

Towards a Predicate-Argument Evaluation for MT

Ondřej Bojar^α, Dekai Wu^β

^α Charles University in Prague, ÚFAL

^β Hong Kong University of Science and Technology,
HKUST

bojar@ufal.mff.cuni.cz, de kai@cs.ust.hk

presented by Rudolf Rosa

Outline

- ▶ Motivating predicate-argument evaluation for MT.
 - ▶ HMEANT Annotation Procedure.
 - ▶ Relation to Tectogrammatical Annotation.
- ▶ Experiment with Czech.
- ▶ Issues Encountered.
- ▶ Summary.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
- ▶ **Ranking of full sentences.**
- ▶ **Ranking of constituents**, i.e. parts of sentences.
- ▶ **Comprehension test:** Blind editing+validation.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
- ▶ **HTER:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
- ▶ **Ranking of constituents**, i.e. parts of sentences.
- ▶ **Comprehension test:** Blind editing+validation.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
- ▶ **HTER:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
Longer sentences hard to rank. Candidates incomparably poor.
- ▶ **Ranking of constituents**, i.e. parts of sentences.
- ▶ **Comprehension test:** Blind editing+validation.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
- ▶ **HTER:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
Longer sentences hard to rank. Candidates incomparably poor.
- ▶ **Ranking of constituents**, i.e. parts of sentences.
Does not evaluate overall coherence.
- ▶ **Comprehension test:** Blind editing+validation.

- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?

- ▶ **HTER:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
Longer sentences hard to rank. Candidates incomparably poor.
- ▶ **Ranking of constituents**, i.e. parts of sentences.
Does not evaluate overall coherence.
- ▶ **Comprehension test:** Blind editing+validation.
Expensive.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
- ▶ **HTER:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
Longer sentences hard to rank. Candidates incomparably poor.
- ▶ **Ranking of constituents**, i.e. parts of sentences.
Does not evaluate overall coherence.
- ▶ **Comprehension test:** Blind editing+validation.
Expensive.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
Preparation expensive. Feels too narrow.
- ▶ **HTEr:** Post-editing effort.

Methods of Manual MT Evaluation

- ▶ Absolute **adequacy** and **fluency** of whole sentences.
Measures correlated. Low agreement.
- ▶ **Ranking of full sentences.**
Longer sentences hard to rank. Candidates incomparably poor.
- ▶ **Ranking of constituents**, i.e. parts of sentences.
Does not evaluate overall coherence.
- ▶ **Comprehension test:** Blind editing+validation.
Expensive.
- ▶ **Task-based:** MT output as useful as the original?
Do I dress appropriately given a translated weather forecast?
Preparation expensive. Feels too narrow.
- ▶ **HTEr:** Post-editing effort.
Expensive. Requires trained people.

HMEANT (Lo and Wu, 2011a)

- ▶ Improved evaluation of adequacy compared to BLEU.
- ▶ Reduced human labour of HTER (Snover et al., 2006).

Essence: Is the basic event structure understandable?
Who did what to whom, when, where and why.

Procedure:

1. SRL: Identify semantic frames and roles in ref & hyp.
2. Align frames and their role fillers.
3. Calculate prec & rec across all frames in the sentence.

HMEANT Illustration: Motivation

It is hard to rank A vs. B (even if we know R is the ref.)



Finally, he stood in the center
of the referee Wolfgang Stark.



At the end of the day, was at the centre
of a referee Wolfgang Stark.



The referee Wolfgang Stark then garnered
some attention.

HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.



Finally, he stood in the center
of the referee Wolfgang Stark.

HMEANT Illustration: SRL


It is easier to mark roles of a single hypothesis.



Finally, he **stood** in the center
Action of the referee Wolfgang Stark.

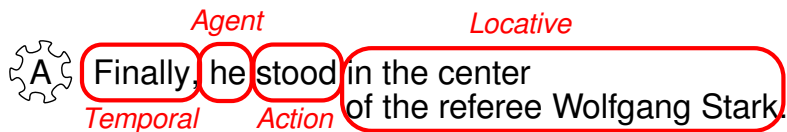
HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.

