Measurement of Progress in Machine Translation

Yvette Graham **Timothy Baldwin** **Aaron Harwood** **Alistair Moffat** **Justin Zobel**
Department of Computing and Information Systems
The University of Melbourne
{ygraham,tbaldwin,aharwood,amoffat,jzobel}@unimelb.edu.au

Abstract

Machine translation (MT) systems can only be improved if their performance can be reliably measured and compared. However, measurement of the quality of MT output is not straightforward, and, as we discuss in this paper, relies on correlation with inconsistent human judgments. Even when the question is captured via "is translation A better than translation B" pairwise comparisons, empirical evidence shows that inter-annotator consistency in such experiments is not particularly high; for intra-judge consistency – computed by showing the same judge the same pair of candidate translations twice – only low levels of agreement are achieved. In this paper we review current and past methodologies for human evaluation of translation quality, and explore the ramifications of current practices for automatic MT evaluation. Our goal is to document how the methodologies used for collecting human judgments of machine translation quality have evolved; as a result, we raise key questions in connection with the low levels of judgment agreement and the lack of mechanisms for longitudinal evaluation.

1 Introduction

Measurement is central to all scientific endeavor. In computing, we rely on impartial and scrutable evaluations of phenomena in order to determine the extent to which progress is being made in that discipline area. We then use those measurements to predict performance on unseen data. That is, we need accurate measurement to know that we have made progress, and we need those measurements to be predictive, so that we can have confidence that we will benefit from the improvements that have been attained. The particular focus of this paper is measurement of translation quality in the field of machine translation (MT).

In some areas of computing, measurement techniques are unambiguous, directly comparable between systems, and enduring over time. For example, a proposed new approach to text compression can be evaluated on a wide range of files, and the criteria to be measured in each case are straightforward: execution time for encoding and decoding; memory space used during encoding and decoding; and, of course, compressed file size. All of these facets are objective, in that, if the same file is compressed a second time on the same hardware, the same measurements (to within some predictable tolerance, in the case of execution speed) will result; and compressing the same file with the same technique on different hardware ten years later should still result in consistent measures of memory use and file size. To compare two approaches to text compression, therefore, the only real complexity is in assembling a collection of documents which is "representative" of utility in general or over some specific domain (for example, compression of microposts from a service such as Twitter). Beyond this, as long as the evaluation is carried out using a fixed computing environment (OS, hardware, and, ideally, programming environment), establishing the superiority of one method over another is clear-cut and predictivity is high.

In other areas of computing, the measurement

Yvette Graham, Timothy Baldwin, Aaron Harwood, Alistair Moffat and Justin Zobel. 2012. Measurement of Progress in Machine Translation. In *Proceedings of Australasian Language Technology Association Workshop*, pages 70−78.

techniques used are, of necessity, more subjective, and predictivity is harder to achieve. Areas that often require subjective human judgments for evaluation are those where the work product is for human consumption, such as natural language processing (NLP) and information retrieval (IR). In IR, systems are measured with reference to subjective human relevance judgments over results for a sample set of topics; a recent longitudinal study has indicated that, despite a considerable volume of published work, there is serious question as to the extent to which actual long-term improvements in effectiveness have been attained (Armstrong et al., 2009). Moreover, while it is possible to achieve predictivity through the use of a fixed set of topics, a fixed document collection, and a static set of relevance judgments (often based on pooling (Voorhees and Harman, 2005)), the set of topics is often small and not necessarily representative of the universe of possible topics, which raises concerns about true predictivity.

The work of Armstrong et al. (2009) raises another important question, one that is relevant in all fields of computing: that any experimentation carried out today should, if at all possible, also lay the necessary groundwork to allow, ten years hence, a retrospective evaluation of "have we made quantifiable progress over the last decade?"

