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

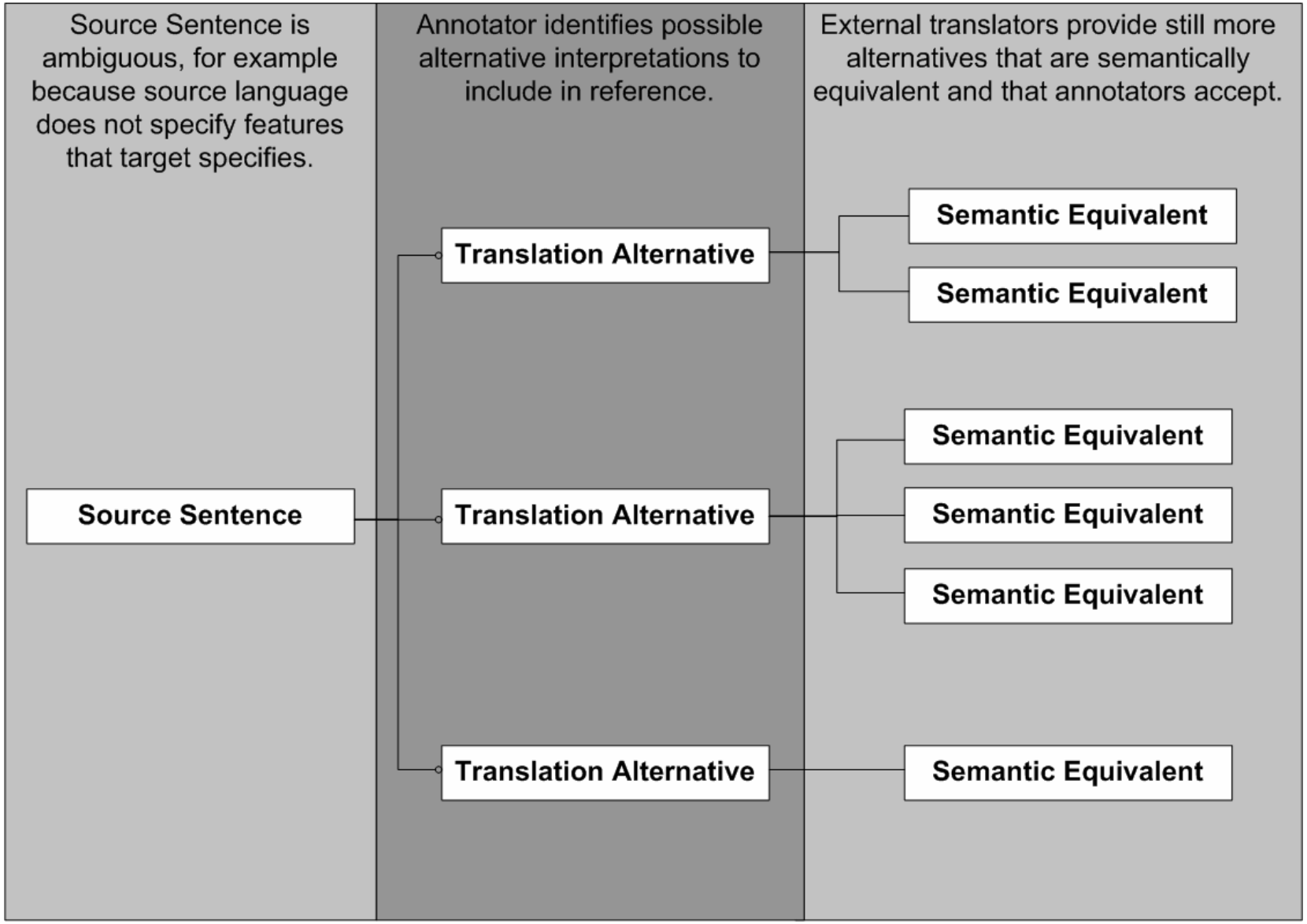
Error	Deduction
Syntactic	4 points
Lexical	2 points
Poor English usage	1 point
Significant spelling/punctuation error	½ points (max 5 points)

