ESSCaSS'15 day 3 Nelijärve, Estonia 2015.08.20

AI = Learning to Translate
Meaningful Transduction

Dekai WU
dekai@cs.ust.hk http://www.cs.ust.hk/~dekai

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

Q. Why bother doing SMT?

A. All cornerstones of machine learning & language acquisition

---

Why do SMT?
Scientific grand challenge

AI + cognitive science: learning to translate encompasses all cornerstone problems of language acquisition + machine learning

- **grammar induction**
- **unsupervised learning**
- **category formation**
- **chunking**
- **relational abstraction**
- **transduction acquisition**
- **contextual disambiguation**
- **inductive bias**
- **semantic generalization**

It's been 25 years since IBM

(Brown et al, COLING 1988)

Which problems have we solved?

None.

☹

The state of SMT
In danger of becoming mired in a plateau

Current SMT models of language acquisition + machine learning
Where are we?

- **stacks of hacks** – system combination, ensembles, hybrids, glueware
- **spaghetti architectures** – long pipelines of mismatched heuristic modules
- **gluttons** – resource-hungry models that are memory, computation, and data guzzlers
- **crammers** – like too many undergrads, just memorize before the test
- **superficial tests** – BLEU, TER don't measure generalization well

SMT today still fails to learn meaningful cross-lingual abstractions

- glorified **translation memory**… instead of true **machine learning**

What will it take?

BLEUaholics Anonymous

BLEUaholics Anonymous
Steps to recover from the hangover

BLEUaholics Anonymous
Steps to recover from the hangover

1 **admit that one cannot control one's addiction or compulsion**
- say "My name is _____ and I am a BLEUaholic."

BLEUaholics Anonymous
Steps to recover from the hangover

1 **admit that one cannot control one's addiction or compulsion**
- say "My name is _____ and I am a BLEUaholic."

2 **recognize a higher power that can give strength**
- science: the wisdom to know the difference

Science: the wisdom to know the difference
Scientific Method

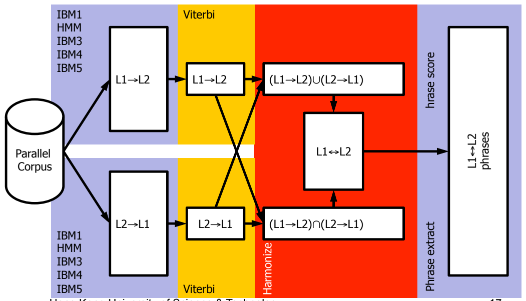
Definition of science

1 **observe** – collect data, do data analysis, error analysis
2 **hypothesize** – hypothesize a model, claim, theory, thesis, …
3 **predict** – make sure your model makes predictions
4 **test** – design and run experiment
5 **go to 1**

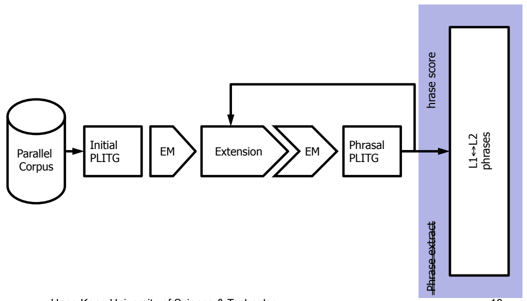
Science: the wisdom to know the difference
Occam's razor

The simplest explanation tends to be the best one.
William of Occam

Occam's razor
Compare this…



Occam's razor
… with this



BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations

Meaningful generalizations?

" التدابير الرامية إلى ||| " measures to ||| 0.111111 4.21087e-05 1 0.215425 2.718 ||| ||| 9 1
" التدابير الرامية إلى القضاء على ||| " measures to eliminate ||| 0.125 3.01811e-06 1 0.0196544 2.718 ||| ||| 8 1
" التدابير الرامية ||| " measures ||| 0.4375 0.00879359 1 0.485767 2.718 ||| ||| 16 7
" التدابير الرامية إلى ||| " measures to ||| 0.777778 0.00144573 1 0.193136 2.718 ||| ||| 9 7
" التدابير الرامية إلى القضاء على ||| " measures to eliminate ||| 0.875 0.000103621 1 0.0176208 2.718 ||| ||| 8 7
" التشريعات الوطنية ||| " national legislation ||| 1 0.158205 1 0.765638 2.718 ||| ||| 2 2
" التشريعات الوطنية المتعلقة ||| " national legislation on ||| 1 0.00403135 1 0.211303 2.718 ||| ||| 2 2
" التشريعات الوطنية المتعلقة بنقل ||| " national legislation on transfer of ||| 1 0.000380316 0.5 0.0313691 2.718 ||| ||| 2 4
" التشريعات الوطنية المتعلقة بنقل ||| " national legislation on transfer ||| 1 0.000380316 0.5 0.158477 2.718 ||| ||| 2 4
" التشريعات الوطنية المتعلقة بنقل الأسلحة ||| " national legislation on transfer of arms ||| 1 0.000224987 1 0.00742604 2.718 ||| ||| 2 2
" التطورات ||| " developments ||| 1 0.506684 1 0.863476 2.718 ||| ||| 7 7
" التطورات في ||| " developments in the ||| 1 0.336365 0.461538 0.129703 2.718 ||| ||| 6 13
" التطورات في ||| " developments in ||| 1 0.336365 0.538462 0.349129 2.718 ||| ||| 7 13
" التطورات في ميدان ||| " developments in the field of ||| 0.857143 0.170757 0.5 0.017437 2.718 ||| ||| 7 12
" التطورات في ميدان ||| " developments in the field ||| 0.857143 0.170757 0.5 0.0880919 2.718 ||| ||| 7 12
" التطورات في ميدان المعلومات ||| " developments in the field of information ||| 0.857143 0.0726128 1 0.0158491 2.718 ||| ||| 7 6
" التطورات في ميـدان ||| " developments in the field of ||| 0.142857 0.000431254 0.5 0.0172767 2.718 ||| ||| 7 2
" التطورات في ميـدان ||| " developments in the field ||| 0.142857 0.000431254 0.5 0.0872822 2.718 ||| ||| 7 2
" التطورات في ميـدان المعلومات ||| " developments in the field of information ||| 0.142857 0.000183387 1 0.0157034 2.718 ||| ||| 7 1
" التعاون ||| " cooperation ||| 0.965517 0.619446 0.982456 0.894784 2.718 ||| ||| 58 57
" التعاون ||| " enhanced ||| 1 0.0100982 0.0175439 0.00079395 2.718 ||| ||| 1 57
" التعاون الإقليمي " ||| " enhanced regional " ||| 1 0.00315024 1 0.000574123 2.718 ||| ||| 1 1
" التعاون الإقليمي " تدرج كلمة " ||| " enhanced regional " , insert " ||| 1 1.28402e-05 1 3.54351e-06 2.718 ||| ||| 1 1
" التعاون الإقليمي " تدرج كلمة ||| " enhanced regional " , insert ||| 1 1.3527e-05 1 3.6022e-06 2.718 ||| ||| 1 1
" التعاون الإقليمي ||| " enhanced regional ||| 1 0.00331873 1 0.000583632 2.718 ||| ||| 1 1
" التعاون الدولي ||| " international cooperation ||| 0.5 0.280061 1 0.842491 2.718 ||| ||| 2 1
" التعاون الدولي من ||| " international cooperation ||| 0.5 0.0285172 1 0.842491 2.718 ||| ||| 2 1
" التعاون الدولي من أجل ||| " international cooperation to ||| 1 0.0004685 1 0.261179 2.718 ||| ||| 1 1
" التعاون الدولي من أجل الحد من ||| " international cooperation to reduce ||| 1 8.01874e-06 1 0.0115687 2.718 ||| ||| 1 1
" التعاون بين ||| " cooperation between the ||| 0.97619 0.573927 0.42268 0.166976 2.718 ||| ||| 42 97
" التعاون بين ||| " cooperation between ||| 0.965517 0.573927 0.57732 0.449459 2.718 ||| ||| 58 97
" التعاون بين الأمم المتحدة ||| " cooperation between the united nations and ||| 0.9 0.498652 0.397059 0.00836641 2.718 ||| ||| 30 68
" التعاون بين الأمم المتحدة ||| " cooperation between the united nations ||| 0.931818 0.498652 0.602941 0.1609 2.718 ||| ||| 44 68
" التعاون بين الأمم المتحدة والمنظمة ||| " cooperation between the united nations and ||| 0.0666667 0.000262789 0.5 0.00928672 2.718 ||| ||| 30 4
" التعاون بين الأمم المتحدة والمنظمة ||| " cooperation between the united nations ||| 0.0454545 0.000262789 0.5 0.178599 2.718 ||| ||| 44 4
" التعاون بين منظومة ||| " cooperation between ||| 0.0172414 5.217e-05 1 0.449459 2.718 ||| ||| 58 1
" التعـــاون ||| " cooperation ||| 0.0172414 0.00164891 1 0.73778 2.718 ||| ||| 58 1

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations
4. **make amends for these errors**
   - design SMT models oriented toward learning the right abstractions

big parallel corpus → **small** transduction grammar

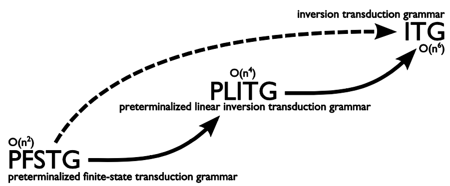
Want:

- Compact generalization of the translation knowledge encoded in the corpus
- Unsupervised learning of transduction grammar rules without Giza, Moses, parsers, or anything else

Why?

