into the languages involved in the IWSLT 2011 evaluation campaign, namely Arabic, Mandarin Chinese (simplified), and French. For training purposes, in addition to TED data, other large amounts of out-of-domain texts were also supplied, collected from a number of different sources.

3.1. In-Domain Parallel Data

For preparing TED parallel corpora, the raw data are first crawled, then the transcripts and the translations of corresponding talks are processed and, finally, sentences within the corresponding talks are aligned.

3.1.1. Crawling

The subtitles of the TED talks were downloaded with the help of a simple crawler based on the Linux command `wget`. From the downloaded `HTML` documents, only subtitles and useful meta data concerning the talks were kept and stored in `XML`. Each talk is enclosed by the tags `<file id="int">` and `</file>`, and includes, among other tags:

<url>	the address of the original `HTML` document of the talk
<speaker>	the name of the speaker
<talkid>	a numeric identifier of the talk
<transcript>	subtitles split in sentences
<date>	the issue date of the talk
<content>	subtitles

The only difference between the `transcript` and `content` field is, that `transcript` contains timestamps that indicate a sentence-based splitting, that make the subtitles readable during playback.

The `talkid` is an integer that univocally identifies the original transcription of a talk and all its translations; thus, it can be used to pair such texts.

Other tags (like `description`, `keywords`, `title`, whose meaning is self-explanatory) provide knowledge about the talks that can be exploited for many purposes, such as clustering, information retrieval, categorization, and adaptation.

Table 1: In-domain bilingual resources.

Task	Data	Lang	Sent	Token	Voc	Talks
MT_{EF}	train	en	107,324	2.07M	46.5K	764
		fr		2.21M	58.1K	
	dev2010	en	934	20.1K	3.4K	8
		fr		20.3K	3.9K	
	tst2010	en	1,664	32.0K	3.9K	11
		fr		33.8K	4.8K	
	tst2011	en	818	14.5K	2.5K	8
		fr		15.6K	3.0K	
MT_{AE}	train	ar	90,590	1.62M	71.1K	672
		en		1.74M	42.4K	
	dev2010	ar	934	18.3K	4.6K	8
		en		20.1K	3.4K	
	tst2010	ar	1,664	29.2K	6.0K	11
		en		32.0K	3.9K	
	tst2011	ar	1,450	25.3K	5.8K	16
		en		27.0K	3.7K	
MT_{CE}	train	zh	107,097	1.95M	56.8K	755
		en		2.07M	46.8K	
	dev2010	zh	934	21.6K	3.7	8
		en		20.1K	3.4K	
	tst2010	zh	1,664	33.3K	4.4K	11
		en		32.0K	3.9K	
	tst2011	zh	1,450	24.8K	3.9K	16
		en		27.0K	3.7K	

3.1.2. Alignment

Given a language pair, it is straightforward to select the talks for which subtitles are available in both languages, by exploiting the `talkid` mentioned in Section 3.1.1.

For each such talk, the sentences in the two languages are extracted from the `transcript` tags and paired in the order of appearance. A number of heuristic checks is performed in order to asses the parallelism: the whole talk is discarded if either the number of sentences in the two documents differs, or the sequences of timestamps differ. A single pair of sentences is discarded if its length ratio is an outlier according to a normal-mean test.

3.1.3. Statistics

The crawled text is not tokenized; Chinese is not even segmented into words. To limit the propagation of errors, we decided to release the text in its original format, leaving the preprocessing task to the participants. To compute significant figures on the size of the data sets, however, we preprocessed the texts before computing statistics with the following tools: for English and French, the tokenizer script released together with Europarl corpus (Koehn, 2005); for Arabic, the AMIRA segmenter (M. Diab and Jurafsky, 2004) (Arianna Bisazza and Federico, 2010); for Mandarin the Standford segmenter (Manning, 2002). Table 1 provides details on the size of the parallel TED resources supplied.

