Linguistic Resources in Support of Various Evaluation Metrics

Christopher Cieri, Stephanie Strassel,
Meghan Lammie Glenn, Lauren Friedman

Linguistic Data Consortium

MT Evaluation

- Criteria
  - adequacy: source and translation provide same information
    - recall:
    - precision: translation should not invent information
  - fluency: translation is grammatical in the target language
    - style is appropriate
  - consistency
  - length: excessive brevity sometimes penalized, excessive wordiness should be too
- MT Evaluation properties
  - fast: facilitates use during system development
  - objective & repeatable: just good science
- Alternatives may be modeled
  - directly, for example by creating multiple references
  - indirectly, for example by permitting alternatives during evaluation

Evaluations & Resources

	Training	Grading	Human Assessment Adequacy	Human Assessment Fluency	BLEU	METEOR	(H)TER	DLPT*
Monolingual Text (t)	✓							
Parallel Text	✓							
Translation Lexicon	✓							
Source Text		✓						
MT Output		✓	✓	✓	✓	✓	✓	
Grading Annotation		✓						
Bilingual, Highly Trained G Annotators		✓						
1-Best Human Translation			✓			✓		
1B HT with Alternatives							✓	✓
Multiple Human Translations					✓	☑	☑	
Adequacy Annotation			✓					
Monolingual Trained Adequacy Annotators			✓					
Fluency Annotation				✓				
Monolingual Trained Fluency Annotators				✓				
Stemmer (t)						☑		
WordNet (t)						☑		
Edit Distance Annotation							✓	
Highly Trained ED Annotators							✓	
ILR Judgments							✓	✓
Comprehension modules							✓	✓
Human subjects								✓

Creation of Reference Translations

Typical Translation Pipeline: Preparing the Data

- Data collection
- Manual or automatic data selection
  - Quick or careful depending on evaluation requirements
- Corpus-wide scans to remove duplicate docs, prevent train/test overlap
- Manual or automatic segmentation of source text into sentence units
- Pre-processing to convert files into translator-friendly format
  - One segment per line, with empty line for translated to input translation

Typical Translation Pipeline: Translating the Data

- Translator-ready files collected into "kits" and distributed to translators
  - Kits customized for individual translation bureaus based on target volume, agency expertise, additional requirements (e.g. source variety, level of difficulty, file length, etc)
- Translation
  - Translators use guidelines originally developed for TIDES, enhanced for GALE and NIST MT that provide detailed instructions and examples
    - Translating/transliterating proper names, speech disfluencies, factual errors, characteristics of newsgroups, typos etc.
  - Multiple translation teams for each language
  - Each team has at least one translator native in the source language and one native in the target language
  - Initial screening and evaluation for all potential translation providers

Typical Translation Pipeline: Validating the Data

- Process incoming translations
- Conduct sanity checks
  - All files have been returned
  - All files are in expected encoding
  - Segment inventory is complete
  - All segments have been translated
  - etc.
- Post-processing to convert files into required evaluation data format
- Manual and/or automatic quality control
- Comprehensive translation database tracks status for each file or data set
  - By language, genre, project, phase, partition, translation agency, due date, QC score, etc.

Regular Translation QC

- An approach to (human) translation evaluation used instead to confirm translation agencies
- 10% of each incoming translation set is reviewed
- Fluent bilinguals review selection deduct points for each error

