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

Christian Federmann

How can we measure machine translation quality?

Abstract

In this opinion paper, we describe our research work on machine translation evaluation approaches that include mechanisms for human feedback and are designed to allow partial adaptation of the translation models which are being evaluated. While there exists a plethora of different automatic evaluation metrics for machine translation, their output in terms of scores, distances, etc. quite often is neither transparent to translators nor shows good correlation with manual evaluation by human experts.

Even worse, machine translation tuning efforts based on these automatic metrics to a certain extent move the research focus into a wrong direction; shifting it from « good » translations to those with a « higher scores ». This further widens the gap between machine translation research and translation producers or users.

We first describe several automatic metrics which are being used in current machine translation research. Afterwards we provide a brief overview on manual evaluation techniques which are used in our machine translation group. As minimum error rate training for tuning of (statistical) machine translation system is an important part of the workflow, we think that a (semi-) automatic implementation of such evaluation tasks would be a helpful extension of current state-of-the-art machine translation systems. We conclude by describing the need to shift from automated metrics to consumer-oriented, semi-automatic evaluation as this seems to be highly important to allow more advanced MT techniques to see wider acceptance and usage in real life applications.

Introduction

Machine translation of written text is a daunting task. The given input text has to be segmented into sentences or smaller units, then handcrafted or automatically induced "rules" transfer these into the target language where some kind of generation or language modeling is applied to ensure fluent output. Even this very simplified view on the machine translation workflow shows that there are several components in any machine translation system which need to be tuned in order to achieve optimal performance.

There exist different approaches to the machine translation task. Initial research and development work has been centered around manually constructed, rule-based systems which modeled the transfer based on linguistic knowledge and human expertise. While, in theory, this paradigm allows for proper handling of unseen input data (as long as it can be properly analysed) and guarantees well-formed translation results, practical implementations of rule-based systems proved to be expensive to create and hard to

maintain. Typical representatives of this class of machine translation systems are e.g. Systran [Senellart, J. (2001)] or Lucy RBMT [Alonso, J. A. (2003)].

Driven by increasing computational power, development of purely statistical machine translation methods started. For these, only minimal human expertise is necessary. Instead, statistical systems rely on huge parallel corpora from which they extract parallel phrase pairs that can then be used to determine possible translation options for any given input text. Due to the fact that many potential translations (including inappropriate or even wrong ones) can be generated using such a statistical decoder approach, it is of utmost importance to rank or score the various translation options both to guarantee computationally feasible translation complexity and good translation output quality. The scoring of phrases directly relates to measuring machine translation quality. One of the most renowned toolkits for statistical machine translation is the Moses decoder [Koehn, P. (2007)].

Next to the aforementioned rule-based and statistical approaches, there also exist other machine translation paradigms; evaluation of translation output quality is however crucial for all of them. A nicely written summary of different methods for (statistical) machine translation of written text can be found in [Lopez, A. (2008)].

The “BLEU Dilemma”

Statistical machine translation systems rely on automated evaluation metrics such as BLEU [Papineni, K. (2001)] which allow them to score different translation options during system training and tuning. While there exists a plethora of different automatic evaluation metrics for machine translation, their output in terms of scores, distances, etc. quite often is neither transparent to translators nor shows good correlation with manual evaluation by human experts. Hence, automated evaluation metrics have been a topic of active research for quite some time now, e.g. [Callison-Burch, C. (2007)].

Even worse, machine translation tuning efforts based on these automatic metrics to a certain extent move the research focus into a wrong direction; shifting it from "good" or "useful" translations to "translation output with higher scores". Theoretically, this further widens the gap between machine translation research and consumers such as translation producers or end users. In a sense, this can be named the "BLEU dilemma".

In order to circumvent the aforementioned problem, machine translation research has to change its evaluation approach and also should take into account actual needs and requirements with regard to machine translation output from customers such as translation professionals, industry and end users alike. Wherever possible, human feedback and evaluation results should be integrated into machine translation systems to ensure that these user-centric requirements can be properly addressed by the tools.

Interestingly, machine translation research faced a similar challenge during the shift to statistical approaches. Before, mostly manual translation evaluation was used to rate translation quality and to improve machine translation systems. As statistical methods

required huge amounts of quick evaluation decisions for training and tuning, it became apparent that automated metrics needed to be designed and implemented to allow efficient systems to be built. While these metrics contributed to the successful evolution of statistical approaches, their (de-facto) exclusive use now seems to become a bottleneck for future improvements.

Evaluation Methods

In this section, we will briefly mention and describe existing automated evaluation metrics and propose several human evaluation methods which have been used for quality estimation within DFKI's language technology lab in our machine translation projects. Finally, we provide a comparison of the different metrics.

