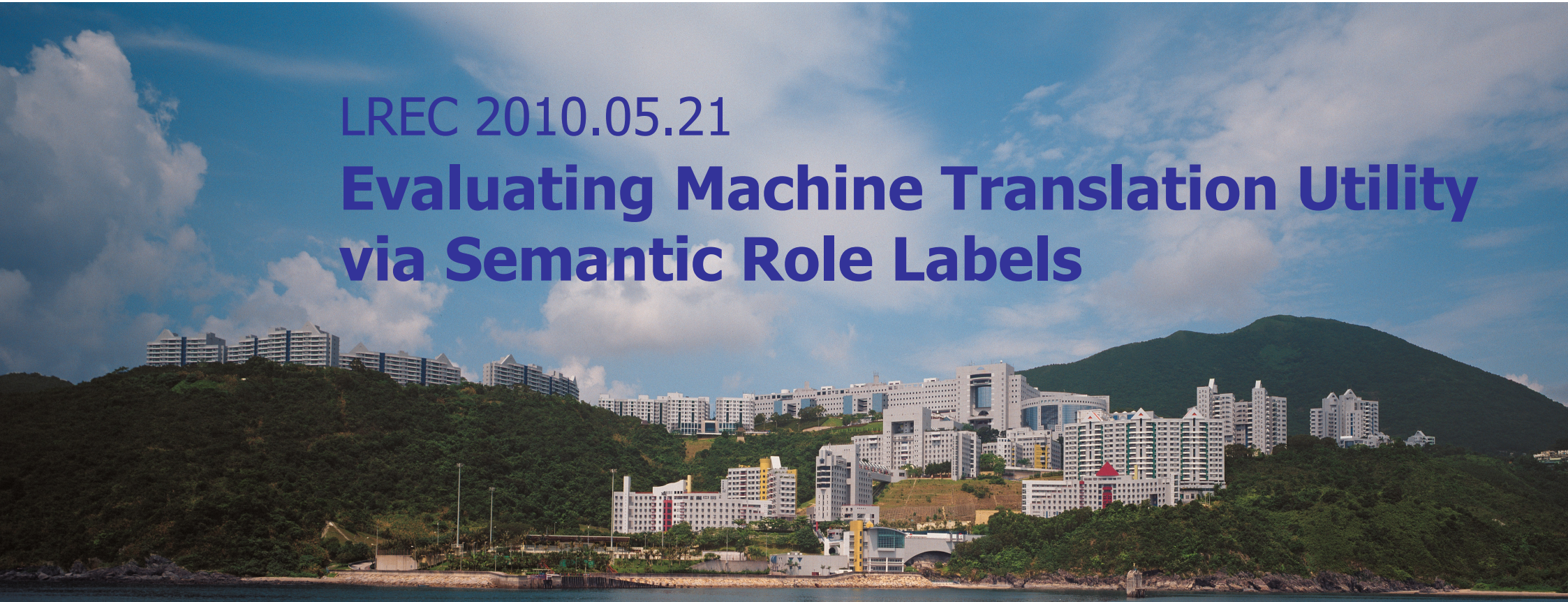


LREC 2010.05.21

Evaluating Machine Translation Utility via Semantic Role Labels

Chi-kiu Lo
jackielo@cs.ust.hk

Dekai Wu
dekai@cs.ust.hk

HKUST
Human Language Technology Center
Department of Computer Science and Engineering
University of Science and Technology, Hong Kong

Q. What makes a translation good?

- Our **utility** perspective:
  - A translation is accurate if it is **useful**

- Can you accurately understand "who did what to whom, when, where and why" after reading the translation?

