Statistical Machine Translation

LECTURE – 10

MT EVALUATION

APRIL 23, 2010

Outline

Reasons for Evaluation

Types of MT evaluation (Manual, Automatic)

Metrics:

- Edit-Distance (SER,WER,PER,TER,RED)
- Precision Based - BLEU/ NIST
- Recall Based - ROUGE(ROUGE-N/L/W/S)
- METEOR

Problems in MT evaluation

Conclusions

Future Scope

Reasons for Evaluation

- Comparison with humans
- Comparison between multiple MT systems
- Decision to use or buy a particular MT system
- Tracking technological process
- Improvement of a particular system

and

- A very interesting Research Topic!!

Manual Evaluation

Criteria for Manual Evaluation - adequacy & fluency

Fluency: A fluent sentence is one that is

- well-formed,
- grammatically correct,
- contains correct spellings,
- adheres to common use of terms, titles, names
- intuitively acceptable,
- can be sensibly interpreted by a native speaker

Adequacy: to what extent the meaning of the source language sentence is conveyed by the generated target language sentence.

These can be rated on a scale of 0-5, say as follows.

Human evaluation

Source Language Sentence : **Je suis fatigué.**

Translated Text	Adequacy	Fluency
Tired is I	5	2
I was fatty!	0	5
I am tired	5	5

Note that TDMT recommended 4 categories:

A - Perfect B – Fair

C - Acceptable D - Nonsense

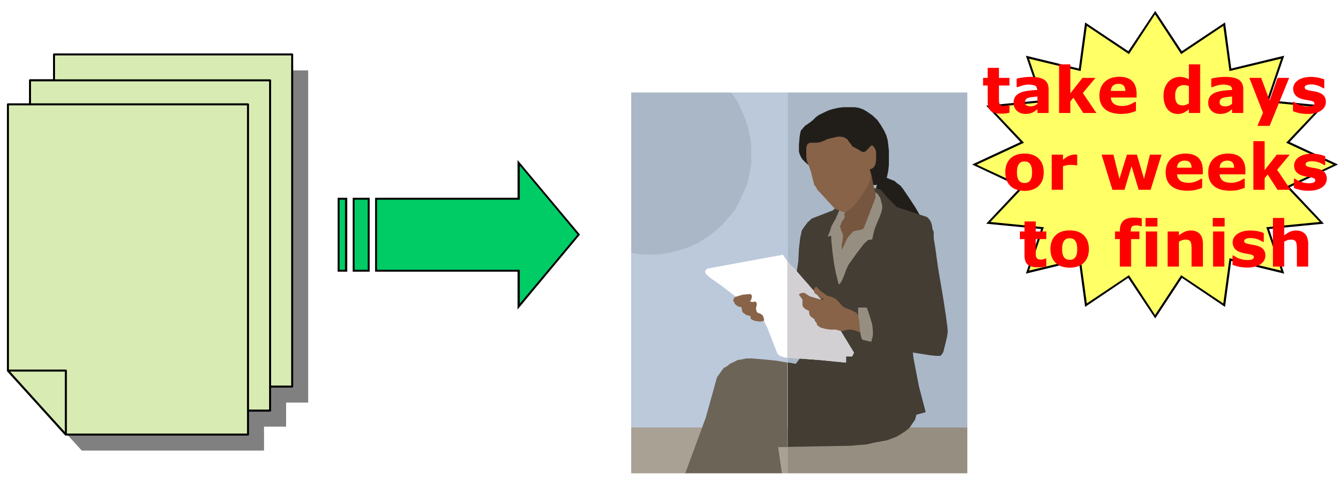
Human evaluation

Human Evaluation is of high quality and is very accurate

Human evaluations of machine translation (MT) Problem: evaluation bottleneck

- Human evaluation is costly, time consuming and non-repetitive.
- Developers need to evaluate daily changes to improve machine translation system