2 Automatic Measurement of MT

The automatic evaluation of MT system output has long been an objective of MT research, with several of the recommendations of the early ALPAC Report (ALPAC, 1966), for example, relating to evaluation:

> 1. Practical methods for evaluation of translations; ... 3. Evaluation of quality and cost of various sources of translations;

In practical terms, improvements are often established through the use of an automatic measure that computes a similarity score between the candidate translation and one or more human-generated reference translations. However it is well-known that automatic measures are not necessarily a good substitute for human judgments of translation quality, primarily because:

- There are different valid ways of translating the same source input, and therefore comparison with a single or even multiple references risks ranking highly those translations that happen to be more reference-like compared to those that made different choices; and
- There are different ways to compute the syntactic similarity between a system output translation and reference translations, and given two possible system translations for a source input, different measures can disagree on which output is more similar to the set of reference translation.

Moreover, with any mechanical method of measurement, there is a tendency for researchers to work to improve their MT system's ability to score highly rather than produce better translations.

To alleviate these concerns, direct human judgments of translation quality are also collected when possible. During the evaluation of MT shared tasks, for example, human judgments of MT outputs have been used to determine the ranking of participating systems. The same human judgments can also be used in the evaluation of automatic measures, by comparing the degree to which automatic scores (or ranks) of translations correlate with them. This aspect of MT measurement is discussed shortly.

One well-known example of an automatic metric is the BLEU (bilingual evaluation understudy) score (Papineni et al., 2002). Computation of a BLEU score for a system, based on a set of candidate translations it has generated, requires only that sets of corresponding reference translations be made available, one per candidate. The ease – and repeatability – of such testing has meant that BLEU is popular as a translation effectiveness measure. But that popularity does not bestow any particular superiority, and, BLEU suffers from drawbacks (Callison-Burch et al., 2006). (As an aside, we note that in all such repeatable scoring arrangements, every subsequent experiment must be designed so that there is clear separation between training and test data, to avoid any risk of hill-climbing and hence over-fitting.)

3 Human Assessment in MT

The standard process by which researchers have tested automatic MT evaluation measures is through analysis of correlation with human judgments of MT quality, as depicted in Figure 1.

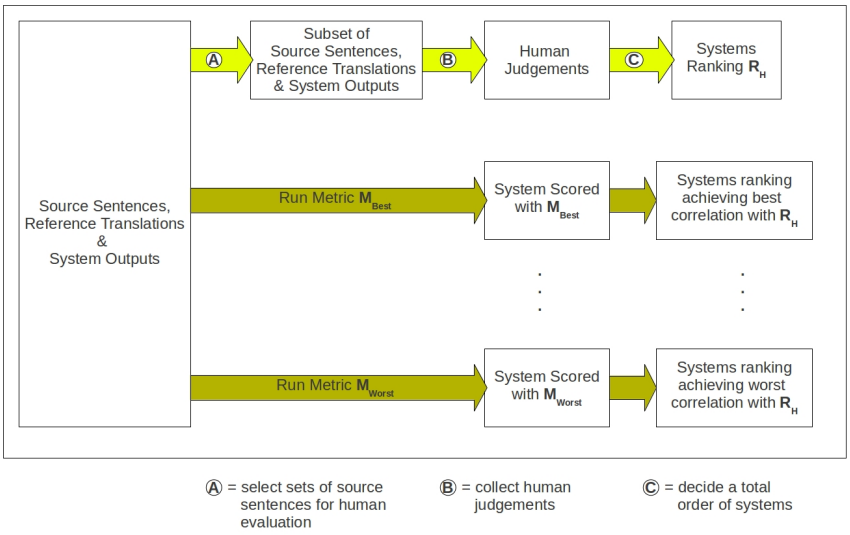

Figure 1: The process by which human assessment is used to confirm (or not) automatic MT evaluation measures.

