Towards Heterogeneous Automatic MT Error Analysis

(6th LREC)

Jesús Giménez and Lluís Màrquez

—

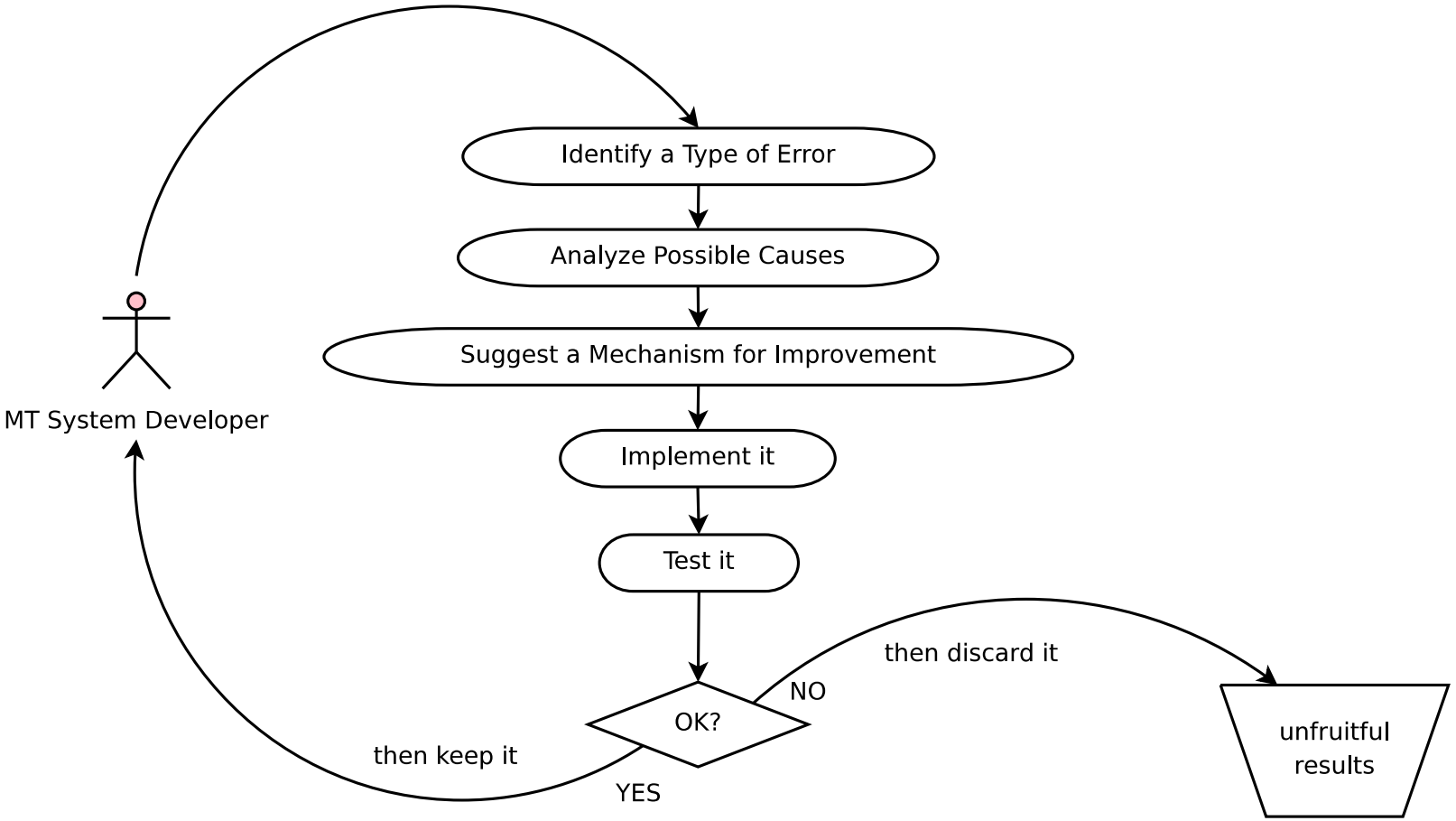
TALP Research Center
Technical University of Catalonia