Machine Translation documents



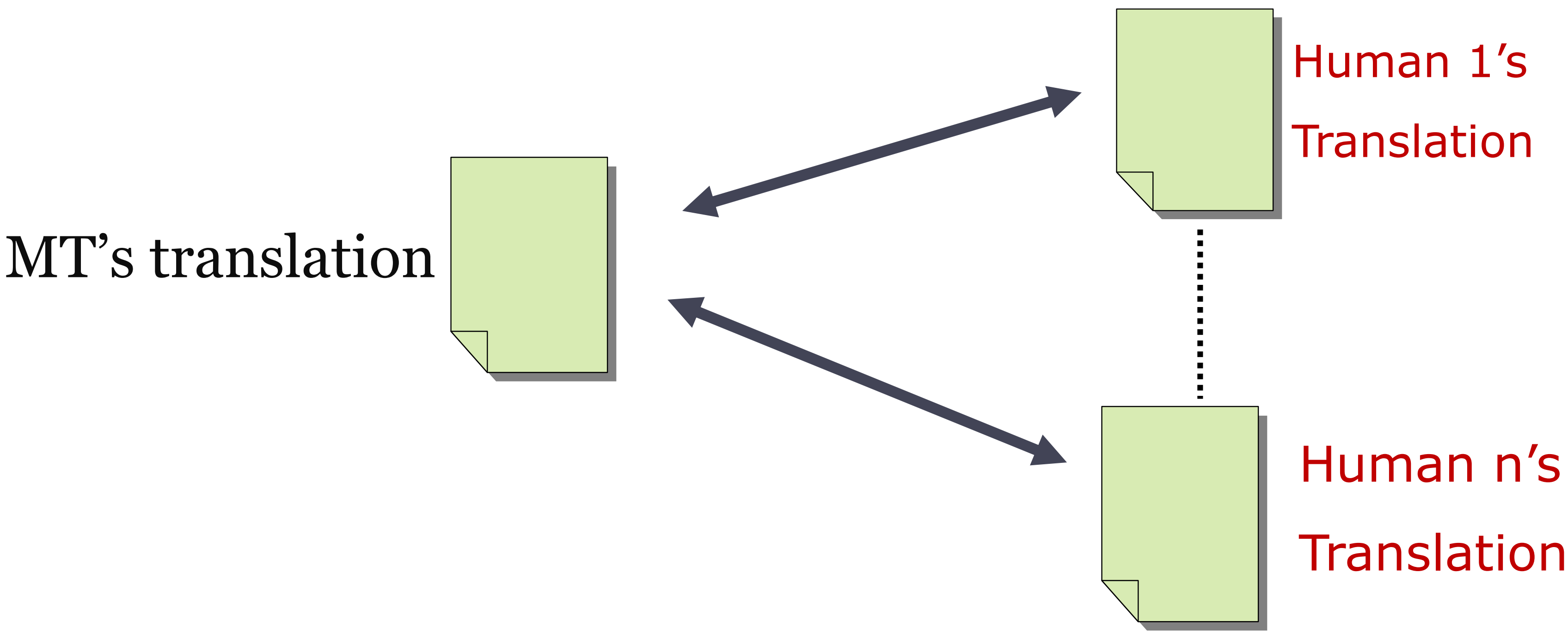
Automatic Evaluation

- Evaluation metric: method for assigning a numeric score to a hypothesized translation

- Automatic evaluation metrics often rely on comparison with previously completed human translations

Automatic Evaluation

The closer a machine translation (MT) is to a professional human translation (HT), the better it is



Measures closeness of one MT to one or more reference HTs according to a numerical metric

Automatic Evaluation

- Pros- inexpensive , quick & unbiased

- Cons-
  - Quality is lower as compared to manual evaluation.
  - Reference translations are needed.

Criteria for MT Evaluation

Completeness

- Lexical completeness: A system is lexically complete if it has source and target language lexicon entries for every word or phrase in the translation domain.
- Grammatical completeness: A system is grammatically complete if it can analyze all of the grammatical structures encountered in the source language, and it can generate all of the grammatical structures necessary in the target language translation.
- Mapping Rule completeness: A system is complete with respect to mapping rules if it assigns an output structure to every input structure in the translation domain, regardless of whether this mapping is direct or via an interlingua.

Criteria for MT Evaluation

Correctness: a system is correct if it assigns a correct output string to every input string it is given to translate.

- Lexical correctness: If each of the words selected in the target sentence is correctly chosen for the concept that it is intended to realize.
- Syntactic correctness: The grammatical structure of each target sentence should be completely correct (no grammatical errors).
- Semantic correctness: Semantic correctness presupposes lexical correctness, but also requires that the compositional meaning of each target sentence should be equivalent to the meaning of the source sentence.