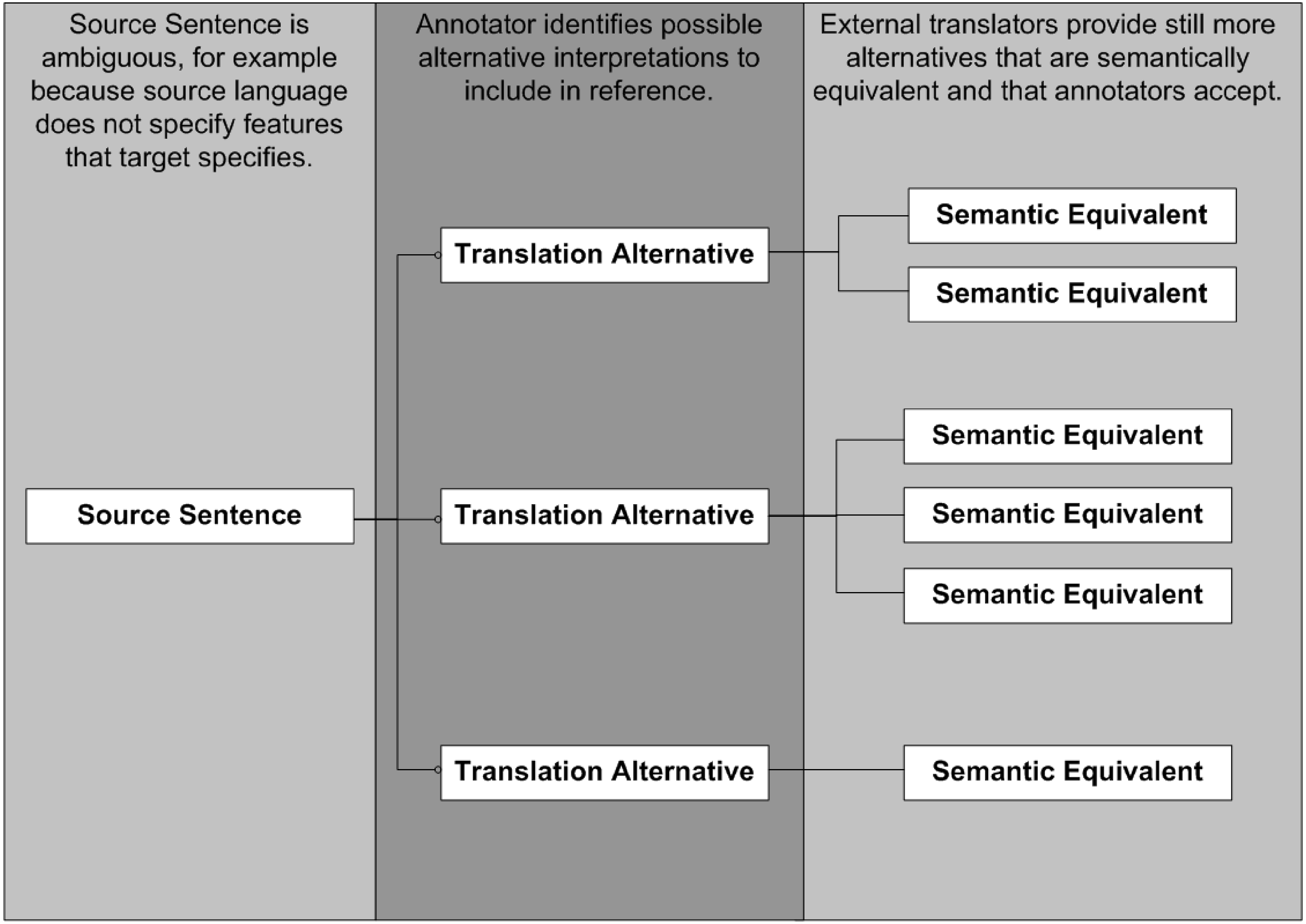
Error	Deduction
Syntactic	4 points
Lexical	2 points
Poor English usage	1 point
Significant spelling/punctuation error	½ points (max 5 points)

- Deliveries that receive a failing score are rejected and returned to the agency to be redone
  - Payment is withheld until corrections are complete