Automated Evaluation Metrics

1. Word error rate (WER) is derived from the Levenshtein distance, working at word level and can be used to estimate the quality of machine translation output.

2. Translation Error Rate (TER) is an error metric for machine translation that measures the number of edits required to change a system output into one of the references. For a more detailed description, see [Snover, M. (2006)].

3. Most papers on machine translation report BLEU scores in their evaluation sections. It has become the de-facto standard for automated evaluation of machine translation quality. In a nutshell, the metric computes the n-gram overlap of the translation result and a given reference translation. There exist other implementations that consider more than one reference translation to allow for a certain flexibility regarding choice of words. BLEU is designed to achieve a good correlation with human judgement on a corpus level.

4. Based on the BLEU score, NIST also computes how informative a particular n-gram in the translation candidate is, giving it more weight if it is rare. Also, the brevity penalty computation differs slightly as small variations in translation length do not impact the overall score as much as in BLEU. The basic problem still remains the same: bad correlation with human judgement on the sentence level.

5. The METEOR metric [Lavie, A. (2007)] for evaluation of machine translation output is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision. It can also make use of features such as stemming and synonymy matching which are not present in other metrics. Contrary to BLEU which aims to achieve good correlation with human judgement at the corpus level, METEOR was designed to produce a good correlation at the sentence or segment level.

While automated metrics are usually fast to compute (which is one of their main advantages over manual evaluation), the interpretation of increased scores is not always a simple task, especially for very small improvements of the respective score. If a translation is a valid sentence but does not contain some of the reference words, it will

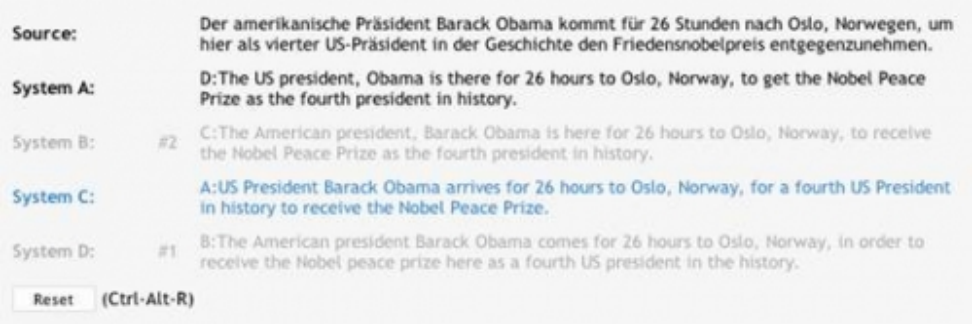
not get an optimal score. As there are many possibilities to generate valid translations for a given source sentence, it seems clear that methods based on n-gram overlap may not be the best evaluation metrics.

Human Evaluation Metrics

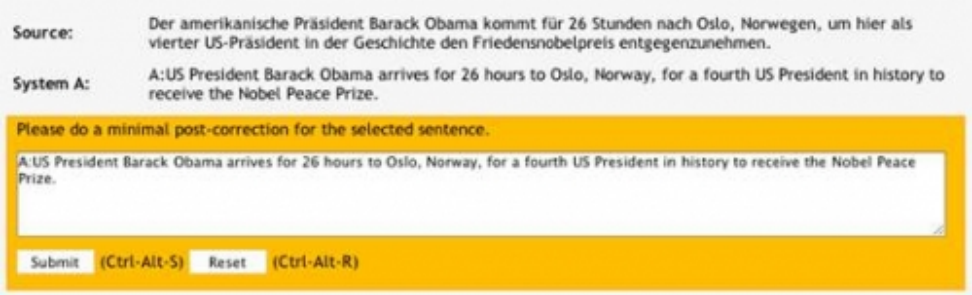
We have defined and implemented several manual evaluation methods in machine translation research at DFKI's language technology lab. Internally, we use an updated version of our annotation tool Appraise [Federmann, C. (2010)] which allows several annotators to quickly create evaluation results for given translation output.

1. **Sentence-Based Ranking**: the most obvious task for manual evaluation of machine translation output is ranking of full sentences. Here, the annotator is shown a set of two (or more) translations for a given source sentence and then has to decide which of then is the best translation. This evaluation method has also previously been used for the shared workshops on machine translation (WMT).

The following screenshot shows the Appraise interface for sentence ranking which allows the user to click on a translation to rank it as #1, #2, etc.



