Machine translation evaluation

Maja Popović

DFKI Berlin

maja.popovic@dfki.de

MT Marathon 2014

Trento, Italy

9 September 2014

outline

- what is the machine translation evaluation?
- why it is important?
- how can be carried out?
  - human evaluation methods
  - automatic evaluation methods
- why is it difficult?

what is translation quality?

once we have a machine translation output

* is it good or bad?

what for?

- MT system development (comparison)
- publishing
- post-editing
- other applications (question answering, information retrieval)

why?

- error classification and analysis

ref: It will be a sort of bridge.
sys1: It is almost as a bridge act.
sys2: It will act as a bridge.
sys3: It will not act as a bridge.
sys4: It will _ sort of bridge be.

system comparison

ranking from the best to the worst:
sys2, sys4, sys1, sys3

error analysis

- sys1: word form error (is), mistranslation (almost), word order (act)
- sys2: no errors
- sys3: insertion (not)
- sys4: omission (_→a), word order (be)

ref: It will be a sort of bridge.
sys1: It is almost as a bridge act.
sys2: It will act as a bridge.
sys3: It will not act as a bridge.
sys4: It will sort of bridge be.

publishing

only sys2 is acceptable

post-editing

sys3 is trivial to correct despite of the severity of the error

preservation of meaning

only sys3 is not acceptable

how to measure those things?

- human evaluators
- automatic methods
  - comparison of translation output with a reference translation
  - relation between translation output and the source sentence: quality estimation (no reference)

human evaluation methods

- adequacy and fluency
  - adequacy: does the translation convey the meaning of the source sentence?
  - fluency: is the output good fluent target language?

  5 = absolutely, ..., 1 = not at all

		adequacy	fluency
ref:	It will be a sort of bridge.		
sys1:	It is almost as a bridge act.	2	1
sys2:	It will act as a bridge.	5	5
sys3:	It will not act as a bridge.	1	5
sys4:	It will sort of bridge be.	4	2

- system ranking
  (basically guided by both adequacy and fluency)

human evaluation methods

- acceptability (estimated post-editing effort)
  - acceptable = no correction needed (1)
  - almost acceptable = little post-editing needed (2)
  - bad = better translate from scratch (3)

		effort
ref:	It will be a sort of bridge.	
sys1:	It is almost as a bridge act.	3
sys2:	It will act as a bridge.	1
sys3:	It will not act as a bridge.	2
sys4:	It will sort of bridge be.	2

- post-editing (implicit error classification)
- error annotation (explicit error classification)

human evaluation methods

disadvantages

- no single objectively correct translation of a given text
- no single correct error class for a number of translation errors

$\Rightarrow$ relatively low inter-annotator agreement

- examples:
  which system is better (worse): sys1 or sys3?
  how to classify each error in sys3?

- resource-intensive and time-consuming

$\Rightarrow$ automatic evaluation and error analysis

automatic evaluation metrics

what is an automatic evaluation metric?

- a computer program which calculates the translation quality
- input: translation output and reference translation(s)
- output: a numerical score related to their similarity

usual methods for comparison

- n-gram matching
  F-score, BLEU, METEOR
- edit (Levenshtein) distance
  WER, TER

n-gram matching: precision and recall

- precision: $\frac{N(matches\_in\_Translation\_Output)}{Translation\_Output\_Length}$
- recall: $\frac{N(matches\_in\_Reference)}{reference\_Length}$

1-gram (word) matches:

ref:	It will be a sort of bridge.	7/8 (87.5%)
sys4:	It will sort of bridge be.	7/7 (100%)

2-gram matches:

ref:	It_will will_be be_a a_sort sort_of of_bridge bridge_.	4/7 (42.8%)
sys4:	It_will will_sort sort_of of_bridge bridge_be be_.	3/6 (50%)

3-gram matches:

ref:	It_will_be will_be_a a_sort_of sort_of_bridge of_bridge_be bridge_be_.	1/6 (16.7%)
sys4:	It_will_sort will_sort_of sort_of_bridge of_bridge_be bridge_be_.	1/5 (20%)

unifying all n-grams, precisions and recalls

- How to put together different n-grams?
  - geometric mean
  - arithmetic mean
    (better, does not penalise too hard unseen n-grams)

- How to put together precision and recall?
  - harmonic mean – F-score:

    $2 \cdot \mathit{precision} \cdot \mathit{recall} / (\mathit{precision} + \mathit{recall})$

n-gram based automatic metrics

* BLEU
  - geometric mean of 1-, 2-, 3- and 4-grams
  - precision + brevity penalty instead of recall
* METEOR
  - flexible unigram matching
  - does not penalise (too hard) common stems, synonyms and paraphrases
* F-score
  - arithmetic mean of 1-,2-,3- and 4-grams
  - standard harmonic mean

edit distance

edit (or Levensthein) distance

- minimum number of edits to transform translation output to the reference
- edit types:
  - substitution: replace one word with another
  - deletion: a word is missing, it should be added
  - insertion: a word is inserted, it should be removed

edit distance based evaluation metrics

- Word Error Rate (WER) – Levenshtein distance itself

$$WER = \frac{N(substitutions)+N(deletions)+N(insertions)}{reference\_Length}$$

ref	It	will	be_{del}	a_{del}	sort	of	bridge		.
sys4	It	will			sort	of	bridge	be_{ins}	.