Gold Standard Translation QC

- **First pass QC**: Bilingual junior annotators correct obvious mistakes
- **Second pass QC**: *Source language-dominant* bilingual senior annotators correct subtler mistakes
  - improve fluency, correct/standardize names, research difficult vocabulary, verify translation against source audio where required
- **Third pass QC**: *Target language-dominant* bilingual senior annotators improve fluency and accuracy and add translation alternatives
- **Fourth pass QC**: *Target-language monolingual* senior annotators read translations for fluency and comprehension, flag problems
- **Corpus wide scans**: Programmers perform multiple manual and automatic scans
  - standardize and validate data format
  - identify any lingering errors in the corpus as a whole
- **Final spot-check**: Team leaders review 10% of all source-translation document pairs to ensure all problems have been resolved

Alternative Translations



Source Sentence is ambiguous, for example because source language does not specify features that target specifies.

Annotator identifies possible alternative interpretations to include in reference.

External translators provide still more alternatives that are semantically equivalent and that annotators accept.

Source Sentence

Translation Alternative

Translation Alternative

Translation Alternative

Semantic Equivalent

Semantic Equivalent

Semantic Equivalent

Semantic Equivalent

Semantic Equivalent

Semantic Equivalent

Assessment of Adequacy and Fluency

Resources Required

- Multiple reference translations
  - Typically 4-5 references for NIST MT evaluations
  - Good quality, but with minimal manual QC
  - No translation alternations included
  - Segment-aligned with source
- Detailed translation guidelines
- Brief assessment guidelines
- Simple assessment GUI
- Assessors have average skill set
  - Typically college students, native speakers of target language
- Limited task-specific training
- 2+ assessors per system

Assessment Process

- NIST selects subset of docs from BLEU evaluation
  - In MT06, every 4th document taken from a list of documents ordered according to each document's average BLEU score
- NIST selects a subset of system outputs for each source language for human assessment
  - In MT06, the systems with the best BLEU score
  - Selected from the "large data" condition
  - Limited to "primary" system submissions
- LDC assigns multiple assessors for each translation of a document
  - In MT06, each doc judged independently by two assessors
  - Each assessor judges all systems
  - No assessor judges the same document more than twice
- As time/budget allow, human translations may also be evaluated against one another for fluency and adequacy

Cost Factors

- Translation of ~100K words
  - 1 week FTE to prepare data and coordinate translators
  - 6-8 weeks calendar time for per “batch” of translation
    - Costs average $0.25/word
  - >1 week FTE for regular QC
- Assessment of ~100K words
  - > 1 week FTE technical, workflow, editor coordination
  - Assessors earn on average $11/hour
    - Realtime rates vary by genre, MT output quality
      - Average 1 minute per segment for fluency
      - Average 2 minutes per segment for adequacy

Edit Distance

The Metric

- HTER: Human Translation Error Rate
  - Skilled monolingual human editors compare MT output against reference translation
    - Modify MT output so that it has the *same meaning* as gold standard translation and is *understandable*
      - Each inserted/deleted/modified word or punctuation mark counts as one edit
      - Shifting a string, of any number of words, by any distance, counts as one edit
- TER: Translation Error Rate
  - No human post-editor
  - Automatic calculation of edit distance
- Edits are counted by automated software
  - Compares the unedited MT output to the edited version (HTER) or to the gold standard translation (TER)
  - Finds the minimum number of edits that will create the edited version (HTER) or reference translation (TER)

Example

HTER

ET: To end conflict , the military began a blockade on October 6 .

MT: To end conflict * *** @ on a a blockade on October 6 .

D D S S SHIFT

HTER Score: 45.45 (5.0/11.0)

TER

RF: ** The military initiated a blockade October sixth to eliminate clashes .

MT: To end conflict on a blockade October ***** 6 on a @.

