Evaluation of Machine Translation Systems

Paula Estrella

Introduction

- MT implies generating text from a SL into a TL
  - Rule-based, example-based, statistical
- Once it's done
  - How can we be sure the generated TL actually is a translation of the SL text?
  - How can we state what is a good MT system?
- MT evaluation attempts to answer these questions

Introduction (cont.)

- MT systems are evaluated according to a set of criteria
- Researchers and developers
  - focus on quality of MT output in a given domain
- Users / buyers
  - Output quality important but also speed, formats handled, adaptability, user-friendliness, etc.

Why is MT evaluation difficult?

- Evaluating MT is a hard task
  - There is no "gold standard"
  - There is (yet) no fully reliable method to evaluate MT
  - Wide range of parameters and users
    - Stakeholders have different priorities
- Today's talk will provide a general overview of the field
  - Slightly biased towards context-based methods

Plan

- Methodologies
  - Evaluation campaigns
  - Task-based evaluation
  - Context-based
- Evaluation of MT output
  - Human-based metrics
  - Automatic metrics
- Context-based evaluation
  - Standards for software evaluation
  - Application of standards: FEMTI

Plan

- Methodologies
  - Evaluation campaigns
  - Task-based evaluation
  - Context-based

System evaluations

- Internal evaluations (e.g during development)
  - Human-based or automatic metrics applied
- External evaluations (e.g during deployment)
  - Include the user's view of quality
- Usually not comparable to other system's results
  - Different methodologies/corpora/metrics used

Evaluation campaigns (1/2)

- Evaluate several systems on a common framework
  - Goals: validate methodologies, compare systems, support and direct future research, etc
- Started in the '90s by Defence Advance Research Projects Agency (DARPA)
- DARPA proposed "standard methodology"
  - Fluency/adequacy/informativeness on 5-points scale
  - Today's de facto standard + automatic metrics

Evaluation campaigns (2/2)

- Many HLT domains: MUC, TIDES, EARS, TREC, NIST *Metrics*MATR
  - Many organizers: NIST, WSMT by Edinburgh SMT group, EVALDA for French text/speech technologies
- + Corpora and evaluation results usually distributed after campaigns
- - Only MT output quality is considered

Task-based evaluation (1/3)

- Measure performance of humans using MT output to accomplish a specific task
- "Good applications for crummy MT" - (Church & Hovy 1991)
  - quality decomposed into e.g. translation of technical terms, correctness of punctuation
  - The relative importance of these parameters varies with the *intended use* of an MT system

Task-based evaluation (2/3)

- (White & Taylor 1998) Propose hierarchy of tasks ordered by difficulty
  - publication > gisting > extraction > … filtering
  - Task Proficiency metric: systems rated as adequate to perform a task and those below it in the hierarchy
- Other task-based metrics: reading comprehension, cloze tests

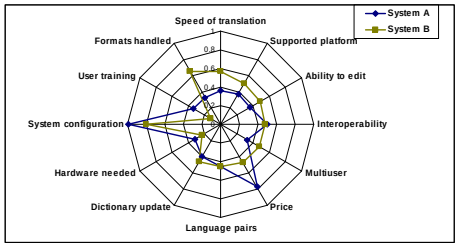
Task-based evaluation (3/3)

- + Tries to measure *utility* of MT output
- - hard to design and difficult to isolate human factors (e.g. ability to "guess" right answers)

Context-based evaluation (1/4)

- Define the intended context of use of the system (e.g task, user, input data) and apply relevant metrics
- JEIDA Report (Nomura & Isahara 1992)
  - objective: to characterize the *intended context of use* and the *performance* of an MT system
    - 3 points of view: economic factors (users), technical (users / developers)
    - 14 dimensions proposed, 2 questionnaires represented as radar charts

Context-based evaluation (2/4)



Context-based evaluation (3/4)