Automatic Evaluation Metrics

- Edit-Distance based:
  SER,WER, PER, RED, TER

- Precision based:
  BLEU, NIST

- F-measure (Precision & Recall) based:
  METEOR, ROUGE

Edit-Distance based Metrics

Edit-Distance (Word Accuracy)

- metric to determine closeness of translations automatically
- the least number of edit operations to turn the translated sentence into the reference sentence
- Advantages
  - fully automatic given a reference set
- Disadvantages
  - penalizes candidates if a synonym is used
  - penalizes swaps of words and block of words too much

EDIT-DISTANCE

$WA = 1 - ((d+s+i)/\max(r,c))$

- d= number of deletions
- s = number of substitutions
- i = number of insertions
- r = reference sentence length
- c = candidate sentence length
- easy to calculate using Levenshtein distance algorithm (dynamic programming)
- various extensions have been proposed

EDIT-DISTANCE Types

- SER (Sentence Error Rate)
- WER (Word Error Rate)
- PER (Position-Independent WER)
- RED (Ranker based on Edit-Distances)-
- TER (Translation Error/Edit Rate

Sentence Error Rate (SER)

- Sentence Error Rate (SER) is a measure of the number of translations produced which exactly match the reference translation.
- To calculate SER for any given test set we simply count the number of output translations which match their corresponding reference translations exactly.
- Express this count as a percentage of the total number of sentences in the original test set. As SER is an error rate, we subtract this percentage from 100 in order to give us our final figure.

Example- SER

English Sentence	Machine Translation	Reference Translation
Did you enjoy reading this book?	*Kyaa aapko yah pustak Padhne mei mazaa aayaa*	*Kyaa aapko yah pustak Padhne mei mazaa aayaa*
I blame myself for not paying attention.	*Dhyaan na de pane ke liye main swayam ko doshi maantaa hoon*	*Dhyaan na de pane ke liye main swayam ko doshi maantaa hoon*
We shall now begin to work.	*Hum karya karnaa ab prarambh honge*	*Hum ab kaam karnaa shuru karenge*
That's going to take hundreds of years.	*Wah kareeb sau ke varsh lene waalaa hai*	*Ismein sainkdhon varsh lagenge*
What is done cannot be undone.	*Kyaa kiyaa jaataa hai rad kiya nahin jaataa hai*	*Jo kuchh ho chukaa hai uss ke bare mein kuchh nahin kiya jaa saktaa*

Example- SER

- Total test set sentences =5
- Sentences matching exactly with standard translations = 2
- Sentences not matching with standard translations = 3

(The machine translation of sentence 3, 4 and 5 do not match exactly with the reference translations)

Hence the SER = 60%.

Word Error Rate (WER)

- Word Error Rate (WER) is a slightly more sophisticated metric, commonly used in the field of speech recognition.
- Based on the Levenshtein distance
- The standard Levenshtein distance is uses for comparison between two individual strings.
- It is a measure of the least amount of insertions, substitutions and deletions that need to be made to transform one string into the other.

Word Error Rate (WER)

- The standard Levenshtein distance gives a penalty of 1 for each insertion, substitution and deletion of a single character that is required for this type of transformation.
- WER is implemented in a similar manner except it considers a word rather than a character as in Levenshtein distance.
- WER = 100 ((#del+#sub+#ins)/ Total # words (in Ref Translation)

Word Error Rate (WER)

- WER: edit distance to reference translation (insertion, deletion, substitution)

- Captures fluency well
- Captures adequacy less well
- Too rigid in matching
  - does not take synonyms into consideration
  - no credit given even when right string but in wrong place is generated.