 Finally, *Agent* **he** *Action* **stood** in the center
of the referee Wolfgang Stark.

HMEANT Illustration: SRL

It is easier to mark roles of a single hypothesis.



HMEANT Illustration: SRL

The same SRL is performed on the reference.



The referee Wolfgang Stark then garnered
some attention.

HMEANT Illustration: SRL

The same SRL is performed on the reference.



The referee Wolfgang Stark then *Action* garnered some attention.

HMEANT Illustration: SRL

The same SRL is performed on the reference.



HMEANT Illustration: Alignment

And finally, frames and role fillers are aligned.



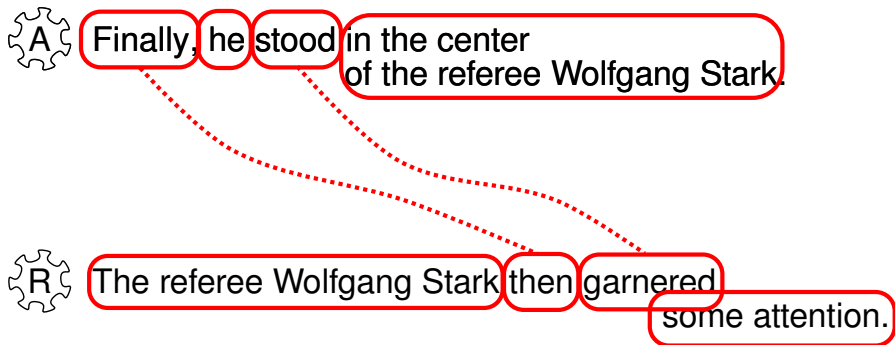
Finally, he stood in the center
of the referee Wolfgang Stark.



The referee Wolfgang Stark then garnered
some attention.

HMEANT Illustration: Alignment

And finally, frames and role fillers are aligned.



HMEANT Illustration

And finally, frames and role fillers are aligned.



Finally, he stood in the center of the referee Wolfgang Stark.

Obviously, the meaning was rather distorted.



The referee Wolfgang Stark then garnered some attention.

HMEANT Illustration

And finally, frames and role fillers are aligned.



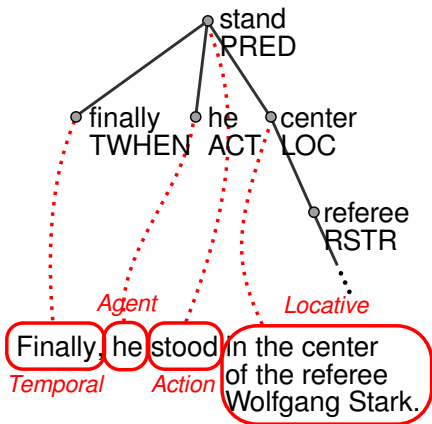
Finally, he stood in the center
of the referee Wolfgang Stark.

...but the annotation was more principled
and we know which parts are wrong.



The referee Wolfgang Stark then garnered
some attention.

Future: Utilize T-Layer of PDT



In terms of pred-arg. formalisms like the tectogrammatical layer of the Prague Dependency Treebank (PDT):

- ▶ HMEANT just checks the match of subtrees under verbs.
- ▶ Tools for English and Czech available to get such trees automatically.
 - ⇒ We could e.g. highlight all words of a subtree at once.

English→Czech Experiment

- ▶ 50 distinct sentences from WMT12 test set.
 - ▶ Selected to have a high overlap with WMT12 manual rankings for future analysis.
- ▶ 13 systems translating from English to Czech.
 - + One reference translation.
- ▶ 14 annotators
 - ▶ No sentence displayed twice to the same person.
- ▶ Unfortunately no overlap in annotation
 - ⇒ No agreement judgments.

Sentence-Level Correlation

HMEANT	0.2833
METEOR	0.2167
WER	0.1708
CDER	0.1375
NIST	0.1167
TER	0.1167
PER	0.0208
BLEU	0.0125

Kendall's τ for
sentence-level correlation
with human rankings.

