

Automatic Metrics for MT Evaluation

11-731:
Machine Translation
Alon Lavie
January 31, 2007

Need for MT Evaluation

- MT Evaluation is important:
 - MT systems are becoming wide-spread, embedded in more complex systems
 - How well do they work in practice?
 - Are they reliable enough?
 - MT is a technology still in research stages
 - How can we tell if we are making progress?
 - Metrics that can drive experimental development
- MT Evaluation is difficult:
 - Human evaluation is subjective
 - How good is “good enough”? Depends on application
 - Is system A better than system B? Depends on specific criteria...
- MT Evaluation is a research topic in itself! How do we assess whether an evaluation method is good?

January 31, 2007

11-731: Machine Translation

2

Dimensions of MT Evaluation

- Human evaluation vs. automatic metrics
- Quality assessment at sentence (segment) level vs. task-based evaluation
- “Black-box” vs. “Glass-box” evaluation
- Adequacy (is the meaning translated correctly?) vs. Fluency (is the output grammatical and fluent?)

January 31, 2007

11-731: Machine Translation

3

Automatic Metrics for MT Evaluation

- Idea: compare output of an MT system to a “reference” good (usually human) translation: how close is the MT output to the reference translation?
- Advantages:
 - Fast and cheap, minimal human labor, no need for bilingual speakers
 - Can be used on an on-going basis during system development to test changes
 - Minimum Error-rate Training (MERT) for search-based MT approaches!
- Disadvantages:
 - Current metrics are very crude, do not distinguish well between subtle differences in systems
 - Individual sentence scores are not very reliable, aggregate scores on a large test set are required
- Automatic metrics for MT evaluation very active area of current research

January 31, 2007

11-731: Machine Translation

4

Similarity-based MT Evaluation Metrics

- Assess the “quality” of an MT system by comparing its output with human produced “reference” translations
- Premise: the more similar (in meaning) the translation is to the reference, the better
- Goal: an algorithm that is capable of accurately approximating this similarity
- Wide Range of metrics, mostly focusing on exact word-level correspondences:
 - Edit-distance metrics: Levenshtein, WER, PIWER, TER & HTER, others...
 - Ngram-based metrics: Precision, Recall, F1-measure, BLUE, NIST, GTM...
- Important Issue: exact word matching is very crude estimate for sentence-level similarity in meaning

January 31, 2007

11-731: Machine Translation

5

Automatic Metrics for MT Evaluation

- 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”
- Possible metric components:
 - Precision: correct words / total words in MT output
 - Recall: correct words / total words in reference
 - Combination of P and R (i.e. $F1 = 2PR / (P+R)$)
 - Levenshtein edit distance: number of insertions, deletions, substitutions required to transform MT output to the reference
- Important Issues:
 - Features: matched words, ngrams, subsequences
 - Metric: a scoring framework that uses the features
 - Perfect word matches are weak features: synonyms, inflections: “Iraq’s” vs. “Iraqi”, “give” vs. “handed over”

January 31, 2007

11-731: Machine Translation

6

Desirable Automatic Metric

- **High-levels** of correlation with quantified human notions of translation quality
- **Sensitive** to small differences in MT quality between systems and versions of systems
- **Consistent** – same MT system on similar texts should produce similar scores
- **Reliable** – MT systems that score similarly will perform similarly
- **General** – applicable to a wide range of domains and scenarios
- **Fast and lightweight** – easy to run

January 31, 2007

11-731: Machine Translation

7

The BLEU Metric

- Proposed by IBM [Papineni et al, 2002]
- Main ideas:
 - Exact matches of words
 - Match against a **set** of reference translations for greater variety of expressions
 - Account for **Adequacy** by looking at word **precision**
 - Account for **Fluency** by calculating **n-gram** precisions for $n=1, 2, 3, 4$
 - **No recall** (because difficult with multiple refs)
 - To compensate for recall: introduce "**Brevity Penalty**"
 - Final score is weighted **geometric average** of the n-gram scores
 - Calculate **aggregate score** over a large test set