- Not ideal for languages with not strict word order (e.g. Hindi)

Example-Word Error Rate (WER)

R: it is a guide to action *which ensures that the military *always *obeys *the - commands *of *the *party

T: it is a guide to action *that ensures that the military *will *forever *heed party commands

4 substitutions + 1 insertion + 3 deletions = 8

No. of words in Reference Text= 18

Hence, the Word Error Rate is

*WER =100*8/18= 44%*

Position-Independent WER (PER)

- PER: similar to WER but uses a position independent Lavenshtein distance (bag-of-word based distance)
- The bag-of-words model is a simplifying assumption used in natural language processing and information retrieval. In this model, a text (such as a sentence or a document) is represented as an unordered collection of words, disregarding grammar and even word order.
- Too flexible in matching
- Captures adequacy at single word (unigram) level
- Does not capture fluency

Position-Independent WER (PER)

Example:

Candidate1 = he saw a man

Candidate2 = a man saw he

Reference= he saw a man

Candidate1 and Candidate2 get **same PER score!!**

Example- PER

R: it is a guide to action *which ensures that the military *always *obeys *the - commands *of *the party

T: it is a guide to action *that ensures that the military *will *forever *heed party commands

No. of words in Reference Text= 18

Edit distance : 3+1 substitutions + 2 deletions = 6

Hence, PER =***100*6/18* = 33.3%**

Ranker-based Edit Distance (RED)

- RED (Akiba et al., 2001) is an automatic ranking method based on edit distances to multiple reference translations.
- Consists of Learning and Evaluation phase with the following steps.
  - Label each machine-translated sentence by the majority rank.
  - Encode each machine-translated sentence into a sixteen dimensional vector.
  - Learn a decision tree from the vectors.
  - Assign a rank to MT output by using the learned decision tree.

Ranker-based Edit-Distance (RED)

Each edit distance is measured by one of sixteen variations of the basic edit distance measure, ED1 with three edit operators-insertion, deletion, replacement.

For ED1 two morphemes are regarded as being matched if and only if the base form of each morpheme is the same and each POS tag is the same.

Morpheme is the smallest linguistic unit that has semantic meaning. E.g - "unbearable" – 3 morphemes

Ranker-based Edit-Distance (RED)

For other edit distances, their definitions are changed due to a combination of the following four changing policy.

- First policy, is whether swap edit operator is additionally used.
- Second policy, is whether semantic codes of content words are referred instead of the base forms of the content words.
- Third, is that whether the editing units are restricted to only content words.
- Fourth, is that whether the editing units are restricted to only keywords*.

*Keywords, are the words that appear in two or more reference translations.

Ranker-based Edit-Distance (RED)

EDIT Distances

	Swap Op.	Content words	Semantic code	Keywords
ED1(Base)	No	No	No	No
ED2	No	No	No	Yes
ED3	No	No	Yes	No
……	……	…….	……	……
……	……	……	……	……
ED14	Yes	Yes	No	Yes
ED15	Yes	Yes	Yes	No
ED16	Yes	Yes	Yes	Yes

RED - Example

- **Source Language - English , Target Language – Hindi**

(S) We shall now begin to work.

(T) *Hum karya karnaa ab prarambh honge*

(H1) *ab hum kaam shuru karenge*

(Now we work start will do)

(H2) *ab hum kaam arambh karenge*

(Now we work start will do)

(H3) *hum ab kaam karnaa shuru karenge*

(We now will start working)

Now – ab
We – hum
Work –
kaam
karya
Begin –
shuru
aarambh
praarambh

RED - Example

Sentence	Surface forms	Base forms	POS	Semantic code	Sentence	Surface forms	Base forms	POS	Semantic code
(T) MT output	hum	main	PRN	Z	(H2)	ab	ab	ADV	
	karya	karya	NN	X		hum	main	PRN	
	karnaa	kar	V			kaam	kaam	NN	
	ab	ab	ADV			arambh	arambh	NN	Y
	prarambh	prarambh	NN	Y		karenge	kar	V	
	honge	ho	V		(H3)	hum	main	PRN	
(H1)	ab	ab	ADV			ab	ab	ADV	
	hum	main	PRN	Z		kaam	kaam	NN	
	kaam	kaam	NN	X		karnaa	kar	V	
	shuru	shuru	NN	Y		shuru	shuru	NN	Y
	karenge	kar	V			karenge	kar	V	

RED - Example

- While calculating **ED1** for T and H3, *karnaa* "do" in T and *karnaa* "do" in H3 are matched as they have the same base form and same POS.

