Multilingual and Parallel Corpora
Evaluation

Amir Zeldes

amir.zeldes@georgetown.edu

Course plan – week of April 10

- Monday, 4/10 – guest lecturer: Laura Vilardell on translation and interpretation (Spanish and Catalan)
- Wednesday 4/12 – remote session (will be announced)

MT evaluation metrics

- The most widespread metrics:
  - BLEU (Papineni et al. 2002)
  - ROUGE (Lin 2004)
  - METEOR (Banerjee & Lavie 2005)

- Attempt to solve precisely this exercise!

BLEU

- **BLEU (bilingual evaluation understudy)**
  - Bounded between 0 and 1
  - Supports **multiple** reference translations
  - Based on **precision:** how many of the TT words are actually in the reference translation?
- But raw precision is a bad metric:

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

- Perfect!

BLEU

- Fixing the problem:
  - We cap each word's allowed occurrences at max(m) appearances in any reference translation (the -> 2)
  - Take proportion of summed *m*'s out of candidate length:
    - the the the the the the the -> 2/7
- Problems:
  - Bias favors short translations:
    - the the -> 1/2
    - the cat -> 2/2!
  - No word sequence/ordering -> correctable by adding metrics for n-grams (in practice, up to n=4)

BLEU

- Additional refinements:
  - Also factor in **recall** (mat, is, on -> not recalled)
  - Can be problematic in the face of **many** reference translations
  - System should produce a sentence containing all words??
    - The cat is on the mat
    - There is a kitty on the mat
    - **Good candidate:** there the cat kitty is on the mat
  - Add a **brevity penalty** based on proportion of reference corpus size to output corpus size

How did BLEU do for you?

- Zum Thema : Ich denke , die Bürger Europas müssen sich darauf verlassen können , dass das , was auf Europas Straßen , Schienen usw. transportiert wird , wenn es denn auch noch gefährliche Güter sind , so sicher wie möglich ist .
- Entro en el tema : creo que los ciudadanos de Europa pueden confiar en que lo que se transporta en Europa por carretera , por ferrocarril o por las vías navegables se transporta con toda la seguridad posible , aun siendo mercancías peligrosas .
- Pour ce qui est du thème proprement dit , je pense que les citoyens d' Europe doivent pouvoir compter sur le fait que les marchandises transportées sur les routes , les voies ferrées et autres voies de transport européennes , qu' il s' agisse ou non de marchandises dangereuses , sont transportées d' une manière aussi sûre que possible .
- Entrando nel merito , ritengo che i cittadini europei debbano poter essere certi che i trasporti in Europa su strada , per ferrovia o con altri mezzi , avvengano in condizioni di massima sicurezza , anche nel caso in cui si tratti di merci pericolose .
- Об этой теме, я думаю, что люди Европы должны быть уверены, что товары перевозимые по европейским дорогам, железным дорогам и т. Д., как бы они опасны, настолько безопасны, насколько это возможно.
- 就这件事来说，我想欧洲人应该相信，在欧洲的公路和铁路上运输的货物，不论是多危险的货物，都有最大的安全保障。
- 현 주제와 관련해서, 아무리 위험한 제품일지라도 유럽의 도로와 철도 등을 통해 가능한 안전하게 수송되는 것에 대하여 유럽인들은 자신을 가져야 한다고 생각합니다.

Alternative 1 - ROUGE

- **Recall-Oriented Understudy for Gisting Evaluation**
- Originally developed to evaluate **automatic summarization**
- Implements five metric variants:
  - ROUGE N (n-gram based; normally bi-gram recall)
  - ROUGE L (longest common subsequence)
  - ROUGE W (weighted LCS)
  - ROUGE S (skip bi-gram)
  - ROUGE SU (skip bi-gram + unigrams)

ROUGE N

- The cat is on the mat
- I see the mat
  - Recalled: “the mat”
  - Not recalled:
    - The cat
    - cat is
    - is on
    - on the
- Recall = 20%

ROUGE L

- Uses f-score of precision and recall based on LCS:
  - The cat is on the mat
  - The dog is on the yellow carpet
- Longest common subsequence - strict:
  - is on the
  - Recall: 3/6
  - Precision: 3/7
  - f measure: 2*P*R/(P+R) = 0.46

ROUGE L and W

- Uses f-score of precision and recall based on LCS:
  - The cat is on the mat
  - The dog is on the yellow carpet
- Longest common subsequence - strict:
  - is on the
- LCS – non consecutive:
  - The … is on the …

ROUGE L and W

- LCS – non consecutive:
  - The … is on the …
- In ROUGE W, we give higher weight to consecutive parts or penalize non-consecutive parts
  - The weight can be exponential with chain length
  - Often weight follows $f(k) = k^2$

Skip bi-grams

- Follows combination of ROUGE N (bi-gram proportion) with non-strict LCS strategy
- Non consecutive but ascending bi-grams count
  - S1. police killed the gunman
  - S2. police kill the gunman
  - S3. the gunman kill police
  - S4. the gunman police killed
- What proportion of skip bigrams do S2-S4 get?
  - There are 6 possible skip bi-grams…
- What kinds of languages does this fit?