- 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



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

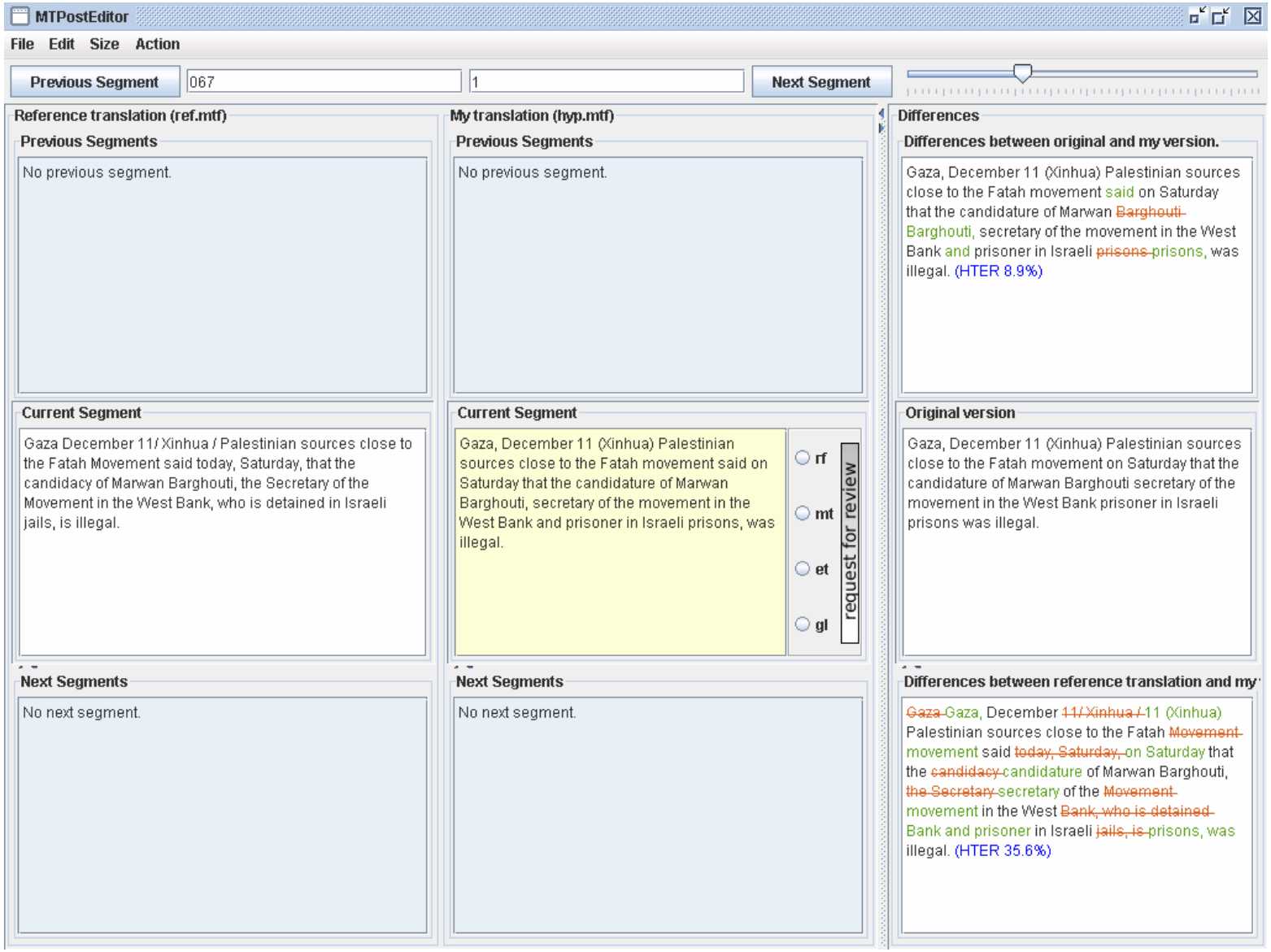
Post-Editor Training

- 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



Cost Factors

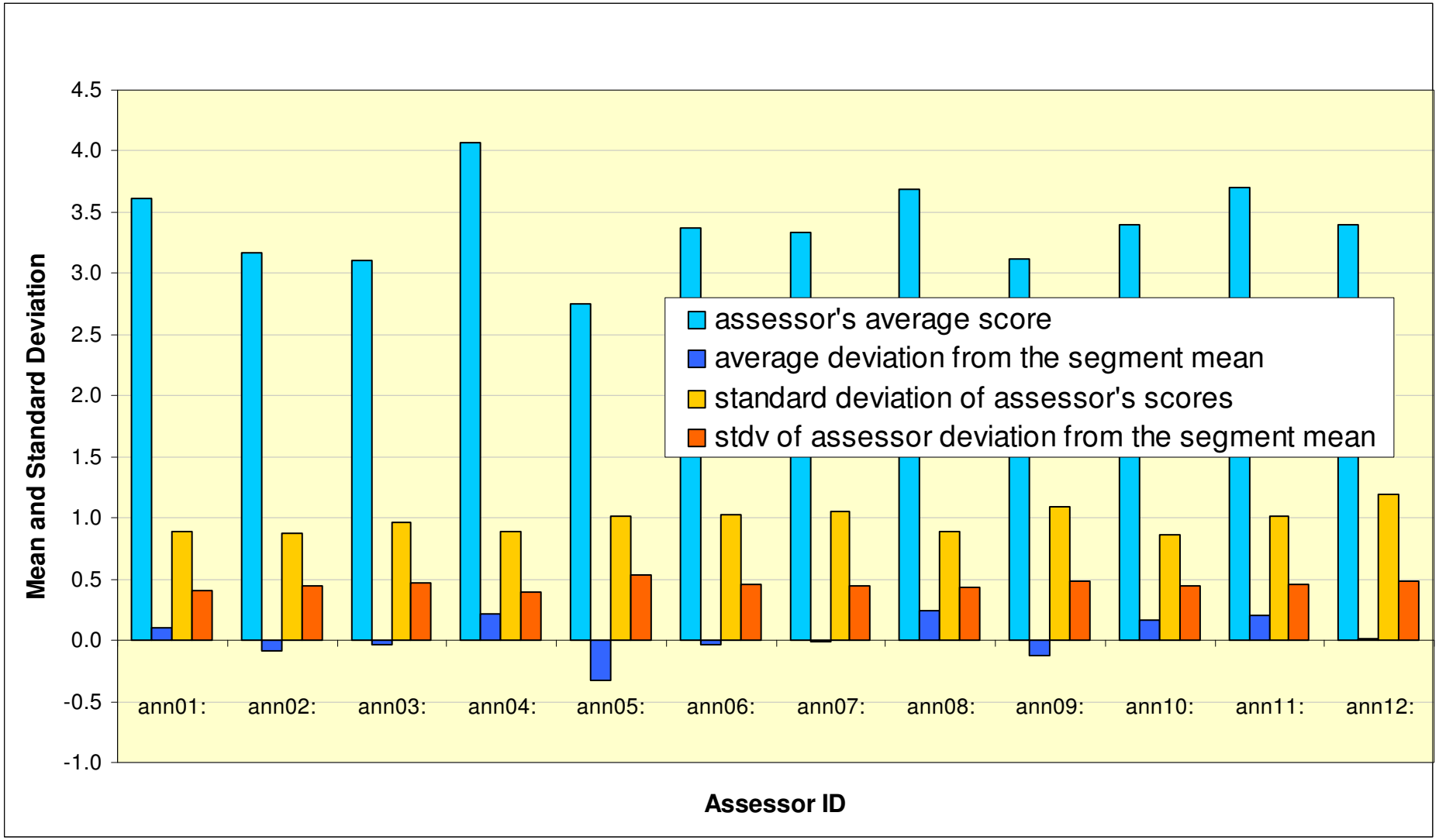
- Translation of ~100K words
  - 1 week FTE to prepare data and coordinate translators
  - 6-8 weeks calendar time for per “batch” of translation
    - Costs average $0.25/word
  - 3 weeks FTE for gold standard QC
- Post-editing of ~100K words
  - 1 week FTE technical, workflow, editor coordination
  - 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