- In case of **ED3** content words having same semantic Codes are matched:
  - karya and kaam (both mean "work" Semantic code X)
  - praarambh, aarambh and shuru (Semantic code Y)

TER(Translation Edit Rate)

TER (Translation Error/Edit Rate)- it measures the amount of editing that a human would have to perform to change the system output so that it exactly matches a ref. translation. (Snover, 2006)

TER = # of edits / average # of reference words

- TER is calculated against best (closest) reference
- Edits include insertions, deletions, substitutions & shifts
- All edits count as 1 edit
- Shift moves a sequence of words within the hypothesis
- Shift of any sequence of words (any distance) is only 1 edit
- Capitalization and punctuation errors are included

TER Example

REF: SAUDI ARABIA denied THIS WEEK information Published in the AMERICAN New York times

MT: THIS WEEK THE SAUDIS denied information published in the ------ New York times

No. of Edits = 4 (1 shift, 2 substitutions, 1 insertion)

- TER score= 4/12.5= 31%
- WER score = ?

BLEU- BiLingual Evaluation Understudy

- BLEU- proposed by IBM's SMT group (Papineni et al, 2002)
- Widely used in MT evaluations
- It combines WER and PER- Trade off between rigid matching of WER and flexible matching of PER.
- BLEU compares the 1,2,3,4-gram overlap with one or more reference translations
- BLEU penalizes generating long strings
- References are usually 1 or 4 translations (done by humans!)
- BLEU correlates well with average of fluency and adequacy at a corpus level but not at a sentence level!

BLEU- BiLingual Evaluation Understudy

- BLEU Metric:

$$BLEU = BP \bullet \exp(\sum_{n=1}^{N} w_n \log p_n)$$

  - p_n: Modified n-gram precision
  - Geometric mean of $p_1, p_2,..p_n$
  - BP: Brevity penalty

c: length of the MT hypothesis
r: effective reference length

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{(1-r/c)} & if \quad c \le r \end{cases}$$

  - Usually, N=4 and w_n=$1/N$.

uni-gram precision

- To calculate, count the number of single word matches.
- If a word of the candidate text appears in the reference text, it is a match.
- The score is

$$0 \leq \frac{\text{number of matches}}{\text{number of words in candidate}} \leq 1$$

- The bigger the score, the better translation

Uni/Multi-gram precision

A translation using same words(1-gram) as in references (professional translation) tends to satisfy adequacy.

- However, different human translators can make different word choice.

- BLEU solves this problem by using a set of different style translations.

- Uni-gram ignores word order.

  - It is dealt by longer-gram precision. (a little)

- **Multi-gram precision:**
- A translation using same n-gram as in references tends to satisfy fluency.

Bad example

Candidate: <u>the</u> <u>the</u> <u>the</u> <u>the</u> <u>the</u> <u>the</u> <u>the</u>

Reference 1: <u>The</u> cat is on the mat.

Reference 2: There is a cat on the mat.

1-gram precision = 7/7 = **1**!!!!!!

Q: How can we fix it?

Modified 1-gram precision

Objective: To ignore excessively used word.

If a word ‘w’ from the candidate sentence is used not more than ‘k’ times in any reference,

- If w is used n times, n-k are redundant.
- We can say we do not need to use word ‘w’ more than ‘k’ times to express the source text.

Modified 1-gram precision- Example

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.

'the' occurs no more than 2 times. So only accept first two 'the' in candidate.

Modified 1-gram precision = 2/7

Does it solve problems?

Modified 1-gram precision- Bad Example

Candidate : I always invariably perpetually do.

Reference:

I always do.

I invariably do.

I perpetually do.

Here modified 1-gram Precision is 1.

Short/ Long sentence problem

Candidate: of the

Reference: It is the guiding principle which guarantees the military forces always being under the command of the Party.

* A bad translation but modified n-gram precision is 1.

- n-gram precision penalizes translations longer than the reference but not translations shorter than the reference.

The short sentence problem is handled by using Brevity Penalty

Example-(4-gram precision)

Candidate 1: It is a guide to action which ensures that the military always obeys the command of the party.

Reference 1: It is a guide to action that ensures that the military will forever heed Party commands.

Reference 2: It is the guiding principle which guarantees the military forces always being under the command of the Party.

Reference 3: It is the practical guide for the army always to heed the directions of the party.

Example-(4-gram precision)

The 4-gram precision is 6/15.

