Statistical Machine Translation

-evaluation-

Artem Sokolov
Computerlinguistik
Universität Heidelberg
Sommersemester 2015

material from P. Koehn

- how good is a given machine translation system?
- hard problem, due to language flexibility
  (what is a correct translation anyway?)
- evaluation metrics
  - ➡ subjective judgments by human evaluators
    (probably the best one, but costly)
  - ➡ automatic evaluation metrics
    (cheap, but only approximates 'true' quality)
  - ➡ task-based evaluation (it depends)
    - post-editing effort
      (time, count of edit operations, mouse clicks, reorderings)
    - grounding
      (task accomplished? CLIR, sales)

这个 机场 的 安全 工作 由 以色列 方面 负责 .

Israeli officials are responsible for airport security.
Israel is in charge of the security at this airport.
The security work for this airport is the responsibility of the Israel government.
Israeli side was in charge of the security of this airport.
Israel is responsible for the airport's security.
Israel is responsible for safety work at this airport.
Israel presides over the security of the airport.
Israel took charge of the airport security.
The safety of this airport is taken charge of by Israel.
This airport's security is the responsibility of the Israeli security officials.

Human judgement:
given: machine translation output and source and/or reference translation
task: assess the quality of the machine translation output
Metrics

- adequacy
  (does the output convey the same meaning as the input sentence?)
- fluency
  (is the output good fluent English?)

➡ slow, costly, inconsistent, confusion between adequacy & fluency, hard to tune

goal: computer program that computes the quality of translations

- pros: low cost, tunable, consistent
- cons: questionable meaningfulness, still low tunability, idiosyncratic (has quirks)

strategy

- given: machine translation output **and** human reference translation(s)
- task: compute some similarity between them

SYSTEM A: Israeli officials ~~responsibility of~~ airport ~~safety~~

REFERENCE: Israeli officials are responsible for airport security

- Precision

$$\frac{\textit{correct}}{\textit{output-length}} = \frac{3}{6} = 50\%$$

- Recall

$$\frac{\textit{correct}}{\textit{reference-length}} = \frac{3}{7} = 43\%$$

- F-measure

$$\frac{\textit{precision} \times \textit{recall}}{(\textit{precision} + \textit{recall})/2} = \frac{.5 \times .43}{(.5 + .43)/2} = 46\%$$

Metric	System A	System B
precision	50%	100%
recall	43%	86%
f-measure	46%	**92%**

flaw: no penalty for wrong ordering

Need a *repeatable* evaluation method that uses:

- a gold standard of human generated **references** (better use many)
- a numerical **translation closeness** metric in order to compare the system output against human references

Motivation

- in MT we are mostly interested in precision
- also in recall, but to a lesser extent

Problem with maximizing precision:

Candidate: the the the the the the the
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.
Unigram precision: $\frac{7}{7} = 1$

Problem with maximizing recall:

Candidate 1: I always invariably perpetually do.
Candidate 2: I always do.
Reference 1: I always do.
Reference 2: I invariably do.
Reference 3: I perpetually do.
Recall(candidate 1) $>$ Recall(candidate 2);

Reminder

$$\text{precision} = \frac{\textit{correct}}{\textit{output-length}}$$

Clipped precision

1. $\forall$ n-gram, $\forall$ hypothesis, count the max number of n-gram matches in a single reference
2. $\forall$ n-gram, $\forall$ hypothesis, clip the total number of matches of a candidate n-gram by the max reference match.
3. $\forall$ n-gram, add up clipped matches over all candidate sentences in corpus.
4. $\forall$ n-gram, divide by the total number of unclipped hypothesis n-gram counts in corpus.

$$p_n = \frac{\sum_{c \in \{\text{candidates}\}} \sum_{\text{n-gram} \in c} \text{count}_{\text{clip}}(\text{n-gram})}{\sum_{c' \in \{\text{candidates}\}} \sum_{\text{n-gram}' \in c'} \text{count}(\text{n-gram}')}$$