Extra Slides

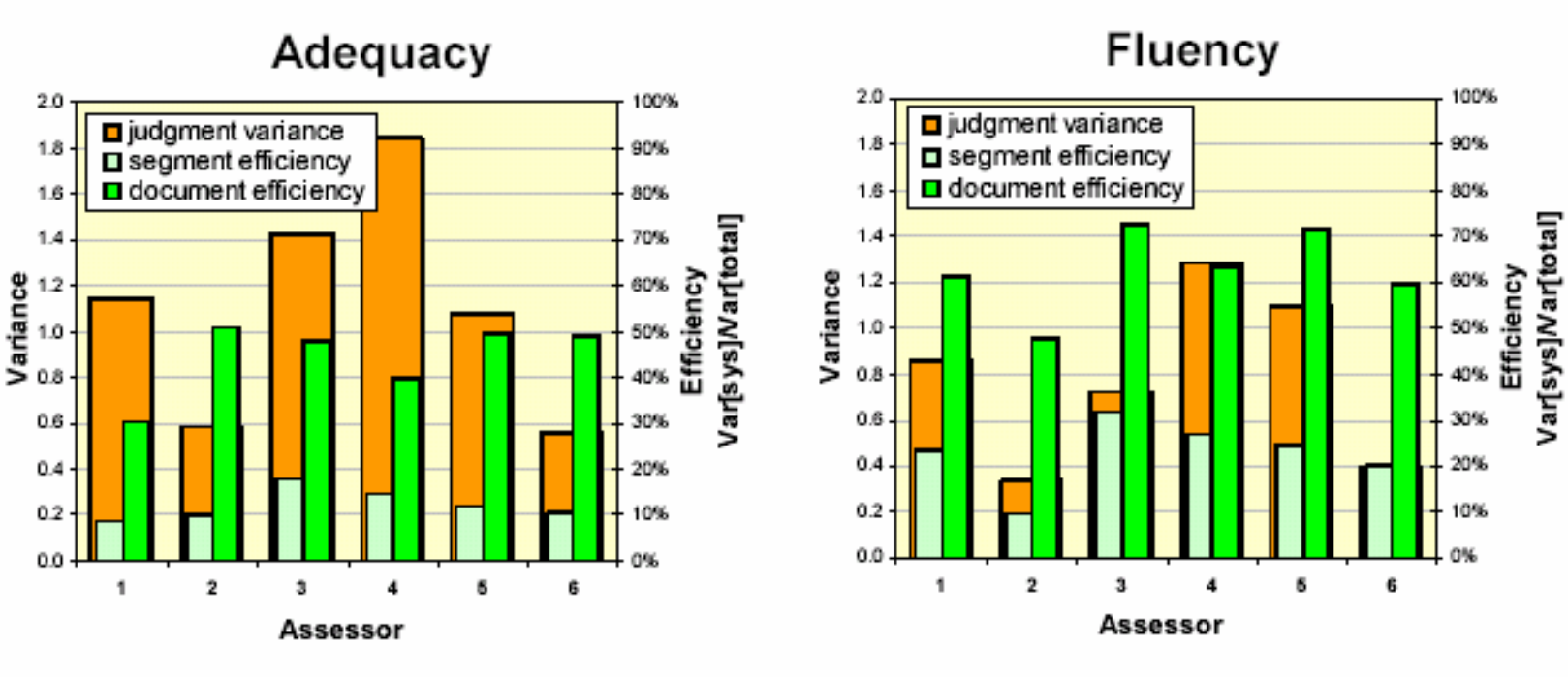
Assessor Consistency



**Thanks to George Doddington for these figures*

Adequacy & Fluency

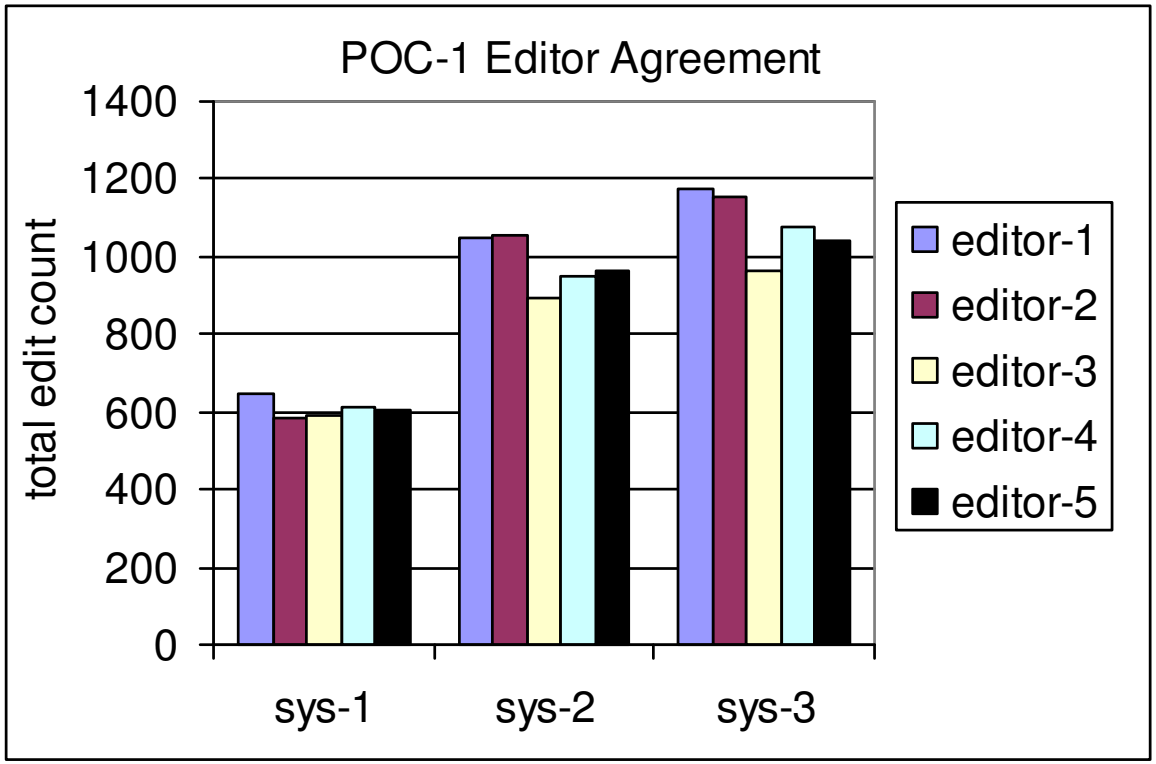
Results from MT05 Arabic to English*



**Thanks to George Doddington for these figures*

Edit Distance

- Pre-GALE proof of concept study*
  - 10 Arabic text documents
  - Translations from 3 MT systems
  - 5 volunteer editors



*Thanks to Greg Sanders at NIST for these figures