I S S S SHIFT D S S S

TER Score: 81.82 (9.0/11.0)

Resources Required

- Single gold standard reference translation
  - Extremely high quality with multiple inputs & manual QC passes
  - Includes translation alternatives to reflect source ambiguity
  - Segment-aligned with source
- Detailed translation guidelines
- Extensive post-editing guidelines
- Customized post-editing GUI
- Highly skilled monolingual target language post-editors
  - Typically professional editors and proofreaders
- Extensive task specific formal training
- In GALE, *four* post-editors per system
  - Two independent first passes (focus primarily on meaning)
  - Followed by second pass over first pass edits (focus primarily on minimizing HTER)
  - Latin square design for file assignment
  - Lowest scoring segments selected as final HTER
- Substantial workflow and tracking infrastructure

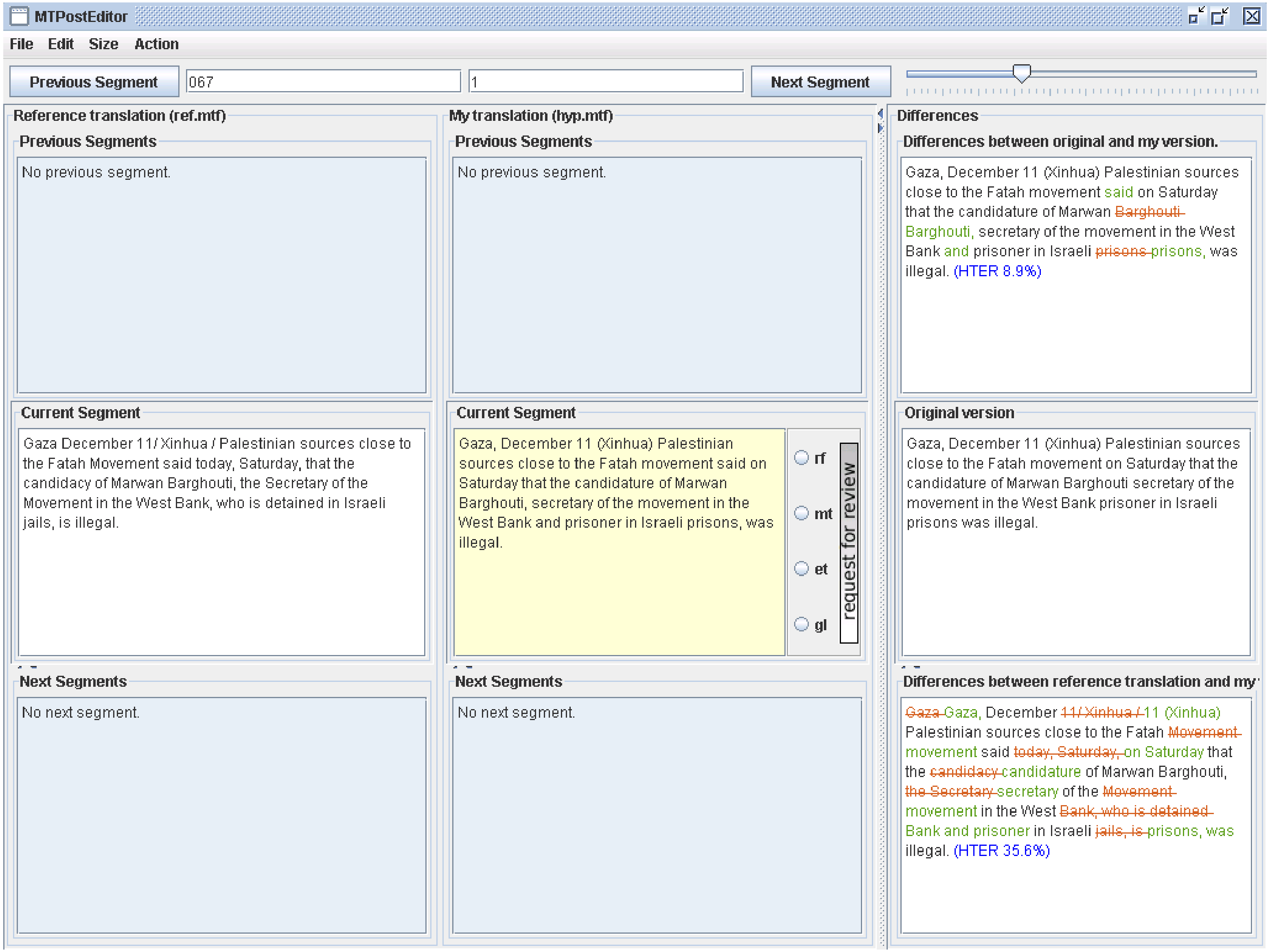
Post-Editor Training

- Initial screening: skills assessment test
  - 10 segments selected for coverage of phenomena
- Half day hands-on training session
  - Guidelines and process covered in detail
  - Group editing of many examples
  - Q&A
- Post-test (repeat of skills test) to gauge improvement
- Completion of "starter kit"
  - Small set of carefully selected data
  - Results reviewed in detail to provide individual feedback on errors, esp. ways to minimize HTER

Post-Editing Guidelines

- Dual emphasis on meaning preservation and edit minimization
- Rules and examples covering
  - Phrasal ordering, POS, grammatical issues
  - Orthography (capitalization, punctuation, numbers)
  - Transliteration of proper names
  - Synonyms
  - Additional info in MT output
  - Ambiguity in reference translation
  - What to do with incomprehensible MT
- Special rules for conversational, spoken genres

Post-Editing Tool



067

1

Reference translation (ref.mtf)

Previous Segments

No previous segment.

Current Segment

Gaza December 11 / Xinhua / Palestinian sources close to the Fatah Movement said today, Saturday, that the candidacy of Marwan Barghouti, the Secretary of the Movement in the West Bank, who is detained in Israeli jails, is illegal.

Next Segments

No next segment.