Candidate: of the
Ref1: It is a guide to action that ensures that the military will forever heed Party commands.
Ref2: It is the guiding principle which guarantees the military forces always being under the command of the Party.
Ref3: It is the practical guide for the army always to heed the directions of the Party.

$$\text{Modified unigram } p = \frac{1 \cdot \text{of} + 1 \cdot \text{the}}{1 \cdot \text{of} + 1 \cdot \text{the}} = \frac{2}{2} = 1$$

$$\text{Modified bigram } p = \frac{1 \cdot \text{of the}}{1 \cdot \text{of the}} = \frac{1}{1} = 1$$

$$\sum_{n=1}^{N} \frac{1}{N} \log p_n = \log \underbrace{(\prod_{n=1}^{N} p_n)^{\frac{1}{N}}}_{\text{geometric mean}}$$

N is (almost) always set to 4.

$$\text{recall} = \frac{\textit{correct}}{\textit{reference-length}}$$

- hypothesis longer than references are already penalized by clipped precision
- use a multiplicative recall-related measure to penalize shorter hypothesis

$$\text{BP} = \begin{cases} 1 & \text{if } c > r, \\ \exp\left(1 - \frac{r}{c}\right) & \text{if } c \leq r. \end{cases}$$

- computed over corpus in order to avoid harsh penalties on short sentences
- **corpus reference length** r is the sum over best match lengths for each candidate*
- **corpus candidate length** c is the total length of candidates in corpus

Definition

- n-gram overlap between machine translation output and reference translation
- account for precision: compute n-gram precision size 1 to n (n is usually 4)
- account for recall (in a way): penalize too short translations (brevity penalty)

