SSST-6 2012.07.12

Unsupervised vs. Supervised Weight Estimation for Semantic MT Evaluation Metrics

Chi-kiu Lo **Dekai Wu**

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

{ jackielo | dekai }@cs.ust.hk

The problem with conventional MT evaluation metrics

This has been our SMT trajectory over the years

- **1993-1995** First unstructured SMT on very different langs (Chinese)
- **1995-now** First syntactic SMT (ITG, BITG, phrasal ITG)
- **2009-now** Recent syntactic SMT (LTG, LITG, PLITG)
- **2005-now** First semantic SMT with WSD-for-SMT (PSD)
- **2007-now** First semantic SMT with SRL-for-SMT

Subjective evaluation shows improvement…

But conventional metrics like BLEU aren't discriminating enough to register it

Serious danger of driving our field astray!

- **2009-now** Semantic MT evaluation with SRL-for-MTE (MEANT)

Background

*Acknowledgments: DARPA GALE, BOLT

- **LREC 2010, SSST 2010**
  - Blueprint HMEANT model, preliminary results
- **ACL 2011**
  - Assesses adequacy via Propbank-style semantic predicates, roles, and fillers
  - Explains MT accuracy with high representational transparency
  - Correlates with human adequacy judgments (HAJ) as well as HTER, BUT at lower cost
- **IJCAI 2011**
  - "Flattened" HMEANT improves correlation with HAJ, by ignoring which frames roles/fillers are associated with (!!)
  - Correlation of individual roles against HAJ
  - Analysis of time cost of evaluation
- **SSST 2011**
  - Back to compositionality – "unflattens" HMEANT and further improves correlation with HAJ
  - Weights the degree of contribution of each frame, according to size of the span it covers

HMEANT

Human semantic MT evaluation via SRL

ARG0

ARGM-ADV ARGM-LOC PRED ARGM-EXT ARG1 PRED ARG1

[IN] 至此， 在 中国 内地 停售 了 近 两 个 月 的 ＳＫ－ＩＩ 全线 产品 恢复 销售 。

ARG0 PRED ARGM-LOC ARGM-TMP ARG1 ARGM-TMP PRED

[REF] Until after their sales had ceased in mainland China for almost two months , sales of the complete range of SK – II products have now be resumed .

Agent Action Experiencer

[MT1] So far , nearly two months sk - ii the sale of products in the mainland of China to resume sales .

Experiencer

Temporal Agent Action Temporal Experiencer Action

[MT2] So far , in the mainland of China to stop selling nearly two months of SK - 2 products sales resumed .

[MT3] So far , the sale in the mainland of China for nearly two months of SK - II line of products .

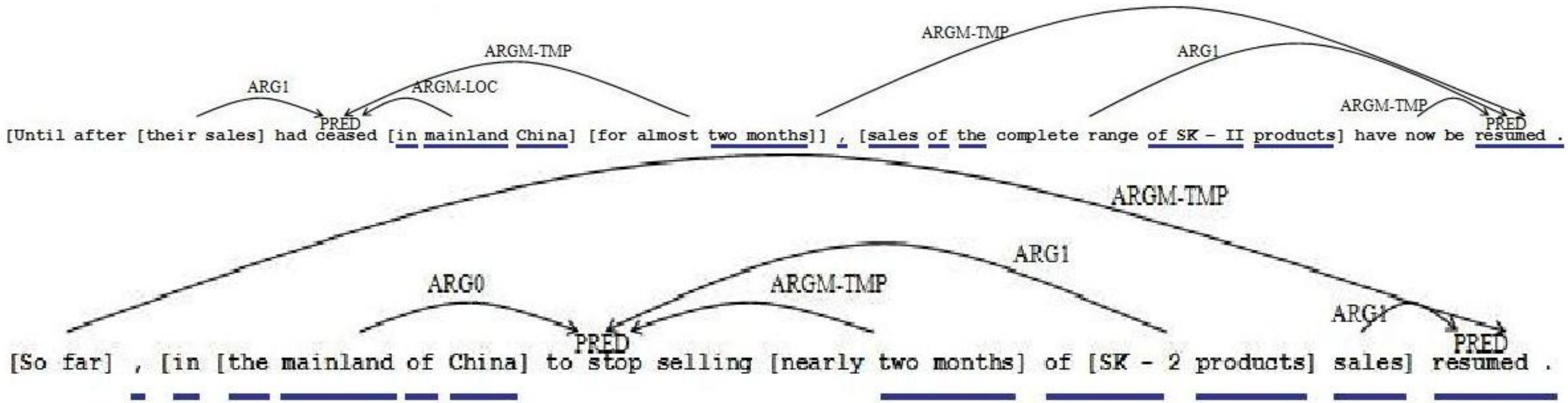
Example: a less useful translation

Fewer SRL matches ☺
but more N-gram and syntax-subtree matches! ☹

N-gram		Syntax-subtree		SRL	
1-gram matches:	15	1-level subtree matches:	34	Predicate matches:	**0**
2-gram matches:	4	2-level subtree matches:	**8**		
3-gram matches:	**3**	3-level subtree matches:	**2**		
4-gram matches:	**1**	4-level subtree matches:	0		

