Evaluating Translation Quality

Presentation: Adam Lopez
Slides: Chris Callison-Burch

Evaluating MT Quality

- Why do we want to do it?
  - Want to rank systems
  - Want to evaluate incremental changes
- How not to do it
  - "Back translation"
  - The vodka is not good

Evaluating Human Translation Quality

- Why?
  - Quality control
  - Decide whether to re-hire freelance translators
  - Career promotion

DLPT-CRT

- Defense Language Proficiency Test/Constructed Response Test
- Read texts of varying difficulty, take test
- Structure of test
  - Limited responses for questions
  - Not multiple choice, not completely open
  - Test progresses in difficulty
  - Designed to assign level at which examinee fails to sustain proficiency

Level 1: Contains short, discrete, simple sentences. Newspaper announcements.
Level 2: States facts with purpose of conveying information. Newswire stories.
Level 3: Has denser syntax, convey opinions with implications. Editorial articles / opinion.
Level 4: Often has highly specialized terminology. Professional journal articles.

Human Evaluation of Machine Translation

- One group has tried applying DLPT-CRT to machine translation
  - Translate texts using MT system
  - Have monolingual individuals take test
  - See what level they perform at
- Much more common to have human evaluators simply assign a scale directly using fluency / adequacy scales
Fluency

• 5 point scale
• 5) Flawless English
  4) Good English
  3) Non-native English
  2) Disfluent
  1) Incomprehensible

Adequacy

• This text contains how much of the information in the reference translation:
• 5) All
  4) Most
  3) Much
  2) Little
  1) None

Relative ranking

• An alternative to absolute scales
• Simply ask
  - Is A better than B?
  - Is B better than A?
  - Or are they indistinguishable?

Consistent-based evaluation

• Rather than ranking the translations of whole sentences, instead have people focus on smaller parts

Human Evaluation of MT v. Automatic Evaluation

• Human evaluation is
  - Ultimately what we're interested in, but
  - Very time consuming
  - Not re-usable
• Automatic evaluation is
  - Cheap and reusable, but
  - Not necessarily reliable
Goals for Automatic Evaluation

- No cost evaluation for incremental changes
- Ability to rank systems
- Ability to identify which sentences we’re doing poorly on, and categorize errors
- Correlation with human judgments
- Interpretability of the score

Methodology

- Comparison against reference translations
- Intuition: closer we get to human translations, the better we’re doing
- Could use WER like in speech recognition

Word Error Rate

- Levenshtein Distance (also "edit distance")
- Minimum number of insertions, substitutions, and deletions needed to transform one string into another
- Useful measure in speech recognition
  - Shows how easy it is to recognize speech
  - Shows how easy it is to wreck a nice beach

Problems with WER

- Unlike speech recognition we don’t have the assumptions of
  - linearity
  - exact match against the reference
- In machine translation there can be many possible (and equally valid) ways of translating a sentence
- Also, clauses can move around, since we’re not doing transcription

Solutions

- Compare against lots of test sentences
- Use multiple reference translations for each test sentence
- Look for phrase / n-gram matches, allow movement

Metrics

- Exact sentence match
- WER
- PI-WER
- Bleu
- Precision / Recall
- Meteor
Bleu

- Use multiple reference translations
- Look for n-grams that occur anywhere in the sentence
- Also has "brevity penalty"
- Goal: Distinguish which system has better quality (correlation with human judgments)

Example Bleu

**R1:** It is a guide to action that ensures that the military will forever heed Party commands.
**R2:** It is the Guiding Principle which guarantees the military forces always being under the command of the Party.
**R3:** It is the practical guide for the army always to heed the directions of the party.

**C1:** It is to insure the troops forever hearing the activity guidebook that party direct.
**C2:** It is a guide to action which ensures that the military always obeys the command of the party.

Automated evaluation

- Because C2 has more n-grams and longer n-grams than C1 it receives a higher score
- Bleu has been shown to correlate with human judgments of translation quality
- Bleu has been adopted by DARPA in its annual machine translation evaluation

Interpretability of the score

- How many errors are we making?
- How much better is one system compared to another?
- How useful is it?
- How much would we have to improve to be useful?
Evaluating an evaluation metric

• How well does it correlate with human judgments?
  - On a system level
  - On a per sentence level
• Data for testing correlation with human judgments of translation quality

NIST MT Evaluation

• Annual Arabic-English and Chinese-English competitions
• 10 systems
• 1000+ sentences each
• Scored by Bleu and human judgments
• Human judgments for translations produced by each system

ACL Workshop on SMT

• Translation between English, French, German, Spanish, Hungarian and Czech
• 30 different systems
• In-domain and out-of-domain test sets
• Scores produced by multiple automatic metrics
• Systems ranked by 100+ human judges

Final thoughts on Evaluation

When writing a paper

• If you’re writing a paper that claims that
  - one approach to machine translation is better than another, or that
  - some modification you’ve made to a system has improved translation quality
• Then you need to back up that claim
• Evaluation metrics can help, but good experimental design is also critical

Experimental Design

• Importance of separating out training / test / development sets
• Importance of standardized data sets
• Importance of standardized evaluation metric
• Error analysis
• Statistical significance tests for differences between systems
Invent your own evaluation metric
- If you think that Bleu is inadequate then invent your own automatic evaluation metric
- Can it be applied automatically?
- Does it correlate better with human judgment?
- Does it give a finer grained analysis of mistakes?

Evaluation drives MT research
- Metrics can drive the research for the topics that they evaluate
- NIST MT Eval / DARPA Sponsorship
- Bleu has lead to a focus on phrase-based translation
- Minimum error rate training
- Other metrics may similarly change the community’s focus

Homework Exercise
- Evaluation exercise for homework
- Examine translations from state-of-the-art systems (in the language of your choice!)
- Manually evaluate quality!
- Perform error analysis!
- Develop ideas about how to improve SMT!