$$\log \texttt{bleu} = \min(1 - \frac{r}{c}, 0) + 0.25 \sum_{n=1}^{4} \log p_n$$

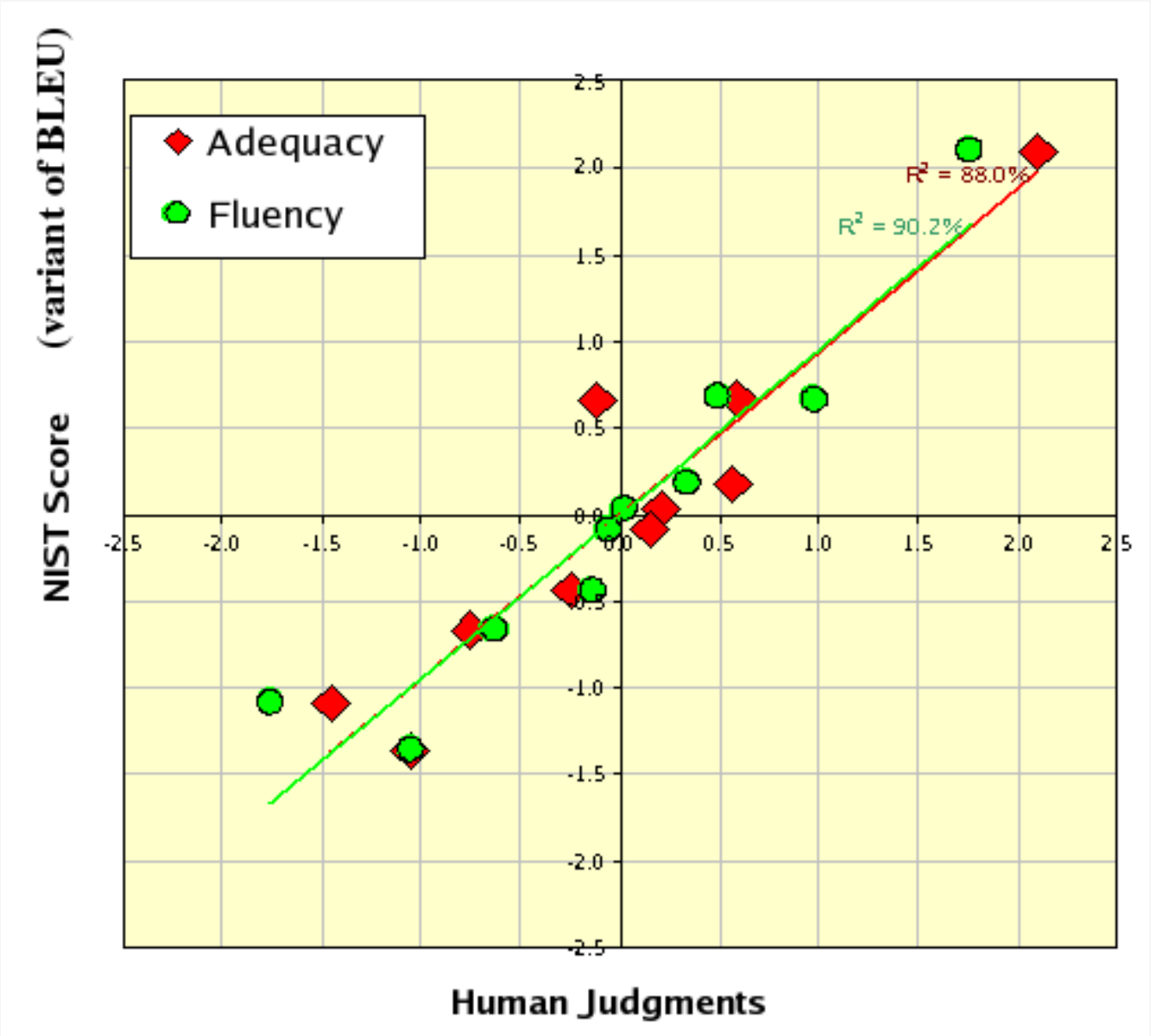
BLEU example

SYSTEM A: Israeli officials (2-GRAM MATCH) responsibility of airport (1-GRAM MATCH) safety

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security (2-GRAM MATCH) Israeli officials are responsible (4-GRAM MATCH)

Metric	System A	System B
precision (1gram)	3/6	6/6
precision (2gram)	1/5	4/5
precision (3gram)	0/4	2/4
precision (4gram)	0/3	1/3
brevity penalty	6/7	6/7
BLEU	0%	52%


NIST Score (variant of BLEU)
Human Judgments
Adequacy
Fluency
R² = 88.0%
R² = 90.2%
2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0
2.5

1. BLEU is **not sufficient** to reflect genuine translation quality. (permutations on unigram or bigram level do not reduce BLEU)
2. BLEU improvement is **not necessary** for improved translation quality. (translations from different systems are not well distinguished by BLEU)

- ignores relevance of words (names and core concepts more important than determiners and punctuation)
- operates on local level (do not consider overall grammaticality of the sentence or sentence meaning)
- scores are meaningless (scores very test-set specific, absolute value not informative)
- human translators score low on BLEU (possibly because of higher variability, different word choices)

Levenstein distance:
minimum number of editing operations to transform output to reference

Operations

- substitution
- insertion
- deletion

Word Error Rate:

$$\text{WER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference-length}}$$

Operations

1. word insertion
2. word deletetion
3. word substitution
4. **block of words move** (phrasal shift)

$$\text{TER} = \frac{\text{substitutions} + \text{insertions} + \text{deletions} + \text{block moves}}{\text{reference-length}}$$

Note: Unlike BLEU, lower TER scores are better.

Example TER

REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times

HYP: THIS WEEK THE SAUDIS denied information published in the new york times

- "this week" in HYP is in a "shifted" position with respect to REF
- "Saudi Arabia" in REF appears as "the Saudis" in HYP (counts as 2 substitutions).
- "American" appears only in REF.