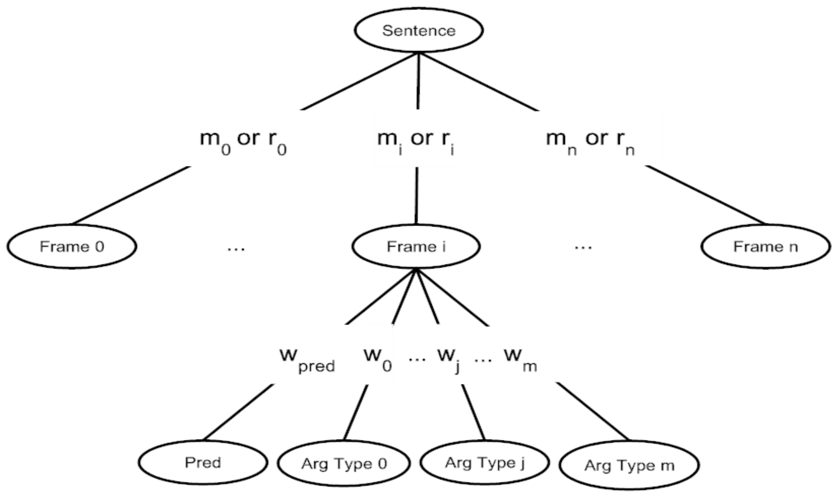
Conversely: a more useful translation

More SRL matches ☺
but fewer N-gram and syntax-subtree matches! ☹



N-gram		Syntax-subtree		SRL	
1-gram matches:	15	1-level subtree matches:	35	Predicate matches:	**2**
2-gram matches:	4	2-level subtree matches:	**6**	Argument matches:	**1**
3-gram matches:	**1**	3-level subtree matches:	**1**		
4-gram matches:	**0**	4-level subtree matches:	0		

HMEANT is just an f-score on semantic frame match (with a tiny number of weights)



$$m_i \equiv \frac{\text{\# word tokens filled in Frame i}}{\text{total\# word tokens in MT}}$$

$$r_i \equiv \frac{\text{\# word tokens filled in Frame i}}{\text{total\# word tokens in REF}}$$

$$\text{precision} \equiv \frac{\sum_i m_i \frac{w_{\text{pred}}+\sum_j w_j (C_{i,j}+w_{\text{partial}} P_{i,j})}{w_{\text{pred}}+\sum_j w_j M_{i,j}}}{\sum_i m_i}$$

$$\text{recall} \equiv \frac{\sum_i r_i \frac{w_{\text{pred}}+\sum_j w_j (C_{i,j}+w_{\text{partial}} P_{i,j})}{w_{\text{pred}}+\sum_j w_j R_{i,j}}}{\sum_i r_i}$$

- **sentence accuracy:** avg translation accuracy over all frames of a <u>sentence</u>
  sentence precision (or recall) = frame precision (or recall) averaged across the total number of frames in MT (or REF)

- **frame accuracy:** avg translation accuracy over all roles of a <u>frame</u>
  frame precision (or recall) = weighted sum of # correctly translated arguments, normalized by the weighted sum of # arguments in MT (or REF)

- **frame importance:** weight each frame by its span coverage ratio

- **role importance:** weight each type of role
  by maximizing HMEANT's correlation with HAJ using a human ranked training corpus

HMEANT is fairly cheap…
… but still requires humans

- Annotation tasks

  1. label semantic predicates, roles, and fillers

  2. align predicates and fillers between the reference and machine translations

- Ranking task
  - label human adequacy judgment
    to form a training corpus for the role importance

Unsupervised weight estimates are needed

- Testing HMEANT on WMT-2012 English-Czech (w/ Bojar *et al.*)
  - Manpower constraint: 14 Czech-speaking annotators
  - Time constraint: within two days
  - Translation of 50 sentences from 13 systems and 1 reference
- What about the labeled training data?
  - No more resources (Czech speakers)
  - Applying the weights learned from English data is obscured
    - linguistic differences between Czech and English, e.g. dropping of pronoun in Czech

Our goal:

- **Further reduce the cost of evaluating MT** by eliminating the dependency on a human adequacy-ranked training corpus for tuning the weights for each semantic role type
- Here, we're mainly targeting the problem of evaluating translation quality for languages with sparse resources