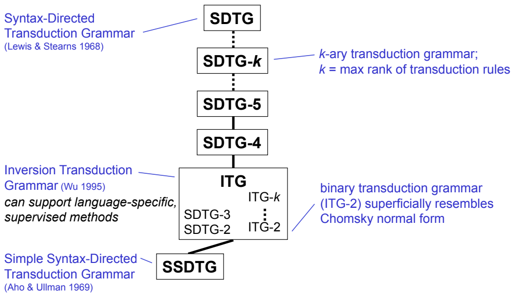
- Rearchitecting the SMT core
  - "Machine Learning 101":
    do training and testing on the same model
  - Core internal representation designed from start for learning semantic frame generalizations
  - Emphasis on generalizing rather than memorizing
  - Minimum description length → Occam's razor for model size
- Evaluated in pure, unadulterated form
  - Not as a preprocessing subroutine (eg, for word alignment) within an off-the-shelf "stack-of-hacks" SMT system
  - ITG decoder matched to ITG learner
  - Better to see lower BLEU scores for now, in order to better understand transduction grammar induction behavior

Bootstrapping

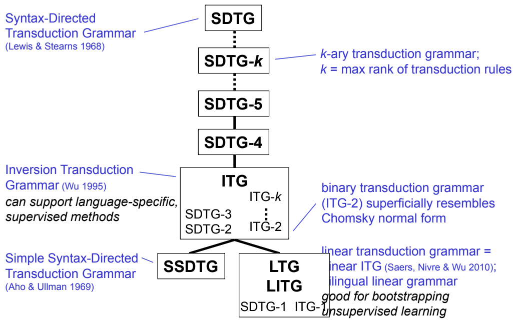


Monolingual **Languages**			Bilingual **Transductions**
regular or finite-state languages **FSA** *or* *CFG that is right or left linear or regular*	$O(n^2)$	$O(n^4)$	regular or finite-state transductions **FST** *or* *SDTG (or synchronous CFG) that is right or left linear or regular*
linear languages **LG** *or* *CFG that is linear or unary*	$O(n^2)$	$O(n^4)$	linear transductions **LTG** *or* *SDTG (or synchronous CFG) that is linear or unary*
context-free languages **CFG**	$O(n^3)$	$O(n^6)$	inversion transductions **ITG** *or* *SDTG (or synchronous CFG) that is binary or ternary or inverting*
		$O(n^{2n+2})$	syntax-directed transductions **SDTG** *(or synchronous CFG)*

The Transduction Grammar Hierarchy



The Transduction Grammar Hierarchy



BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations
4. **make amends for these errors**
   - design SMT models oriented toward learning the right abstractions

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations
4. **make amends for these errors**
   - design SMT models oriented toward learning the right abstractions
5. **learn to live a new life with a new code of behavior**
   - evaluate your MT models against semantically meaningful metrics

HMEANT
human semantic MT metric

Semantic SMT – Part III
Evaluation metrics and objective functions

- Semantic MT evaluation metrics based on semantic frame agreement
- Deeply integrating semantic frames into MT evaluation metrics
- Desirable characteristics to maintain:
  - simplicity
  - inexpensiveness
  - representational transparency for scientific error analysis
- Human evaluated semantic MT evaluation metric **HMEANT** significantly outperforms even the state-of-the-art expensive HTER used by DARPA
- Fully automatic semantic MT evaluation metric **MEANT** significantly outperforms BLEU, NIST, METEOR, WER, PER, CDER, and even the state-of-the-art expensive TER used by DARPA
- Exploiting MEANT as the objective function for tuning SMT robustly increases translation accuracy

The problem with conventional MT evaluation metrics

This has been our trajectory toward semantic SMT over the years

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

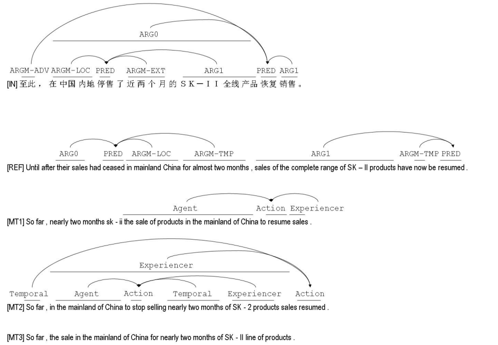
- **2009-** Semantic MT evaluation with SRL-for-MTE (MEANT)

HMEANT history

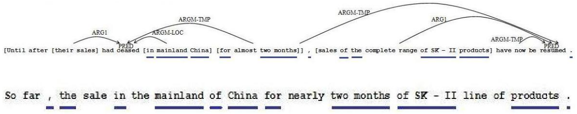
*Acknowledgments: DARPA GALE, BOLT

- **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

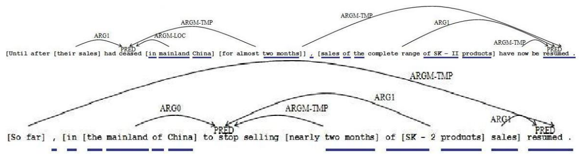


Example: a less useful translation
Fewer SRL matches ☺
but more N-gram and syntax-subtree matches! ☹



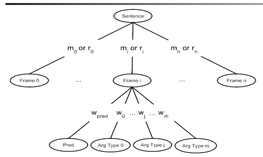
N-gram		Syntax-subtree		SRL	
1-gram matches:	15	1-level subtree matches:	34	Predicate matches:	0
2-gram matches:	4	2-level subtree matches:	8		
3-gram matches:	3	3-level subtree matches:	2		
4-gram matches:	1	4-level subtree matches:	0		

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 sentence
  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 frame
  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

HMEANT is fairly cheap…
… but still requires humans

- Annotation tasks

1. label semantic predicates, roles, and fillers
   replace humans with automatic SRL?
2. align predicates and fillers between the reference and machine translations

HMEANT is fairly cheap…
… but still requires humans

- Annotation tasks

1. label semantic predicates, roles, and fillers
   replace humans with automatic SRL?
2. align predicates and fillers between the reference and machine translations
   replace humans with automatic alignment?

toward eliminating humans

- UMEANT: unsupervised approach to estimating MEANT's parameters (SSST-6, at ACL 2012; WMT-8, at ACL 2013)
  - further reduce the evaluation cost by eliminating the dependency on a human adequacy-ranked training corpus for tuning the weights for each semantic role type
  - good for evaluating resources sparse language
- Fully automated MEANT (WMT-7, at NAACL 2012; IWSLT 2014)
  - 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

HMEANT
Further reducing 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.

UMEANT
Unsupervised weight estimates for HMEANT

- 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.

UMEANT
Unsupervised parameter estimation for HMEANT

- 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
  - **performed extremely well at WMT-2013** metrics evaluation

MEANT
automatic semantic MT metric

HMEANT vs. MEANT - SRL and alignment algorithm

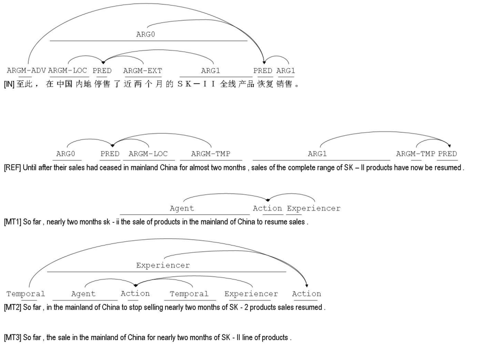
HMEANT

1. Human annotators annotate the shallow semantic structures of both the references and MT output.
2. Human judges align the semantic frames between the references and MT output by judging the correctness of the predicates.
3. For each pair of aligned semantic frames,
   (a) Human judges determine the translation correctness of the semantic role fillers.
   (b) Human judges align the semantic role fillers between the reference and MT output according to the correctness of the semantic role fillers.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers.