January 31, 2007

11-731: Machine Translation

8

The BLEU Metric

- 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**"
- **BLUE** metric:
 - 1-gram precision: 4/8
 - 2-gram precision: 1/7
 - 3-gram precision: 0/6
 - 4-gram precision: 0/5
 - **BLEU score** = 0 (weighted geometric average)

January 31, 2007

11-731: Machine Translation

9

The BLEU Metric

- Clipping precision counts:
 - Reference1: "**the** Iraqi weapons are to be handed over to **the** army within two weeks"
 - Reference2: "**the** Iraqi weapons will be surrendered to **the** army in two weeks"
 - MT output: "**the the the the**"
 - Precision count for "**the**" should be "clipped" at **two**: max count of the word in any reference
 - Modified unigram score will be 2/4 (not 4/4)

January 31, 2007

11-731: Machine Translation

10

The BLEU Metric

- Brevity Penalty:
 - **Reference1**: "the **Iraqi weapons** are to be handed over to the army within two weeks"
 - **Reference2**: "the **Iraqi weapons** will be surrendered to the army in two weeks"
 - **MT output**: "the **Iraqi weapons** will"
 - **Precision score**: 1-gram 4/4, 2-gram 3/3, 3-gram 2/2, 4-gram 1/1 → BLEU = 1.0
 - MT output is much too short, thus boosting precision, and BLEU doesn't have recall...
 - An **exponential Brevity Penalty** reduces score, calculated based on the aggregate length (not individual sentences)

January 31, 2007

11-731: Machine Translation

11

Formulae of BLEU

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

Then,

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

$$\log BLEU = \min \left(1 - \frac{r}{c}, 0 \right) + \sum_{n=1}^N w_n \log p_n$$

January 31, 2007

11-731: Machine Translation

12

Weaknesses in BLEU

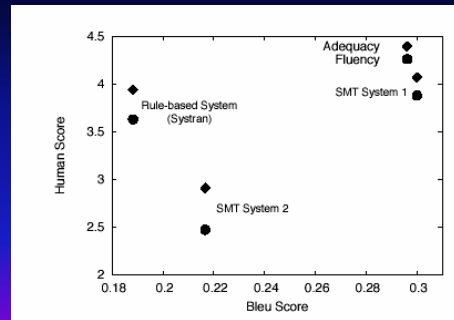
- BLEU matches word ngrams of MT-translation with **multiple** reference translations **simultaneously** → Precision-based metric
 - Is this better than matching with each reference translation separately and selecting the best match?
- BLEU Compensates for Recall by factoring in a “**Brevity Penalty**” (BP)
 - Is the BP adequate in compensating for lack of Recall?
- BLEU’s ngram matching requires **exact** word matches
 - Can stemming and synonyms improve the similarity measure and improve correlation with human scores?
- All matched words **weigh equally** in BLEU
 - Can a scheme for weighing word contributions improve correlation with human scores?
- BLEU’s **higher order ngrams** account for fluency and grammaticality, ngrams are **geometrically averaged**
 - Geometric ngram averaging is volatile to “zero” scores. Can we account for fluency/grammaticality via other means?

January 31, 2007

11-731: Machine Translation

13

BLEU vs Human Scores



January 31, 2007

11-731: Machine Translation

14

The METEOR Metric

- New metric under development at CMU/LTI: METEOR = **Metric for Evaluation of Translation with Explicit Ordering**
- Main new ideas:
 - Reintroduce Recall and combine it with Precision as score components
 - Look only at **unigram** Precision and Recall
 - Align MT output with **each** reference individually and take score of **best pairing**
 - Matching takes into account **word inflection** variations (via stemming)
 - Address fluency via a direct penalty: how **fragmented** is the matching of the MT output with the reference?