- Not measured by current MT evaluation metrics
  - … which tend to reward fluency more than adequacy

Recent trends toward **Semantic SMT**

- WSD for SMT
  - Carpuat & Wu (2007, 2008)
  - Giménez & Màrquez (2007)
  - Chan *et al.* (2007)
- SRL for SMT
  - Wu & Fung (2009)

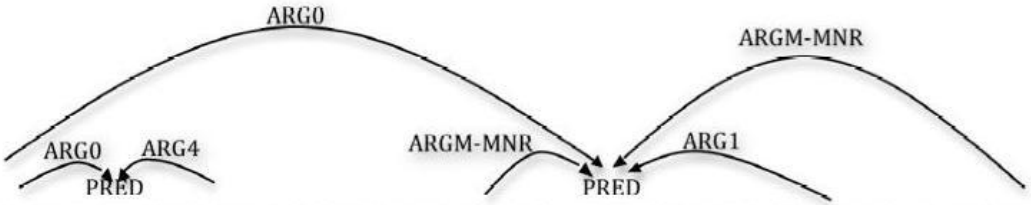
- Translation quality improves more than reflected by current MT evaluation metrics!
- Are BLEU, HTER the wrong objective function to drive this type of work?

Toward **Semantic MT Evaluation**

- Hypothesis:
  - MT **utility** can best be evaluated via semantic role labeling
- We aim to measure:
  - How accurately can readers of MT output reconstruct the semantic frames of the source sentences and/or reference translations?
- Should reflect translation utility better than:
  - automated n-gram precision based MT evaluation metrics, like BLEU
  - non-automated MT evaluation metrics like HTER

Example: a lower-utility translation

Fewer SRL matches, but more N-gram matches!



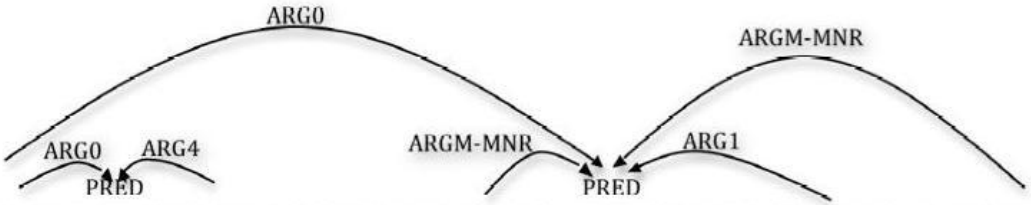
REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.

MT Kerry, who **arrived** in Cairo with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and active **consultations** [on bilateral relations and other issues]]

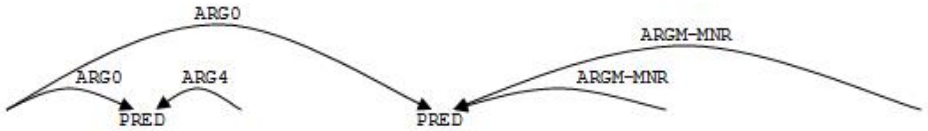
1-gram matches: 13
2-gram matches: 7
3-gram matches: 4
4-gram matches: 3
5-gram matches: 2
6-gram matches: 1

Example: a higher-utility translation

More SRL matches, but fewer N-gram matches!



REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.



MT Kerry, who **arrived** in Cairo active **consultations** with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and on bilateral relations and other issues]

1-gram matches: 13
2-gram matches: 6
3-gram matches: 2
4-gram matches: 0
5-gram matches: 0
6-gram matches: 0

Corpus

- Data drawn from DARPA GALE program Phase 2.5 evaluation
  - Parallel corpus of source sentences and reference translations
  - Annotated with gold standard semantic role labels in Propbank style
  - 3 state-of-the-art MT systems' outputs

Annotation protocol

- Human annotators are given simple, minimal instructions and examples on what they should label
  - "who did what to whom, when, where and why"

Agent (who)	Action (did)	Experiencer (what)
Patient (whom)	Temporal (when)	Location (where)
Purpose (why)	Manner (how)	Degree or Extent (how)
Other adverbial argument (how)		

  - Aim: capture the key semantic roles

"Sanity check" experiments

- Normal condition
  - **Output** = annotators see English translations only
  - Two sub-variants:
    - Annotators are English monolinguals
    - Annotators are bilinguals (controls for the degree to which MT users can "guess" based on knowledge of source language)

- Control conditions
  - **Input** = annotators see foreign source sentences only
  - **Input-output** = annotators see English translations plus foreign source sentences
    - Annotators must be bilinguals
  - Provides baselines for comparison with the normal conditions