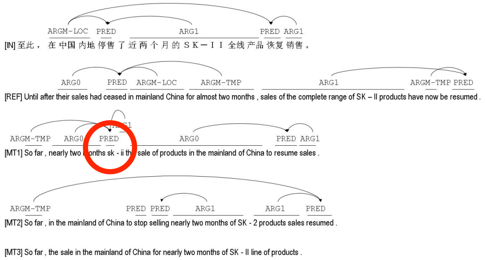
MEANT

1. *Apply an automatic shallow semantic parser* on both the references and MT output.
2. *Apply maximum weighted bipartite matching algorithm to* align the semantic frames between the references and MT output by the *lexical similarity of the predicates*.
3. For each pair of aligned semantic frames,
   (a) *Lexical similarity scores* determine the *similarity* of the semantic role fillers.
   (b) *Apply maximum weighted bipartite matching algorithm to* align the semantic role fillers between the reference and MT output according to their *lexical similarity*.
4. Compute the weighted f-score over the matching role labels of these aligned predicates and role fillers.

MEANT vs. HMEANT
Human SRL



MEANT vs. HMEANT
Automatic SRL errors can create problems



MEANT vs. HMEANT
Auto alignment judgments can be more precise

HMEANT

REF roles	REF	MT2 roles	MT2	decision
PRED	ceased	Action	stop	match
ARG0	their sale	—	—	incorrect
ARGM-LOC	in mainland China	Agent	the mainland of China	correct*
ARGM-TMP	for almost two months	Temporal	nearly two months	correct
—	—	Experiencer	SK - 2 products	incorrect
PRED	resumed	Action	resume	match
ARG0	sales of complete range of SK - II products	Experiencer	in the mainland of China to stop selling nearly two months of SK - 2 products sales	incorrect
ARGM-TMP	Until after , their sales had ceased in mainland China for almost two months	Temporal	So far	partial
ARGM-TMP	now	—	—	incorrect

MEANT

REF roles	REF	MT2 roles	MT2	similarity
PRED	ceased	PRED	stop	0.0377
ARG0	their sales	—	—	—
ARGM-LOC	in mainland China	—	—	—
ARGM-TMP	for almost two months	—	—	—
—	—	PRED	selling	—
—	—	ARG1	nearly two months of SK	—
PRED	resumed	PRED	resumed	1.0
ARG1	sales of complete range of SK - II products	ARG1	2 products sales	0.0836
ARGM-TMP	now	ARGM-TMP	So far	0.0459

MEANT vs. HMEANT
Calculation of scores

HMEANT

$$m_i \equiv \frac{\#\text{tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}$$
$$r_i \equiv \frac{\#\text{tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}$$
$$M_{i,j} \equiv \text{total \# ARG j of aligned frame i in MT}$$
$$R_{i,j} \equiv \text{total \# ARG j of aligned frame i in REF}$$
$$C_{i,j} \equiv \text{\# correct ARG j of aligned frame i in MT}$$
$$P_{i,j} \equiv \text{\# partially correct ARG j of aligned frame i in MT}$$

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

MEANT

$$m_i \equiv \frac{\#\text{tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}$$
$$r_i \equiv \frac{\#\text{tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}$$
$$M_{i,j} \equiv \text{total \# ARG j of aligned frame i in MT}$$
$$R_{i,j} \equiv \text{total \# ARG j of aligned frame i in REF}$$
$$S_{i,\text{pred}} \equiv \text{sim. of pred of REF and MT in aligned frame i}$$
$$S_{i,j} \equiv \text{sim. of ARG j of REF and MT in aligned frame i}$$

$$\text{precision} = \frac{\sum_i m_i \frac{w_{\text{pred}} S_{i,\text{pred}}+\sum_j w_j S_{i,j}}{w_{\text{pred}}+\sum_j w_j M_{i,j}}}{\sum_i m_i}$$
$$\text{recall} = \frac{\sum_i r_i \frac{w_{\text{pred}} S_{i,\text{pred}}+\sum_j w_j S_{i,j}}{w_{\text{pred}}+\sum_j w_j R_{i,j}}}{\sum_i r_i}$$

Challenges in automating HMEANT

- A wide range of lexical similarity scoring models are available
- We experimented on
  - BLEU
  - METEOR
  - Similarity measures based on word context vectors
    - Cosine similarity
    - MinMax-MI (Dagan, 2000)
    - and many more…

Results
MEANT outperforms all automatic metrics

	GALE-A (training)	GALE-B (testing)
Human metrics		
HMEANT	0.49	0.27
HTER	0.43	0.20
Automatic metrics		
MEANT	—	—
- with MinMax-MI on context vector model of window size 3	**0.37**	0.19
- with MinMax-MI on context vector model of window size 5	0.37	0.17
- with Cosine on context vector model of window size 3	0.32	0.13
- with Cosine on context vector model of window size 5	0.30	0.08
- with METEOR	0.17	—
- with BLEU	0.00	—
METEOR	0.20	**0.21**
NIST	0.29	0.09
TER	0.20	0.10
BLEU	0.20	0.12
PER	0.20	0.07
WER	0.10	0.11
CDER	0.12	0.10

- Statistical anomaly: METEOR is exceptionally high when testing on GALE-B (even higher than human HTER!!)

Why are word context vector similarities more suitable for judging role filler similarity than BLEU and METEOR?

- High variation between alternative paraphrasing of relatively short role
  - Makes the number of matching n-grams quite small, which hurts BLEU and METEOR
- Easy trainability of word context vectors
  - Can readily be trained using any publicly available large monolingual corpus

More on the first batch of results

- MinMax-MI is better than cosine similarity
- Context vector models using a window size of 3 appear to be as good or better than those using a window size of 5

Q: Does auto semantic frame alignment perform as well as human?

- MEANT vs. semi-automatic version of HMEANT (2011)
  - SRL in both metrics is performed automatically
  - Semantic frame alignment in HMEANT is done manually

Results
Don't align semantic frames manually

Semantic frame alignment	GALE-A	GALE-B
Automatic	0.37	0.19
Manual	0.35	0.17

- Automatic semantic frame alignment is as good or even better than doing the alignment manually

Why?

- Automatic alignment is finer grained
  - **Human:** Only a 3-point scale of translation correctness (correct / partial / incorrect)
  - **Automatic:** Continuous points scale of lexical similarity between semantic role fillers
- The lexical similarity metric appears highly reliable
  - at least, when the candidates for role fillers are restricted to the fairly small set defined by the sentence pairs

Q: Use fillers to help align frames?

- **Background:** in HMEANT, semantic frames are aligned only if predicates match
  - Reduces mental challenge for lay annotators to compare and keep in mind all the semantic role fillers at the same time
  - BUT… easy for an algorithm to do!
- Good idea to align by maximizing the match of the semantic role fillers (in addition to the predicate)?
- 2 obvious, natural ways of aggregating the lexical similarity of aligned semantic role fillers:
  - linear average
  - inverse of sum of negative log

Results
Match predicates when aligning frames

Frame alignment	GALE-A	GALE-B
Predicate only	0.37	0.19
Linear average	0.35	0.10
Inverse of sum of neg. log	0.30	0.17

- Using only the predicates to align semantic frames is more robust than two natural ways to aggregate role filler match

Why?

- Lexical similarities are aggregated with uniform weight across different types of role fillers
  - Ignores the fact that different role types contribute to a widely varying degree to the meaning of an entire semantic frames as studied by Lo and Wu (2011c)
- What about adding weights for each semantic role type?

Cons:
  - the complexity of MEANT would be greatly increased
  - unlikely to be worthwhile as the automatic alignment is already performing as well as human alignment

Q: Would word-aligning the tokens within role fillers help?

- Hypothesis:
  - summing the lexical similarities
    only between word aligned tokens in the role filler strings
    (instead of all pairwise combinations of tokens)
    should reduce the level of noise
    and thus improve MEANT performance
- Method: variant of competitive linking (Melamed 1996)
  - aim: avoid the danger of aligning a token in one segment
    to excessive numbers of tokens in the other segment

Results
Don't word align tokens in semantic role fillers

Semantic role filler similarity	GALE-A	GALE-B
All pairwise tokens	0.37	0.19
Only aligned tokens	0.36	0.17

- Surprisingly, word aligning the role fillers' tokens does not help!
- Why?
  - Word alignments over-constrain calculation of segment similarities
  - Individual lexical similarities are already fairly accurate
    - ➩ similarities between words that do not correspond do not hurt (since they are already close to zero)
    - ➩ BUT… when word alignment is ambiguous, strictly obeying a hard word alignment undesirably forces dropping of some individual lexical similarity scores that are actually relevant

Fully automatic
MEANT

- Surprisingly, when aligning semantic frames
  - automatic algorithm is as good as manual alignment
  - using only the similarities of the predicates is better than aggregating that of all semantic role fillers
- and surprisingly, when judging similarity between semantic role fillers
  - aggregating similarity of all pairwise combination of word tokens is more accurate than considering only the similarity of the tokens that obey word alignment

UMEANT & MEANT ranked top 3-4 in WMT 2013 metrics evaluation

- UMEANT ranked the highest in evaluating Czech-English
- MEANT and UMEANT are recommended
  for evaluating MT into English (Macháček and Bojar, 2013)
- UMEANT is better at adapting to the linguistic differences for evaluating translation from different languages