Document Assessment

An assessor reviews 1 document at a time. Each segment is judged first for *Fluency* and then for *Adequacy*, according to a 5-point scale.

Fluency – done without a "correct" reference:

How do you judge the fluency of this translation? It is:

1. **Flawless English**
2. **Good English**
3. **Non-native English**
4. **Disfluent English**
5. **Incomprehensible**

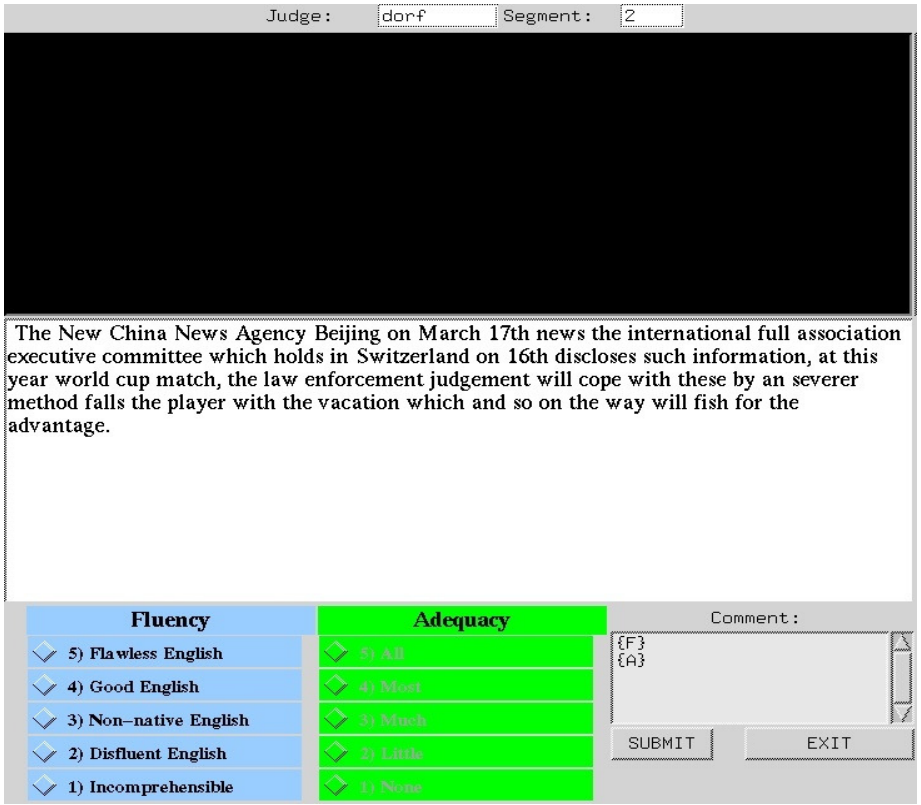
Adequacy – compared to a "correct" reference:

