[From: *Tralogy*, Paris, 3-4 March 2011]

Philipp Koehn

What is a Better Translation? Reflections on Six Years of Running Evaluation Campaigns

Abstract

We have been actively involved in the development of machine translation systems and the organization of evaluation campaigns for such systems. We report on the challenges and our experience with automatic and manual metrics used in research.

Introduction

Machine translation, and even more so human translation, are long standing efforts that aim to re-create a document in a different language that contains the same meaning as the original language document. Since the task involves *meaning* at its core, we are confronted with all the unsolved problems of representation, equivalence, and similarity.

To illustrate the problem, see Figure 1, where a short Chinese sentence was translated into English by ten different human translators. Each came up with a different translation. This is a very typical example. Given any sentence of non-trivial length, a group of ten translators will come up with ten different translations. In fact, if presented with the same sentence the next day, they will come up with even more different translations.

The task of assessing what is a *correct* translation is hence rather difficult. But it is an essential task not only for assessing human translation quality, but also for evaluating machine translation systems. In this paper, we will discuss our efforts to create metrics and methods to evaluate the quality of machine translation systems which participated in an annual evaluation campaign organized around a workshop of the annual conferences of the Association of Computational Linguistics, namely the Workshop of Statistical Machine Translation (Koehn and Monz, 2005, 2006; Callison-Burch *et al.*, 2007, 2008, 2009, 2010).

Figure 1: Ten translations of a Chinese sentence
(typical example from the 2001 NIST evaluation set)

We will discuss the metrics adequacy and fluency, sentence ranking, and a sentence understanding metric. We are concerned with inter-evaluator agreement and also efforts to gather a large number of judgments by crowd-sourcing.

Goals

We may want to set ourselves the following goals for developing evaluation metrics — taken from (Koehn, 2010b):

Correct : metric must prefer better systems
Consistent : repeated use of metric should give same results
Low cost : little time and money spent on carrying out evaluation
Tunable : possible to automatically optimize system performance towards metric
Meaningful : score should give intuitive interpretation of translation quality

There are other evaluation criteria when deploying machine translation systems that go beyond the quality of translations, such as speed (we prefer faster machine translation systems), size (do they fit into the memory of available machines, e.g., handheld devices), integration (can they be integrated into existing workflow), and customization (can they be adapted to user's needs).

However, in this paper, we are only concerned with measures of quality.

Evaluation in Machine Translation Research

Current research in machine translation (and especially in statistical machine translation) is driven by constantly measuring and aiming at improving translation quality.

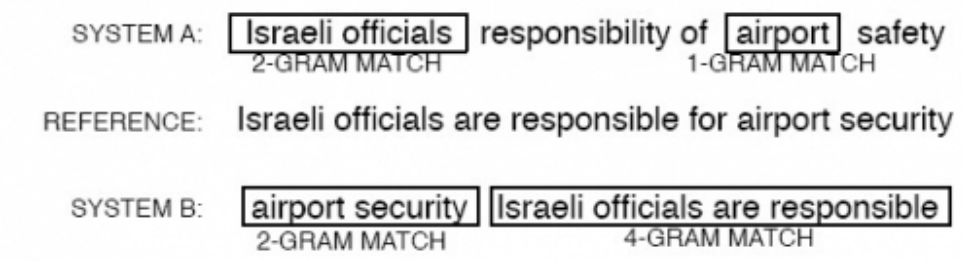
The basic methodological paradigm consists of building a baseline system, implementing a new idea and testing the resulting system. If the change led to an improvement in translation quality, it is maintained, otherwise it is dropped or refined, and the cycle continues.

Automatic Metrics

Evaluation is at the core of research and development and may be carried out multiple times a day. What this methodology requires is a very fast and typically fully automatic evaluation metric.

The key insight in developing evaluation metric for machine translation research was the idea to compare the system output against one or more human reference translations.

As we have seen in Figure 1, machine translation systems cannot be expected to match human translations, so what is used instead is a measure of *similarity* between the machine translation output and the human reference translations.



Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%

Figure 2: BLEU score: measuring n-gram overlap

The most widely used metric, BLEU (Papineni *et al.*, 2002), measures n-gram overlap between the machine translation output and the reference translation. For each n-gram of size 1–4 in the machine translation output, we check if it also occurs in the reference translation. From this, we compute the precision for each n-gram size, i.e., the ratio of machine translation n-grams of a certain size that occur in the reference translation.