3.1.4. Insights

When introducing a new task such as the TALK translation task, one of the most interesting questions is how difficult the new task is. One way to do this is to look at the performance that was achieved by the systems participating in the

evaluation of that task.
For the ASR evaluation track the best system in the 2011 evaluation achieved a WER of 15.3%, a number low enough for the transcription result to be generally considered usable for speech translation. In contrast the worst system yielded a WER of 27.3% while the average WER was 19.6%.
When looking at the translation scores obtained by participants in the IWSLT 2010 and 2011 evaluation campaigns, one sees the quality is quite high, but with a high dependence on the languages involved: in 2011, the best systems achieved a score of about 38% BLEU score on the English-to-French MT track, 26 on Arabic-to-English and 17 on Chinese-to-English, all computed over a single reference (Table 2).

Table 2: Ranges of official scores ("case sensitive+punctuation" mode) from IWSLT 2011 evaluation campaign on the evaluation set `tst2011`.

tst2011	BLEU	Meteor	TER
ar-en	19.56–26.32	54.66–61.10	64.65–55.81
zh-en	11.90–16.89	45.91–52.84	70.66–62.80
en-fr	34.39–37.65	24.46–27.14	45.69–41.70

Beyond looking at the performance of systems on it, the difficulty of a translation task is typically also measured in terms of perplexity (PP) and out-of-vocabulary rate (OOV). When such figures are computed on in-domain data, they provide an insight into how hard the task is intrinsically; when they are computed on out-of-domain texts, they provide a cue on how useful that resource could be in building TED models.
Hence, as a case study, we analyzed the English-to-French translation track of the 2011 IWSLT evaluation campaign. First, 5gram LMs were estimated on a number of French texts made available for training purposes (see Tables 5 and 4), namely:

- TED: the monolingual French corpus consisting of TED talks; it is the only in-domain text
- NC: the French side of the parallel English-French News Commentary corpus
- EPPS: the French side of the parallel English-French Europarl corpus
- MultiUN: the French side of the parallel English-French MultiUN corpus.

The PP/OOV of the target side of the 2011 English-to-French test set were then computed on them and collected in Table 3.
Looking at the numbers, the following conclusions can be drawn:

- The in-domain language model yields the lowest PP, even though it is trained on a much smaller corpus; this shows that even if the topics covered by the TED talks are rather different, the common environment induces the speakers to use a somehow similar language.
- The TED talks are quite far from all the other types of text considered here, that is news, proceedings of the European Parliament and resolutions of the General Assembly of the United Nations. It is quite unexpected that EPPS is closer to talks than news, thus the difference in PP could be due to the size of the two corpora rather than their nature
- The OOV rate with respect to out-of-domain corpora seems to be mainly related to their size; it is worth noting that the OOV can be more than halved if out-of-domain corpora are added to the in-domain one (see entry `all`), showing that the proper exploitation of all available data can indeed be very beneficial.

Data	PP	%OOV
TED	103.8	1.67
NC	266.8	2.83
EPPS	200.3	1.79
MultiUN	288.2	1.21
all	150.8	0.72

Table 3: PP and %OOV of IWSLT 2011 test set with respect to four 5gram LMs estimated on in- and out-of-domain different sized corpora. Values are also reported for the LM built on the union of all corpora.

Table 4: Other bilingual resources.

Task	Data	Lang	Sent	Token	Voc
MT_{EF}	NC	en	115.6K	2.87M	57.9K
		fr		3.36M	64.6K
	EPPS	en	1.83M	50.6M	129.0K
		fr		56.2M	148.7K
	MultiUN	en	12.3M	345.5M	729.2K
		fr		402.8M	621.9K
MT_{AE}	MultiUN	ar	8.21M	248.6M	508.9K
		en		244.5M	520.3K
MT_{CE}	MultiUN	zh	8.82M	229.4M	800.3K
		en		250.8M	544.7K

3.2. Other Parallel Data

In addition to the TED texts, several out-of-domain parallel corpora could also be used by the participants to train their systems. In particular, the organizers of the *6th Workshop on Statistical Machine Translation* (WMT) kindly provided French-English texts from the *United Nations*, *Giga French-English*, *European Parliament* (EPPS) and *News Commentaries* (NC) corpora; parallel data for Arabic, Chinese, English and French languages from the *multi United Nations* (MultiUN) corpus were supplied by the *EuroMatrixPlus* project. Table 4 collects statistics on some of them; for details on the others, please refer to the specific IWSLT web page.