Comment:

- A big negative with BLEU
- It picks matches from the different reference translations.
- Hence the precision will be quite high.
- Though as a whole it is a bad translation

BLEU-Example

Candidate : ***the gunman was shot dead by police .***

Ref 1: The gunman was shot dead by the police .

Ref 2: The gunman was shot to death by the police .

Ref 3: The gunman was shot to death by the police .

Ref 4: The Police has killed the gunman .

- Precision: p_1=1.0(8/8) p_2=0.86(6/7) p_3=0.67(4/6) p_4=0.6 (3/5)
- Brevity Penalty: $c = 8$, $r = 9$, $BP = 0.8825$
- Final Score:

$$\sqrt[4]{1 \times 0.86 \times 0.67 \times 0.6} \times 0.8825 = 0.68$$

Is BLEU Okay?

Sample BLEU performance

Reference: George Bush will often take a holiday in Crawford Texas

1. George Bush will often take a holiday in Crawford Texas (1.000)
2. Bush will often holiday in Texas (0.4611)
3. Bush will often holiday in Crawford Texas (0.6363)
4. George Bush will often holiday in Crawford Texas (0.7490)
5. George Bush will not often vacation in Texas (0.4491)
6. George Bush will not often take a holiday in Crawford Texas (0.9129)

Do you notice something very interesting??

Sample BLEU performance

Reference: George Bush will often take a holiday in Crawford Texas

1. George Bush will often take a holiday in Crawford Texas (1.000)
2. Bush will often holiday in Texas (0.4611)
3. Bush will often holiday in Crawford Texas (0.6363)
4. George Bush will often holiday in Crawford Texas (0.7490)
5. George Bush will not often vacation in Texas (0.4491)
6. George Bush will **not** often take a holiday in Crawford Texas (0.9129)

1 & 6 have very high score but opposite semantics

Problems with *using* BLEU

- For longer n-grams (n>=4) score is mostly 0.

- Semantics not taken into consideration. two sentences though semantically opposite could at times be given very high score.

- Recall measure cannot be directly used due to multiple reference translations. Though, Recall score predicts translation quality better than BLEU [Banerjee,2005].

Problems with *using* BLEU

- The BLEU score reliability depends on the number and quality of reference translations. So more the reference translations, higher will be the reliability of the score. Its difficult to arrange large number of reference translations.

- For free order languages it cannot capture reordering . E.g. Hindi . Being free-order could have two sentences ordered differently but equally & grammatically correct. BLEU scores for both the sentences could be very different.

- Yet it is the most used metric, though needs a lot of improvements

NIST

Weight more heavily those n-grams that are more informative

$$Info(w_1....w_n) = \log_2\left(\frac{\#\ of\ \ occurrences\ of\ w_1...w_{n-1}}{\#\ of\ \ occurrences\ of\ w_1...w_n}\right)$$

Use a geometric mean of the n-gram score

$$NIST = BP \bullet \sum_{n=1}^{N}\left\{\frac{\sum_{all\_w_1...w_n\_that\_co-occur} Info(w_1...w_n)}{\sum_{all\_w_1...w_n\_in\_hyp}(1)}\right\}$$

NIST

- Pros: more sensitive than BLEU
- Cons:
  - Info gain for 2-gram and up is not meaningful
    - 80% of the score comes from unigram matches
    - Most matched 5-grams have info gain 0 !
  - Score increases when the testing set size increases

ROUGE

- ROUGE- Recall-Oriented Understudy for Gisting Evaluation (Lin, C.Y.,2004)
- Developed by Chin-Yew Lin at ISI, USC
- Measures quality of a summary by comparison with ideal summaries and generally used evaluation of summaries but can also be used for MT evaluation.

Variations of ROUGE

- ROUGE-N: N-gram co-occurrence statistics
- ROUGE-L: Based on longest common subsequence
- ROUGE-W: weighted longest common subsequence, favours consecutive matches
- ROUGE-S: Skip-Bigram recall metric. Arbitrary in-sequence Bigrams are computed
- ROUGE-SU adds unigrams to ROUGE-S

ROUGE -N

- ROUGE-N: N-gram co-occurrence statistics is a recall oriented metric

$$\text{ROUGE - n} = \frac{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}_{\text{match}}(\text{n - gram})}{\sum_{S \in \{\text{Refs}\}} \sum_{\text{n-gram} \in S} \text{count}(\text{n - gram})}$$