My translation (hyp.mtf)

Previous Segments

No previous segment.

Current Segment

Gaza, December 11 (Xinhua) Palestinian sources close to the Fatah movement said on Saturday that the candidature of Marwan Barghouti, secretary of the movement in the West Bank and prisoner in Israeli prisons, was illegal.

○ rf
○ mt
○ et
○ gl

request for review

Next Segments

No next segment.

Differences

Differences between original and my version.

Gaza, December 11 (Xinhua) Palestinian sources close to the Fatah movement said on Saturday that the candidature of Marwan ~~Barghouti~~ Barghouti, secretary of the movement in the West Bank and prisoner in Israeli ~~prisons~~ prisons, was illegal. (HTER 8.9%)

Original version

Gaza, December 11 (Xinhua) Palestinian sources close to the Fatah movement on Saturday that the candidature of Marwan Barghouti secretary of the movement in the West Bank prisoner in Israeli prisons was illegal.

Differences between reference translation and my

~~Gaza~~ Gaza, December ~~11 / Xinhua /~~ 11 (Xinhua) Palestinian sources close to the Fatah ~~Movement~~ movement said ~~today, Saturday,~~ on Saturday that the ~~candidacy~~ candidature of Marwan Barghouti, ~~the Secretary~~ secretary of the ~~Movement~~ movement in the West ~~Bank, who is detained~~ Bank and prisoner in Israeli ~~jails, is~~ prisons, was illegal. (HTER 35.6%)

Cost Factors

- Translation of ~100K words
  - 1 week FTE to prepare data and coordinate translators
  - 6-8 weeks calendar time for per "batch" of translation
    - Costs average $0.25/word
  - 3 weeks FTE for gold standard QC
- Post-editing of ~100K words
  - 1 week FTE technical, workflow, editor coordination
  - Editors earn on average $15-20/hour
    - Realtime rates vary by genre, MT output quality, editor experience
      - New editors: 3-4 wpm
      - Experienced editors: 7+ wpm
    - Additional financial incentives for quality, productivity

Conclusions

- Resources required vary depending on (explicit or implicit) assumptions of the various metrics
- Translation variation in the reference may be directly modeled or it may be assumed
- Consistency in application of manual metrics is influenced by both of these factors