May 29, 2008

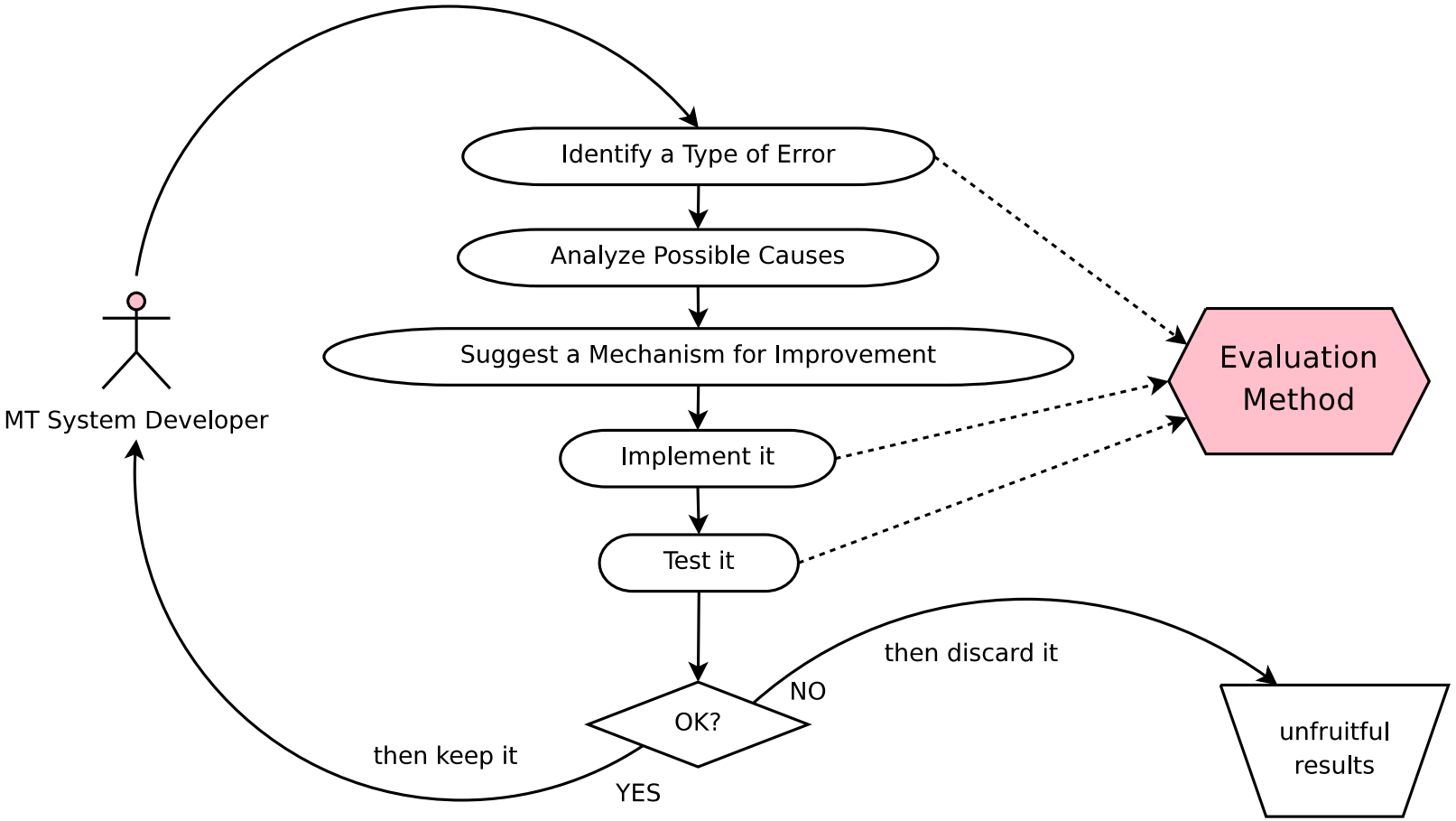
Outline

Outline

Development Cycle of MT systems



Development Cycle of MT systems



Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations

Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations

Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations

Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations

Error Analysis Today

- Error analyses are conducted manually
  - **low-level** analysis related to the linguistic analysis of translation quality (i.e., what?)
  - **high-level** analysis involving knowledge about the system architecture (i.e., why?)
- Error analyses require intensive human labor
- Automatic metrics are used only as quantitative evaluation measures
  - to identify high/low quality translations

Metrics Based on Lexical Similarity

- **Edit Distance**
  WER, PER, TER
- **Precision**
  BLEU, NIST, WNM
- **Recall**
  ROUGE, CDER
- **Precision/Recall**
  GTM, METEOR, BLANC, SIA

Outline

Extending the Reference Lexicon

- Lexical variants
  - Morphological variations (i.e., stemming) $\rightarrow$ ROUGE and METEOR
  - Synonymy lookup $\rightarrow$ METEOR (based on WordNet)
- Paraphrasing support
  - Zhou et al. [ZLH06]
  - Kauchak and Barzilay [KB06]
  - Owczarzak et al. [OGGW06]

Beyond the Lexical Level

Syntactic Similarity

- Shallow Parsing
  - Popovic and Ney [PN07]
  - Giménez and Màrquez [GM07]
- Constituency Parsing
  - Liu and Gildea [LG05]
- Dependency Parsing
  - Liu and Gildea[LG05]
  - Amigó et al. [AGGM06]
  - Mehay and Brew [MB07]
  - Owczarzak et al. [OvGW07]

Beyond the Lexical Level

Semantic Similarity

- Semantic Roles
  - Giménez and Màrquez [GM07]
- Named Entities
  - Reeder et al. [RMDW01]
  - Giménez and Màrquez [GM07]
- Discourse Representations
  - Giménez and Màrquez [GM08b]

Outline

Rely on Automatic Metrics

