

Human Language Technology: Applications to Information Access

Lesson 8: Evaluation of Machine Translation & Applications

November 17, 2016

EPFL Doctoral Course EE-724
Andrei Popescu-Belis, Idiap Research Institute

Plan of the lesson

- The problem of evaluating MT
 - practical exercise
- Metrics of MT output quality
 - metrics applied by human judges
 - automatic metrics
- A user-oriented view of MT evaluation
 - a multitude of quality aspects & FEMTI
- Applications of MT
 - main types of use & examples

Examples of online MT output: which one is better?

- **Source sentence**
Les résultats d'études récentes le démontrent clairement : plus la prévention commence tôt, plus elle est efficace.
- **Systran Pure Neural MT (NMT, on Nov. 16, 2016)**
The results of recent studies clearly demonstrate this: the more prevention starts early, the more effective it is.
- **Google Translate (PBSMT?, on Nov. 16, 2016)**
The results of recent studies clearly demonstrate this: the earlier the prevention begins, the more effective it is.
- **Systran box (*direct*)**
The results of recent studies show it clearly: the more the prevention starts early, the more it is effective.
- **Metal / L&H T1 / Compendium (*transfer*)**
The results of recent studies demonstrate it clearly: the earlier the prevention begins, the more efficient it|she is.

What does “better” mean?

- Hands-on exercise: rate four FR/EN translations
 - look at the human translation if you don't understand French
- 1. Intuitive approach: “feeling of translation quality”
- 2. Analytical approach: estimate sentence by sentence
 - translation quality
 - fluency in English
 - effort required for correction
- Synthesis of observed results
 - is there a general agreement among us on ranking?

- Like you, we are convinced that the prevention of dependence begins at home, through the relationship between adults and children. This is done through reinforcing the child's self-esteem.
- The findings of recent studies clearly show that the earlier prevention starts, the more efficient it will be.
- You do not necessarily need to be an expert in drug dependence to talk about this issue with your children.

Evaluation of MT by human judges (1)

- Fluency
 - is the sentence acceptable (well-formed) in the target language?
 - i.e. is it good French, English, etc?
 - rated e.g. on a 5-point scale
 - monolingual judges are sufficient, no reference needed
- Adequacy
 - does the translated sentence *convey the same meaning* as the source sentence? (e.g. on a 5-point scale)
 - requires bilingual judges or a reference translation
- Informativeness
 - is it possible to answer *a set of pre-defined questions* using the translation, with the same accuracy as using the source or a reference translation?

Evaluation of MT by human judges (2)

- Reading time
 - people read more quickly a well-formed text
- Cloze test
 - ask a human to restore missing words from MT output: easier if the text is well-formed
- Post-editing time / HTER
 - time required to turn MT into a good translation
 - HTER: human-targeted translation error rate
 - how many editing operations are required for a human to change MT output into an acceptable translation (not necessarily the reference one)

A quick-and-dirty method (and an old joke)

- Compare a **sentence** and its **retroversion** (back translation)
 - only if systems are available for both translation directions
- Anecdotal example of the 1960s
 - EN: “**The spirit is willing but the flesh is weak.**”
 - translate into X (e.g. Russian) → then back into English →
 - EN’: “**The vodka is strong and the meat is rotten.**”
- Advantage
 - easier to compare EN/EN than EN/RU, e.g. edit distance
- Important idea: **monolingual comparison can be automated**
 - a candidate translation vs. a reference translation

Automatic metrics for MT evaluation

Principles of automatic metrics

- Compute a similarity score between a candidate translation and one or more reference translations
 - references: done by human experts, e.g. professional translators
 - note that human translations may also vary in quality...
 - several references: account for variability of good translations
- Typically: $\text{Average}_{i=1..k} (\text{Sim}(\text{Ref}_i, \text{Cand}))$ with $1 \leq k \leq 4$
 - where **Sim** is a similarity metric between sentences
 - **Sim** can use a variety of properties: string distance, word precision/recall, syntactic similarity, semantic distance, etc.
- Criterion for validating automatic metrics: *automatic scores must correlate with human ones on test data*