ROUGE-N-Example

- N-gram co-occurrences between reference and candidate translations.
- Similar to BLEU in MT

- Example:

  Ref: police killed the gunman

  MT1: police kill the gunman

  MT2: the gunman kill police

- ROUGE-N: MT1=MT2 (“police”, “the gunman”)

ROUGE-L

Longest Common Subsequence (LCS)

- Given two sequences X and Y, LCS of X and Y is a common subsequence with maximum length.
- The longer the LCS of two translations is, the more similar the two translations are.
- Use LCS-based recall score (ROUGE-L) to estimate the similarity between two translations.
- It doesn't require consecutive matches but checks in-sequence matches.
- It automatically includes longest in-sequence common n-grams, therefore no predefined n-gram length is necessary.

ROUGE-L

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

R_{lcs} = Recall

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

P_{lcs} = Precision

$$F_{lcs} = \frac{(1+\beta^2)R_{lcs}P_{lcs}}{R_{lcs}+\beta^2 P_{lcs}}$$

F_{lcs} = ROUGE-L

X is the Reference translation of length m

Y is the candidate translation of length n

LCS(X,Y) is the Longest Common Subsequence of X and Y

$\beta > 0$

Often β is taken as Precision / Recall

ROUGE-L-Example

Example:

Ref : police killed the gunman

MT1: police kill the gunman

MT2: the gunman kill police

- ROUGE-N: MT1=MT2 (“police”, “the gunman”)
- ROUGE-L:

MT1=3/4 (“police the gunman”)← LCS for MT1

MT2=2/4 (“the gunman”) ← LCS for MT2

MT1>MT2

ROUGE-L

Problem with LCS is that it does not differentiate LCSs of different spatial relations within their embedded sequences.

Example:

Ref: [The boy who came here is my student]

MT1: [The boy who came here studies with me]

MT2: [The is my boy who came study here]

ROUGE-L for (MT1) = ROUGE-L (MT2) although MT1 should be scored higher as compared to MT2

ROUGE-W

Stands for Weighted Longest Common Subsequence

- ROUGE-W favors strings with consecutive matches.

Example:

Ref: [A B C D E F G]

MT1: [A B C D H I K]

MT2: [A H B K C I D]

ROUGE-W for (MT1) > ROUGE-W (MT2)

It can be computed efficiently using dynamic programming.

ROUGE-S

This metric is based on the Skip Bi-gram co-occurrence statistics:

A Skip-Bigram is: Any pair of words in their sentence order, allowing for arbitrary gaps.

It considers long distance dependency.

It allows gaps in matches as LCS but count all in-sequence pairs; while LCS only counts the longest subsequences.

ROUGE-S

$$R_{skip2} = \frac{SKIP2(X,Y)}{C(m,2)}$$

$$P_{skip2} = \frac{SKIP2(X,Y)}{C(n,2)}$$

$$F_{skip2} = \frac{(1+\beta^2)R_{skip2}P_{skip2}}{R_{skip2} + \beta^2 P_{skip2}}$$

X is the Reference translation of length m

Y is the candidate translation of length n

C is the combination function

SKIP2(X,Y)- number of skip Bi-gram matches between X and Y.

$\beta > 0$

F_{skip2} = ROUGE-S

ROUGE-S example

Example:

Ref: police killed the gunman

MT1: police kill the gunman

MT2: the gunman kill police

MT3: the gunman police killed

- ROUGE-N: MT3>MT1=MT2
- ROUGE-L: MT1>MT2=MT3
- ROUGE-S:
- Skip Bi-grams for Ref are: ("police killed", "police the", "police gunman", "killed the", "killed gunman", "the gunman")

ROUGE-S example

- Skip Bi-grams for MT1 are: ("police kill", "police the", "police gunman", "kill the", "kill gunman", "the gunman")
- Skip Bi-grams for MT2 are: ("the gunman", "the kill", "the police", "gunman kill", "gunman police", "kill police")
- Skip Bi-grams for MT3 are: ("the gunman", "the police", "the killed", ","gunman police", "gunman killed", "police killed")

The skip Bi-grams that match with that of the Ref are considered

- MT1=3/6 ("police the", "police gunman", "the gunman")
- MT2=1/6 ("the gunman")
- MT3=2/6 ("the gunman", "police killed")