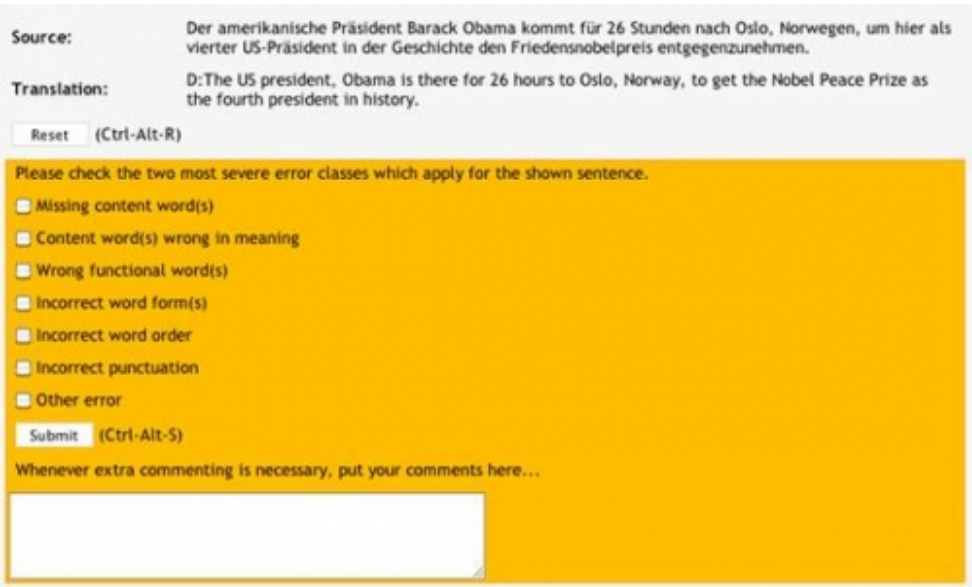
2. Post-Editing: machine translation output can also be post-edited. The resulting data then allows to improve the underlying machine translation system or can be used for more detailed error analysis. A post-editing task shows a translation and the corresponding reference translation to the annotator who then has to transform the given translation into an "acceptable" paraphrase of the reference translation. The amount of post-editing depends on task specific parameters, sometimes it is enough to create a rough translation which conveys the meaning of the sentence, sometimes post-editing even includes stylistic changes and grammar improvements. Our post-editing interface is shown in the following figure.



2. **Phrase-Based Ranking**: depending on the diversity of the translation candidates, ranking on sentence level can become a time consuming and difficult task. Often, machine translation output is neither well structured nor grammatical which makes it hard to determine which of the given two candidates is the better one. For more than two sentences, the complexity gets even worse.

To alleviate the effects of this, we have implemented so called phrase-based ranking in our evaluation tool. The translation candidates are first aligned to each other and then segmented into shared and different phrases. The annotator is shown the source sentence, a reference translation and the two candidates. Instead of ranking on the sentence level, the annotator then has to rank the phrasal differences which allows to gain insights on local problems within the machine translation systems. In our experiments we found that these local ranking decisions seemed to be easier to undertake for annotators.

4. Error Classification: for a more detailed error analysis, we have also created an evaluation task in which annotators are requested to classify the errors in the given translation. We use a task specific classification scheme and allow the submission of free text comments in case of more complex problems. A screenshot of the error classification interface is shown below.



Improved Systems by Advanced Evaluation

Better evaluation of machine translation output can directly feed into the improvement of machine translation. In this section, we briefly sketch how advanced evaluation techniques could be used to increase the translation quality of existing systems.

Integrating Human Feedback into the Workflow

Very often, machine translation systems are trained and tuned for specific tasks only. A re-training of the same system using slightly different training data can already result in vastly different scores and hence a different machine translation system. In a sense, the two systems are not comparable to each other. We believe that machine translation models have to be altered in a way that allows to continuously integrate feedback into them. This would result in a better comparability and also allow to include human feedback into the update process.

Manual evaluation efforts can be used to give higher scores to phrases which achieved a good ranking or to exclude translation candidates which did not match the source phrase. Even shallow semantics and context modeling could be implemented provided suitable training data collected from human annotators could be made available. While it is clear that full manual evaluation of the huge corpora used by statistical machine translation systems is not at all feasible, the development of incremental models for machine translation which can make best use of small sets of annotated information seems a promising research goal.

Human-aided bootstrapping of annotated training data for machine translation via crowd-sourcing or networked cloud computing applications for large social networks could also help to further improve state-of-the-art machine translation systems. After all, human evaluation is still the best possible training material.

Challenges for (Semi-) Automatic Tuning

As we have already stated, full manual evaluation is not possible. For training and tuning of machine translation, however, the inclusion of human annotations is an extremely desirable extension as automated metrics still may have questionable correlation with human judgements. For further improvement, statistical machine translation systems need to get access to context-based semantics and task specific tuning.

Hence, research on semi-automatic metrics which integrate statistical methods and human annotation is an important area of future work. The resulting hybrid evaluation metrics can then be used to improve existing machine translation systems and may also lead to advanced models for machine translation.