ROUGE S vs. SU

- ROUGE S does badly in this situation:
  - S5. gunman the killed police -> 0
  - S6. banana spam spam spam -> 0

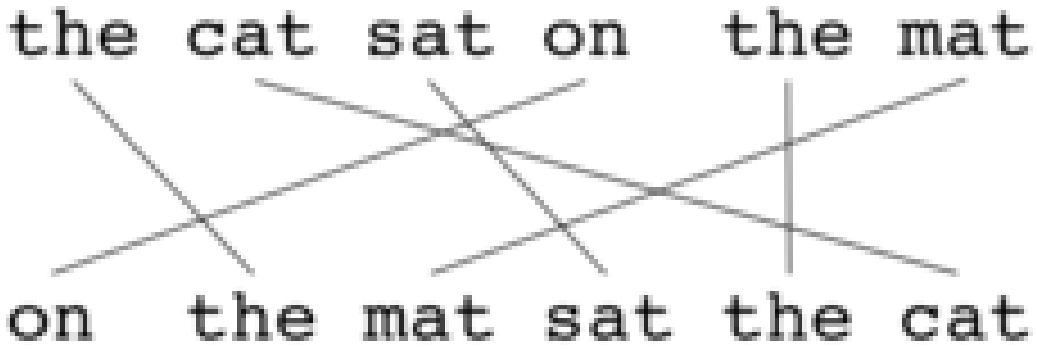
- ROUGE SU adds some score for unigrams

Alternative 2 - METEOR

- **Metric for Evaluation of Translation with Explicit Ordering**

METEOR

- **Metric for Evaluation of Translation with Explicit Ordering**
- Based on **unigram alignment** between candidate and reference translation
- Values for Recall are weighted 9 * Precision



METEOR

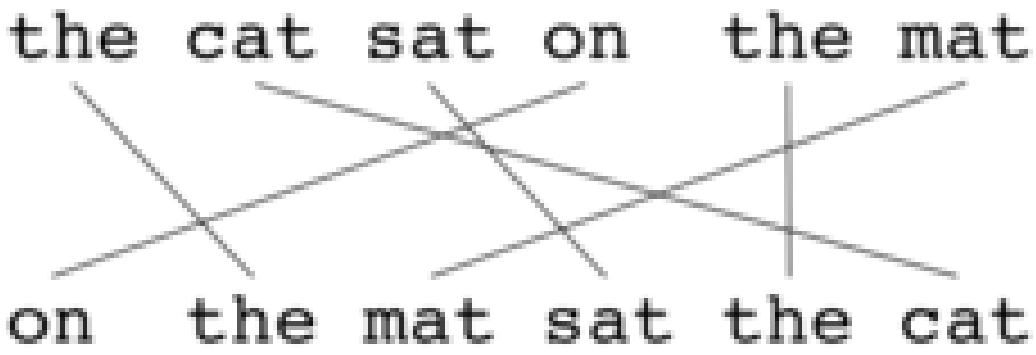
- Calculating unigram f-score:
  - How many words are actually in the translation?
  - How many words from the translation have been found?

    - A farmer grew cereals in his field.
    - The farmer grew the grain in the garden.

  - F = (10*P*R) / (R + 9*P)

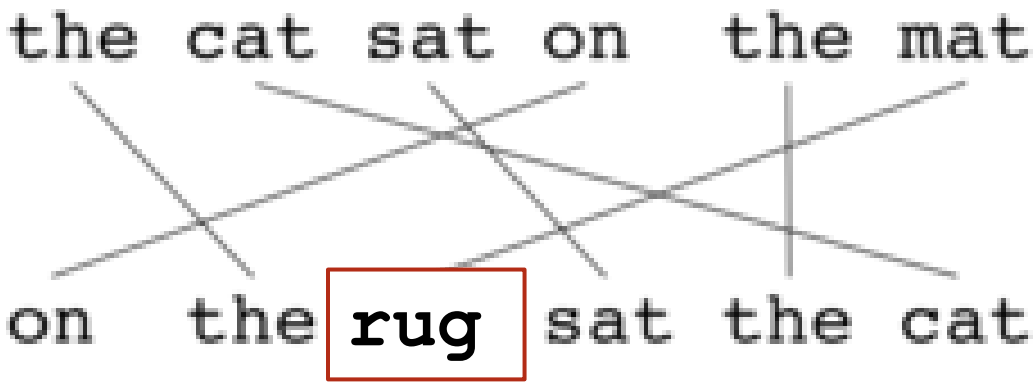
METEOR

- Penalty *p* for number of **consecutive chunks**
- How many consecutive chunks here?



METEOR

- Possible to add (near) synonyms?



METEOR

- Calculating the penalty:
  - *P = 0.5 * (chunks/matched_unigrams)*
  - *METEOR = f *(1-p)*
- This means the f-score can be weighted down by as much as 50%, if there are as many chunks as unigrams (= no bigram matches)

Evaluation (of evaluations)

- From Banerjee & Lavie (2005):

System ID	Correlation
BLEU	0.817
NIST	0.892
Precision	0.752
Recall	0.941
F1	0.948
Fmean	0.952
METEOR	0.964

Table 1: Comparison of human/METEOR correlation with BLEU and NIST/human correlations