3.3. Monolingual Data

In addition to the parallel data described above, monolingual data was provided to train language models for the speech recognition and machine translation systems. The data was collected from different sources. First, all English transcriptions and French translations of TED talks were supplied as in-domain monolingual data for language

modelling. Furthermore, the WMT organizers also provided us with a huge amount of web-crawled news data, as well as monolingual data from the European Parliamentary speeches. In addition, the Google Books ngrams, copyright of Google Inc. and distributed under a Creative Commons Attribution 3.0 license, were also made available to the participants to train the language models.

3.3.1. Statistics

As for parallel data, these texts are provided without preprocessing, but for generating the statistics, the Europarl tokenizer script was used. Table 5 shows the statistics of the respective corpora: amount of sentences, tokens and vocabulary size.

Table 5: Monolingual resources.

Data	Lang	Sent	Tokens	Voc
TED	en	123,814	2.42M	51.3K
	fr	111,431	2.36M	60.3K
NC	en	180,657	4.32M	70.1K
	fr	147,251	4.17M	71.4K
EPPS	en	2,015,440	54.73M	134.4K
	fr	1,897,429	59.55M	153.9K
News 2007	en	13,984,262	339.52M	586.3K
	fr	946,684	23.68M	242.0K
News 2008	en	34,737,842	839.78M	1,086.0K
	fr	9,295,932	235.52M	800.7K
News 2009	en	44,041,422	1,022.07M	1,281.1K
	fr	9,544,953	234.41M	810.6K
News 2010	en	17,676,013	398.33M	695.7K
	fr	3,720,213	88.94M	492.2K
News 2011	en	2,466,169	54.57M	223.9K
	fr	1,455,577	33.79M	172.7K

3.4. ASR Development Sets

For ASR system development two test sets were provided: the 2010 development set (*dev2010*) and the 2010 evaluation set (*tst2010*) from the IWSLT 2010 evaluation campaign. The sets included the respective audio files and a scoring package.

For both data sets a segmentation was given which was mandatory to use, in order to conduct the speech translation evaluation (see Section 3.5.). The segmentation was derived from the punctuation and segmentation of the TED subtitles.

It turned out that the time markers provided with the subtitles are not suited for scoring recognizer output against them, as they are tuned for displaying the subtitles, but do not indicate the exact timing of the their respective speech in the talk. Therefore, the time markers of the segmentation were automatically created, by calculating a forced alignment (Viterbi alignment) of the complete subtitles against the audio of the talk with the help of an English speech recognition system.

The resulting segmentation of the talks was provided in the form of an *UEM* file, while the reference transcriptions of the talks where provided in *STM* format. The scoring package was complemented with a *GLM* file for performing some normalizations on the output and references before calculating the word error rate. Given this scoring package, participants were then able to score their system output using the NIST SCTK Scoring Toolkit.[3]

3.5. Speech Transcriptions

For the SLT task, the participants of the ASR task provided their outputs for the following three data sets: dev2010, tst2010 and tst2011. In order to evaluate its quality, the SLT output needs to be aligned to the reference translation. For the evaluation of ASR output, an alignment to the reference transcription is not necessarily required, but in order to facilitate the scoring for the SLT task, the speech data for the ASR task was pre-segmented according to the reference transcriptions and translations. For dev2010 and tst2010, three participants of the ASR task provided their transcriptions for SLT. Two of these provided ASR lattices in addition. For tst2011, the official test set, the ASR output of five different systems was provided, two of which were mixed-cased and three were lower-cased hypotheses. For three ASR outputs, lattices were also available. In addition, a *ROVER* system combination (Fiscus, 1997) of four participants was provided, which led to a significant reduction in WER. No case information was considered during this combination of ASR outputs. For all these automatic transcriptions, text files as well as CTM files were made available and no post processing was applied. The SLT participants were free to choose any of the provided data sets for optimizing their systems and for generating their official translations for evaluation.