In this process, a suite of different MT systems are each given the same corpus of sentences to translate, across a variety of languages, and required to output a 1-best translation for each input in the required target language. Since the total number of translations in the resulting set is too large for exhaustive human assessment, a sample of translations is selected, and this process is labeled A in Figure 1. To increase the likelihood of a fair evaluation, translations are selected at random, with some number of translations repeated, to facilitate later measurement of consistency levels.

Label B in Figure 1 indicates the assessment of the sample of translations by human judges, where judges are required to examine translated sentences, perhaps several at a time, and assess their quality. It is this issue in particular that we are most concerned with: to consider different possibilities for acquiring human judgments of translation quality in order to facilitate more consistent assessments.

Once sufficient human judgments have been collected, they are used to decide a best-to-worst ranking of the participating machine translation systems, shown as R_H in Figure 1. The process of computing that ranking is labeled C. The best approach to process C, that is, going from raw human judgments to a total-order rank, to some degree still remains an open research question (Bojar et al., 2011; Lopez, 2012; Callison-Burch et al., 2012), and is not considered further in this paper.

Once the suite of participating systems has been ordered, any existing or new automatic MT evaluation metric can be used to construct another ordered ranking of the same set. The ranking generated by the metric can then be compared with the ranking R_H generated by the human assessment, using statistics such as Spearman's coefficient, with a high correlation being interpreted as evidence that the metric is sound.

Since the validity of an automatic MT evaluation measure is assessed relative to human judgments, it is vital that the judgments acquired are reliable. In practice, however, human judgments, as evaluated by intra- and inter-annotator agreement, can be inconsistent with each other. For example, inter-annotator agreement for human assessment of translation quality, as measured using Cohen's Kappa coefficient (Cohen, 1960), in recent WMT shared tasks are reported to be at as low levels as $k = 0.44$ (2010), $k = 0.38$ (2011) and $k = 0.28$ (2012), with intra-annotator agreement levels not faring much better: $k = 0.60$ (2010), $k = 0.56$ (2011) and $k = 0.46$ (2012) (Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012). This lack of coherence amongst human assessments then forces the question: *are assessments of MT evalu-*

ation metrics robust, if they are validated via low-quality human judgments of translation quality?

While one valid response to this question is that the automatic evaluation measures are no worse than human assessment, a more robust approach is to find ways of increasing the reliability of the human judgments we use as the yard-stick for automatic metrics by endeavoring to find better ways of collecting and assessing translation quality. Considering just how important human assessment of translation quality is to empirical machine translation, although there is a significant amount of research into developing metrics that correlate with human judgments of translation quality, the underlying topic of finding ways of increasing the reliability of those judgments to date has received a limited amount of attention (Callison-Burch et al., 2007; Callison-Burch et al., 2008; Przybocki et al., 2009; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Denkowski and Lavie, 2010).

4 Human Assessment of Quality

To really improve the consistency of the human judgments of translation quality, we may need to take a step back and ask ourselves *what are we really asking human judges to do when we require them to assess translation quality?* In the philosophy of science, the concept of *translation quality* would be considered a (hypothetical) *construct*. MacCorquodale and Meehl (1948) describe a construct as follows:

> ... constructs involve terms which are not wholly reducible to empirical terms; they refer to processes or entities that are not directly observed (although they need not be in principle unobservable); the mathematical expression of them cannot be formed simply by a suitable grouping of terms in a direct empirical equation; and the truth of the empirical laws involved is a necessary but not a sufficient condition for the truth of these conceptions.

Translation quality is an abstract notion that exists in theory and can be observed in practice but cannot be measured directly. Psychology often deals with the measurement of such abstract notions, and provides established methods of measurement and validation of those measurement techniques. Although "translation quality" is not a psychological construct as such, we believe these methods of measurement and validation could be used to develop more reliable and valid measures of translation quality.

Psychological constructs are measured indirectly, with the task of defining and measuring a construct known as *operationalizing* the construct. The task requires examination of the mutual or commonsense understanding of the construct to come up with a set of items that together can be used to indirectly measure it. In psychology, the term *construct validity* refers to the degree to which inferences can legitimately be made from the operationalizations in a study to the theoretical constructs on which those operationalizations were based.