- EAGLES EWG (1993-1996) attempted to create standards for NLP systems
  - Applied to spell/grammar chequers, speech recognition, dialogue systems
- ISLE Project : International Standards for Language Engineering (1999–2002)
  - apply the EAGLES guidelines to MT
  - ensure compatibility with the ISO/IEC standards for software evaluation

Context-based evaluation (4/4)

- FEMTI implements hierarchies for context of use and quality characteristics
  - Explained in detail later
- + Considers broad range of other factors than MT output quality
- - More expensive to design/execute and probably less reusable

Summary

Methodology	Goal	Relevant work
Evaluation campaigns	Compare several systems in terms of output quality	DARPA, NIST, EVALDA
Task-based	Utility of MT output to perform a task	Task proficiency, reading comprehension
Context-based	Consider wide range of features for intended context of use	JEIDA, TEMAA, FEMTI

Plan

- Methodologies
  - Evaluation campaigns
  - Task-based evaluation
  - Context-based
- Evaluation of MT output
  - Human-based metrics

Human-based metrics

- Rating-based assign a score from a given scale
  - E.g. intelligibility on 100-point scale, fluency on 9- 5-points, fidelity on 9- 7- 4- 3- points
  - Aspect of quality evaluated requires bilingual or monolingual judges
- + Complex aspects can be assessed (e.g. register, style, etc)
- - Difficult to decide what counts as an error and how to penalize the translation
- - It is not clear how scale influences results

Human-based metrics

- Comprehension-based metrics
  - Cloze tests, reading comprehension tests
  - - Hard to design: e.g. control deletion of content words in cloze tests
  - + useful if main variables are controled, e.g. [Miller 2000]
- Post-editing metrics
  - Post-editor editing time/effort measures, HTER: Human Translation Edit Rate
  - + attempt to measure utility of MT output
  - - difficult to correctly edit if done without context, constrained if a reference is shown (HTER)

Example of scales

- ALPAC's scale for intelligibility on 9 points
  - "1 = hopelessly unintelligible" to "9 = perfectly clear and intelligible"
  - Middle points include "5 = between 4 and 6"
- Van Slype's scale on 4 points
  - 3 = Very intelligible: all the content of the message is comprehensible
  - 2: Fairly intelligible: the major part of the message passes
  - 1: Basely intelligible: a part only of the content is understandable, representing less than 50% of the message
  - 0: Unintelligible: nothing or almost nothing of the message is comprehensible

Meta-evaluation

- Human-based evaluation is difficult
  - Painful to read long or low quality MT output
- Reliability and Consistency: difficulty in obtaining high-levels of agreement
  - Intra-judge agreement: consistency of same human judge
  - Inter-judge agreement: judgment agreement across multiple judges of quality
- Even so, human-based metrics remain most reliable evaluators

Meta-evaluation

- WSMT focuses evaluation on human-based metrics
  - Relative ranking of sentences (2008)
    - how frequently is a system judged better than or equal to other systems
  - MT output post-editing (2009)
    - edit the translation without seeing the reference; after that indicate whether the edited output is equivalent to the reference
    - Showed higher inter/intra-judge agreement than relative ranking

Example

- Sentences from CESTA evaluation campaign FR-EN
- Human evaluation with metrics
  - Fluency on 5-points
    - Looking at the translation, how grammatical and fluent is it?
  - Adequacy on 5-points
    - how much meaning of the original source text is present in the translation?

Example

Src	the creation of an enormous quantity of refuse.

S1	la création d'une énorme quantité de déchets.
S2	la création d'une énorme quantité de la refuser.
S3	la création d'une énorme quantité d'ordures.

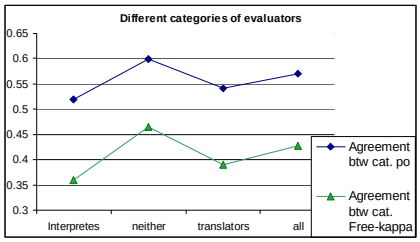
	S1	S2	S3
Ade	0.87	0	0.62
Flu	0.87	0.25	0.62

Ref1	la création d'une quantité énorme de déchets.
Ref2	la création d'une énorme quantité de déchets.
Ref3	la création d'une énorme quantité d'ordures.