Since we use precision, we may cheat by generating too short translations or even dropping difficult sentences. Hence, a brevity penalty is added to penalize too short translations. Formally, the BLEU score is computed as:

$$\text{BLEU} = \min\left(1, \frac{output - lenght}{reference - lenght}\right)\left(\prod_{i=1}^{4} precision_i\right)^{\frac{1}{4}} \quad (1)$$

For an example, please see Figure 2. The BLEU score is computed over the entire corpus, not single sentences. Sometimes multiple reference translations are used to account for the variability in translation, but such data are not always available.

Recent research into machine translation, such as the exemplified by the METEOR metric (Banerjee and Lavie, 2005), aims at relaxing the matching criteria. Partial credit may be given for matching stems when output and reference words differ in their morphological properties, matching synonyms — which requires resources such as WordNet (Miller *et al.*, 1993) —, or matching paraphrases of the reference translation.

Criticism

There is widespread criticism of the use of automatic evaluation metrics, or specifically BLEU. To summarize some of the arguments (Koehn, 2010b):

BLEU ignores the relative relevance of different words: Some words matter more than others. One the most glaring exampling is the word *not* that, if omitted, will cause very misleading translations. Names and core concepts are also important words, much more so than, e.g., determiners and punctuation are often irrelevant. However, all words are treated the same way.

BLEU operates only on a very local level and does not address overall grammatical coherence. System output may look good on an n-gram basis, but very muddled beyond that. There is a suspicion that this biases the metric in favour of phrase-based statistical systems, which are good at producing good n-grams, but less able to pro-duce grammatically coherent sentences.

The actual BLEU scores are meaningless. Nobody knows what a BLEU score of 30% means, since the actual number depends on many factors, such as the number of reference translations, the language pair, the domain, and even the tokenization scheme used to break up the output and reference into words.

Recent experiments computed so-called human BLEU scores, where a human reference translation scored against other human reference translations. Such human BLEU scores are barely higher (if at all) than BLEU scores computed for machine translation output, even though the human translations are better.