Given some data, it is possible then to examine each pair of constructs within the semantic net, and evidence of convergence between theoretically similar constructs supports the inclusion of both constructs (Campbell and Fiske, 1959). To put it more simply, when two theoretically *similar* constructs that *should* (in theory) relate to one another do in fact *highly correlate* on the data, it is evidence to support their use. Similarly, when a *lack of correlation* is observed for a pair of constructs that theoretically *should not* relate to each, this also validates their use. This is just one example of a range of methods used in psychology to validate techniques used in the measurement of psychological constructs (see Trochim (1999) for a general introduction to construct validity).

5 Past and Current Methodologies

The ALPAC Report (ALPAC, 1966) was one of the earliest published attempts to perform cross-system MT evaluation, in determining whether progress had been made over the preceding decade. The (somewhat anecdotal) conclusion was that:

> (t)he reader will find it instructive to compare the samples above with the results obtained on simple, selected, text 10 years earlier ... in that the earlier samples are more readable than the later ones.

The DARPA Machine Translation Initiative of the 1990s incorporated MT evaluation as a central tenet, and periodically evaluated the three MT

systems funded by the program (CANDIDE, PANGLOSS and LINGSTAT). It led to the proposal of *adequacy* and *fluency* as the primary means of human MT evaluation, in addition to human-assisted measurements. For instance, the DARPA initiative examined whether post-editing of MT system output was faster than simply translating the original from scratch (White et al., 1994). Adequacy is the degree to which the information in the source language string is preserved in the translation,[1] while fluency is the determination of whether the translation is a well-formed utterance in the target language and fluent in context.

Subsequently, many of the large corporate machine translation systems used regression testing to establish whether changes or new modules had a positive impact on machine translation quality. Annotators were asked to select which of two randomly-ordered translations (one from each system) they preferred (Bond et al., 1995; Schwartz et al., 2003), and this was often performed over a reference set of translation pairs (Ikehara et al., 1994). While this methodology is capable of capturing longitudinal progress for a given MT system, it is prohibitively expensive and doesn't scale well to multi-system comparison.

The annual workshop for statistical machine translation (WMT) has, over recent years, been the main forum for collection of human assessment of translation quality, despite this not being the main focus of the workshop (which is to provide a regular cross-system comparison over standardized datasets for a variety of language pairs by means of a shared translation task) (Koehn and Monz, 2006; Callison-Burch et al., 2007; Callison-Burch et al., 2008; Callison-Burch et al., 2009; Callison-Burch et al., 2010; Callison-Burch et al., 2011; Callison-Burch et al., 2012). Figure 2 shows the approaches used for human judgments of translation quality at the annual workshops.

To summarize, across the field of machine translation human judges have been asked to assess translation quality in a variety of ways:

- Single-item or two-items (for example, *fluency* and *adequacy* being a two-item assessment);
- Using different labels (for example, asking which translation is *better* or asking which is more *adequate*);
- Ordinal level scales (ranking a number of translations from best-to-worst) or interval-level scales (for example, interval-level fluency or adequacy judgments);
- Different lexical units (for example, whole sentences rather than sub-sentential constituents);
- Different numbers of points on interval-level scale;
- Displaying interval-level scale numbering to judges or not displaying it;
- Simultaneously judging fluency and adequacy items or separating the assessment of fluency and adequacy;
- Displaying a reference translation to the judge or not displaying it;
- Including the reference translation present among the set being judged or not including it;
- Displaying a preceding and following context of the judged translation or not displaying any surrounding context;
- Displaying session/overall participation meta-information to the judge (for example, the number of translations judged so far, the time taken so far, or the number of translations left to be judged) or not displaying session meta-information;
- Allowing judges to assess translations that may have originated with their own system versus holding out these translations;
- Including crowd-sourced judgments or not.