4. Evaluation Tools

The organizers set up online evaluation servers for the TED development data sets (dev2010, tst2010) that could be used by the participants to tune and test their systems prior to the official run submission phase. After the official run submission period, an additional evaluation server for the official IWSLT 2011 evaluation data set (tst2011) could be used by the participants to carry out additional experiments.

- usage: http://iwslt2011.org/doku.php?id=071_evaluation_server
- dev2010: https://$nictpath/devset_IWSLT10
- tst2010: https://$nictpath/testset_IWSLT10
- tst2011: https://$nictpath/testset_IWSLT11

where
$nictpath=mastarpj.nict.go.jp/EVAL/IWSLT11/automatic

The evaluation of all primary run submissions was carried out by the organizers using standard automatic evaluation metrics. For the English ASR task, the *word error rate* (WER) which calculates the edit distance between the system output and a reference transcription was used. For all translation tasks, the automatic evaluation of all run submissions was carried out using the standard automatic evaluation metric *BLEU* which calculates the geometric mean of n-gram precision by the system output with

[3] http://www.itl.nist.gov/iad/mig//tools/

respect to reference translations (Papineni et al., 2002). In addition and as contrastive numbers six additional standard metrics—METEOR (Lavie and Agarwal, 2007), WER (Niessen et al., 2000), PER (Och, 2003), TER (Snover et al., 2006), GTM (Turian et al., 2003), and NIST (Doddington, 2002)—were calculated offline and reported in the evaluation overview paper (Federico et al., 2011).

- WER: $nistpath/pub/sctk-2.4.0-20091110-0958.tar.bz2
- BLEU: $nistpath/mt/resources/mteval-v13a.pl

where

$nistpath=ftp://jaguar.ncsl.nist.gov

5. Human Evaluation

The IWSLT 2011 subjective evaluation was carried out on all primary runs submitted by participants to the SLT, MT and MT^{SC} tracks. Regarding all MT tasks, individual systems were jointly evaluated with the SC runs and additional *online* system runs prepared by the organizers. For each task, systems were evaluated on an evaluation set composed of 400 sentences randomly taken from the test set used for automatic evaluation.

Traditionally, subjective evaluation has been focusing on *System Ranking*, which aims at producing a complete ordering of the systems participating in a given task. In IWSLT 2011, the ranking evaluation was carried out using the paired-comparison method, where judges were given two MT outputs of the same input sentence as well as a reference translation and had to decide which of the two translation hypotheses was better, taking into account both content and fluency of the translation. Judges were also given the possibility to assign a tie, in case both translations were equally good or bad. Full coverage of paired comparisons between systems was achieved by adopting a round-robin tournament structure, which is the the most complete way to determine system ranking.

In IWSLT 2011, subjective evaluation was not carried out by hired expert graders but by relying on crowd-sourced data. All the pairwise comparisons to be evaluated were posted to Amazon's Mechanical Turk (MTurk) through the CrowdFlower interface.

In order to ensure the quality of the collected data, a number of actions was taken. Firstly, we exploited the mechanisms offered by CrowdFlower, namely locale qualifications and gold units. Gold units are items with known labels which allow to distinguish between trusted contributors (those who correctly replicate the gold units) and untrusted contributors (those who fail the gold units). In our task, gold units were paired comparisons in which one system output was clearly better than the other.[4] In order to be considered trusted in a job, contributors are required to judge a minimum of four gold units and to be above an accuracy threshold of 70%. Untrusted contributors are automatically blocked and not paid, and their labels are filtered out from the final data.

[4]For a detailed description of the quality control mechanisms implemented in CrowdFlower see (Bentivogli et al., 2011).

Furthermore, for each pairwise comparison we requested three redundant judgements from different MTurk contributors. This means that for each task we collected three times the number of the necessary judgements. Redundant judgement collection is a typical method to ensure the quality of crowd-sourced data. In fact, instead of relying on a single judgement, label aggregation is performed by applying majority voting. Moreover, agreement information is systematically collected for each pairwise comparison.