How much of the meaning expressed in the reference translation is also expressed in the target translation?

1. **All**
2. **Most**
3. **Much**
4. **Little**
5. **None**

MT Assessment GUI

For Fluency Judgments



For Adequacy Judgments

LDC Translation Team

- 1 FT senior administrator (linguist)
- 1 FT project manager responsible for translation agency management and translation QC
  - 2 FT lead annotators responsible for translation QC
    - 3-5 PT fluent bilingual translation QC assistants per language
- 1 FT project manager responsible for editor & assessor training & supervision
  - 2 PT assistants responsible for editor coordination and payment
- 1 FT programmer responsible for workflow system and translation tracking database
- 1 FT programmer responsible for data formatting and delivery processing

Data Management

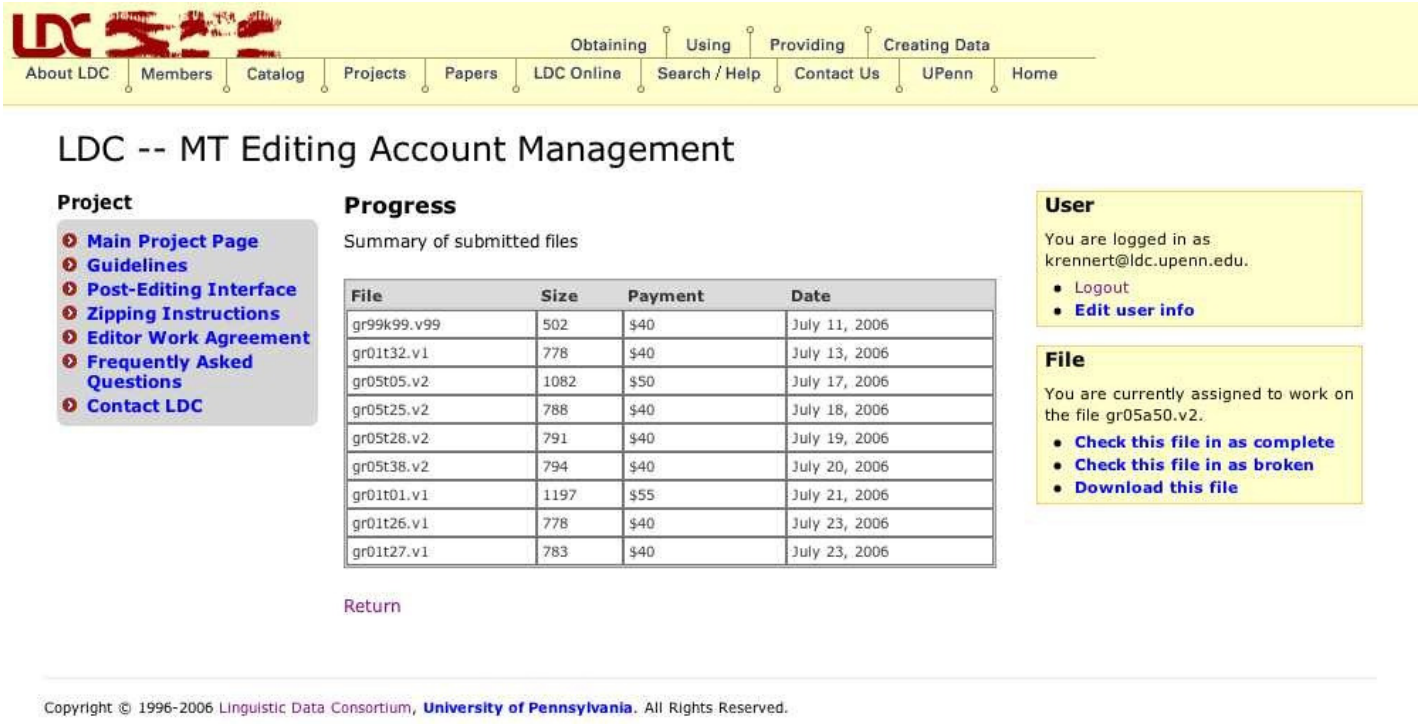
- MT Editing Workflow System Web Interface
  - Database backend tracks kit assignments and progress
  - Editors check out one kit at a time
    - Must submit completed kit before checking out another
    - First kit for each editor “frozen” until reviewed and approved
- Scripts control processing of completed kits
  - Workflow System runs script continually to search for newly submitted kits
    - Runs HTER scorer
    - Flags problems, automatically freezes kit and sends to manager for review
      - 20% or more segments have a high TER score
      - Unedited segment(s)
    - For any problem, manager reviews kit and leaves feedback for editor
    - For severe problems, manager returns kit to editor
- Web system logs problems, emails managers
  - Logs comments on kit reports
    - Time checked in/out
    - HTER scores for each stage
  - Daily progress reports per user, per kit, overall
  - Detailed statistics and graphical summary
  - HTER for each submitted kit (overall and per-segment)
  - Alerts for kits designated as problematic or needing further review