Example inter-judge agreement

- Evaluation of EN-SP medslt system
- K = P(A) – P(E) / 1- P(E)
  - P(A) proportion of times annotators agree
  - P(E) proportion of times they would agree by chance

Example inter-judge agreement



Plan

- Methodologies
  - Evaluation campaigns
  - Task-based evaluation
  - Context-based
- Evaluation of MT output
  - Human-based metrics
  - Automatic metrics

Automatic metrics

- Ideally automatic evaluation would avoid humans in the loop
- Generally need several *reference translations* to account for *translation variation*
  - References ~ gold standards (?)
- Can be used on ongoing basis during development

Distance-based

- mWER: number of insertions/deletions/ substitutions to convert a translation into a reference
- mPER: does not consider word order
- Translation Edit Rate (TER): allows moving blocks of words (count as 1 edit)

N-gram-based metrics

- BLEU: n-gram precision + brevity penalty (BP)
  - BP: very short candidates don not get too high a score
- NIST: variation of BLEU, BP has less impact on overall score, averages n-grams on arithmetic mean, weights rare n-grams heavily to account for informativeness
- WNM: variation of BLEU, weights n-grams according to their frequency in a test monolingual corpus, precision/recall on n-grams, 1 reference

Precision/recall-based

- Precision: correct words / total words in MT output
- Recall: correct words / total words in reference
- General text matcher (GTM): maximum subset of non-repeated words, higher weight to longer matches and matches in the right order, the weight is a parameter to the metric.
- METEOR: unigram precision/recall, consider best score again each reference, allow stemming/synonymy

Meta-evaluation

- These metrics need to be validated
  - Must be applied to large nr of language pairs, domains, etc
  - Must be applied at sentence/document/system level
- Automatic metrics are evaluated against human-based metrics
  - Pearson correlation between scores
  - Spearman rank correlation of scores
  - Applied to human-generated texts

Meta-evaluation

- Several results show low correlations between human-based and automatic metrics
  - At different levels, on different corpora
  - Not clear what automatic metrics measure
- [Callison-Burch 2006] shows that some automatic metrics fail for RB systems
  - biased towards SMT → MERT with BLEU
  - Automatic metrics fail as output quality improves

Example from CESTA FR-EN track

Src	the creation of an enormous quantity of refuse.

S1 soft	la création d'une énorme quantité de déchets.
S2 rali	la création d'une énorme quantité de la refuser.
S3 syst	la création d'une énorme quantité d'ordures.

Ref1	la création d'une quantité énorme de déchets.
Ref2	la création d'une énorme quantité de déchets.
Ref3	la création d'une énorme quantité d'ordures.

	S1	S2	S3
Ade	0.87	0	0.62
Flu	0.87	0.25	0.62
BLEU	0.95	0.55	1
NIST	4.74	3.67	4.95
WER	0.13	0.40	0

Example II

- Taken from [Gimenez 2008]

Src	la casa verde estaba situada justo delante del lago .

Ref1	the green house was right in front of the lake .
S1	the green house was by the lake shore .
S2	the green potato right in front of the lake was right .
S3	a green house was by the lake shore .