Table 6 presents an overview of data collection through crowd-sourcing for each IWSLT task. The number of collected judgements is given, together with the time and cost required to collect them. Moreover, the number of different trusted contributors for each task is presented, together with their average level of trustworthiness, calculated by CrowdFlower for each contributor as the proportion of correctly judged gold units on the total gold units seen.

Comparing the tasks where the target language was English with those for which it was French, we can notice some expected trends.[5] In fact, data collection for French was more expensive and required a longer time than data collection for English. Moreover, less contributors carried out the French tasks but it is worthwhile to note that a high average trustworthiness of contributors is recorded in all the tasks.

Table 6: Summary of the crowd sourced data collection.

Task	# teams	#collected judgements	MTurk Time	Cost ($)	# contributors	Average trust level
SLT$_{EF}$	5	12,000	10d+13h	217	51	95%
MT$_{EF}$	9	43,200	20d+10h	750	87	93%
MT$_{AE}$	6	18,000	7d+23h	140	176	94%
MT$_{CE}$	6	18,000	8d+15h	140	124	94%

The inter-annotator agreement for the various IWSLT tasks was calculated using *Fleiss' kappa coefficient* κ (Siegel and Castellan, 1988; Fleiss, 1971). The agreement rates resulted to be 0.20 for the SLT$_{EF}$ task, 0.39 for the MT$_{EF}$ task, 0.29 for the MT$_{AE}$ task and 0.22 for the MT$_{CE}$ task, corresponding to "slight" agreement for the SLT$_{EF}$ task and "fair" agreement for all the other tasks.

For each task, we also calculated the statistical significance of the pairwise head-to-head comparisons between systems. We applied the Approximate Randomization Test (Noreen, 1989) to find statistically significant differences between pairs of systems in the same tasks, i.e. differences that cannot be attributable to chance. The results, presented in Appendix A, show that the great majority of the differences between systems in all tasks are significant at $p \geq 0.01$.

[5]The actual response of the MTurk workforce to a task is beyond the task requester's control: cost, time and number of workers for a given task are not always predictable as they usually change over time.

The comparison of inter-annotator agreement rates with statistical significance scores leads to an interesting observation. Although the collected data are noisy, as highlighted by the low agreement rates, the huge number of assessments collected with the round robin tournament structure allows significant differences between systems to emerge, showing an overall evaluation which clearly goes beyond chance.

6. Participation and Awards

In this year's evaluation, 11 different sites participated. Distributed over the different evaluation tracks a total of 30 primary runs and 51 contrastive runs was submitted. Considering the novelty of the TALK translation task and its increased difficulty compared to the previous tasks of the tourism domain, this constitutes a satisfactory participation which shows the acceptance of the TALK translation task within the speech translation community.

6.1. Translation task

The translation quality of the participating systems was evaluated on the manual transcripts in the MT task as well as on real ASR output in the SLT task. When using the ASR output instead of the manual transcripts additional problems arise. First, errors made by the ASR systems will lead to additional translation errors. To be able to deal with this errors two participants experimented with using multiple ASR hypotheses as input for the MT system.

Furthermore, the ASR output does not have any punctuation and is often not capitalized. To be able to generate case and punctuation information on the target side, the participants tried to generate this information at different steps of the translation process. While some tried to add the punctuation and case information before the actual translation process, other did it during or after the translation process.

When comparing the results of the automatic evaluation metrics on both tasks, we see that the additional errors of the ASR output decreases the scores. The BLEU scores of the best system on the MT task is 37 points while the best one on the SLT track is 28. So the scores drop by around 9 BLEU points. For the TER metric the error rate increases from 41.7 to 53.7.

6.2. Award

Every participant to the evaluation was also required to hand in a paper describing the systems with which he took part in the evaluation. While the a paper review process secured a minimum quality of the system papers submitted, we wanted to encourage participants to make an extra effort to produce high quality papers. In order to do so we offered a best system paper award for the best system description. With sponsoring by Microsoft Research, the award was offered with a price money of 1,000USD. In the end the award was given to the system paper by LIUM (Rousseau et al., 2011) for its comprehensive work and experiments on all aspects of the TALK task, including tight coupling between ASR and MT, acoustic modelling tailored specifically to the TED domain, and original work on system combination.