More "sanity check" experiments

- Control conditions
  - Annotators see reference translations (not MT)
    - under **Output** condition (without the source sentence)?
    - under **Input-Output** condition (with source sentence)?
  - Provides baselines for comparing how well humans can reconstruct semantic frames from machine translations instead

How annotators are assigned sentences

[Note: a sentence may be either a source sentence, machine translation, or reference translation]

- Each sentence is annotated by at least two human annotators
  - Helps reduce the effect of personal bias
- Each human annotator annotates only one sentence from any source-MT-reference set
  - Avoids contamination in annotators' judgments

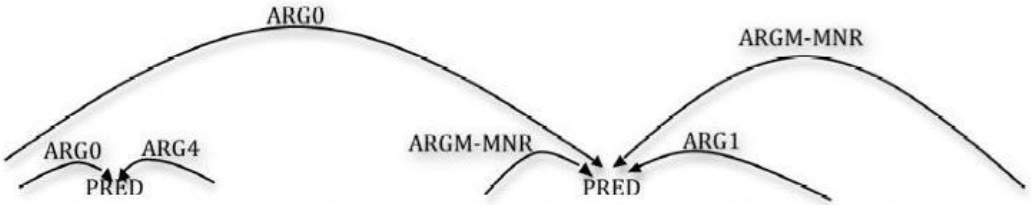
How partially correct reconstructions of a semantic frame can be counted

- For each predicate in the source or reference
  - find the matching predicate in the annotated sentence
- For each argument in a matched predicate
  - **Correct** = expresses the exact same content as that in the source or reference
  - **Incorrect** = expresses content that belongs in other arguments
  - **Partial** = expresses part of the correct content
    - note: extra correct content is not penalized, unless it belongs in other arguments
- Facilitates a finer-grained measurement of utility

- The relative utility of MT against human translation can then be measured via precision/recall as follows…

Example: a lower-utility translation

Fewer SRL matches, but more N-gram matches!



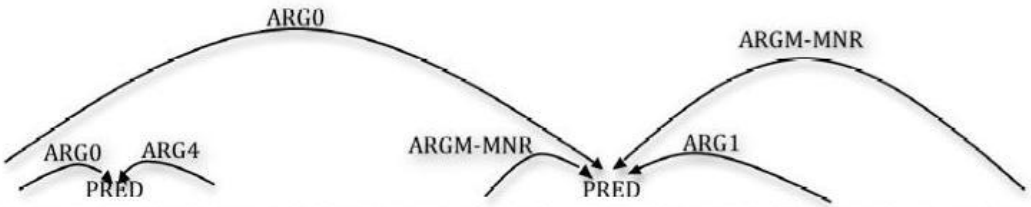
REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.

MT Kerry, who **arrived** in Cairo with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and active **consultations** [on bilateral relations and other issues]]

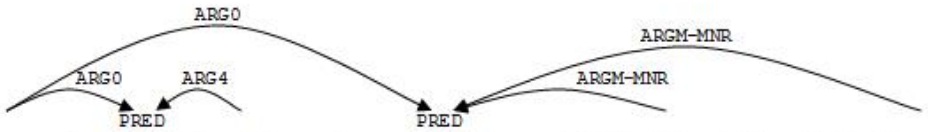
# matched predicates	1 (arrived)
# **Correct** arguments	2 (Kerry, in Cairo)
# **Incorrect** arguments	2 (the two ARGM)
total # predicates in MT	2
total # predicates in reference	2

Example: a higher-utility translation

More SRL matches, but fewer N-gram matches!



REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.



MT Kerry, who **arrived** in Cairo active **consultations** with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and on bilateral relations and other issues]

# matched predicates	1 (arrived)
# **Correct** arguments	2 (Kerry, in Cairo)
# **Incorrect** arguments	0
total # predicates in MT	2
total # predicates in reference	2

What do the measurements mean?