Correlation coefficient	Spearman's ρ Correlation Coefficient					
Directions	fr-en	de-en	es-en	cs-en	ru-en	Average
Considered systems	12	22	11	10	17	
METEOR	.984 ± .014	.961 ± .020	**.979** ± .024	.964 ± .027	.789 ± .040	**.935** ± .012
DEPREF-ALIGN	**.995** ± .011	**.966** ± .018	.965 ± .031	.964 ± .023	.768 ± .041	.931 ± .012
UMEANT	.989 ± .011	.946 ± .018	.958 ± .028	**.973** ± .032	.775 ± .037	.928 ± .012
MEANT	.973 ± .014	.926 ± .021	.944 ± .038	**.973** ± .032	.765 ± .038	.916 ± .013
SEMPOS	.938 ± .014	.919 ± .028	.930 ± .031	.955 ± .018	**.823** ± .037	.913 ± .012
DEPREF-EXACT	.984 ± .011	.961 ± .017	.937 ± .038	.936 ± .027	.744 ± .046	.912 ± .015
SIMPBLEU-RECALL	.978 ± .014	.936 ± .020	.923 ± .052	.909 ± .027	.798 ± .043	.909 ± .017
BLEU-MTEVAL-INTL	.989 ± .014	.902 ± .017	.895 ± .049	.936 ± .032	.695 ± .042	.883 ± .015
BLEU-MTEVAL	.989 ± .014	.895 ± .020	.888 ± .045	.936 ± .032	.670 ± .041	.876 ± .015
BLEU-MOSES	.993 ± .014	.902 ± .017	.879 ± .051	.936 ± .036	.651 ± .041	.872 ± .016
CDER-MOSES	**.995** ± .014	.877 ± .017	.888 ± .049	.927 ± .036	.659 ± .045	.869 ± .017
SIMPBLEU-PREC	.989 ± .008	.846 ± .020	.832 ± .059	.918 ± .023	.704 ± .042	.858 ± .017
NLEPOR	.945 ± .022	.949 ± .025	.825 ± .056	.845 ± .041	.705 ± .043	.854 ± .018
LEPOR V3.100	.945 ± .019	.934 ± .027	.748 ± .077	.800 ± .036	.779 ± .041	.841 ± .020
NIST-MTEVAL	.951 ± .019	.875 ± .022	.769 ± .077	.891 ± .027	.649 ± .045	.827 ± .020
NIST-MTEVAL-INTL	.951 ± .019	.875 ± .022	.762 ± .077	.882 ± .032	.658 ± .045	.826 ± .021
TER-MOSES	.951 ± .019	.833 ± .023	.825 ± .077	.800 ± .036	.581 ± .045	.798 ± .021
WER-MOSES	.951 ± .019	.672 ± .026	.797 ± .070	.755 ± .041	.591 ± .042	.753 ± .020
PER-MOSES	.852 ± .027	.858 ± .025	.357 ± .091	.697 ± .043	.677 ± .040	.688 ± .024
TERRORCAT	.984 ± .011	.961 ± .023	.972 ± .028	n/a	n/a	**.972** ± .012

training SMT against MEANT

What's new in MEANT research?

- MEANT and UMEANT ranked top 3 and 4 in WMT 2013 metrics evaluation track
  - UMEANT ranked the highest in evaluating Czech-English
  - MEANT and UMEANT are recommended for evaluating MT into English (Macháček & Bojar 2013)
- Training SMT on MEANT (Lo Addanki Saers Wu, ACL 2013)
  - First ever SMT system to be trained on a purely semantic objective
  - MEANT-tuned Chinese-English system outperforms BLEU-tuned or TER-tuned systems, across the commonly used automatic evaluation metrics and human adequacy evaluation
  - Forthcoming: same consistent improvement also for English-Chinese using new automatic Chinese MEANT (IWSLT 2013)

Tuning MT against MEANT more robustly produces adequate translations than tuning against BLEU or TER!

- MEANT-tuned systems achieve the best scores across nearly all other metrics
- MEANT-tuned systems maintain a fine balance between lexical choice and word order, performing well as measured by:
  (a) n-gram metrics that reward lexical matching
  (b) edit distance metrics that penalize incorrect word order

newswire	BLEU	NIST	METEOR no_syn	METEOR	WER	CDER	TER	MEANT
BLEU-tuned	29.85	8.84	52.10	55.42	67.88	55.67	**58.40**	0.1667
TER-tuned	25.37	6.56	48.26	51.24	66.18	52.58	56.96	0.1578
MEANT-tuned	**25.91**	**7.81**	**50.15**	**53.60**	**67.76**	**54.56**	58.61	**0.1676**

Table 1: Translation quality of MT system tuned against MEANT, BLEU and TER on newswire data

Baseline: Moses hierarchical MT
Corpus: (dev) NIST 02-06 (test) NIST 08

but won't informal genres break MEANT's semantic parsing?

Q. Are semantic frames less useful on informal genres translation?

- Automatic shallow semantic parsing fares worse on informal genres
  - Accuracy drops
    - around 10% on speech data (Favre et al., 2010)
    - more than 30% on tweets data (Liu et al., 2010)
  - Why?
    - Robustness of the POS tagging and syntactic parsing that the automatic semantic parser depends on suffers
      - Data demonstrates a large variety of grammar issues, such as disfluencies, incomplete sentences and misspellings (Mei and Kirchhoff, 2010)
- So previous work on informal text machine translation mostly focused on
  1. fixing the grammar issues in the input or
  2. addressing the training data sparsity problem using domain adaptation techniques
- **Can informal genres be better translated by tuning against MEANT?**

Can informal genres be better translated by tuning against MEANT?

- Informal genres
  1. IWSLT 2012 Chinese-English TED talk
  2. BOLT P1 web forum data
- Baselines (common practices)
  1. Tuning against BLEU
  2. Tuning against TER

Cross evaluation using automatic metrics

Table 1: Translation quality of MT system tuned against MEANT, BLEU and TER on TED talk data

TED talk	BLEU ↑	NIST ↑	METEOR no_syn ↑	METEOR ↑	WER ↓	CDER ↓	TER ↓	MEANT ↑
BLEU-tuned	12.09	4.36	38.14	41.28	83.87	68.55	80.83	22.70
TER-tuned	9.63	3.67	32.75	35.19	74.00	59.24	72.31	20.41
MEANT-tuned	**11.24**	**4.22**	**38.57**	**41.96**	**80.97**	**66.21**	**78.10**	**22.74**

Table 2: Translation quality of MT system tuned against MEANT, BLEU and TER on web forum data

forum	BLEU ↑	NIST ↑	METEOR no_syn ↑	METEOR ↑	WER ↓	CDER ↓	TER ↓	MEANT ↑
BLEU-tuned	9.58	4.10	31.77	34.63	80.09	64.54	76.12	17.11
TER-tuned	6.94	2.21	28.55	30.85	76.15	57.96	74.73	15.39
MEANT-tuned	**7.92**	**3.11**	**30.40**	**33.08**	**77.32**	**61.01**	**74.64**	**17.27**

- Tuning against BLEU achieves the highest BLEU, but overfits
- MEANT-tuned systems outperform BLEU- or TER-tuned systems across the commonly used metrics
  - if we ignore the similar metrics that the MT systems are trained on,
- MEANT-tuned systems maintain a fine balance between lexical choices and word order
  - as it performs well on both n-gram metrics that reward lexical matches and edit distance metrics that penalize incorrect word order

Human evaluators more frequently prefer MEANT-tuned systems over BLEU- or TER-tuned systems

- MEANT-tuned system are ranked **the most adequate more frequently** than BLEU- or TER-tuned systems
- MEANT-tuned systems are more adequate
  - than TER-tuned systems at 95% significance level (even at 99% level)
  - than BLEU-tuned systems at 95% significance level

Table 3: No. of sentences ranked the most adequate by human evaluators for each system in the web forum experiment.