Idea: Let automatic metrics do most of the *low-level* analysis, so system developers may concentrate on *high-level* analysis.

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects:** lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects:** lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects:** lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)

Heterogeneous Error Analysis

- as automatic as possible
- as heterogeneous as possible
  - **Quality Aspects:** lexical, syntactic, semantic, etc.
  - **Granularity**
    - fine aspects → transfer of specific linguistic elements (e.g., what proportion of singular nouns are correctly translated?)
    - coarse aspects → overall linguistic structure (e.g., what proportion of the semantic role structure is correctly translated?)

Outline

Linguistic Similarities

- More than *500* metric variants operating at different *linguistic* levels:
  - Lexical
  - Shallow Syntactic (Lemmatization, PoS Tagging, and Base Phrase Chunking)
  - Syntactic (Constituency and Dependency Parsing)
  - Shallow Semantic (Semantic Roles and Named Entities)
  - Semantic (Discourse Representations)

Shallow Syntactic Level

SP-O$_p$-$\star$ Average overlapping between words belonging to the same PoS.

SP-O$_c$-$\star$ Average overlapping between words belonging to the same phrase chunk type.

SP-NIST$_l$ NIST score over sequences of lemmas.

SP-NIST$_p$ NIST score over PoS sequences.

SP-NIST$_{iob}$ NIST score over chunk IOB sequences.

SP-NIST$_c$ NIST score over sequences of chunks.

Syntactic Level (i)

- **Dependency Overlapping**
  - **DP-O$_l$-$\star$** Average overlapping between words hanging at the same level.
  - **DP-O$_c$-$\star$** Average overlapping between words hanging from terminal nodes (i.e., grammatical categories).
  - **DP-O$_r$-$\star$** Average overlapping between words ruled by non-terminal nodes (i.e., grammatical relations).