WER = 3/7 (37.5%)

- Translation Edit Rate (TER)

$$TER = \frac{N(substitutions)+N(deletions)+N(insertions)+N(block\_shifts)}{reference\_Length}$$

ref	It	will	be	a_{del}	sort of bridge			.
sys4	It	will			sort of bridge	be_{shift}		.

TER = 2/7 (28.6%)

properties of automatic evaluation metrics

desirable characteristics

- $+$ fast and cheap
- $+$ consistent: repeated use should always give same results
- $\pm$ informative: the score should give intuitive interpretation of translation quality
- $\pm$ correct: better systems should be ranked higher

evaluation of automatic evaluation metrics

is an automatic metric good?

- yes, if it is fast, cheap and consistent
  (and it almost certainly is!)
- and if it is correct,
  i.e. if its system ranking correlates with human ranking
  (is it?)

how to measure correctness?

- correlation coefficients

evaluation of evaluation metrics – correlations

correlation coefficients between human and automatic ranks

- 1 $\Rightarrow$ absolute correlation (-1 $\Rightarrow$ inverse correlation)
- 0 $\Rightarrow$ no correlation

- document level
  - Spearman's correlation coefficient
  - takes only rank into account
  - Pearson's correlation coefficient
  - takes into account both rank and linearity
- sentence level
  - Kendall's Tau coefficient
  - compares pairwise sentence rankings

- widely used metrics correlate reasonably (BLEU, TER) or rather well (METEOR) with human rankings

metric research

- WMT shared evaluation task
  `http://www.statmt.org/wmt14/metrics-task/`
  - develop a metric
  - check its correlations with human ranks
- a number of new metrics have shown high correlations
  - semantic equivalence (MEANT, HMEANT)
  - syntactic similarity (POS n-grams)
  - linguistic features
  - combination of metrics
  - ...
- many of them have (significantly) higher correlations than BLEU and TER
- however...
  - many of them are rather complex
  - no improvements for system tuning

F-score for MT evaluation

- word-level F-score correlates better than BLEU (and TER, not better than METEOR)
- arithmetic n-gram averaging better than geometric
- optimal n-gram length is 4
- even better correlations for morpheme and POS based F-scores, especially
  - on the sentence level
  - for translation from English
  - however: complex (external tools needed)

- rgbF tool:
  calculates the F-score averaged on all n-grams (default=4) of an arbitrary set of distinct units such as words, morphemes, POS tags or whatever, aligned on the sentence level

`http://www.dfki.de/~mapo02/rgbF/`

automatic evaluation metrics – summary

advantages and issues

+ fast and cheap
+ consistent
± not fully able to rank different types of systems (especially on the sentence level)
  - research on extended and new metrics
- scores do not give any details about actual translation errors
  - error classification and analysis
- require some kind of human reference translation
  - evaluation without references – quality estimation

error classification

what evaluation scores cannot answer?

- what is a particular strength/weakness of the system?
- what does a certain modification of a system exactly improve?
- does a worse-ranked system outperform a better-ranked one in any aspect?

$\Rightarrow$ error classification and analysis is needed

Two main goals:

- distribution of errors over the error classes within an output
- distribution of errors over translation outputs within a class

human error classification (MQM scheme)

- adequacy (accuracy)
  - mistranslation
  - omission
  - addition
  - untranslated
- fluency
  - grammar
    - morphology (word form)
      - part of speech
      - agreement
      - tense/aspect/mood
    - word order
    - function words
  - spelling
    - capitalisation
  - typography
    - punctuation
  - unintelligible

automatic error classification

Hjerson tool:

- compares raw machine translation output with the reference translation
- based on edit distance in combination with precision and recall
- distinguishes five error classes:
  - inflectional errors
  - reordering errors
  - missing words
  - extra words
  - incorrect lexical choice

`http://www.dfki.de/~mapo02/hjerson/`

evaluation of automatic error classification

- good correlation (Spearman and Pearson) with human error classification distributions
  - * both over error classes and over translation outputs
- high recall (except for extra words)
- low precision
  N(automatic_errors) >> N(human_errors)
- * better precision when post-edited output is used as a reference

evaluation without reference translations

- both automatic evaluation and error classification require a reference translation

! but

  - there is not much reference translations in "real life"!
  - if we already have a (high quality) translation, why would we need a machine translation output?

$\Rightarrow$ evaluate without a reference

  - naive approach: IBM-1 scores (on different levels) for each source sentence and its translation output
  - quality estimation system

quality estimation

- provides a metric which estimates quality of unseen translations
- main components of a QE system:
  - definition of quality – what to predict
  - human labelled data
  - features
  - machine learning algorithm

what to predict?

- absolute scores for adequacy/fluency
- absolute scores for post-editing effort
- average post-editing time per word
- relative rankings
- percentage of edits for the given sentence
- word-level edits and its types
- BLEU or other scores for document

features

- number of words in source and target sentences
- average source word length
- average number of word occurrences in the target sentence
- number of punctuation marks in source and target sentences
- LM probabilities of source and target sentences
- average number of translations per source word
- ...

machine translation evaluation – summary

- machine translation evaluation
  - important task
  - difficult task

  $\rightarrow$ still an open problem
- different aspects, goals, users
- human evaluation
  - time and resource extensive
  - not easily repeatable
- automatic methods
  - crucial for MT system development
  - good correlations with human results but it can be better
  - human knowledge is, one way or another, necessary
    - human references or annotations
    - human judgments for development/improvement

  $\Rightarrow$ human evaluations are needed too

Questions?

?