	Eval 1	Eval 2
BLEU-tuned (B)	47	42
TER-tuned (T)	28	23
MEANT-tuned (M)	59	68
B=T	0	0
M=B	8	9
M=T	4	4
M=B=T	4	4

Error analysis When the shallow semantic parser fails

- The shallow semantic parser fails to output a parse for
  - over 14% of the sentences in the TED talk data
  - on average over 8% of the sentences in the web forum data
- Why further investigate these cases?
  - failure of the shallow semantic parser to provide any parse automatically results in a zero MEANT score

Table 4: Number of sentences with no automatic semantic parsing output in each data set

dataset	genre	#sentences	#no semantic parse	%no semantic parse
TED-dev	public talk	934	138	14.78%
TED-test	public talk	1664	237	14.24%
BOLT P1-dev	forum	2000	229	11.45%
BOLT P1-test	forum	1697	100	5.89%
MetricsMaTr 08	broadcast news	221	9	4.07%

Failure to label the "be" semantic frame

- Surprisingly: ungrammatical sentences are not the biggest cause!
- Rather:
  the major source of errors is failing to identify the semantic frame for copulas or existential sense of "be" in grammatical sentences
  - Up to 11% of the sentences in informal genres have the copula or the existential sense of "be" as a predicate

Table 5: Detailed breakdown of the sentences with no semantic frame identified by the automatic semantic parser. (#"be" is the number of sentences that has at least one grammatical and valid semantic frame of the copula or existential sense of "be"; #no verb (≤ 10) and #no verb (> 10) are the number of sentences that has no verb in the sentence with the sentence length is "less than or equal to 10" or "greater than 10" respectively; #other is the number of sentences that do not fall into any of the previous categories.)

dataset	genre	#no parse	#"be"	#no verb ($\leq$10)	#no verb ($>$ 10)	#others
TED-dev	public talk	138	110	20	3	5
TED-test	public talk	237	191	38	3	5
BOLT P1-dev	forum	229	166	56	6	1
BOLT P1-test	forum	100	81	4	5	10
MetricsMaTr 08	broadcast news	9	9	0	0	0

Why does automatic semantic parsing fail to label the "be" semantic frame?

- Propbank framesets definition of the predicate "be"
  Roleset *be.01*: *copula*
  Roles: ARG1-*topic*, ARG2-*comment*
  Roleset *be.02*: *existential*
  Roles: ARG1-*thing that is*
  Roleset *be.03*: *auxiliary*
  Roles: **do not tag**
- Examples in TED talk or web forum data
  - Copula: A language is a flash of the human spirit .
  - Existential: There is no feed .
  - Auxiliary: [ARG0 The sun] is [PRED rising] .
- Shallow semantic parsers are trained on formal text
  - where "be" is more often used as auxiliary verb together with the present or past participle to realize different tenses or voices in grammar

what makes a translation useful

how well is

who did what to whom, for whom, when, where, why and how

preserved in translation?

surface MT metrics (BLEU, NIST, …)

how well do
n-grams
match

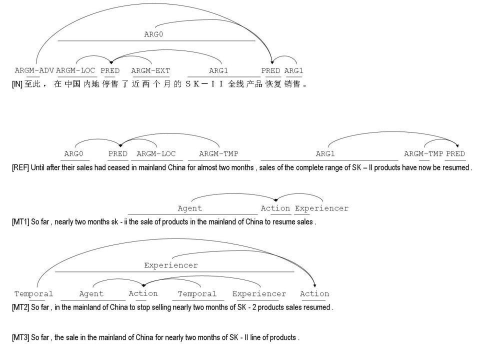
between reference and machine translations?

semantic MT metrics (MEANT, …)

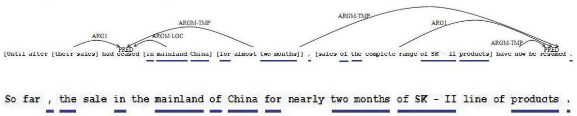
how well do
semantic frames
match

between reference and machine translations?

HMEANT
Human semantic MT evaluation via SRL

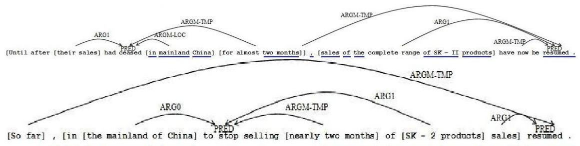


Example: a less useful translation
Fewer SRL matches ☺
but more N-gram and syntax-subtree matches! ☹



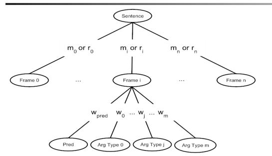
N-gram		Syntax-subtree		SRL	
1-gram matches:	15	1-level subtree matches:	34	Predicate matches:	0
2-gram matches:	4	2-level subtree matches:	8		
3-gram matches:	3	3-level subtree matches:	2		
4-gram matches:	1	4-level subtree matches:	0		

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 sentence
  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 frame
  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, MEANT, UMEANT a family of semantic frame based MT evaluation metrics

- **HMEANT** human [Lo & Wu, ACL, IJCAI, SSST 2011]
  - assesses MT utility via semantic frames with high representational transparency
  - needs only unskilled humans to annotate and align semantic frames
  - correlates with human adequacy judgment better than HTER at lower labor cost
  - applies easily on any language pair
- **MEANT** automatic [Lo, Tumuluru & Wu, WMT 2012]
  - outperforms all commonly used automatic MT evaluation metrics
    - replaces human SRL with automatic shallow semantic parsing
    - replaces human semantic frame alignment with automatic alignment
  - simple & transparent – preserves Occam's razor spirit of HMEANT
  - now in both English and Chinese
  - top 4 in WMT2013 metrics track evaluation
- **UMEANT** unsupervised automatic [Lo & Wu, SSST 2012]
  - eliminates any dependency on a corpus with human ranked MT output in training the weights of semantic role labels by estimating them via the relative frequency of the labels in the reference
  - good for resource-sparse languages
  - top 3 in WMT2013 metrics track evaluation

the first ever **directly semantically trained SMT systems**

- **why tune MT against MEANT?**
  - produces more robustly adequate translations than tuning against BLEU or TER
    - across genres (newswire, web forum, TED)
    - across output languages (English, Chinese)
    - accros MT paradigms (phrase based, hierarchical phrase based)
  - constrains the MT system to make more accurate lexical and reordering choices
    - preserving the meaning of the translation as captured by semantic frames right in the training process
- the first time in 25 years of history that SMT has ever been directly trained **to maximize preserving who did what to whom, for whom, when, where, how, why** (a bit scary!)

XMEANT a cross-lingual semantic frame based MT evaluation metric

- **XMEANT** cross-lingual MEANT [Lo, Beloucif, Saers & Wu, ACL 2014]
  - eliminates the need for expensive reference translations … yet correlates with human adequacy judgment even more closely than MEANT!
  - since words come from different vocabularies for input and output languages, can't use MEANT's word vector similarities to align role fillers any more; instead use translation probabilities plus **language-independent BITGs constraints** (Wu 1997; Zens & Ney 2003; Saers & Wu 2009)
  - a new generation of Wu & Fung's (NAACL, EAMT 2009) cross-lingual score … that exploits all our recent advances on monolingual MEANT
- well, if BITG constraints work so well for cross-lingual XMEANT… could they also improve ordinary monolingual MEANT?

IMEANT *new!* an ITG-based semantic frame based MT evaluation metric

- **further improves** MEANT's correlation with human adequacy judgment which was already high
- achieved by using **bracketing ITGs** to biparse the semantic role fillers in both reference and machine translations
- shows that ITGs
  - appropriately constrain the allowable permutations between the compositional segments across the reference and machine translations
  - score the phrasal similarity of the semantic role fillers more accurately than the simple heuristics like bag-of-word alignment or maximum alignment

MEANT

1. apply automatic shallow semantic parsing to the reference and machine translations
2. apply maximum weighted bipartite matching to align the semantic frames between the reference translation and the machine translation, according to the lexical similarity of the semantic predicates
3. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the reference translation and the machine translation, according to the lexical similarity of the semantic role fillers
4. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

MEANT

1. apply automatic shallow semantic parsing to the reference and machine translations
2. apply maximum weighted bipartite matching to align the semantic frames between the reference translation and the machine translation, according to the lexical similarity of the semantic predicates
3. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the reference translation and the machine translation, according to the lexical similarity of the semantic role fillers
4. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT}\\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF}\\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}\\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}\\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates}\\
w_j &\equiv \text{weight of similarity of ARG j}\\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i}\\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}\\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i}\\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i}\\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{aligned}
$$

MEANT **IMEANT**

MEANT

1. apply automatic shallow semantic parsing to the reference and machine translations
2. apply maximum weighted bipartite matching to align the semantic frames between the reference translation and the machine translation, according to the lexical similarity of the semantic predicates
3. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the reference translation and the machine translation, according to the lexical similarity of the semantic role fillers
4. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

IMEANT

1. apply automatic shallow semantic parsing to the reference and machine translations
2. apply maximum weighted bipartite matching to align the semantic frames between the reference translation and the machine translation, according to the lexical similarity of the semantic predicates
3. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the reference translation and the machine translation, according to the lexical similarity of the semantic role fillers **aggregated under ITG-constrained alignments**
4. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