5.1 Pre 2007 Methodologies

A widely used methodology for human evaluation of MT output up to 2007 was to assess translations under the two items, fluency and adequacy, each on a five-point scale (Callison-Burch et al., 2007). Fluency and adequacy had originally been part of the US Government guidelines for assessment of manually produced translations and was adopted by DARPA for the evaluation of machine translation output, as the fact that these established criteria had originally been designed for the more general purpose of grading translators helped validate their use (White et al., 1994).

When WMT began in 2006 the fluency and adequacy measures were again adopted, as had also

[1] Or, in the case of White et al. (1994), the degree to which the information in a professional *translation* can be found in the translation, as judged by monolingual speakers of the target language.

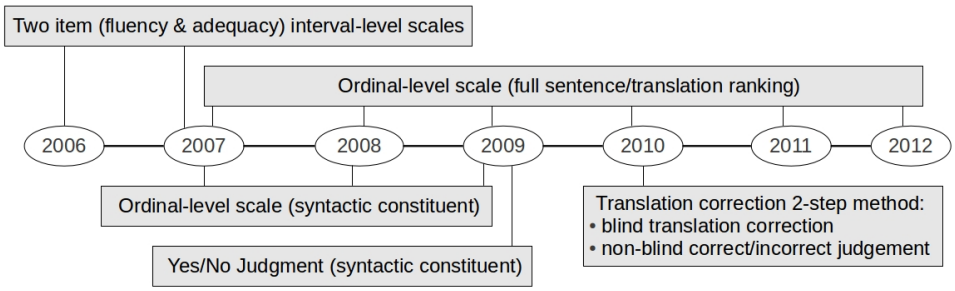

Figure 2: Methodologies of human assessment of translation quality at statistical machine translation workshops

been used in LDC (2005), to assess output of shared task participating systems in the form of a two item interval-level scale. Too few human assessments were recorded in the first year to be able to estimate the reliability of human judgments (Koehn and Monz, 2006). In 2007, the workshop sought to better assess the reliability of human judgments in order to increase the reliability of results reported for the shared translation task. Reliability of human judgments was estimated by measuring levels of agreement as well as adding two new supplementary methods of human assessment. In addition to asking judges to measure the fluency and adequacy of translations, they were now also requested in a separate evaluation set-up to rank translations of full sentences from best-to-worst (the method of assessment that has been sustained to the present), in addition to ranking translations of sub-sentential source syntactic constituents.[2] Both of the new methods used a single item ordinal-level scale, as opposed to the original two item interval-level fluency and adequacy scales.

Highest levels of agreement were reported for the sub-sentential source syntactic constituent ranking method ($k_{\text{inter}} = 0.54$, $k_{\text{intra}} = 0.74$), followed by the full sentence ranking method ($k_{\text{inter}} = 0.37$, $k_{\text{intra}} = 0.62$), with the lowest agreement levels observed for two-item fluency ($k_{\text{inter}} = 0.28$, $k_{\text{intra}} = 0.54$) and adequacy ($k_{\text{inter}} = 0.31$, $k_{\text{intra}} = 0.54$) scales. Additional methods of human assessment were trialled in subsequent experimental rounds; but the only method still currently used is ranking of translations of full sentences.

When the WMT 2007 report is revisited, it is difficult to interpret reported differences in levels of agreement between the original fluency/adequacy method of assessment and the sentence ranking method. Given the limited resources available and huge amount of effort involved in carrying out a large-scale human evaluation of this kind, it is not surprising that instead of systematically investigating the effects of individual changes in method, several changes were made at once to quickly find a more consistent method of human assessment. In addition, the method of assessment of translation quality is required to facilitate speedy judgments in order to collect sufficient judgments within a short time frame for the overall results to be reliable, an inevitable trade-off between bulk and quality must be taken into account. However, some questions remain unanswered: *to what degree was the increase in consistency caused by the change from a two item scale to a single item scale and to what degree was it caused by the change from an interval level scale to an ordinal level scale?* For example, it is wholly possible that the increase in observed consistency resulted from the combined effect of a reduction in consistency (perhaps caused by the change from a two item scale to a single item scale) with a simultaneous increase in consistency (due to the change from an interval-level scale to an ordinal-level scale). We are not suggesting this is in fact what happened, just that an overall observed increase in consistency resulting from multiple changes to method cannot be interpreted as each individual alteration causing an increase in consistency. Although a more consistent method of human assessment was indeed found, we cannot be at all certain of the reasons behind the improvement.