- TER $= \frac{4}{13} = \mathbf{0.13}$ (not bad)

Example TER

REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times

HYP: THIS WEEK THE SAUDIS denied information published in the new york times

- "this week" in HYP is in a "shifted" position with respect to REF
- "Saudi Arabia" in REF appears as "the Saudis" in HYP (counts as 2 substitutions).
- "American" appears only in REF.

- TER $= \frac{4}{13} = \mathbf{0.13}$ (not bad)
- breakdown of n-grams precision: 0.833/0.545/0.300/0.111, brevity 0.920
- $1 - \text{BLEU} = 1 - 0.32 = 0.68 \gg 0.13$ (**fail**)

Example TER

REF: SAUDI ARABIA denied THIS WEEK information published in the AMERICAN new york times

HYP: THIS WEEK THE SAUDIS denied information published in the new york times

- "this week" in HYP is in a "shifted" position with respect to REF
- "Saudi Arabia" in REF appears as "the Saudis" in HYP (counts as 2 substitutions).
- "American" appears only in REF.

- TER $= \frac{4}{13} = \mathbf{0.13}$ (not bad)
- breakdown of n-grams precision: 0.833/0.545/0.300/0.111, brevity 0.920
- $1 - \text{BLEU} = 1 - 0.32 = 0.68 \gg 0.13$ (**fail**)

Reason: bad accounting for phrasal shift!

- if we had operations w/o shifts, the Levenstein distance is $\mathcal{O}(n^2)$
- adding shifts makes the problem NP-hard, so an approximation must be used

I min. number of insertions, deletions, and substitutions is calculated using dynamic programming.

II a greedy search is used to find the set of shifts, by repeatedly selecting the shift that most reduces the number of basic edits, until no more beneficial shifts remain.

III dynamic programming is reused to optimally calculate the remaining edit distance using a minimum-edit-distance over 3 basic operations

Require: hypothesis h, references R

$E \leftarrow \infty$
for all $\forall r \in R$ **do**
  $h' \leftarrow h$
  $e \leftarrow 0$
  repeat
    {Find shift, s, that most reduces min-edit-distance(h', r)}
    if s reduces edit distance **then**
      $h' \leftarrow$ apply s to h
      $e \leftarrow e + 1$
    end if
  until no distance-reducing shifts remain
  $e \leftarrow e+$ min-edit-distance(h', r)
  if $e < E$ **then**
    $E \leftarrow e$
  end if
end for

- Situation
  - ➡ system A has score x on a test set
  - ➡ system B has score y on the same test set
  - ➡ $x > y$
- Is system A really better than system B?
- In other words:
  Is the difference in score **statistically significant**?

- Null hypothesis: The two systems are equal and observed difference is random.
- p-value: probability of incorrectly rejecting null hypothesis. A small p-value (≤ 0.05) means that observed difference is statistically significant, i.e., difference is not random.

p-value is:

- the probability to the same result if H_0 were true
- NOT the probability that H_0 is true
- NOT the "probability that the results are due to chance"
- NOT whether the experiment is reliable

- Non-parametric:
  - ➡ Sign test/binomial test
  - ➡ Wilcoxon signed rank test
- Parametric:
  - ➡ Student's t-test
- Distribution-free:
  - ➡ Randomization test
  - ➡ Bootstrap test

1: Set $c = 0$
2: Compute actual statistic of score differences $|S_X - S_Y|$ on test data for system X, Y
3: **for all** random shuffles $r = 0, \ldots, R$ **do**
4: **for all** sentences in test set **do**
5: Shuffle variable tuples between system X and Y with probability 0.5
6: **end for**
7: Compute pseudo-statistic $|S_{X_r} - S_{Y_r}|$ on shuffled data
8: **if** $|S_{X_r} - S_{Y_r}| \geq |S_X - S_Y|$ **then**
9: c ++
10: **end if**
11: **end for**
12: $p = (c + 1)/(R + 1)$
13: Reject null hypothesis if p is less than or equal to specified rejection level.