MEANT **IMEANT**

MEANT

$$
\begin{aligned}
q^0_{i,j} &= ARG\ j \text{ of aligned frame } i \text{ in MT}\\
q^1_{i,j} &= ARG\ j \text{ of aligned frame } i \text{ in REF}\\
w^0_i &= \frac{\text{\# tokens filled in aligned frame } i \text{ of MT}}{\text{total \# tokens in MT}}\\
w^1_i &= \frac{\text{\# tokens filled in aligned frame } i \text{ of REF}}{\text{total \# tokens in REF}}\\
w_{pred} &= \text{weight of similarity of predicates}\\
w_j &= \text{weight of similarity of } ARG\ j\\
\boldsymbol{e}_{i,pred} &= \text{pred string of the aligned frame } i \text{ of MT}\\
\boldsymbol{f}_{i,pred} &= \text{pred string of the aligned frame } i \text{ of REF}\\
\boldsymbol{e}_{i,j} &= \text{role fillers of } ARG\ j \text{ of the aligned frame } i \text{ of MT}\\
\boldsymbol{f}_{i,j} &= \text{role fillers of } ARG\ j \text{ of the aligned frame } i \text{ of REF}\\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f
\end{aligned}
$$

$$
\begin{aligned}
prec_{\boldsymbol{e},\boldsymbol{f}} &= \frac{\sum_{e\in\boldsymbol{e}} \max_{f\in\boldsymbol{f}} s(e,f)}{|\boldsymbol{e}|}\\
rec_{\boldsymbol{e},\boldsymbol{f}} &= \frac{\sum_{f\in\boldsymbol{f}} \max_{e\in\boldsymbol{e}} s(e,f)}{|\boldsymbol{f}|}\\
s_{i,pred} &= \frac{2 \times prec_{\boldsymbol{e}_{i,pred},\boldsymbol{f}_{i,pred}} \times rec_{\boldsymbol{e}_{i,pred},\boldsymbol{f}_{i,pred}}}{prec_{\boldsymbol{e}_{i,pred},\boldsymbol{f}_{i,pred}} + rec_{\boldsymbol{e}_{i,pred},\boldsymbol{f}_{i,pred}}}\\
s_{i,j} &= \frac{2 \times prec_{\boldsymbol{e}_{i,j},\boldsymbol{f}_{i,j}} \times rec_{\boldsymbol{e}_{i,j},\boldsymbol{f}_{i,j}}}{prec_{\boldsymbol{e}_{i,j},\boldsymbol{f}_{i,j}} + rec_{\boldsymbol{e}_{i,j},\boldsymbol{f}_{i,j}}}
\end{aligned}
$$

$$
\begin{aligned}
precision &= \frac{\sum_i w^0_i \frac{w_{pred} s_{i,pred} + \sum_j w_j s_{i,j}}{w_{pred} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i}\\
recall &= \frac{\sum_i w^1_i \frac{w_{pred} s_{i,j} + \sum_j w_j s_{i,j}}{w_{pred} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i}\\
MEANT &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned}
$$

IMEANT

$$
\begin{aligned}
q^0_{i,j} &= ARG\ j \text{ of aligned frame } i \text{ in MT}\\
q^1_{i,j} &= ARG\ j \text{ of aligned frame } i \text{ in REF}\\
w^0_i &= \frac{\text{\# tokens filled in aligned frame } i \text{ of MT}}{\text{total \# tokens in MT}}\\
w^1_i &= \frac{\text{\# tokens filled in aligned frame } i \text{ of REF}}{\text{total \# tokens in REF}}\\
w_{pred} &= \text{weight of similarity of predicates}\\
w_j &= \text{weight of similarity of } ARG\ j\\
\boldsymbol{e}_{i,pred} &= \text{pred string of the aligned frame } i \text{ of MT}\\
\boldsymbol{f}_{i,pred} &= \text{pred string of the aligned frame } i \text{ of REF}\\
\boldsymbol{e}_{i,j} &= \text{role fillers of } ARG\ j \text{ of the aligned frame } i \text{ of MT}\\
\boldsymbol{f}_{i,j} &= \text{role fillers of } ARG\ j \text{ of the aligned frame } i \text{ of REF}\\
s(e,f) &= \text{lexical similarity of token } e \text{ and } f
\end{aligned}
$$

$$
\begin{aligned}
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A} \rangle\\
\mathcal{R} &\equiv \{\mathrm{A} \to [\mathrm{A\ A}], \mathrm{A} \to \langle \mathrm{A\ A} \rangle, A \to e/f\}\\
p([\mathrm{A\ A}]|\mathrm{A}) &= p(\langle \mathrm{A\ A} \rangle|\mathrm{A}) = 1\\
p(e/f\,|\mathrm{A}) &= s(e,f)\\
s_{i,pred} &= lg^{-1}\left(\frac{\lg\left(P\left(A \xrightarrow{*} \boldsymbol{e}_{i,pred}/\boldsymbol{f}_{i,pred}\,|G\right)\right)}{max(|\boldsymbol{e}_{i,pred}|, |\boldsymbol{f}_{i,pred}|)}\right)\\
s_{i,j} &= lg^{-1}\left(\frac{\lg\left(P\left(A \xrightarrow{*} \boldsymbol{e}_{i,j}/\boldsymbol{f}_{i,j}\,|G\right)\right)}{max(|\boldsymbol{e}_{i,j}|, |\boldsymbol{f}_{i,j}|)}\right)
\end{aligned}
$$

$$
\begin{aligned}
precision &= \frac{\sum_i w^0_i \frac{w_{pred} s_{i,pred} + \sum_j w_j s_{i,j}}{w_{pred} + \sum_j w_j |q^0_{i,j}|}}{\sum_i w^0_i}\\
recall &= \frac{\sum_i w^1_i \frac{w_{pred} s_{i,j} + \sum_j w_j s_{i,j}}{w_{pred} + \sum_j w_j |q^1_{i,j}|}}{\sum_i w^1_i}\\
IMEANT &= \frac{2 \times precision \times recall}{precision + recall}
\end{aligned}
$$

IMEANT outperforms previous versions of MEANT

- IMEANT shows a 3 point improvement over MEANT on GALE-A
- IMEANT is tied with MEANT in correlation with HAJ on GALE-B

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

	GALE-A	GALE-B
HMEANT	0.53	0.37
IMEANT	**0.51**	**0.33**
XMEANT	**0.51**	0.20
MEANT	0.48	**0.33**
METEOR 1.5 (2014)	0.43	0.10
NIST	0.29	0.16
METEOR 0.4.3 (2005)	0.20	0.29
BLEU	0.20	0.27
TER	0.20	0.19
PER	0.20	0.18
CDER	0.12	0.16
WER	0.10	0.26

IMEANT
outperforms cross-lingual XMEANT

- IMEANT is tied with XMEANT on GALE-A
- IMEANT correlates with HAJ much better than XMEANT on GALE-B

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

	GALE-A	GALE-B
HMEANT	0.53	0.37
IMEANT	**0.51**	**0.33**
XMEANT	**0.51**	0.20
MEANT	0.48	**0.33**
METEOR 1.5 (2014)	0.43	0.10
NIST	0.29	0.16
METEOR 0.4.3 (2005)	0.20	0.29
BLEU	0.20	0.27
TER	0.20	0.19
PER	0.20	0.18
CDER	0.12	0.16
WER	0.10	0.26

IMEANT
outperforms any of the others

- IMEANT produces much higher HAJ correlations than any of the other metrics on both GALE-A and GALE-B

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

	GALE-A	GALE-B
HMEANT	0.53	0.37
IMEANT	**0.51**	**0.33**
XMEANT	**0.51**	0.20
MEANT	0.48	**0.33**
METEOR 1.5 (2014)	0.43	0.10
NIST	0.29	0.16
METEOR 0.4.3 (2005)	0.20	0.29
BLEU	0.20	0.27
TER	0.20	0.19
PER	0.20	0.18
CDER	0.12	0.16
WER	0.10	0.26

IMEANT
even closes the gap with HMEANT

- IMEANT even comes within a few points of the human upper bound established by HMEANT