The BLEU metric

(BiLingual Evaluation Understudy)

- Proposed by K. Papineni et al. (2001) (IBM for NIST)
 - see ‘mteval’ at <http://www.itl.nist.gov/iad/mig/tools/>
 - also included with Moses: `scripts/generic/multi-bleu.perl`
- Principle
 - compare **n-gram overlap** between candidate and references
 - originally proposed with 4 references, but often used with one
 - mean of **n-gram precisions** (e.g. $n \leq 4$) \times brevity penalty
- Validation
 - shown to correlate well with human adequacy and fluency
- Variant proposed by NIST (G. Doddington 2002)
 - considers information gain of each n-gram over (n-1)-grams

Formula for the BLEU metric

(can be applied at sentence or corpus level)

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N w_n \log p_n\right)$$

$$BP = \min(1, \exp(1 - r/c))$$

$$p_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} count_{in_ref, bound}(ngram)}{\sum_{C \in \{\text{candidates}\}} \sum_{ngram \in C} count(ngram)}$$

r = length of reference translation
 c = length of candidate translation

$count_{in_ref, bound}(\cdot)$ = number of n grams in common with reference(s), bound by maximum number of occurrences in reference

- 2-4 reference translations (concatenated)
- n -grams from 1 to N (often $N=4$), weighted (often $1/N$)

Other automatic metrics

- METEOR
 - harmonic mean of unigram precision and recall, plus stemming and synonymy matching (if exact matching is impossible)
- Weighted N-gram Metric: model legitimate translation variation
 - considers tf.idf score of words to weigh their contribution to BLEU
- Word error rate, minimal string edit distance etc.
- Translation error rate (TER)
 - minimum number of edits needed to change a hypothesis so that it exactly matches one of the references (normalize: avg length of refs)
 - insertion, deletion, substitution of words; shifting phrases → all same cost
- Human-targeted TER = HTER
 - ask human judges to create reference translations which are as close as possible to a system translation (typically by editing system's hypotheses), then measure TER

Significance testing

- Problem
 - does a 0.1% BLEU increase show that a system is “really better”?
 - i.e. that it will also increase BLEU on a different data set
 - or is the variation due to randomness?
- Solution: split the test data into several folds
 - average scores over folds, compute *confidence intervals*
 - is the improvement larger than the c.i.? (similar to *t-test*)
 - another solution for pairwise ranking: *sign test*
- What if we do not have enough data to split?
 - generate the different folds by bootstrap resampling
 - for an N-sentence data set, draw N sentences with repetition (-> `multeval`)
- *Note*: tuning Moses is non-deterministic
 - results with several MERT runs should be averaged for confidence

Comparison of automatic and human metrics



- Cost-effective, fast
- Deterministic, “objective”
 - easy to reproduce
- Imperfect correlation with reference human metrics
 - holds mainly for data similar to setup data
- Need several high-quality reference translations
- Mainly applicable to English and weakly-inflected languages
- Very useful for MT development



- Human appreciation of translation quality is the ultimate reference
- Able to detect acceptable variations in translation
- Accurate on all system types
- Expensive
- “Subjective”
 - different judges, different scales
 - a judge might have different appreciations depending on what they saw before + fatigue
- Still the reference, especially for end-users, who do not care so much about BLEU

Trends of automatic metrics

- Finding automatic metrics = optimization problem
 - apply machine learning over training data as:
{(source sentence, imperfect translation, human score)}
- Increasing risk of over-fitting an MT system for BLEU
 - BLEU scores improve if a better language model is used, but “real quality” does not necessarily improve
- BLEU favors statistical over rule-based systems
 - ranked higher by BLEU than by human judgments
 - do not apply BLEU to human translations!
 - and maybe not to MT when it reaches human-like level