Post-Editing QC

- Manual
  - Detailed review of starter kit and first production kit
    - Feedback on problems and strategies to minimize HTER
  - Spot check for all remaining kits
  - Additional checks for flagged kits
  - Spell check on all kits
- Automatic warnings to managers & editors on check-in
  - Too-high HTER (suspicious)
  - Unedited segments
  - Poorly formatted kits
    - XML formatting errors
    - UTF-8 encoding errors
  - File ID or content mismatches

Post-Editor Management

- Editing supervisor, trouble ticket system for questions
- Editor website
  - Links to guidelines, tool manual, FAQ, editor help
  - Click to check out, check in files
  - Summary of progress and payment info



Obtaining | Using | Providing | Creating Data

About LDC | Members | Catalog | Projects | Papers | LDC Online | Search / Help | Contact Us | UPenn | Home

LDC -- MT Editing Account Management

Project

- Main Project Page
- Guidelines
- Post-Editing Interface
- Zipping Instructions
- Editor Work Agreement
- Frequently Asked Questions
- Contact LDC

Progress

Summary of submitted files

File	Size	Payment	Date
gr99k99.v99	502	$40	July 11, 2006
gr01t32.v1	778	$40	July 13, 2006
gr05t05.v2	1082	$50	July 17, 2006
gr05t25.v2	788	$40	July 18, 2006
gr05t28.v2	791	$40	July 19, 2006
gr05t38.v2	794	$40	July 20, 2006
gr01t01.v1	1197	$55	July 21, 2006
gr01t26.v1	778	$40	July 23, 2006
gr01t27.v1	783	$40	July 23, 2006

Return

User

You are logged in as krennert@ldc.upenn.edu.

- Logout
- Edit user info

File

You are currently assigned to work on the file gr05a50.v2.

- Check this file in as complete
- Check this file in as broken
- Download this file

- ***"Fluency" and "Adequacy"***: developed at PRC in 1993 to ***measure research progress***
- ***BLEU***: developed at IBM in 2001 to ***support MT research***
- ***DLPT****: developed at MIT Lincoln Lab in 2004 to ***measure operational readiness***
- ***GALE post-editing***: to be developed at NIST in 2005 to … *(deferred to Joe Olive)*