Table 1. Sent-level correlation with HAJ on GALE P2.5 data

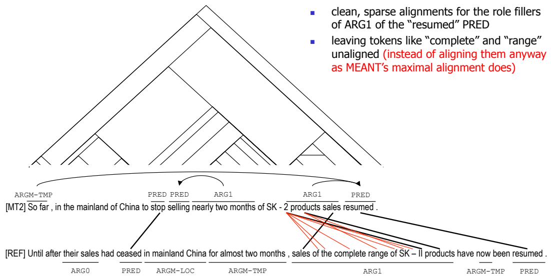
	GALE-A	GALE-B
HMEANT	0.53	0.37
IMEANT	**0.51**	**0.33**
XMEANT	**0.51**	0.20
MEANT	0.48	**0.33**
METEOR 1.5 (2014)	0.43	0.10
NIST	0.29	0.16
METEOR 0.4.3 (2005)	0.20	0.29
BLEU	0.20	0.27
TER	0.20	0.19
PER	0.20	0.18
CDER	0.12	0.16
WER	0.10	0.26

observation
how ITG constraints help IMEANT

- empirically, we see
  - ITGs produce significantly more accurate phrasal similarity aggregation
  - compared to MEANT's standard bag-of-words based heuristics
- **permutation** and **bijectivity** constraints enforced by the ITG
  - offer better leverage to reject inappropriate token alignments
  - compared to the maximal alignment approach which tends to be rather promiscuous

example
how ITG constraints help IMEANT

- clean, sparse alignments for the role fillers of ARG1 of the "resumed" PRED
- leaving tokens like "complete" and "range" unaligned (instead of aligning them anyway as MEANT's maximal alignment does)



semantic MT evaluation
the MEANT viewpoint

- **simple** Occam's razor: easy to define, easy to implement, easy to use
- **representationally transparent** can look at a score and understand scientifically why it was high or low
  - eg, MEANT's degree of match between semantic frames
  - who did what to whom, for whom, when, where, why and how
- **tunable** support fast scoring of massive numbers of hypotheses for tuning/training
- **discriminating** fine-grained scores (not just ranking or "good/bad" binary classification)
- **language independent** methodology that works across all language pairs
  - eg, IMEANT and XMEANT's incorporation of language universal ITG biases
- **stable** high HAJ correlations without retraining

lessons from IMEANT

- **IMEANT** – our newest 2014 version of MEANT is based on ITGs
- **achieves highest correlation** with HAJ among all variants of MEANT as well as other common MT evaluation metrics
- aligns and scores frames via a simple, consistent BITG which provides **informative permutation and bijectivity biases**
  - replaces MEANT's maximal alignment and bag-of-words heuristics
- retains MEANT's characteristics of **Occam's Razor style simplicity** and **representational transparency**

XMEANT *new!* a cross-lingual
semantic frame based MT evaluation metric

- **XMEANT** cross-lingual [Lo, Beloucif, Saers & Wu, ACL 2014]
  - eliminates the need for expensive reference translations …
    yet correlates with human adequacy judgment even more closely than MEANT!
  - aligns role fillers by leveraging language-independent BITGs constraints (Wu 1997; Zens & Ney 2003; Saers & Wu 2009)
  - a new generation of Wu & Fung's (NAACL, EAMT 2009) cross-lingual score …
    that exploits all our recent advances on monolingual MEANT

challenges
cross-lingual semantic frame based MT evaluation

- is it possible to improve HAJ correlation with structural semantics?
- is it possible to do so while avoid losing representational transparency?
- is it possible to have a fine-grained metric – not just "good/bad" binary classification?
- is it possible to preserve accuracy while supporting fast scoring of massive numbers of hypotheses for tuning/training?
  - (sophisticated high-dimensional classification is too costly)
- is it possible to do all this in a metric that works well across different languages without retraining?

monolingual

MEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

monolingual

MEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT}\\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF}\\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}\\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}\\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates}\\
w_j &\equiv \text{weight of similarity of ARG j}\\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i}\\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}\\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^0_{i,j}\right|}}{\sum_i w^0_i}\\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^1_{i,j}\right|}}{\sum_i w^1_i}\\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{aligned}
$$

monolingual

MEANT vs. cross-lingual XMEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w^1_i **and** $q^1_{i,j}$

monolingual

MEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT}\\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF}\\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}\\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}\\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates}\\
w_j &\equiv \text{weight of similarity of ARG j}\\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i}\\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}\\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^0_{i,j}\right|}}{\sum_i w^0_i}\\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^1_{i,j}\right|}}{\sum_i w^1_i}\\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{aligned}
$$

monolingual

MEANT vs. cross-lingual XMEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

IN

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT}\\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF}\\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}\\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}\\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates}\\
w_j &\equiv \text{weight of similarity of ARG j}\\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i}\\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}\\
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^0_{i,j}\right|}}{\sum_i w^0_i}\\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^1_{i,j}\right|}}{\sum_i w^1_i}\\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{aligned}
$$

monolingual

MEANT vs. cross-lingual XMEANT

1. apply automatic shallow semantic parsing to the **reference translation**, in the **output** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **reference translation** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **reference translation** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w^1_i **and** $q^1_{i,j}$

cross-lingual

XMEANT

IN

$$
\begin{aligned}
q^0_{i,j} &\equiv \text{ARG j of aligned frame i in MT}\\
q^1_{i,j} &\equiv \text{ARG j of aligned frame i in REF}\\
w^0_i &\equiv \frac{\text{\#tokens filled in aligned frame i of MT}}{\text{total \#tokens in MT}}\\
w^1_i &\equiv \frac{\text{\#tokens filled in aligned frame i of REF}}{\text{total \#tokens in REF}}\\
w_{\text{pred}} &\equiv \text{weight of similarity of predicates}\\
w_j &\equiv \text{weight of similarity of ARG j}\\
s_{i,\text{pred}} &\equiv \text{predicate similarity in aligned frame i}\\
s_{i,j} &\equiv \text{ARG j similarity in aligned frame i}
\end{aligned}
$$

???

$$
\begin{aligned}
\text{precision} &= \frac{\sum_i w^0_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^0_{i,j}\right|}}{\sum_i w^0_i}\\
\text{recall} &= \frac{\sum_i w^1_i \frac{w_{\text{pred}}\, s_{i,\text{pred}} + \sum_j w_j s_{i,j}}{w_{\text{pred}} + \sum_j w_j \left|q^1_{i,j}\right|}}{\sum_i w^1_i}\\
\text{MEANT} &= \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} \cdot \text{recall}}
\end{aligned}
$$

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w^1_i **and** $q^1_{i,j}$

cross-lingual

XMEANT

role filler similarity approach 1
apply MEANT's f-score approach within semantic role fillers as well

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i\\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i\\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i\\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i\\
p(e,f) &= \sqrt{t(e|f)\,t(f|e)}\\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|}\\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|}\\
s_{i,\text{pred}} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} \cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}\\
s_{i,j} &= \frac{2 \cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} \cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w^1_i **and** $q^1_{i,j}$

cross-lingual

XMEANT

role filler similarity approach 0
apply MEANT 2013's approach
(Mihalcea, Corley & Strapparava 2006)

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i\\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i\\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i\\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i\\
p(e,f) &= \sqrt{t(e|f)\,t(f|e)}\\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|}\\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|}\\
s_{i,\text{pred}} &= \cancel{\frac{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}} + \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{2}}\\
s_{i,j} &= \cancel{\frac{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}} + \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{2}}
\end{aligned}
$$

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w^1_i **and** $q^1_{i,j}$

cross-lingual
XMEANT

role filler similarity approach 1
apply MEANT's f-score approach within semantic role fillers as well

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i\\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i\\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i\\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i\\
p(e,f) &= \sqrt{t(e|f)\,t(f|e)}\\
\text{prec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{e\in\mathbf{e}} \max_{f\in\mathbf{f}} p(e,f)}{|\mathbf{e}|}\\
\text{rec}_{\mathbf{e},\mathbf{f}} &= \frac{\sum_{f\in\mathbf{f}} \max_{e\in\mathbf{e}} p(e,f)}{|\mathbf{f}|}\\
s_{i,\text{pred}} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}\cdot \text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}{\text{prec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}+\text{rec}_{\mathbf{e}_{i,\text{pred}},\mathbf{f}_{i,\text{pred}}}}\\
s_{i,j} &= \frac{2\cdot \text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}\cdot \text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}{\text{prec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}+\text{rec}_{\mathbf{e}_{i,j},\mathbf{f}_{i,j}}}
\end{aligned}
$$

1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w_i^1 **and** $q_{i,j}^1$

cross-lingual
XMEANT

role filler similarity approach 2
apply MEANT's ITG bias on reordering within semantic role fillers as well

$$
\begin{aligned}
\mathbf{e}_{i,\text{pred}} &\equiv \text{the output side of the pred of aligned frame } i\\
\mathbf{f}_{i,\text{pred}} &\equiv \text{the input side of the pred of aligned frame } i\\
\mathbf{e}_{i,j} &\equiv \text{the output side of the ARG } j \text{ of aligned frame } i\\
\mathbf{f}_{i,j} &\equiv \text{the input side of the ARG } j \text{ of aligned frame } i\\
G &\equiv \langle \{\mathrm{A}\}, \mathcal{W}^0, \mathcal{W}^1, \mathcal{R}, \mathrm{A}\rangle\\
\mathcal{R} &\equiv \{\mathrm{A}\rightarrow[\mathrm{AA}], \mathrm{A}\rightarrow\langle\mathrm{AA}\rangle, \mathrm{A}\rightarrow e/f\}\\
p([\mathrm{AA}]|\mathrm{A}) &= p(\langle\mathrm{AA}\rangle|\mathrm{A}) = 0.25\\
p(e/f|\mathrm{A}) &= \frac{1}{2}\sqrt{t(e|f)\,t(f|e)}\\
s_{i,\text{pred}} &= \frac{1}{1-\frac{\ln\left(P\left(\mathrm{A}\stackrel{*}{\Rightarrow}\mathbf{e}_{i,\text{pred}}/\mathbf{f}_{i,\text{pred}}|G\right)\right)}{\max(|\mathbf{e}_{i,\text{pred}}|,|\mathbf{f}_{i,\text{pred}}|)}}\\
s_{i,j} &= \frac{1}{1-\frac{\ln\left(P\left(\mathrm{A}\stackrel{*}{\Rightarrow}\mathbf{e}_{i,j}/\mathbf{f}_{i,j}|G\right)\right)}{\max(|\mathbf{e}_{i,j}|,|\mathbf{f}_{i,j}|)}}
\end{aligned}
$$