Syntactic Level (ii)

- **Head-word Chain Matching** (Liu and Gildea [LG05])

  DP-HWC$_w$ Average head-word chain matching up to length-4 word chains.

  DP-HWC$_c$ Average head-word chain matching up to length-4 category chains.

  DP-HWC$_r$ Average head-word chain matching up to length-4 relation chains.

Syntactic Level (iii)

- **Syntactic Overlapping**
  - **CP-O_p-$\star$** Average overlapping between words belonging to the same PoS (similar to '*SP-O_p-$\star$*').
  - **CP-O_c-$\star$** Average overlapping between words belonging to the same phrase type (similar to '*SP-O_c-$\star$*').
- **Syntactic Tree Matching** (Liu and Gildea [LG05])
  - **CP-STM** Constituent tree matching averaged up to length-9 syntactic subpaths.

Shallow Semantic Level (i)

- **Named Entity Overlapping/Matching**
  - **NE-O_e-$\star$** Average lexical overlapping between named entities of the same type (excluding type 'O', i.e., Not-a-NE).
  - **NE-O_e-$\star\star$** Average lexical overlapping between named entities of the same type (including 'O').
  - **NE-M_e-$\star$** Average lexical matching between named entities of the same type.

Shallow Semantic Level (ii)

- **Semantic Role Overlapping/Matching**
  - **SR-O_r-$\star$** Average lexical overlapping between semantic roles (arguments and adjuncts) of the same type.
  - **SR-M_r-$\star$** Average lexical matching between semantic roles of the same type.
  - **SR-O_r** Role overlapping independently from the lexical realization.

Semantic Level

- **Discourse Overlapping**
  - **DR-O_r-$\star$** Average lexical overlapping between DR structures of the same type.
  - **DR-O_{rp}-$\star$** Average morphosyntactic overlapping between DR structures of the same type.
- **Semantic Tree Matching**
  - **DR-STM** Matching between discourse representations averaged up to length-9 semantic subpaths.

Linguistic Features at Work

ACL'07 MT Workshop (French/German/Spanish/Czech-to-English)