Some MT evaluation campaigns

- DARPA 1993-1995: adequacy, fluency, informativeness
- TIDES (~2000): BLEU + human judgments
- CESTA (2003-2006): MT into FR, human + automatic
- GALE (~2005-2009): HTER
- TRANSTAC (~2006-2009): concept transfer, WER
- NIST Open MT (2006, 2008, ...): continues TIDES
- MADCAT: same metrics as GALE
- MATR (2007, 2009, ...): competition among metrics
- Workshop on MT (2006-today): BLEU + human judges

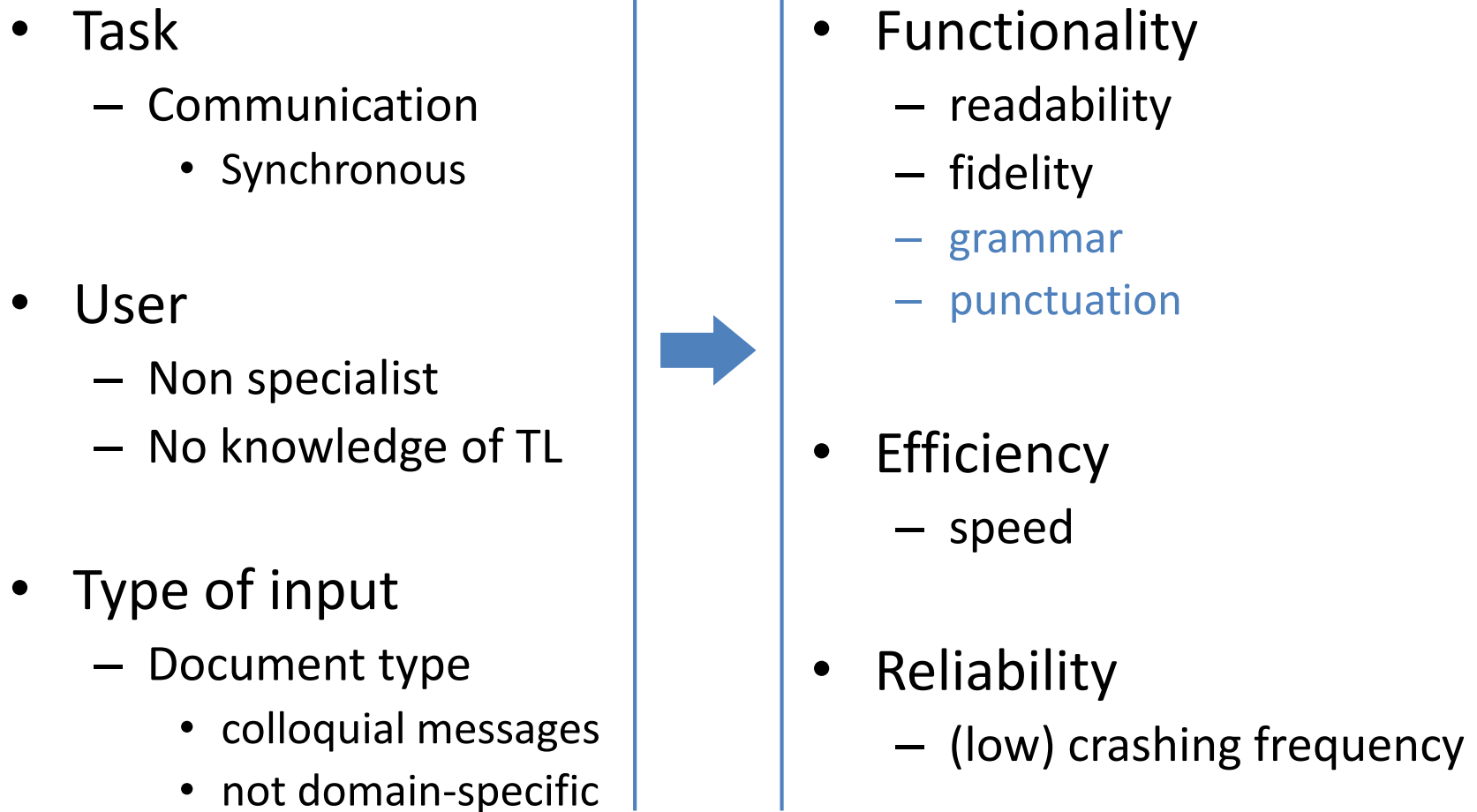
Evaluation of MT software: two views

- NLP researchers / developers
 - focus on the core functionality of their system, i.e. quality of machine translated text
- NLP users / buyers
 - are sensitive to a much larger range of qualities
 - core functionality (translation quality) still important
 - plus: speed, translation of technical terms, adaptability (e.g. facility to update dictionaries), user-friendliness, ...
 - ➔ indicators of quality depend on the intended use
- See: Ken Church & Ed Hovy, “Good applications for crummy MT”, *Machine Translation*, 8:239-258, 1993

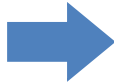
Complete evaluation of commercial MT software

- **FEMTI: Framework for MT Evaluation in ISLE**
 - encyclopedia of potential qualities with metrics
 - possible characteristics of the context of use
 - context characteristics related to qualities
- ➔ **FEMTI** helps evaluators specify an intended context of use and provides them with a quality model, i.e. a weighted list of qualities (+ metrics)
- <http://www.issco.unige.ch/femti>

Example 1: contextual evaluation of an instant messaging translation system



Example 2: contextual evaluation of routing systems for multilingual patents

- | | | |
|--|---|---|
| <ul style="list-style-type: none">• Task<ul style="list-style-type: none">– Assimilation<ul style="list-style-type: none">• Doc. routing• User<ul style="list-style-type: none">– Specialist– Knowledge of TL• Type of input<ul style="list-style-type: none">– Doc. type<ul style="list-style-type: none">• patent-related doc.