Using relative frequency to estimate MEANT's parameters

- Basic assumption:
  - Roles that are more important for humans to understand should appear more often in the language

- We propose an <u>unsupervised</u> approach:
  - Use the relative frequency of how often a type of semantic role appears in reference translations, to estimate the degree of contribution of that role type

$$c_j \equiv \text{\# count of ARG j in REF of the test set}$$

$$w_j = \frac{c_j}{\sum_j c_j}$$

Correctness of the proposed unsupervised approach

- Problem: No ground truth on which role type contributes more to the overall meaning
- Solution: Evaluate how closely the unsupervised weight of each role type approximates the weight obtained from supervised training

Results

- Relative frequency of each semantic role type closely approximates the supervised weight of that type

Role	Deviation (GALE-A)	Deviation (GALE-B)	Deviation (WMT12)
Agent	-0.09	-0.05	0.03
Experiencer	0.23	0.05	0.02
Benefactive	0.02	0.04	-0.01
Temporal	0.11	0.08	0.03
Locative	-0.05	-0.05	-0.07
Purpose	-0.01	0.03	-0.01
Manner	-0.01	0.00	-0.01
Extent	-0.02	0.00	-0.01
Modal	—	0.04	0.01
Negation	—	0.01	-0.01
Other	-0.12	0.05	-0.01

Table 1: Deviation of relative frequency from optimized weight of each semantic role in GALE-A, GALE-B and WMT12

Estimating the weight for the predicate

- Treating predicate the same way as the arguments
  - Using relative frequency of the predicate
    in addition to all semantic arguments

$$c_{\text{pred}} \equiv \#\text{ count of PRED in REF of the test set}$$

$$\text{Method (i)} = \frac{c_{\text{pred}}}{c_{\text{pred}} + \sum_j c_j}$$

- BUT, predicates are fundamentally different from arguments
  - Every semantic is defined by one predicate,
    and arguments are defined relative to the predicate

- In the supervised weights,
  predicate is usually one-fourth as important as the agent role

$$\text{Method (ii)} = 0.25 \cdot w_{\text{agent}}$$

Results

- The heuristic of one-fourth of the agent's weight closely approximates the weight of the predicate

PRED estimation	Deviation (GALE-A)	Deviation (GALE-B)	Deviation (WMT12)
Method (i)	0.16	0.16	0.31
Method (ii)	0.02	0.01	0.01

Table 2: Deviation from optimized weight in GALE-A, GALE-B and WMT12 of the predicate's weight as estimated by (i) frequency of predicates in frames, relative to predicates and arguments; and (ii) one-fourth of agent's weight.

HMEANT using unsupervised weight estimates

- Unsupervised approach closely approximates the weights obtained from supervised approach

- Then, comparing to other MT evaluation metrics, how does HMEANT using unsupervised weights perform?

Results

- Unsupervised HMEANT correlates with HAJ comparably to supervised HMEANT

Metrics	GALE-A	GALE-B	WMT12
HMEANT (supervised)	0.49	0.27	0.29
HMEANT (unsupervised)	0.42	0.23	0.20
NIST	0.29	0.09	0.12
METEOR	0.20	0.21	0.22
TER	0.20	0.10	0.12
PER	0.20	0.07	0.02
BLEU	0.20	0.12	0.01
CDER	0.12	0.10	0.14
WER	0.10	0.11	0.17

Table 3: Average sentence-level correlation with human adequacy judgments of HMEANT using supervised and unsupervised weight scheme on GALE-A, GALE-B and WMT12, (with baseline comparison of commonly used automatic MT evaluation metric.

Conclusion

- Using relative frequency of semantic roles (unsupervised) to estimate HMEANT's parameters:
  - **further reduces the evaluation cost** by eliminating the dependency on a human adequacy-ranked training corpus for tuning the weights for each semantic role type
  - **correlates with HAJ** comparably to supervised HMEANT on all three data set, including WMT-2012 English-Czech
  - **is well suited to sparse languages** for evaluating translation

Progress toward automating HMEANT…

- Fully automated MEANT (WMT-2012, at NAACL, in June 2012)
  - First fully automated semantic MT evaluation metric
    - Replaces human SRL with automatic shallow semantic parsing
    - Replaces human semantic frame alignment with a simple maximum weighted bipartite matching algorithm based on the lexical similarity between semantic frames
  - Preserves the spirit of Occam's razor of HMEANT
  - Outperforms all commonly used automatic metrics

- Training SMT with MEANT as the objective function
  - Minimum error rate training runs completed two weeks ago
  - Highly competitive results
  - In progress: Human quality evaluation on MT output tuned on MEANT vs. BLEU vs. TER