Evaluation of Translation Quality

ESSLLI 2005

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

Ways to Evaluate

- Task-based evaluation
  - Reading comprehension
  - Assemble something
  - Navigate a map
- Assign a value using a quality scale
  - US military categorization system
  - Fluency / Adequacy

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

DLPT-CRT

- 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

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 referece
- 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.

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.

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.

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

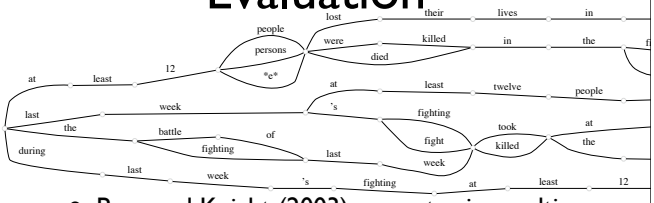
NIST MT Evaluation

- 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

Tricks with automatic evaluation

- Learning curves -- show how increasing training data improves statistical MT
- Euromatrix -- create translation systems for every pair of European languages, and give performance scores
- Finite state graphs from multi-reference translations -- Exact sentence matches possible?

Multi-reference Evaluation



- Pang and Knight (2003) suggest using multi-reference evaluation to do *exact match*
- Combine references into a word graph with hundreds of paths through it
- Most paths correspond to good sentences

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

Other Uses for Parallel Corpora

Chris Callison-Burch
ESSLLI 2005

Statistical NLP and Training Data

- Most statistical natural language processing applications require training data
  - Statistical parsing requires treebanks
  - WSD requires text labeled w/ word senses
  - NER requires text w/named entities
  - SMT requires parallel corpora

Cost of Creating Training Data

- Creating this training data is usually time consuming and expensive
- As a result the amount of training data is often limited
- SMT is different
  - Parallel corpora are created by other human industry
  - For some language pairs huge data sets are available

Exploiting Parallel Corpora

- Can we use this abundant resource for tasks other than machine translation?
- Can we use it to alleviate the cost of creating training data?
- Can we use it to port resources to other languages?

Three Applications of Parallel Corpora

- Automatic generation of paraphrases
- Creating training data for WSD
- "Projecting" annotations through parallel corpora so that they can be applied to new languages

Paraphrasing with Bilingual Parallel Corpora

Paraphrasing

- Paraphrases are alternative ways of conveying the same information
- Useful in NLP application such as:
  - *Generation:* more varied and fluent text
  - *Multidocument summarization*: allows repeated information to be condensed
  - *Question answering*: paraphrases of same answer provide evidence of correctness

Previous Approaches

- Used *monolingual* parallel corpora
- Multiple translations of the same thing
  - Multiple translations of classic French novels into English
  - Evaluation data for Bleu MT Eval metric
- People have also used *comparable* corpora (encyclopedia articles on the same topic)

Paraphrasing with monolingual parallel data

- Methodology:
  - Align sentences across translations
  - Identify similar contexts in aligned sentences
  - Phrases that appear in similar contexts may be paraphrases

Emma burst into tears **and he tried to** comfort **her,** saying things to make her smile.
Emma cried, **and he tried to** console **her,** adorning his words with puns.

burst into tears = cried, console = comfort

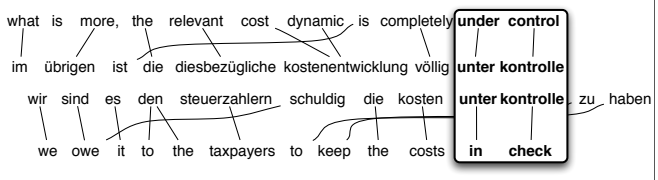
Problems

- Monolingual parallel corpora are very uncommon
- This fact might limit what paraphrases we are able to generate

Paraphrasing with bilingual parallel data

- Our Methodology:
  - Use MT techniques to align English-German parallel text
  - Get German phrase aligned with the English phrase we want to paraphrase
  - Find other English phrases that German phrases align with
  - Treat those English phrases as paraphrases, and rank them

Example



Extracted Paraphrases

- military force → armed forces, defence, force, forces, peace-keeping personnel, military forces
- sooner or later → at some point, eventually
- great care → a careful approach, greater emphasis, particular attention, specific attention, special attention, very careful
- at work → at the workplace, employment, held, holding, in the work sphere, organised, operate, taken place, took place, working