7. Acknowledgments

The IWSLT organizers would like to acknowledge the support by the EuroMatrixPlus project (IST-231720) and the T4ME network of excellence (IST-249119), which are both funded by the European Commission under the Seventh Framework Programme for Research and Technological Development. 'Research Group 3-01' received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

8. References

Mauro Cettolo Arianna Bisazza, Ioannis Klasinas and Marcello Federico. 2010. FBK @ IWSLT 2010. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2–3. December.

Luisa Bentivogli, Marcello Federico, Giovanni Moretti, and Michael Paul. 2011. Getting Expert Quality from the Crowd for Machine Translation Evaluation. In *Proceedings of the MT Summmit XIII*, pages 521–528, Xiamen, China.

George Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In *Proceedings of the Second International Conference on Human Language Technology (HLT)*, pages 257–258, San Diego, USA.

Marcello Federico, Luisa Bentivogli, Michael Paul, and Sebastian Stüker. 2011. Overview of the IWSLT 2011 Evaluation Campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 347–354, Santa Barbara, CA, USA, December. IEEE.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5).

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the Tenth Machine Translation Summit (MT X)*, pages 79–86, Phuket, Thailand.

Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation (WMT)*, pages 228–231, Prague, Czech Republic.

K. Hacioglu M. Diab and D. Jurafsky. 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *Proceedings of the HLT-NAACL 2004: Short Papers*, pages 149–152, Boston, MA, USA.

Huihsin Tseng; Pichuan Chang; Galen Andrew; Daniel Jurafsky; Christopher Manning. 2002. A conditional random field word segmenter for sighan bakeoff 2005. In *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, pages 311–318, Philadelphia, PA, USA.

Sonja Niessen, Franz J. Och, Georg Leusch, and Hermann

Ney. 2000. An Evaluation Tool for Machine Translation: Fast Evaluation for Machine Translation Research. In *Proceedings of the Second International Conference on Language Resources & Evaluation (LREC)*, pages 39–45, Athens, Greece.

Eric W. Noreen. 1989. *Computer-intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons Inc.

Franz J. Och. 2003. Minimum Error Rate Training in SMT. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 160–167, Sapporo, Japan.

Sebastian Padó, 2006. *User's guide to* `sigf`*: Significance testing by approximate randomisation*.

K. Papineni, S. Roukos, T. Ward, and W. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, pages 311–318, Philadelphia, PA, USA.

Michael Paul, Marcello Federico, and Sebastian Stüker. 2010. Overview of the iwslt 2010 evaluation campaign. In *Proc. of the International Workshop on Spoken Language Translation (IWSLT)*, Paris, France, 2–3. December.

Anthony Rousseau, Fethi Bougares, Paul Delglise, Holger Schwenk, and Yannick Estve. 2011. LIUM's systems for the IWSLT 2011 Speech Translation Tasks. In *International Workshop on Spoken Language Translation (IWSLT)*, San Francisco, CA.

Sidney Siegel and N. John Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In *Proceedings of the The Seventh Conference of the Association for Machine Translation in the Americas (AMTA)*, pages 223–231, Cambridge, USA.

T. Takezawa, G. Kikui, M. Mizushima, and E. Sumita. 2007. Multilingual spoken language corpus development for communication research. *Computational Linguistics and Chinese Language Processing*, 12(3):303–324.

Joseph P. Turian, Luke Shen, and I. Dan Melamed. 2003. Evaluation of Machine Translation and its Evaluation. In *Proceedings of the MT Summmit IX*, pages 386–393, New Orleans, USA.

Appendix A. Human Evaluation - Pairwise System Comparisons

The following tables show pairwise comparisons between systems for each task. Wins read column by row, i.e. the numbers in the table cells indicate the percentage of times that the system in that column was judged to be better than the system in that row. The winner of the two systems is indicated in bold. The difference between 100 and the sum of the complementary cells corresponds to the percentage of ties.