- ▶ Better correlation than automatic metrics (expected).
- ▶ Overall quite low. Possible reasons:
 - ▶ Evaluated 13 systems.
 - ▶ Gold standard ranks overall quality, not just adequacy as Lo and Wu (2011b) who achieve 0.49.
 - ▶ HMEANT problems discovered by our experiment, see below.
 - ▶ Gold standard disputable.

Gold Standard

Interpretation Sentences	Ties Ignored
	50
cu-depfix	72.5
onlineB	61.4
uedin-wmt12	60.3
cu-tamch-boj	54.6
cu-bojar_2012	53.2
CU_TectoMT	54.9
onlineA	61.4
pctrans2010	54.1
commercial2	51.3
cu-poor-comb	41.6
uk-dan-moses	33.2
SFU	31.0
jhu-hiero	26.7

Somewhat Shaky Gold Standard

Interpretation Sentences	Ties Ignored	
	All	50
cu-depfix	66.4	72.5
onlineB	63.0	61.4
uedin-wmt12	55.8	60.3
cu-tamch-boj	55.6	54.6
cu-bojar_2012	54.3	53.2
CU_TectoMT	53.1	54.9
onlineA	52.9	61.4
pctrans2010	47.7	54.1
commercial2	46.0	51.3
cu-poor-comb	44.1	41.6
uk-dan-moses	43.5	33.2
SFU	36.1	31.0
jhu-hiero	32.2	26.7

Somewhat Shaky Gold Standard

Interpretation Sentences	Ties Ignored		\geq Others		$>$ Others	
	All	50	All	50	All	50
cu-depfix	66.4	72.5	73.0	77.5	53.3	59.4
onlineB	63.0	61.4	70.5	69.3	50.3	49.0
uedin-wmt12	55.8	60.3	63.6	66.3	46.0	51.1
cu-tamch-boj	55.6	54.6	64.7	62.1	44.2	45.7
cu-bojar_2012	54.3	53.2	64.1	62.2	42.6	43.0
CU_TectoMT	53.1	54.9	60.5	59.8	44.6	49.0
onlineA	52.9	61.4	60.8	66.7	44.0	53.0
pctrans2010	47.7	54.1	55.1	60.1	40.9	47.1
commercial2	46.0	51.3	54.6	59.5	38.7	42.7
cu-poor-comb	44.1	41.6	54.7	50.5	35.7	35.2
uk-dan-moses	43.5	33.2	53.4	44.2	35.9	27.7
SFU	36.1	31.0	46.8	43.0	30.0	25.6
jhu-hiero	32.2	26.7	43.2	36.0	27.0	23.3

Problems of HMEANT Annotation

- ▶ Vague SRL Guidelines:
 - ▶ Complex predicates.
 - ▶ PP-attachment.
 - ▶ Unclear or insufficient role labels.
 - ▶ Co-reference.
- ▶ Problems in the Alignment Phase:
 - ▶ Correctness of the Predicate.
 - ▶ Need for M:N Frame and Slot Alignment.

Complex Predicates

HMEANT tool requires exactly 1 word to serve as Action.

- ▶ Modals have a separate “role” label.

In Czech:

- ▶ It is the modal that conjugates \Rightarrow disputable.

Czech (made up)	<u>Představení</u>	<u>musí</u>	<u>pokračovat.</u>
Gloss	Show	must	go on.
English-Like Labels	Agent	Modal	Action
Natural for Czech	Agent	Action	Action
Forced to 1 Word	Agent	Action	Experiencer

- ▶ Copula “to be” is frequent.

Czech (made up)	<u>Řidič</u>	<u>byl</u>	<u>unaven.</u>
Gloss	The driver	was	tired.
English-Like Labels	Agent	Action	Experiencer?
Natural for Czech	Agent	Action	Action

\Rightarrow We suggest allowing more words to denote an Action.

Prepositional Phrase (PP) Attachment

Reference	Oblečky	musíme	vystříhat	z časopisů
Gloss	clothes	we-must	cut	from magazines
Roles	Experiencer	Modal	Action	Locative
Meaning	We must cut the clothes (paper toys) from magazines			
Hypothesis	Musíme	vyříznout	oblečení	z časopisů
Gloss	We-must	cut	clothes	from magazines
Roles	Modal	Action	Experiencer	

- ▶ The PP “from magazines” in the hypothesis can be annotated as:
 - ▶ a separate Locative.
 - ▶ or a part of Experiencer.(Sometimes the separate annotation is forced by word order.)