January 31, 2007

11-731: Machine Translation

15

METEOR vs BLEU

- **Highlights of Main Differences:**
 - METEOR word matches between translation and references includes semantic equivalents (inflections and synonyms)
 - METEOR combines *Precision and Recall* (weighted towards recall) instead of BLEU’s “brevity penalty”
 - METEOR uses a direct word-ordering penalty to capture fluency instead of relying on higher order n-grams matches
- **Outcome:** METEOR has significantly better correlation with human judgments, especially at the segment-level

January 31, 2007

11-731: Machine Translation

16

METEOR Components

- **Unigram Precision:** fraction of words in the MT that appear in the reference
- **Unigram Recall:** fraction of the words in the reference translation that appear in the MT
- $F1 = P \cdot R / 0.5 \cdot (P + R)$
- $Fmean = P \cdot R / (0.9 \cdot P + 0.1 \cdot R)$
- **Generalized Unigram matches:**
 - Exact word matches, stems, synonyms
- Match with each reference **separately** and select the **best match** for each sentence

January 31, 2007

11-731: Machine Translation

17

The Alignment Matcher

- Find the best word-to-word alignment match between two strings of words
 - Each word in a string can match at most one word in the other string
 - Matches can be based on generalized criteria: word identity, stem identity, synonymy...
 - Find the alignment of highest cardinality with minimal number of crossing branches
- Optimal search is NP-complete
 - Clever search with pruning is very fast and has near optimal results
- Greedy three-stage matching: exact, stem, synonyms

January 31, 2007

11-731: Machine Translation

18

Matcher Example

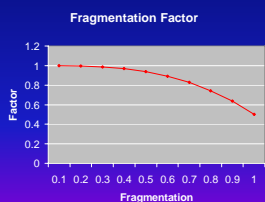
the sri lanka prime minister criticizes the leader of the country
 President of Sri Lanka criticized by the country's Prime Minister

The Full METEOR Metric

- Matcher explicitly aligns matched words between MT and reference
- Matcher returns fragment count (frag) – used to calculate average fragmentation
 - $(\text{frag} - 1) / (\text{length} - 1)$
- METEOR score calculated as a discounted Fmean score
 - Discounting factor: $\text{DF} = 0.5 * (\text{frag}^{**3})$
 - Final score: $\text{Fmean} * (1 - \text{DF})$
- Scores can be calculated at sentence-level
- Aggregate score calculated over entire test set (similar to BLEU)

METEOR Metric

- Effect of Discounting Factor:



METEOR Example

- 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"
- Matching: Ref: Iraqi weapons army two weeks
 MT: two weeks Iraq's weapons army
- $P = 5/8 = 0.625$ $R = 5/14 = 0.357$
- $\text{Fmean} = 10 * P * R / (9P + R) = 0.3731$
- Fragmentation: 3 frags of 5 words = $(3-1)/(5-1) = 0.50$
- Discounting factor: $\text{DF} = 0.5 * (\text{frag}^{**3}) = 0.0625$
- Final score:
 $\text{Fmean} * (1 - \text{DF}) = 0.3731 * 0.9375 = 0.3498$

BLEU vs METEOR

- How do we know if a metric is better?
 - Better correlation with human judgments of MT output
 - Reduced score variability on MT outputs that are ranked equivalent by humans
 - Higher and less variable scores on scoring human translations against the reference translations

Correlation with Human Judgments

- Human judgment scores for adequacy and fluency, each [1-5] (or sum them together)
- Pearson or spearman (rank) correlations
- Correlation of metric scores with human scores at the system level
 - Can rank systems
 - Even coarse metrics can have high correlations
- Correlation of metric scores with human scores at the sentence level
 - Evaluates score correlations at a fine-grained level
 - Very large number of data points, multiple systems
 - Pearson correlation
 - Look at metric score variability for MT sentences scored as equally good by humans

Evaluation Setup

- Data: LDC Released Common data-set (DARPA/TIDES 2003 Chinese-to-English and Arabic-to-English MT evaluation data)
- Chinese data:
 - 920 sentences, 4 reference translations
 - 7 systems
- Arabic data:
 - 664 sentences, 4 reference translations
 - 6 systems
- Metrics Compared: BLEU, P, R, F1, Fmean, METEOR (with several features)