	S1	S2	S3
BLEU	0.30	**0.52**	0
NIST	2.29	**2.90**	1.96
GTM	0.70	**0.87**	0.60

Other metrics (1/2)

- X-scores [Rajman&Hartley 2001]
  - distribution of POS tags is compared with a reference corpus fluency-annotated
- [Liu and Gildea 2005] proposed
  - Syntactic Tree Matching (STM) metric compares parse trees of hypothesis and references
- [Gimenez 2008] focuses on metric combination operating at different linguistic levels (e.g., lexical, syntactic and semantic)
  - IQMT framework available online
- Many other metrics exist
  - **32 new metrics** presented at MetricsMATR 2008

Other metrics (2/2)

- + This type of metrics seem to be neutral to different types of systems
  - Except for X-scores which do not correlate well
- + They also outperformed some metrics based on lexical similarity (e.g. BLEU, etc)
- - Parsers introduce additional noise
- - Might need special development before application
  - Necessary resources not always available for all language pairs
  - Exact algorithms for metrics not always available

Automatic vs. Human-based metrics

- Time-consumig vs. quick evaluation
  - Not if automatic metrics must be adapted to new language pairs
- Costly vs. cheap to apply
  - *Cheap* if set of reference/human scores available
- Subjective vs. objective
  - Automatic metrics seem *too objective*
- How do we choose the metrics or design an evaluation?

Plan

- General overview
  - Brief history
  - Different methodolgies
- Evaluation of MT output
  - Human-based metrics
  - Automatic metrics
- Context-based evaluation
  - Standards for software evaluation
  - Application of standards: FEMTI
- Conclusion

EAGLES 7-step recipe

1. Why is the evaluation being done?
2. Elaborate a task model
3. Define top level quality characteristics
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3
5. Devise the metrics to be applied to the system for the requirements produced under 4.
6. Design the execution of the evaluation
7. Execute the evaluation

- Steps 1- 5 implemented in FEMTI

What is FEMTI?

- FEMTI is
  - Set of context-based evaluation guidelines
  - Repository of evaluation metrics, references
  - Web-based tool to generate evaluation plans
- Implements hierarchies for
  - Context of use
    - Environment where the system is to be used
  - ISO-based quality characteristics
    - Attributes that constitute software quality
- Note: it doesn't cover execution of the evaluation or quality in use

Plan

- General overview
  - Brief history
  - Different methodolgies
- Evaluation of MT output
  - Human-based metrics
  - Automatic metrics
- Context-based evaluation
  - Standards for software evaluation
  - Application of standards: FEMTI
- Conclusion

Standards for software evaluation

- ISO 14958 – product evaluation process
  - quality in the software life cycle
  - process for developers, acquirers and evaluators
- ISO 9126 – product quality
  - model for software product quality
  - defines six main quality characteristics
    - functionality, reliability, usability, efficiency, maintainability, portability
  - further subdivided into subcharacteristics
    - terminal nodes of this hierarchy (*quality model*) can be measured using internal or external metrics

Application of standards

ISO generic quality characteristics

- functionality
- reliability
- usability
- efficiency
- maintainability
- portability

Quality characteristics particular to MT

- Functionality
  - Suitability
    - Accuracy
      - Fidelity
        - BLEU, NIST…
      - Consistency
      - Terminology
      - … … …

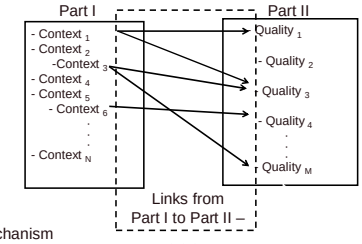
Plan

- General overview
  - Brief history
  - Different methodolgies
- Evaluation of MT output
  - Human-based metrics
  - Automatic metrics
- Context-based evaluation
  - Standards for software evaluation
  - Application of standards: FEMTI
- Conclusion

Components of FEMTI

- Part I: classification of the characteristics of the translation *task / user / input / purpose* of the evaluation
  - E.g.: document routing, email translation, information extraction
- Part II: classification of *MT software* quality characteristics
  - Examples: fidelity, readability, terminological correctness, speed
- Result of using FEMTI: evaluation plan
  - Characteristics of context of use + quality characteristics + related metrics
- Recent developments
  - Linking mechanism, support tools for evaluators and experts

Relating context of use to qualities



Linking mechanism to suggest qualities according to context

Example: evaluation of an MT system for instant messaging

- Task
  - Communication
    - Synchronous
- User
  - Non specialist
  - No knowledge of TL
- Type of input
  - Document type
    - colloquial messages
    - not domain-specific

→

- Functionality
  - readability
  - fidelity
  - grammar
  - punctuation
- Efficiency
  - speed
- Reliability
  - (low) crashing frequency

« Part 1 » « Links » « Part 2 »

How does the linking mechanism work?