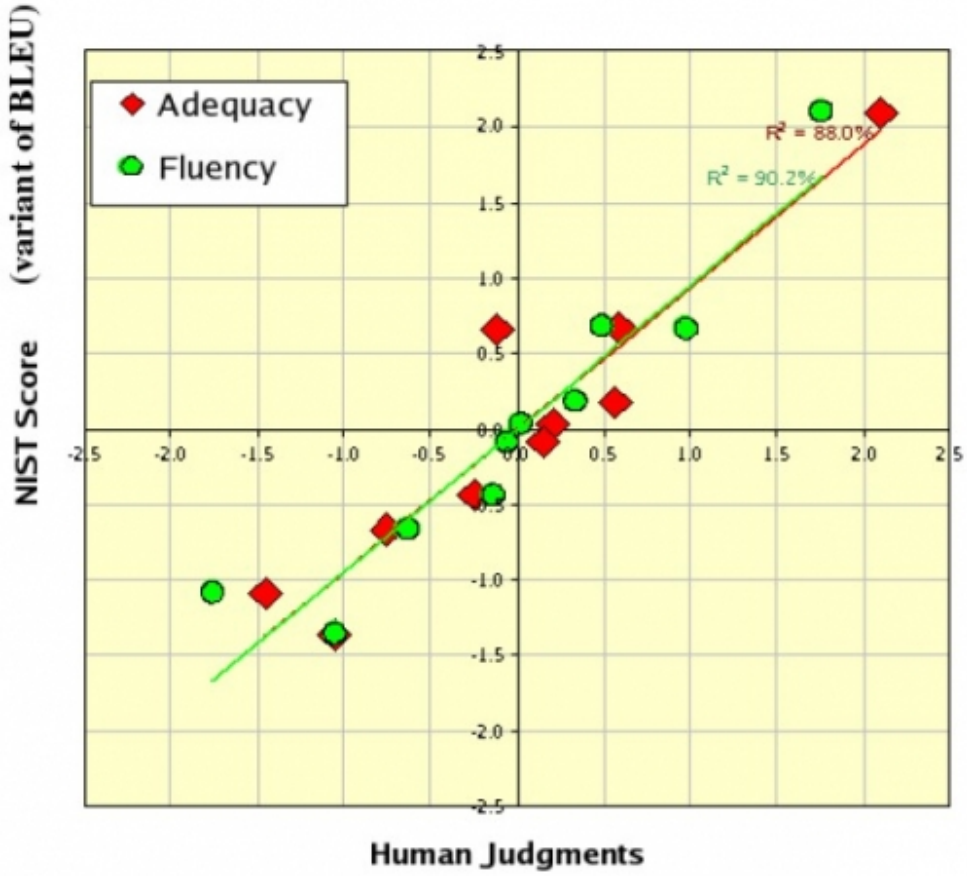


Figure 3: Correlation of automatic metrics with human judgement
(from George Doddington, NIST)

Evaluation of Automatic Metrics

Recalling our original goals (Section 2) for evaluation metrics, automatic metrics are low cost, tunable, and consistent. But are they correct? We can assess this, by checking correlation with human judgement.

An influential graph has been a finding by NIST (see Figure 3), which showed strong correlation between an automatic metric and human judgment in terms of adequacy and fluency.

Correlation between metrics and human judgement can be computed using Pearson's correlation coefficient. Given two variables, the automatic score x and human judgment y

for multiple systems on the same data set ((*x*1,*y*1), (*x*2,*y*2), ...), Pearson's correlation coefficient is computed from the means ($\bar{x}$, $\bar{y}$) and variances (s_x^2 , s_y^2) as

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y} \quad (2)$$

Recent evaluation campaigns have revealed short-comings of automatic metrics, in particular two cases (Callison-Burch *et al.*, 2006): When comparing statistical machine translation system against translations that were obtained by post-editing machine translation output, the latter did score much better according to human judgment, but not on automatic scores. In another example, rule-based systems received a much lower automatic score than human judgement would warrant.

Current Research

Current research into automatic metrics tries to address these challenges. Metrics have been proposed that use syntactic similarity, semantic equivalence or entailment, metrics targeted at reordering, trainable metrics and so on. There are ongoing evaluation campaigns on evaluation metrics that foster this research.

At this point, automatic metrics are an essential tool for system development — a tool that like others needs to be constantly refined. It is recognized that automatic metrics are not fully suited to rank systems of different types. The development of better evaluation metrics is still an open challenge.

Manual Evaluation

The idea behind manual evaluation is straightforward: Ask human evaluators to assess machine translation quality. But how?

Over the years, we have experimented with a number of different manual evaluation metrics: quality metrics such adequacy and fluency, asking manual evaluators to rank translation of the same sentence from different systems against each other, and also developed a sentence understanding test.

Evaluation Campaign

Carrying out large-scale manual evaluation is a very labor-intensive activity, spanning from the creation of a test set, over gathering output from different machine translation systems and collecting judgments from human evaluators, all the way to extensive analysis of the obtained data.

We were able to carry this out as part of the ACL Workshop on Statistical Machine Translation (WMT), where, with funding from the EU-sponsored EuroMatrix and EuroMatrixPlus projects, we organized an open evaluation campaign.

Every year since 2005, we posted training data on a web site and prepared a test set of news stories (2,000–3,000 sentences). Participants were given 5 days to translate the test set with their machine translation systems and we score the resulting output. In the most recent 2010 campaign, we addressed eight language pairs (Czech, German, French, Spanish into English and back), but we also used Finnish and Hungarian as well as language pairs not involving English in prior campaigns.

We were lucky to attract a large number of participants. In the most recent 2010 campaign, 29 institutions participated (21 from Europe, 7 from North America and 1 from Asia). Some institutions fielded multiple teams, so in total 33 groups took part. A total of 153 system translations were submitted (for all language pairs), and we also included two popular online translation systems and rule-based systems for English–Czech.

Adequacy and Fluency

The basic scenario for manual evaluation is that a human evaluator is given the machine translation output, paired with the source or a human reference translation (or both), and is asked to assess the quality of the machine translation output.