Word Sense Disambiguation with Parallel Corpora

Word Sense Disambiguation

- Define a set of senses for words, or draw them from a dictionary
- $plant_1$=foliage, $plant_2$= factory, $plant_3$= to put something in the ground
- Develop an algorithm to assign a sense to a word in context

Data for WSD

- Statistical approaches to WSD generally use a large set of labeled training data, where each of the words to be disambiguated has been labeled with its sense
- The Senseval competitions create such data, for training and for evaluation
- Generally label 1,000s of instances of around 30 vocabulary items

Problem

- Data is costly to create
- Consequently only a few vocabulary items get labeled
- There is a bottleneck in the process of creating statistical WSD systems

Breaking the Bottleneck

- Rather than doing WSD with explicitly labeled data we could do it with parallel corpora
- Treat words as polysemous when they translate to more than one foreign word

$drugs_1$	regards classification as medicinal products or **drugs** , *en particulier par rapport aux* ***médicaments*** *et autres produits pharmaceutiques ,*
$drugs_2$	in my country we have 2 million people on illegal **drugs** , *dans mon pays , 2 millions de personnes consomment des* ***drogues*** *illicites ,*

Just say no to hard drugs!

Senseless Violence

- Two options:
  - Either use the foreign words as the senses
  - Or keep using the senses as given dictionary and use the foreign words as a way of acquiring more labeled data
- When a particular dictionary sense only occurs with one of the foreign word then use sentence pairs which have that contain that word as additional data points

Annotation Projection for Parsing

Statistical Parsing

- Statistical parsers are trained on treebanks containing sentences annotated with parse trees
- Developing treebanks requires linguistics expertise and is a time consuming, expesive process
- Treebanks exist only for a limited number of languages, and are often constrained in size

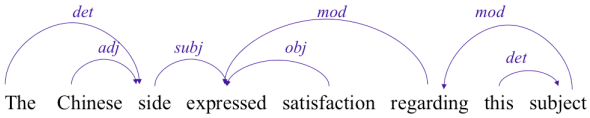
Development time for Treebanks

Language	Treebank	Dev Time	Size of corpus	Parser Performance
English	Penn Treebank	5 Years	1M words 40k sentences	90%
Chinese	Chinese Treebank v2	2 Years	100k words 4k sentences	75%
Chinese	Chinese Treebank v2	4 Years	400k words 15k sentences	~80%
Others (Farsi, Hindi)	?	?	?	?

Exploit Resources for New Languages

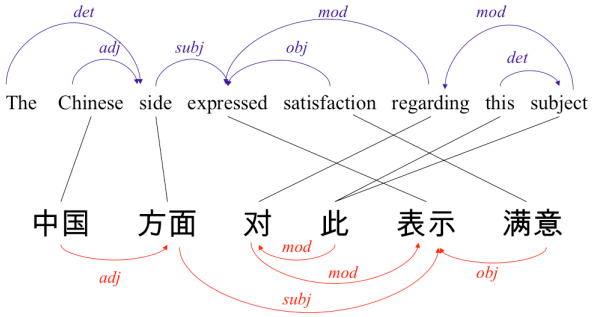
- Use high accuracy English parsers to parse English section of a parallel corpus
- Align the parallel corpus
- Project the trees onto the other language
- Use that to train a foreign language parser

Create an English Dependency Parse



中国 方面 对 此 表示 满意

Project the Dependency Parse



Automatically Create a Foreign Treebank

- Continue repeating those steps for all of the sentences in the parallel corpus
- Use the project trees to train a Chinese parser

Potential Problems

- Low quality word alignments may cause problems
- Mis-matches between the syntax of languages complicate things
- Differences between text that English parser was trained on (newswire) and parallel corpus (gov't) may exacerbate things

Results

- With manually created alignments, a 67% accuracy was achieved
- With automatic alignments, a 57% accuracy was achieved
- Equivalent to manually creating a treebank containing 3000 sentences

Conclusions

- Parallel data can be exploited to supplement data for a number of statistical NLP tasks
- Many different tasks can be treated using annotation projection
- Increases the feasibility of developing NLP technologies for other languages
- As quality of word alignment algorithms improve, these techniques will become more viable