- Generic Contextual Quality Model (GCQM)
  - Data structure to store links (matrix)
  - Vector representation of classifications
- Context vectors
  - Represent user description of context
- Quality vectors: context vector x GCQM
  - Represent a customized quality model

Computing quality vectors

Part I

1. Evaluation requirements
 1.1 Characteristics of the translation task
  1.1.1 Assimilation
   1.1.1.1 Document routing or sorting
   1.1.1.2 Information extraction √
   1.1.1.3 Search

Part II

2. System characteristics
 2.1 Functionality
  2.1.1 Accuracy
   2.1.1.1 Terminology
   2.1.1.2 Fidelity
   2.1.1.3 Consistency

Context vector: 0 0 0 0 1 0

X

GCQM	2	2.1	2.1.1	2.1.1.1	2.1.1.2	2.1.1.3
1						
1.1		0.5				0.4
1.1.1			0.4			
1.1.1.1					0.7	
1.1.1.2				0.5		0.6
1.1.1.3				0.3		

GCQM

= [0, 0, 0, **0.5**, 0, **0.6**]

quality vector

Summary

- FEMTI is a rich source of information
  - Improved with linking mechanism
  - Some activities carried out to enrich GCQM
  - Guidelines converted into support tools
- But …
  - Complex framework, specially for "beginners"
    - More guidance via chatbots (Prof. Volk's idea!)
  - Content needs re-work
    - Add/delete metrics/characteristics, develop/enrich characteristics
  - Use cases or templates would increase FEMTI's usability
- Context-based evaluation is "real-world" oriented (?)
  - Researchers do not use FEMTI - "real users" use other methods
  - Future work: Improve FEMTI to encourage its use

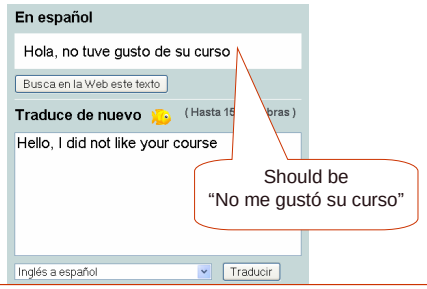
Exercise

- Build a quality model for an MT system in the following context (no need to use FEMTI)
  - Task: You have to send me an email in Spanish with some feedback about this course
  - You have basic knowledge of Spanish
  - You do not want to spend a lot of money

Exercise (cont.)

- Cost of the system has highest importance in my model → online translation (e.g. babelfish)
- It doesn't provide German-Spanish → language pairs handled is also an important feature (< cost)
  - Solution: Babelfish English-Spanish

Exercise (cont.)



Exercise (cont.)

- Cost of the system has highest importance in my model → online translation (e.g. babelfish)
- It doesn't provide German-Spanish → language pairs handled is also an important feature (< cost)
  - Solution: Babelfish English-Spanish
- Output quality is also important → we tune the model

Resources

- Alpac report
  http://www.nap.edu/openbook.php?isbn=ARC000005
- Van Slype report [1979]
  www.issco.unige.ch/en/research/projects/isle/van.slype.pdf
- Summaries of metrics in [Van Slype 1979] & chapter 3 [Estrella 2008]
- EAGLES 7-step recipe
  http://www.issco.unige.ch/en/research/projects/eagles/ewg99/7steps.html
- Mt-archive www.mt-archive.info/
- JC Sager. Quality and standards – the evaluation of translations (1989) in The translator's handbook - ASLIB