We applied the Approximate Randomization Test to measure the significance of result differences for each pairwise comparison. In the following tables † indicates statistical significance at $p \geq 0.10$, ‡ indicates statistical significance at $p \geq 0.05$, and $\star$ indicates statistical significance at $p \geq 0.01$, according to the Approximate Randomization Test based on 10,000 iterations. To carry out the significance test we used the package available at: http://www.nlpado.de/ sebastian/software/sigf.shtml (Padó, 2006).

SLT English-French (SLT$_{EF}$)

	FBK	KIT	LIG	LIUM	RWTH
FBK	-	**35.26***	**34.25***	**41.25***	**34.92***
KIT	21.66*	-	24.06	27.75	24.00
LIG	21.50*	**28.82**	-	**30.08**‡	26.88
LIUM	20.50*	**29.25**	21.55‡	-	21.55†
RWTH	21.11*	**27.75**	**29.65**	**28.82**†	-
$>$	21.19	30.26	27.38	**31.98**	26.83
$\geq$	63.57	75.63	73.18	**76.78**	73.17

MT English-French (MT$_{EF}$)

	DFKI	FBK	KIT	LIG	LIMSI	MIT	RWTH	ONLINE	DFKI^{SC}
DFKI	-	32.00*	**43.00**	37.25	**46.50**‡	**42.50**	39.50	**57.00***	**35.00**
FBK	**50.25***	-	**49.75***	**43.25**†	**53.50***	**48.25***	**50.75***	**64.25***	**42.00***
KIT	42.25	33.00*	-	36.50‡	**43.50**	39.75	38.25	**58.25***	32.75
LIG	**42.25**	35.75†	**46.50**‡	-	**51.00***	**42.50**	**45.00**†	**61.50***	**32.75**‡
LIMSI	35.00‡	33.50*	41.25	32.50*	-	37.75*	37.00*	**53.50***	31.50‡
MIT	40.00	35.00*	**44.75**	39.75	**51.00***	-	37.75‡	**59.00***	**36.50***
RWTH	**42.25**	34.50*	**45.25**	37.25†	**50.50***	**47.25**‡	-	**61.25***	**28.00**
ONLINE	28.50*	25.75*	26.25*	25.00*	30.50*	27.00*	26.00*	-	21.50*
DFKI^{SC}	28.25	24.00*	**35.50**	25.00‡	**40.00**‡	23.75*	27.75	**54.50***	-
$>$	38.59	31.69	41.53	34.56	45.81	38.59	37.75	**58.66**	32.50
$\geq$	58.41	49.75	59.47	55.34	62.25	57.03	56.72	**73.69**	67.66

MT Arabic-English (MT$_{AE}$)

	DCU	FBK	MIT	RWTH	ONLINE	DFKI^{SC}
DCU	-	**55.00***	**29.75**‡	**57.50***	**62.81***	**41.75***
FBK	08.75*	-	17.00*	**33.25***	**43.18***	14.50*
MIT	21.00‡	**44.25***	-	**53.75***	**58.04***	**31.00***
RWTH	08.00*	21.00*	08.25*	-	**36.59***	09.00*
ONLINE	07.79*	19.44*	11.31*	24.56*	-	18.48*
DFKI^{SC}	08.75*	**25.50***	08.25*	**31.50***	**50.89***	-
$>$	10.86	33.07	14.91	40.12	**50.30**	22.96
$\geq$	50.65	76.70	58.41	83.44	**83.69**	75.09

MT Chinese-English (MT$_{CE}$)

	DCU	MSR	NICT	RWTH	ONLINE	MSR^{SC}
DCU	-	**69.75***	**55.25***	**53.75***	**75.00***	**66.00***
MSR	10.00*	-	24.00*	27.75*	**48.75***	**21.50**
NICT	15.25*	**46.50***	-	**42.75***	**67.75***	**50.50***
RWTH	12.75*	**48.00***	30.75*	-	**56.50***	**42.00***
ONLINE	08.25*	27.00*	14.50*	21.75*	-	26.75*
MSR^{SC}	10.00*	21.25	21.50*	20.50*	**51.75***	-
$>$	11.25	42.50	29.20	33.30	**59.95**	41.35
$\geq$	36.05	73.60	55.45	62.00	**80.35**	75.00