[2]Ties are allowed for both methods.

A high correlation between fluency and adequacy across all language pairs included in the evaluation is also reported, presented as follows (Callison-Burch et al., 2007):

> ..., in principle it seems that people have a hard time separating these two aspects (referring to fluency and adequacy) of translation. The high correlation between people's fluency and adequacy scores ... indicates that the distinction might be false.

The observed high correlation between fluency and adequacy is interpreted as a negative. However, in the field of psychology according to construct validity, an observed high correlation between two items that in theory should relate to each other is interpreted as evidence of the measure in fact being valid (see Section 4), and there is no doubt that in theory the concepts of fluency and adequacy do relate to each other. Moreover, in general in psychology, a measure that employs more items as opposed to fewer (given the validity of those items), is regarded as better.

In addition, human judges were asked to assess fluency and adequacy at the same time, and this could have inflated the observed correlation. A fairer examination of the degree to which fluency and adequacy of translations correlate, would have judges assess the two criteria of translations on separate occasions, so that each judgment could be made independently of the other. Another advantage of judging fluency and adequacy separately might be to avoid revealing the reference translation to judges before they make their fluency assessment. A fluency judgment of translations without a reference translation would increase the objectivity of the assessment and avoid the possibility of a bias in favor of systems that produce reference-like translations.

Confusion around how well fluency and adequacy can be used to measure translation quality, to some degree may stem from the implicit relationship between the two notions. For instance, does the adequacy of a translation imply its fluency, and, if so, why would we want to assess translations under both these criteria? However, the argument for assessing adequacy on its own and dropping fluency, only stands for translations that are fully fluent. The fluency of a translation judged to be fully adequate can quite rightly be assumed. However, when the adequacy of a translation is less than perfect, very little can be assumed from an adequacy judgment about the fluency of the translation. Moving from a two-item fluency/adequacy scale to a single-item scale loses some information that could be useful for analyzing the kinds of errors present in translations.

5.2 Post 2007 Methodologies

Since 2007, the use of a single item scale for human assessment of translation quality has been common, as opposed to the more traditional two item fluency/adequacy scale, sometimes citing the high correlation reported in WMT 2007 as motivation for its non-use other times not (Przybocki et al., 2009; Denkowski and Lavie, 2010). For example, Przybocki et al. (2009) use (as part of their larger human evaluation) a single item (7-point) scale for assessing the quality of translations (with the scale labeled *adequacy*) and report inter-annotator agreement of $k = 0.25$, lower than those reported for the two item fluency/adequacy scales in WMT 2007. Although caution needs to be taken when directly comparing such agreement measurements, this again raises questions about the validity of methodologies used for human assessment of translation quality.

When we look at the trend in consistency levels for human assessments acquired during the three most recent WMT shared tasks, where the only surviving method of human assessment of translation quality is full sentence ranking (or translation ranking as it is also known), we unfortunately see ever-decreasing consistency levels. Agreement levels reported in the most recent 2012 WMT using translation ranking are lower than those reported in 2007 for the two item fluency and adequacy interval-level scales. Although caution must again be taken when making direct comparisons, this may cause us to revisit our motivation for moving away from more traditional methods. In addition, due to the introduction of the new kind of shared task, quality estimation, the traditional ordinal-level scale has again resurfaced for human assessment of translation quality, although on this occasion in the guise of a 4-point scale (Callison-Burch et al., 2012). This causes us to pose the question *is the route we have chosen in the search of more reliable human assessment of translation quality really going to lead to*

an optimal method? Machine translation may benefit from a systematic investigation into which methods of human assessment of translation quality are in fact most reliable and result in most consistent judgments.