Two metrics have been used in the past:

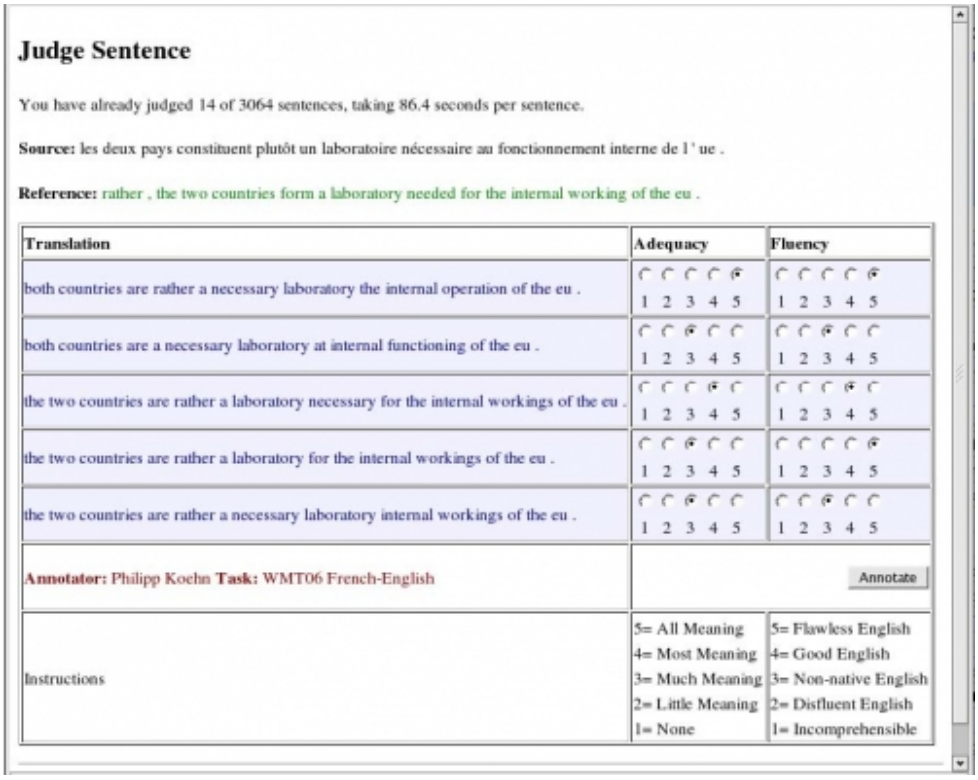
Adequacy: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?

Fluency: Is the output good fluent English? This involves both grammatical correctness and idiomatic word choices.

These assessments have to be made as numerical scores, given five choices:

	Adequacy
5	all meaning
4	most meaning
3	much meaning
2	little meaning
1	none

	Fluency
5	flawless English
4	good English
3	non-native English
2	disfluent English
1	incomprehensible



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Figure 4: Evaluation tool to assess machine translation quality in terms of adequacy and fluency

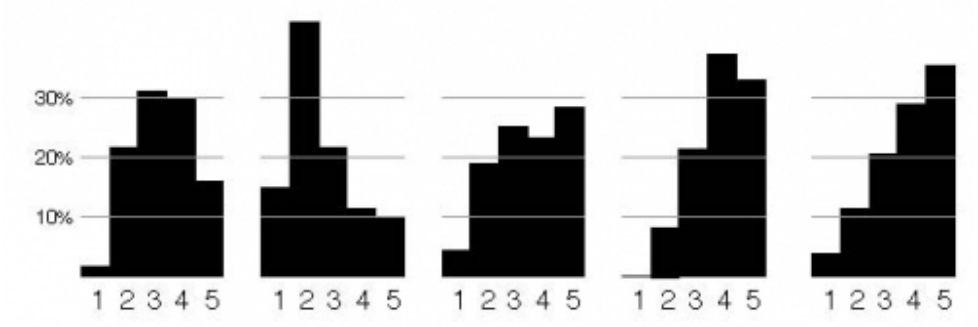


Figure 5: Histogram of adequacy judgments by different human evaluators
(from WMT 2006 evaluation)

The evaluation tool used in our evaluation campaign is shown in Figure 4.

Evaluator Agreement

Human evaluators are using the adequacy and fluency scales differently from each other. Figure 5 shows that some evaluators hardly ever assign a score of 1. Some are generally more generous, while one of the evaluators predominately assigns a score of 2.