January 31, 2007

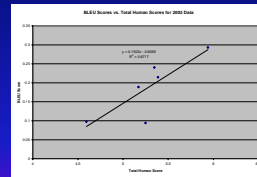
11-731: Machine Translation

25

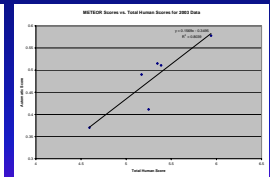
METEOR vs. BLEU: 2003 Data, System Scores

R=0.8196

R=0.8966



BLEU



METEOR

January 31, 2007

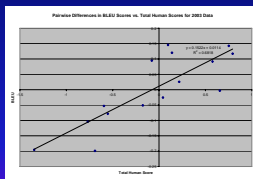
11-731: Machine Translation

26

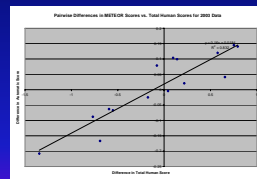
METEOR vs. BLEU: 2003 Data, Pairwise System Scores

R=0.8257

R=0.9121



BLEU



METEOR

January 31, 2007

11-731: Machine Translation

27

Evaluation Results: System-level Correlations

	Chinese data	Arabic data	Average
BLEU	0.828	0.930	0.879
Mod-BLEU	0.821	0.926	0.874
Precision	0.788	0.906	0.847
Recall	0.878	0.954	0.916
F1	0.881	0.971	0.926
Fmean	0.881	0.964	0.922
METEOR	0.896	0.971	0.934

January 31, 2007

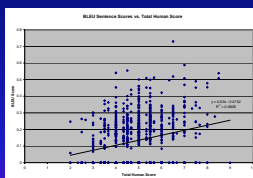
11-731: Machine Translation

28

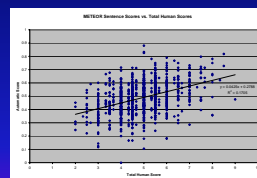
METEOR vs. BLEU Sentence-level Scores (CMU SMT System, TIDES 2003 Data)

R=0.2466

R=0.4129



BLEU



METEOR

January 31, 2007

11-731: Machine Translation

29

Evaluation Results: Sentence-level Correlations

	Chinese data	Arabic data	Average
BLEU	0.194	0.228	0.211
Mod-BLEU	0.285	0.307	0.296
Precision	0.286	0.288	0.287
Recall	0.320	0.335	0.328
Fmean	0.327	0.340	0.334
METEOR	0.331	0.347	0.339

January 31, 2007

11-731: Machine Translation

30

Adequacy, Fluency and Combined: Sentence-level Correlations Arabic Data

	Adequacy	Fluency	Combined
BLEU	0.239	0.171	0.228
Mod-BLEU	0.315	0.238	0.307
Precision	0.306	0.210	0.288
Recall	0.362	0.236	0.335
Fmean	0.367	0.240	0.340
METEOR	0.370	0.352	0.347

January 31, 2007

11-731: Machine Translation

31

METEOR Mapping Modules: Sentence-level Correlations

	Chinese data	Arabic data	Average
Exact	0.293	0.312	0.303
Exact+Pstem	0.318	0.329	0.324
Exact+WNste	0.312	0.330	0.321
Exact+Pstem +WNsyn	0.331	0.347	0.339

January 31, 2007

11-731: Machine Translation

32

Normalizing Human Scores

- Human scores are noisy:
 - Medium-levels of intercoder agreement, Judge biases
- MITRE group performed score normalization
 - Normalize judge median score and distributions
- Significant effect on sentence-level correlation between metrics and human scores

	Chinese data	Arabic data	Average
Raw Human Scores	0.331	0.347	0.339
Normalized Human Scores	0.365	0.403	0.384