Conclusion and Outlook

In this paper, we have described and discussed automated and manual metrics for the evaluation of machine translation output. We believe that a deeper integration of manual evaluation techniques into the tuning and re-training of statistical machine translation systems or system combination approaches is a desirable extension of current state-of-the-art machine translation approaches.

Local Model Adaptation from Human Feedback

We plan to further investigate how human feedback from the manual evaluation of both machine translation output and related information such as phrase table and alignment data can be utilised to achieve local adaptation of the translation models, hopefully resulting in improved translation performance. Especially for huge training corpora, it is however clear that manual efforts can only support the automated processes. It will be a challenging task to alter the translation model underlying current statistical machine translation systems such that small but focused contributions of human feedback can be exploited in a way that preserves existing translation quality and helps to overcome the respective model's deficiencies.

Previous work on stream-based translation [Levenberg, A. (2010)] and language modeling [Levenberg, A. (2009)] did already show that incremental models for machine translations can be used with comparable performance and translation quality. By adding results from manual evaluation to such a model, we want to avoid the re-training of machine translation systems and instead move to directed (in a sense "local") improvements of their knowledge base. By doing so, we also think that the resulting "versions" of the incremental translation model could be compared in a more meaningful manner, hopefully making the use of statistical machine translation tools more interesting in areas where traditionally rule-based systems have been used so far.

Focus Shifts: Research to Industry, Automated to Semi-Automatic

In our introduction we stated that machine translation research has moved from creating good translations to improving the scores of the phrases inside their knowledge bases using automated metrics. We think that this is a problem for future research efforts and usage of machine translation tools by industry and end users alike. Hence we want to shift focus from pure research activities to a better understanding and consideration of consumer desiderata. Even statistical approaches can benefit from the inclusion of linguistic information or translation techniques from human translation professionals as recent work on hierarchical models or more linguistically driven models has shown. As the underlying translation models are improved, there is also the need to improve and adapt methods for evaluation of their quality.

Machine translation has reached a good level of quality and acceptance by industry and end users. However, translation quality is not going to improve as long as automated metrics are considered the main choice for evaluation of translation output. While the actual scores might (and likely will) improve, their meaning and correlation to the actual translation quality suggest that the reported improvements may not relate to human

judgement of the translation output. As robust (and meaningul) evaluation metrics are needed in order to increase the usage of machine translation technology in industry, the usage of automated metrics has to be replaced by hybrid approaches that bring in as much human knowledge as possible.

The integration of manual evaluation efforts into the machine translation workflow represents one of the most important challenges within the next years. We are confident that continued work in this area will eventually result in better machine translation approaches and to improved translation quality.

Bibliography

Alonso, Juan A. and Thurmair, Gregor (2003) « The Comprendium Translator System » in Proceedings of the Ninth Machine Translation Summit.

Callison-Burch, Chris and Fordyce, Cameron and Koehn, Philipp and Monz, Christof and Schroeder, Josh (2007), « (Meta-) Evaluation of Machine Translation » in Proceedings of the Second Workshop on Statistical Machine Translation, pp. 136-158.

Federmann, Christian (2010), « Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations » in Proceedings of the Seventh Conference on International Language Resources and Evaluation.

Koehn, Philipp and Hoang, Hieu and Birch, Alexandra and Callison-Burch, Chris and Federico, Marcello and Bertoldi, Nicola and Cowan, Brooke and Shen, Wade and Moran, Christine and Zens, Richard and Dyer, Chris and Bojar, Ondrej and Constantin, Alexandra and Herbst, Evan (2007) « Moses: Open Source Toolkit for Statistical Machine Translation » in Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session.

Lavie, Alon and Agarwal, Abhaya (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, pp. 228-231.

Levenberg, Abby and Osborne, Miles (2009), « Stream-based Randomised Language Models for SMT » in Proceedings EMNLP-2009.

Levenberg, Abby and Callison-Burch, Chris and Osborne, Miles (2010), « Stream-based Translation Models for Statistical Machine Translation » in Proceedings NAACL-2010.

Lopez, Adam (2008), « Statistical Machine Translation » in ACM Computing Surveys 40(3): Article 8, pp. 1-49.

Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing (2001), « BLEU: a Method for Automatic Evaluation of Machine Translation » in IBM Research Report RC22176(W0109-022).

Senellart, Jean and Dienes, Péter and Váradi, Tamás (2001), « New generation Systran Translation System » in Proceedings of the Ninth Machine Translation Summit.

Snover, Matthew and Dorr, Bonnie and Schwartz, Richard and Micciulla, Linnea and Makhoul, John (2006), « A Study of Translation Edit Rate with Targeted Human Annotation » in Proceedings of Association for Machine Translation in the Americas, pp. 223-231.