We can measure the agreement between multiple evaluators with the Kappa coefficient, which is defined as

$$\mathrm{K} = \frac{p(A) - p(E)}{1 - p(E)} \quad (3)$$

where $p(A)$ is proportion of times that the evaluators agree and $p(E)$ is the proportion of time that they would agree by chance. For instance on a 5-point scale, chance agreement is $p(E)= 1/5$

In the 2007 WMT evaluation campaign (Callison-Burch *et al.*, 2007), we found a Kappa of .250 for fluency and .226 for adequacy. These are considered rather low numbers.

Ranking Translations

In an evaluation campaign, we are primarily interested in which systems have better translations than others. In other words, we would like to rank the system translations against each other. Hence, we asked the evaluators: *Is translation X better than translation Y?* The choices were *better, worse,* or *equal*.

In this task, evaluators are more consistent. In the WMT 2007 evaluation campaign we mea sured a Kappa of .373, versus .250 for fluency and .226 for adequacy. See Table 1 for details.

Evaluation type	P (*A*)	P (*E*)	K
Fluency	.400	.2	.250
Adequacy	.380	.2	.226
Sentence ranking	.582	.333	.373

Table 1: Inter-evaluator agreement in WMT 2007 evaluation campaign

Stricter Guidelines?

One may look at these numbers and argue for stricter guidelines for human evaluators.

For instance, a point system could be introduced that, for instance, penalizes disfluency due to omitted function word with one 1 point, a mistranslated word with 2 points, a reversal of meaning with 4 points, and so on.

However, any such rules would be arbitrary, and very time-consuming to apply. Our experience with similar fine-grained evaluation methods (Vilar *et al.*, 2006) indicates that human evaluators are not very consistent with them.

Hence, we do not have much hope for guidelines along these lines.

Task-Oriented Evaluation

Machine translation is a means to an end, so the usefulness should also be evaluated by the question, if its output helps to accomplish a given task.

The main applications of machine translation are producing high-quality translations postediting machine translation and information gathering from foreign language sources (there are others, such as assisting communication).

The criteria for success for machine translation in assisting human translation efforts is the reduction in time spent on post-editing machine translation versus translation from scratch. However, carrying out such an evaluation is time consuming, and depends on very much on the skills of translator and post-editor. There are also significant user interface issues on how the translation is presented and how it can be edited.

Some automatic metrics are inspired by this task. The TER score (Snover *et al.*, 2006) is based on number of editing steps: the Levenshtein operations (insertion, deletion, substitution) plus movement. This metric was has been by the DARPA GALE program (2005, 2011), where human evaluators also created reference translation translations as close to the machine translation output as possible (using the metric this way is referred to as HTER).

Content Understanding Tests

If the application of machine translation is the understanding of foreign language content, then an evaluation method may be framed as: Given machine translation output, can a mono-lingual target side speaker answer questions about it?

Questions may be grouped according to the level of understanding:

- basic facts: who? where? when? names, numbers, and dates
- actors and events: relationships, temporal and causal order
- nuance and author intent: emphasis and subtext

It is not easy to come up with good questions (that in turn do not give away too much information), and to calibrate their difficulty. We used a variant of this method, which asked a very straightforward question about each sentence: *What does the translation mean?*

Applying this metric involves two annota-tors: Person A edits the translation (without access neither to source nor reference). She is given the instruction:

Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."

Person B assesses the correctness of the edited translation (with full access to source and reference in context):

Language pair	Reference	Best system
French-English	85 %	52 %
English-French	79 %	49 %
German-English	83 %	47 %
English-German	85 %	47 %
Spanish-English	88 %	41 %
English-Spanish	69 %	52 %
Czech-English	98 %	25 %
English-Czech	91 %	32 %
Hungarian-English	93 %	22 %

Table 2: Ratio of how many edited sentences were judged as correct in WMT 2009 evaluation campaign (note: 95% confidence interval is about ±10%.)

Evaluation type	P (*A*)	P (*E*)	K
Sentence ranking	.549	.333	.323
Yes/no to edited output	.774	.5	.549

Table 3: Inter-evaluator agreement for content understanding metric (WMT 2009)

Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is bold.

Table 2 shows the quality of the best machine translation system. Note that the human reference translation was also included in this evaluation (it may have been edited), and that it is not always judged as being correct. This may be due to actual translation errors of the human translator, or simple testament to the fact that not only any translator will come up with a different translation, she will also often judge other's translations as wrong.