Planning for the future: Two major components of evaluation are not catered for by current approaches. The first is the value of longitudinal evaluation, the ability to measure how much improvement is occurring over time. Mechanisms that could be used include: capture of the output of systems that is not evaluated at the time; strategic re-use of evaluation data in different events; probably others. In the TREC context, a long-held belief that systems were measurably improving is not supported by longitudinal study, demonstrating the value of such mechanisms. In other contexts, longitudinal mechanisms allow meta studies that yield insights that would not otherwise be available.

Context for judgments: The other omitted component is sufficient consideration of what might be called "role", the persona that the assessor is expected to adopt as they make their decisions. An MT system that is used to determine, for example, whether a statement in another language is factually correct may be very different from one that is used to translate news for a general audience. Without understanding of role, assessors can only be given very broad instructions, and may have varying interpretations of what is expected of them. The design of such instructions needs to be considered with extreme caution, however, as a seemingly unambiguous instruction inevitably has the potential to bias judgments in some unexpected way.

6 Open Questions

Our review of approaches to MT system evaluation illustrates that a range of questions need to be asked:

- What are the effects of context and specificity of task on human assessment of translation quality?
- Can we identify the key "components" annotators draw on in evaluating translation quality? Could insights allow us to develop more reliable evaluation methodology?
- Should we reconsider how conclusions are drawn from results by taking into account the degree to which automatic metrics correlate with human judgments as well as levels of consistency of those judgments? How do these factors effect the practical significance of a result?
- What can be done to enhance the reusability of previous experimental data? Can current regimes be adapted to testing of new systems that did not originally participate in particular experimental rounds?
- Is data being collected now that would allow retrospective evaluation in ten years, to know if the state of the art has changed? Similarly, is it possible to demonstrate with the evaluation data that MT systems today are better than they were ten years ago?

7 Summary

Regular competitive evaluation of systems in a common framework has become widespread in computing, in areas as diverse as message understanding and genome assembly. However, at core these evaluations are dependent on principled, robust measurement of systems and their ability to solve particular tasks. Our review has established that there are significant issues with current approaches to measurement in MT, and should provide the basis of development of new approaches that will allow researchers to be confident of the value of different MT technologies.

Acknowledgments

This work was funded by the Australian Research Council.

References

ALPAC. 1966. Languages and machines: Computers in translation and linguistics. A report by the Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Sciences, National Research Council.

T. G. Armstrong, A. Moffat, W. Webber, and J. Zobel. 2009. Improvements that don't add up: Ad-hoc retrieval results since 1998. In *Proc. 18th ACM CIKM Int. Conf. on Information and Knowledge Management*, pages 601–610, Hong Kong, China, November.

O. Bojar, M. Ercegovčevic, M. Popel, and O. Zaidan. 2011. A grain of salt for the WMT manual evaluation. In *Proc. 6th Wkshp. Statistical Machine Translation*, pages 1–11, Edinburgh, Scotland. Assoc. Computational Linguistics.

F. Bond, K. Ogura, and S. Ikehara. 1995. Possessive pronouns as determiners in Japanese-to-English machine translation. In *Proc. 2nd Conf. Pacific Assoc. Computational Linguistics*, pages 32–38, Brisbane, Australia.

C. Callison-Burch, M. Osborne, and P. Koehn. 2006. Re-evaluating the role of BLEU in machine translation research. In *Proc. 11th Conf. European Chapter of the Assoc. Computational Linguistics*, pages 249–256, Trento, Italy, April. Assoc. for Computational Linguistics.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2007. (Meta-) evaluation of machine translation. In *Proc. 2nd Wkshp. Statistical Machine Translation*, pages 136–158, Prague, Czech Republic. Assoc. Computational Linguistics.