– Author type<ul style="list-style-type: none">• domain specialist |  | <ul style="list-style-type: none">• Functionality<ul style="list-style-type: none">– accuracy<ul style="list-style-type: none">• terminological correctness– readability– style• Amount of linguistic resources<ul style="list-style-type: none">– size/type of dictionaries• Maintainability<ul style="list-style-type: none">– Changeability<ul style="list-style-type: none">• Ease of dictionary updating |
|--|---|---|

Examples of applications

- Translation on the Web (assimilation): millions+ of words/day
 - Google Translate, Systran, Reverso, PROMT, LINGUATECH, etc.
 - also as a showcase for their corporate systems
- Cross-language information retrieval: IR + MT
 - retrieve documents in a language different from the query
 - useful if results in a foreign language can be understood
- Spoken translation: ASR + MT (communication)
 - e.g. for European Parliament, or on handheld devices
 - also: visual translation on smartphones: OCR + MT
- Aids for professional translators (dissemination)
 - translation memories (e.g. Trados) are the mainstream tools
 - MT followed by post-editing might be useful (e.g. at Autodesk)

References

- Philipp Koehn, *Statistical Machine Translation*, Cambridge University Press, 2010 – chapter 8
- Hovy E., King M. & Popescu-Belis A. (2002) - Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17:43-75.
- Papineni, K. and Roukos, S. and Ward, T. and Zhu, W.J. (2002) - BLEU: a method for automatic evaluation of machine translation, *Proceedings of ACL 2002*, p. 311-318.
- Michael Denkowski and Alon Lavie, "Meteor 1.3: Automatic Metric for Reliable Optimization and Evaluation of Machine Translation Systems", *Proc. of the EMNLP 2011 Workshop on Statistical Machine Translation*.
- Ying Zhang and Stephan Vogel, "Significance tests of automatic machine translation evaluation metrics", *Machine Translation*, 24:51-65, 2010

Goal of practical work

- Choose a language pair and a (small) corpus
- Train Moses on 2-3 subsets of increasing sizes
 - keeping time reasonable, e.g. 0.1k-1k-10k
 - tune it, if possible, on a small separate set
- Evaluate Moses on 1-2 fixed test sets
 - how do BLEU scores vary with size of training set?
 - how do scores vary with test set?