We achieve higher inter-evaluator agreement with this metric, a kappa of .549 versus a kappa of .323 for sentence ranking (see Table 3).

We have used *sentence correctness* as a metric for other evaluation tasks, such as the evaluation of manual translations by novice translators (Koehn, 2009) and the evaluation of mono-lingual translators who constructed translations of sentences in from unknown source language with various types of assistance (Koehn, 2010a).

We found that human evaluators vary significantly how strictly they apply the standard of *correctness*. Some may assess only 60% of professional translations as correct.

Crowd-Sourcing Evaluation

In order to collect a large number of human judgments for manual evaluation, we experimented with using a crowd-sourcing platform, Mechanical Turk (Callison-Burch *et al.*, 2010). We are able to present the same web interface that we used for our internal manual evaluation to users of Mechanical Turk.

The main problem with crowd-sourcing is quality control. We required some basic qualifications (existing approval rating of at least 85%, must have at least performed 5 tasks, and resides in a country where target language is spoken) and developed methods for detecting and filtering out bad workers.

Indicators for low quality workers are a low **reference preference rate**, i.e., preference of MT output often over references, and low agreement with experts. If we filter out the bad workers, we can achieve inter-evaluator agreement comparable to experts. In our experience, very few workers have to removed for better quality (two worst offenders responsible for most damage in WMT 2010).

Conclusion

Where are we now in terms of the goals that we set for evaluation metrics?

Correctness of evaluation metrics is a very difficult question. For automatic metrics, we aim to correlate with human judgment. But it is very hard to even assess if what we measure with manual metrics corresponds to some mythical notion of translation quality.

Our main guiding light for developing manual metrics is to increase *consistency*, as measured by inter-evaluator agreement. The *cost* of manual metrics is also a problem so that we are only able to carry out an extensive study once a year with significant funding. Automatic metrics do very well in terms of consistency and cost[1], and they also have the advantage of being *tunable*, i.e., they can used to automatically optimize translation system performance.

It is unclear *meaningful* any of the metrics are. Even seemingly straight-forward measures such as the ratio of correct or understandable translation hinges on the subject judgment of a human evaluator, which may be swayed by the last ten translations she has seen. The harshest critic may never like anybody else's translations and may not even like her own translations on the next day.

The development of both automatic and manual metrics still has a way to go and we will continue to pursue further research.

Acknowledgment

This work was supported by the EuroMatrix-Plus project funded by the European Commission (7th Framework Programme).

Bibliography

Satanjeev Banerjee and Alon Lavie. METEOR : An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics*, Trento, Italy, April 2006.

Chris Callison-Burch, Cameron Shaw Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio, June 2008. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, Athens, Greece, March 2009. Association for Computational Linguistics.

Chris Callison-Burch, Philipp Koehn, Christof Monz, Kay Peterson, Mark Przybocki, and Omar Zaidan. Findings of the 2010 joint workshop on statistical machine translation and metrics for machine translation. In *Proceedings of the Joint Fifth Workshop on Statistical Machine Translation and MetricsMATR*, pages 17–53, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. Shared task: Statistical machine translation between European languages. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*, pages 119–124, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics.

Philipp Koehn. A process study of computeraided translation. *Machine Translation*, 23(4):241–263, November 2009.

Philipp Koehn. Enabling monolingual translators: Post-editing vs. options. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of*

the Association for Computational Linguistics, pages 537–545, Los Angeles, California, June 2010. Association for Computational Linguistics.

Philipp Koehn. *Statistical Machine Translation*. Cambridge University Press, 2010.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An online lexical database. Technical Report CSL 43, Cognitive Science Laboratory Princeton University, 1993.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association of Computational Linguistics (ACL)*, 2002.

Matthew Snover, Bonnie J. Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. A study of translation edit rate with targeted human annotation. In *5th Conference of the Association for Machine Translation in the Americas (AMTA)*, Boston, Massachusetts, August 2006.

D. Vilar, J. Xu, L. F. D'Haro, , and H. Ney. Error analysis of machine translation output. In *International Conference on Language Resources and Evaluation (LREC)*, pages 697–702, May 2006.

Notes

1 https://www.mturk.com/