C. Callison-Burch, C. Fordyce, P. Koehn, C. Monz, and J. Schroeder. 2008. Further meta-evaluation of machine translation. In *Proc. 3rd Wkshp. Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Assoc. Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, and J. Schroeder. 2009. Findings of the 2009 Workshop on Statistical Machine Translation. In *Proc. 4th Wkshp. Statistical Machine Translation*, pages 1–28, Athens, Greece. Assoc. Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, K. Peterson, M. Przybocki, and O. Zaidan. 2010. Findings of the 2010 Joint Workshop on Statistical Machine Translation and Metrics for Machine Translation. In *Proc. 5th Wkshp. Statistical Machine Translation*, Uppsala, Sweden. Assoc. Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, and O. Zaidan. 2011. Findings of the 2011 Workshop on Statistical Machine Translation. In *Proc. 6th Wkshp. Statistical Machine Translation*, pages 22–64, Edinburgh, Scotland. Assoc. Computational Linguistics.

C. Callison-Burch, P. Koehn, C. Monz, M. Post, R. Soricut, and L. Specia. 2012. Findings of the 2012 Workshop on Statistical Machine Translation. In *Proc. 7th Wkshp. Statistical Machine Translation*, pages 10–51, Montreal, Canada, June. Assoc. Computational Linguistics.

D. T. Campbell and D. W. Fiske. 1959. Convergent and discriminant validation by the muultitrait-multimethod matrix. 56:81–105.

J. Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

M. Denkowski and A. Lavie. 2010. Choosing the right evaluation for machine translation: an examination of annotator and automatic metric performance on human judgement tasks. In *Proc. 9th Conf. Assoc. Machine Translation in the Americas (AMTA)*.

S. Ikehara, S. Shirai, and K. Ogura. 1994. Criteria for evaluating the linguistic quality of Japanese to English machine translations. *J. Japanese Soc. Artificial Intelligence*, 9. (in Japanese).

P. Koehn and C. Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proc. Wkshp. Statistical Machine Translation*, pages 102–121.

LDC. 2005. Linguistic data annotation specification: Assessment of fluency and adequacy in translations. Technical report, Linguistic Data Consortium. Revision 1.5.

A. Lopez. 2012. Putting human machine translation systems in order. In *Proc. 7th Wkshp. Statistical Machine Translation*, pages 1–9, Montreal, Canada, June. Assoc. Computational Linguistics.

K. MacCorquodale and P.E. Meehl. 1948. On a distinction between hypothetical constructs and intervening variables. *Psychological Review*, 55(2):307–321.

K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proc. 40th Ann. Meet. Assoc. Computational Linguistics*, pages 311–318, Philadelphia, PA, July. Assoc. Computational Linguistics.

M. Przybocki, K. Peterson, S. Bronsart, and G. Sanders. 2009. The NIST 2008 metrics for machine translation challenge – overview, methodology, metrics and results. *Machine Translation*, 23(2-3):71–103.

L. Schwartz, T. Aikawa, and C. Quirk. 2003. Disambiguation of English PP attachment using multilingual aligned data. In *Proc. 9th Machine Translation Summit (MT Summit IX)*, New Orleans, LA.

William M.K. Trochim. 1999. *The Research Methods Knowledge Base*. Cornell University Custom Publishing, Ithaca, New York.

E. M. Voorhees and D. K. Harman. 2005. *TREC: Experiment and evaluation in information retrieval*. MIT Press, Cambridge, MA.

J. S. White, T. O'Connell, and F. O'Mara. 1994. The ARPA MT evaluation methodologies: Evolution, lessons, and future approaches. In *Proc. 1st Conf. Assoc. for Machine Translation in the Americas (AMTA'94)*, pages 193–205.