January 31, 2007

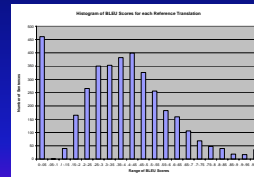
11-731: Machine Translation

33

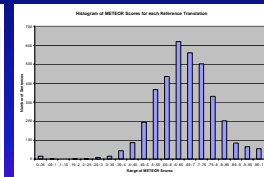
METEOR vs. BLEU Histogram of Scores of Reference Translations 2003 Data

Mean=0.3727 STD=0.2138

Mean=0.6504 STD=0.1310



BLEU



METEOR

January 31, 2007

11-731: Machine Translation

34

Using METEOR

- METEOR software package freely available and downloadable on web: <http://www.cs.cmu.edu/~alavie/METEOR/>
- Required files and formats identical to BLEU → if you know how to run BLEU, you know how to run METEOR!!
- We welcome comments and bug reports...

January 31, 2007

11-731: Machine Translation

35

Conclusions

- Recall more important than Precision
- Importance of focusing on *sentence-level* correlations
- Sentence-level correlations are still rather low (and noisy), but significant steps in the right direction
 - Generalizing matchings with stemming and synonyms gives a consistent improvement in correlations with human judgments
- Human judgment normalization is important and has significant effect

January 31, 2007

11-731: Machine Translation

36

The BLANC Metric Family

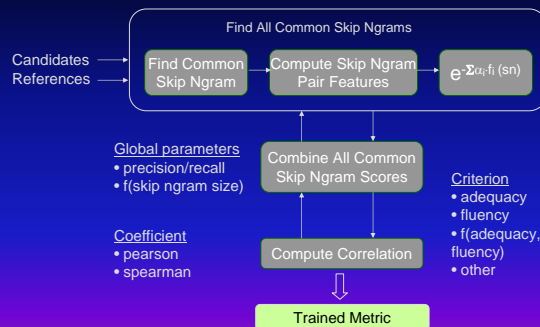
- Generalization of established evaluation metrics
 - N-gram features used by BLEU and ROUGE
- Trainable parameters
 - Skip n-gram contiguity in C
 - Relative importance of n (i.e. bigrams vs. trigrams)
 - Precision-recall balance
- Adaptability to different:
 - Translation quality criteria, languages, domains
- Allow additional processing/features (e.g. METEOR matching)

January 31, 2007

11-731: Machine Translation

37

BLANC Overview



January 31, 2007

11-731: Machine Translation

38

Advantages of BLANC

- Consistently good performance
- Candidate evaluation is fast
- Adaptable
 - fluency and adequacy
 - languages, domains
- Help train MT systems for specific tasks
 - e.g. information extraction, information retrieval
- Model complexity
- Can be optimized for specific MT system performance levels

January 31, 2007

11-731: Machine Translation

39

Summary

- MT Evaluation is important for driving system development and the technology as a whole
- Different aspects need to be evaluated – not just translation quality of individual sentences
- Human evaluations are costly, but are most meaningful
- New automatic metrics are becoming popular, but are still rather crude, can drive system progress and rank systems
- New metrics that achieve better correlation with human judgments are being developed

January 31, 2007

11-731: Machine Translation

40

References

- 2002, Papineni, K, S. Roukos, T. Ward and W.-J. Zhu, BLEU: a Method for Automatic Evaluation of Machine Translation, in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002), Philadelphia, PA, July 2002.
- 2005, Banerjee, S. and A. Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", In Proceedings of Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization at the 43th Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, Michigan, June 2005. Pages 65-72.
- 2004, Iwano, A., K. Suresh and S. Agarwal, "The Significance of Exact N-gram Automatic Metrics for MT Evaluation", In Proceedings of the 6th Conference of the Association for Machine Translation in the Americas (AMTA-2004), Washington, DC, September 2004.
- 2005, Iida, T., S. H. Ogino and A. Iwano, "Learning Evaluation", In Proceedings of the Joint Conference on Human Language Technologies and Empirical Methods in Natural Language Processing (HLT/EMNLP-2005), Vancouver, Canada, October 2005. Pages 740-747.

January 31, 2007

11-731: Machine Translation

41

Questions?

January 31, 2007

11-731: Machine Translation

42