⇒ Impossible to align 2 to 1 role fillers.

⇒ Translation quality underestimated.

Unclear or Insufficient Role Labels

- ▶ HMEANT requires role labels to match to give credit.
- ▶ HMEANT set of labels is sufficiently simple:
 - ▶ So the disagreement is hopefully kept low.
 - ▶ Sometimes still hard to use, e.g. in passive constructions.
⇒ Disagreement ⇒ Translation quality underestimated.
- ▶ On the other hand, the set feels too small in some cases:

Czech	Byl převezen	do nemocnice	ve vrtulníku.
Gloss	He was transported	to the hospital	in a helicopter.
SRL	Action	Locative	Locative

 - ▶ One of our annots. actually joined the two Locatives into one.
⇒ 2:1 alignment problem.

⇒ We suggest experimenting with no role labels altogether.

Co-reference

Consider annotation of the frame of “wins”:

English (made up)	<u>It</u>	is	<u>the man</u>	<u>who</u>	<u>wins.</u>
Roles	Agent?		Agent?	Agent?	Action

- ▶ Three candidates for the Agent.
- ▶ In Czech, some can be pro-dropped.
 - ⇒ Risk of no Agent annotated at all.
 - ⇒ 0:1-alignment problem.
 - ⇒ Translation quality underestimated.

⇒ We suggest giving more examples in the guidelines.

Correctness of the Predicate

Reference	Opilý řidič	těžce	zraněn
Gloss	A drunken driver	seriously	injured (passive form)
Roles	Agent	Extent	Action
Meaning	A drunken driver is seriously injured.		

Hypothesis	Opilý řidič	vážně	zranil
Gloss	A drunken driver	seriously	injured (active form)
Roles	Agent	Extent	Action
Meaning	A drunken driver seriously injured (someone).		

- ▶ All role fillers match exactly.
- ▶ The Action's form reverses the meaning.
- ▶ Current HMEANT does not allow to mark Action as mistranslated.

⇒ We suggest judging the quality of predicate match as well.

Need for M:N Frame and Slot Alignments

HMEANT aligns first frames and then slots within them.

- ▶ But the frames do not always match 1-1, e.g. due to:
 - ▶ inconsistent annotation of modals, phasic verbs (“to begin”)
 - ▶ or simply not quite literal but correct translation.

⇒ Cannot align fillers across frames.

⇒ Translation quality underestimated.

PP-attachment ambiguity:

- ▶ Happens in the SRL phase.
- ▶ Causes a 2:1 problem in the alignment phase.

⇒ Translation quality underestimated.

⇒ We suggest allowing M:N ali. for both frames and fillers.

Summary

We applied HMEANT to Czech.

- ▶ Overall positive experience.
 - ▶ Annotators know what they are doing when, where and why.
- ▶ Multiple issues identified:
 - ▶ Some can be solved by more examples to current guidelines.
 - ▶ Some require an update of the interface.
 - ▶ Multiple (non-adjacent) words forming the Action.
 - ▶ Indication of the correctness of the predicate.
 - ▶ Some need changes to prec/rec formulas.
 - ▶ M:N alignments of predicates and slots.

Future: Use t-layer tools to:

- ▶ Speed up SRL (highlight more words at once).
- ▶ Fully automate HMEANT \rightsquigarrow MEANT.

References

Chi-kiu Lo and Dekai Wu. 2011a. Meant: An inexpensive, high-accuracy, semi-automatic metric for evaluating translation utility based on semantic roles. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 220–229, Portland, Oregon, USA, June. Association for Computational Linguistics.

Chi-kiu Lo and Dekai Wu. 2011b. Structured vs. flat semantic role representations for machine translation evaluation. In Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation, SSST-5, pages 10–20, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A Study of Translation Edit Rate with Targeted Human Annotation. In Proceedings AMTA, pages 223–231, August.