- Counts of **Correct**, **Partial** and all arguments associated with a matched predicate

$$
\begin{aligned}
N_{\mathrm{ci}} &= \text{no. of Correct ARG of PRED i in MT} \\
N_{\mathrm{pi}} &= \text{no. of Partial ARG of PRED i in MT} \\
N_{i} &= \text{total no. of ARG of PRED i in MT}
\end{aligned}
$$

- Sum of **Correct, Partial** predicate-argument structures in a sentence level

$$
N_c = \sum_{\text{all matched predicates}} \frac{N_{\mathrm{ci}}}{N_i}
$$

$$
N_p = \sum_{\text{all matched predicates}} \frac{N_{\mathrm{pi}}}{N_i}
$$

What do the measurements mean?

- Sentence-level precision-recall accuracy of predicate-argument structure

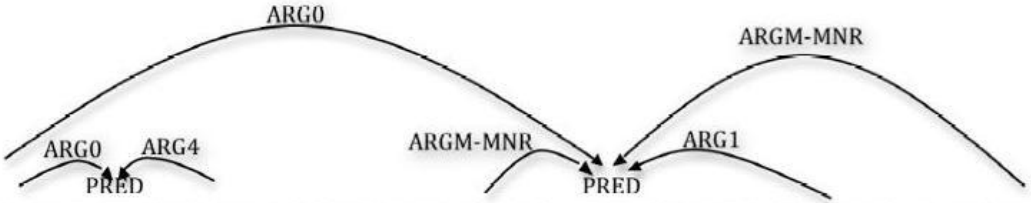
$$P = \frac{N_c + (0.5 * N_p)}{\text{total no. of predicates in reference}}$$

$$R = \frac{N_c + (0.5 * N_p)}{\text{total no. of predicates in MT output}}$$

$$F - \text{measure} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Example: a lower-utility translation

Fewer SRL matches, but more N-gram matches!



REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.

MT Kerry, who **arrived** in Cairo with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and active **consultations** [on bilateral relations and other issues]]

$N_c = 2/4 = 0.5$

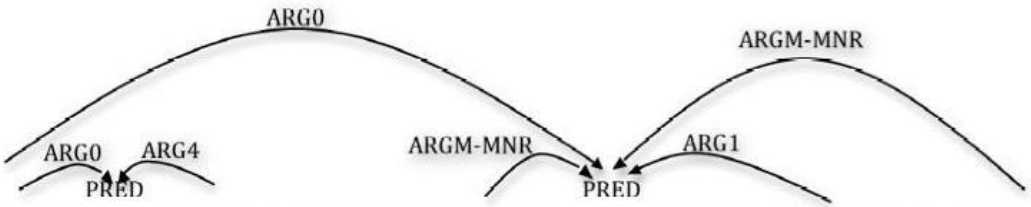
$P = 0.5/2 = 0.25$

$R = 0.5/2 = 0.25$

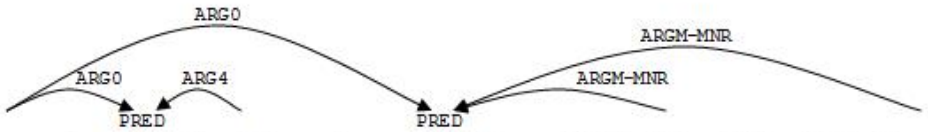
F-measure = 0.25

Example: a higher-utility translation

More SRL matches, but fewer N-gram matches!



REF Kerry, **arrived** in Cairo with [President Mubarak] **engaged** in active discussions on such topics as Iraq, Lebanon, and Darfur, Sudan as well as bilateral relations.



MT Kerry, who **arrived** in Cairo active **consultations** with [President Mubarak] [on Iraq, Lebanon, Sudan's Darfur issue and on bilateral relations and other issues]

$N_c = 2/2 = 1$

$P = 1/2 = 0.5$

$R = 1/2 = 0.5$

F-measure = 0.5

Conclusion

- A new **semantic MT evaluation** methodology
  - Aims at evaluating **translation utility**
  - Measures the accuracy with which users of MT can correctly reconstruct the semantic frames

- In progress
  - Human evaluators currently annotating semantic frames in Chinese-English MT data from GALE P2.5