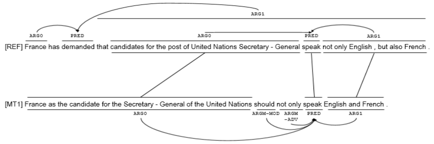
1. apply automatic shallow semantic parsing to the **foreign input sentence**, in the **input** language
2. apply automatic shallow semantic parsing to the machine translation, in the output language
3. apply maximum weighted bipartite matching to align the semantic frames between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic predicates
4. for each pair of aligned semantic frames, apply maximum weighted bipartite matching to align arguments between the **foreign input sentence** and the machine translation, according to the lexical similarity of the semantic role fillers
5. compute the weighted f-score over the matching role labels of these aligned predicates and role fillers **except replacing the reference translation with the foreign input when calculating** w_i^1 **and** $q_{i,j}^1$

in a nutshell
how XMEANT differs from MEANT

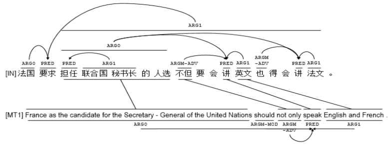
- rather than monolingual word vectors to score lexical similarities, instead substitute simple cross-lingual lexical translation probabilities
- try aggregating these cross-lingual lexical translation probabilities by comparing two natural ways to generalize MEANT's biases:
  - **approach 1** f-scores
  - **approach 2** bracketing ITGs constraints

XMEANT vs MEANT
[example 1]

MEANT

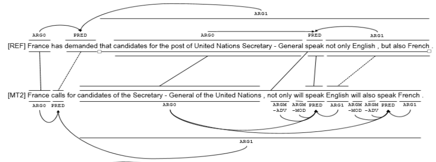


XMEANT

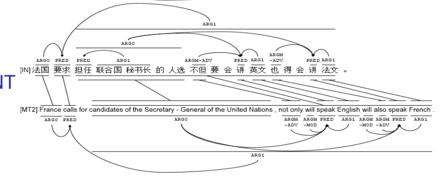


XMEANT vs MEANT
[example 2]

MEANT



XMEANT



comparative results

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

comparative results

- **setup**
  - English SRL: ASSERT
  - Chinese SRL: C-ASSERT

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

comparative results

- **new state-of-the-art** XMEANT correlates with human adequacy judgments more closely than other monolingual automatic MT metrics

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

comparative results

- **new state-of-the-art** XMEANT correlates with human adequacy judgments more closely than other monolingual automatic MT metrics (even MEANT 2013!)

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

comparative results

- **new state-of-the-art** XMEANT correlates with human adequacy judgments more closely than other monolingual automatic MT metrics (even MEANT 2013!)
- **f-score aggregation helps** the new f-score based method of aggregating lexical similarities between role fillers even improves monolingual MEANT

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

comparative results

- **new state-of-the-art** XMEANT correlates with human adequacy judgments more closely than other monolingual automatic MT metrics (even MEANT 2013!)
- **f-score aggregation helps** the new f-score based method of aggregating lexical similarities between role fillers even improves monolingual MEANT
- **ITG aggregation helps even more** lexical similarity between cross-lingual role fillers is more accurately estimated via bracketing ITGs than f-scores

sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

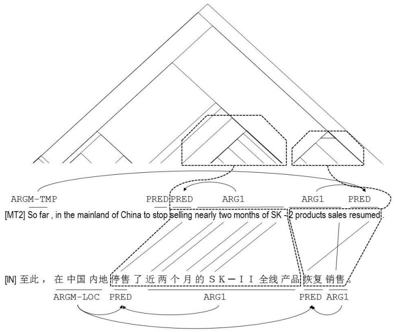
comparative results

- **new state-of-the-art** XMEANT correlates with human adequacy judgments more closely than other monolingual automatic MT metrics (even MEANT 2013!)
- **f-score aggregation helps** the new f-score based method of aggregating lexical similarities between role fillers even improves monolingual MEANT
- **ITG aggregation helps even more** lexical similarity between cross-lingual role fillers is more accurately estimated via bracketing ITGs than f-scores
- **closing the gap with humans** XMEANT is nearly as accurate as HMEANT!

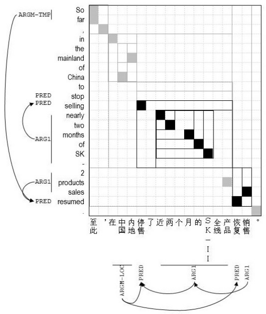
sentence-level correlations with HAJ (GALE phase 2.5 evaluation data)

Metric	*Kendall*
HMEANT	0.53
XMEANT (BITG)	**0.51**
MEANT (f-score)	0.48
XMEANT (f-score)	**0.46**
MEANT (2013)	0.46
NIST	0.29
BLEU/METEOR/TER/PER	0.20
CDER	0.12
WER	0.10

example
ITG based XMEANT



example
ITG based XMEANT



this makes a translation useful

this makes a translation useful

how well is

who did what to whom, for whom, when, where, why and how

preserved in translation?

Conclusion
MEANT

- The first **purely semantic MT metric**
  - cheap enough to tune SMT against

Conclusion
MEANT

- The first **tunable!** **purely semantic MT metric**
  - cheap enough to tune SMT against

Conclusion
MEANT

- The first tunable! **purely semantic MT metric**
  - cheap enough to tune SMT against
- Tuning MT against MEANT more robustly produces adequate translations than tuning against BLEU or TER
  - not only on formal genres like newswire
  - **but also on informal genres like TED lectures and web forums**

Conclusion
MEANT

- The first tunable! **purely semantic MT metric**
  - cheap enough to tune SMT against
- Tuning MT against MEANT more robustly produces adequate translations than tuning against BLEU or TER
  - not only on formal genres like newswire
  - **but also on informal genres like TED lectures and web forums**
- Latest work: further improvements to MEANT and MEANT-tuned systems
  - eg, problem of missing semantic frames for "be"
  - in top group of forthcoming WMT 2015 shared task for tuning metrics

Fully automatic
MEANT

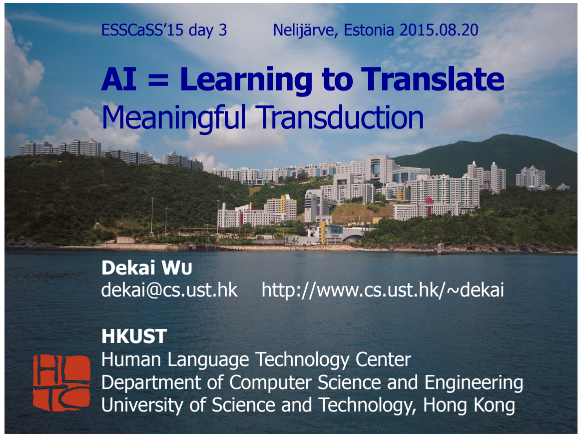
- First fully automatic semantic MT evaluation metric to succeed at correlating with HAJ better than all surface metrics
  - **replaces human SRL** with automatic shallow semantic parsing
  - **replaces human semantic frame alignment** with a simple maximum weighted bipartite matching algorithm based on the lexical similarity between semantic frames
- Preserves the spirit of HMEANT
  - **Occam's razor** simplicity
  - **representational transparency**
- Tunable! (ACL 2013, IWSLT 2013, WMT 2015)
  - **the most robust objective function** for tuning SMT

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations
4. **make amends for these errors**
   - design SMT models oriented toward learning the right abstractions
5. **learn to live a new life with a new code of behavior**
   - evaluate your MT models against semantically meaningful metrics

BLEUaholics Anonymous
Steps to recover from the hangover

1. **admit that one cannot control one's addiction or compulsion**
   - say "My name is _____ and I am a BLEUaholic."
2. **recognize a higher power that can give strength**
   - science: the wisdom to know the difference
3. **examine past errors with the help of an experienced member**
   - analyze if your MT model learns meaningful generalizations
4. **make amends for these errors**
   - design SMT models oriented toward learning the right abstractions
5. **learn to live a new life with a new code of behavior**
   - evaluate your MT models against semantically meaningful metrics
6. **help others who suffer from the same addictions or compulsions**



ESSCaSS'15 day 3 Nelijärve, Estonia 2015.08.20

AI = Learning to Translate
Meaningful Transduction

Dekai Wu
dekai@cs.ust.hk http://www.cs.ust.hk/~dekai

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