ROUGE-S: MT1>MT3>MT2

METEOR

Metric for Evaluation of Translation with Explicit ORdering

METEOR: metric developed at CMU

\- (Lavie & Banerjee,2005)

Improves upon BLEU metric developed by IBM

Main ideas:

- Assess the similarity between a machine-produced translation and (several) human reference translations

METEOR

- Similarity is based on word-to-word matching that matches:
  - Identical words
  - Morphological variants of same word (stemming)
  - Synonyms
- Similarity is based on weighted combination of Precision and Recall
- Address fluency/grammaticality via a direct penalty: how well-ordered is the matching of the MT output with the Ref?

Example:

- Reference: "the Iraqi weapons are to be handed over to the army within two weeks"
- MT output: "in two weeks Iraq's weapons will give army"

METEOR-Example

Matching: Ref: Iraqi weapons army two weeks

MT: two weeks Iraq's weapons army

- P = 5/8 =0.625 R = 5/14 = 0.357
- Fmean = 10*P*R/(9P+R) = 0.3731
- Fragmentation: 3 frags of 5 words = (3-1)/(5-1) = 0.50
- Discounting Factor: DF = 0.5 * (frag**3) = 0.0625
- Final score: Fmean * (1- DF) = 0.3731*0.9375 = 0.3498

Conclusions

- Automatic scores are:
  - Very useful in development cycle of MT systems
  - Useful when comparing different MT systems
  - may prove useless to compare systems of different nature
- Subjective scores are:
  - Very useful to assess general level of performance
  - Useful when comparing systems of different nature
  - Slightly more informative than automatic scores

Future Scope

Subjective evaluation should be more efficient:

- Use trained and expert graders only
- Avoid analyzing long (awful) MT outputs
- Focus on specific parts of the sentence:
  - a portion, clause, or syntactic constituent
- Use large test sets to be able to extract interesting parts only

Future Scope

- MT research needs new automatic scores:
  – Informative: to profile system behavior
  – Discriminative: to tell if and where improvements are
  – Effective: to be computed quickly and often
- We need more deep insight into system behavior:
  – More complex and informative benchmarks
  – Encourage development of open tools for MT output profiling

Further Reading

- Alon Lavie and Abhaya Agarwal. 2007. METEOR: An Automatic Metric for MT Evaluation with High Levels of Correlation with Human Judgments. In Proceedings of the Second ACL Workshop on Statistical Machine Translation, pages 228–231, Prague, Czech Republic, June.

- Niladri Chatterjee, Anish Johnson and Madhav Krishna. 2007. Some improvements over the BLEU metric for measuring the translation quality for Hindi. In proc. Of the International Conference on Computing: Theory and Applications (ICCTA'07)

- Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul, "A Study of Translation Edit Rate with Targeted Human Annotation," Proceedings of Association for Machine Translation in the Americas (AMTA-2006),pages 223-231,Cambridge,MA, August.

- S. Banerjee and Alon Lavie. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments. In Proc. of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization.

- Lin, C.Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*, Barcelona, Spain.

Further Reading

- Lin, C.-Y. and E.H. Hovy 2003. Automatic Evaluation of Summaries Using N-gram Co-occurrence Statistics. In *Proceedings of 2003 Language Technology Conference* (HLT-NAACL 2003), Edmonton, Canada.
- Lin, C.-Y. and F. J. Och. 2004. ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation. In *Proceedings of 20th International Conference on Computational Linguistic* (COLING 2004), Geneva, Switzerland.
- George Doddington. 2002. Automatic Evaluation of Machine Translation Quality using N-gram Co-Occurrence Statistics. In Proceedings of 2nd Human Language Technologies Conference (HLT-02). San Diego, CA. pp. 128-132.
- Kishore Papineni, Salim Roukos, Todd Ward and WeiJing Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In Proc. 40th Annual Meeting of the ACL, July 2002, pp. 311-318.
- Akiba, Y., K. Imamura, and E. Sumita. 2001. Using Multiple Edit Distances to Automatically Rank Machine Translation Output. In *Proceedings of the MT Summit VIII*, Santiago de Compostela, Spain.pp. 15-20

Thank You