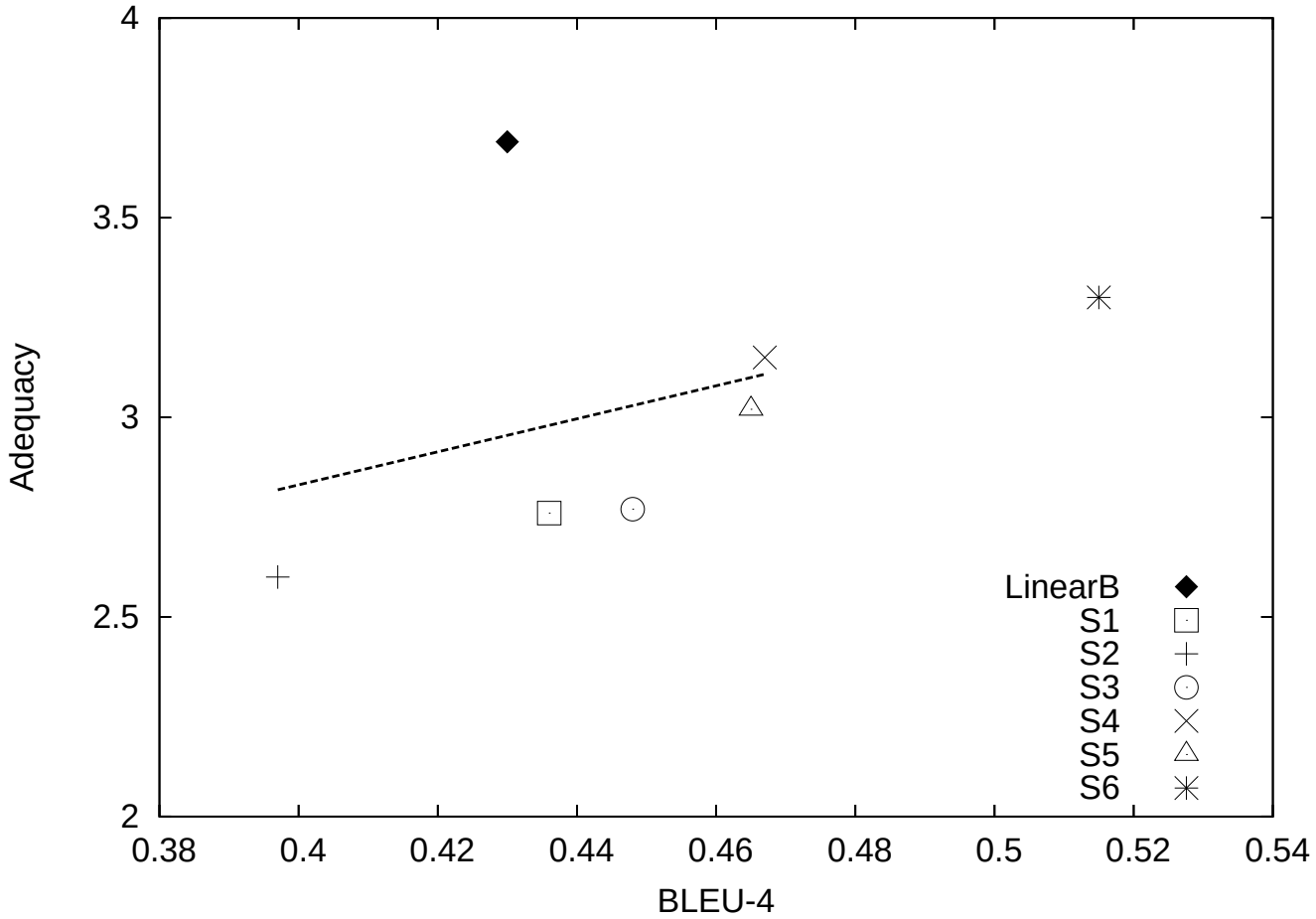
Metric	**Adeq.**	**Fluen.**	**Rank**	**Const.**	**all**
SR-O$_r$-$\star$	**.774**	**.839**	**.803**	.741	**.789**
ParaEval-Recall	.712	.742	.768	**.798**	.755
METEOR	.701	.719	.745	.669	.709
BLEU	.690	.722	.672	.602	.671
1-TER	.607	.538	.520	.514	.644
Max Adeq. Corr.	.651	.657	.659	.534	.626
Max Fluen. Corr.	.644	.653	.656	.512	.616
GTM	.655	.674	.616	.495	.610

Outline

Outline

NIST 2005 MT Evaluation Puzzle

- Arabic-to-English Translation Exercise [LP05]



Linguistic Features Solved the Puzzle

- Giménez and Màrquez [GM07]

Feature	Metric	R_{sys}
Lexical	BLEU	0.06
	GTM	0.03
Syntactic	SP-NIST$_p$	0.42
	DP-HWC$_r$	**0.88**
	CP-STM	0.74
Semantic	SR-O_r-$\star$	0.61
	SR-M_r-$\star$	0.72
	DR-O_r-$\star$	0.92
	DR-O_{rp}-$\star$	**0.97**

Linguistic Features Solved the Puzzle

- Giménez and Màrquez [GM07]

Feature	Metric	R_{sys}
Lexical	BLEU	0.06
	GTM	0.03
Syntactic	SP-NIST$_p$	0.42
	DP-HWC$_r$	**0.88**
	CP-STM	0.74
Semantic	SR-O_r-$\star$	0.61
	SR-M_r-$\star$	0.72
	DR-O_r-$\star$	0.92
	DR-O_{rp}-$\star$	**0.97**

Outline

A Note on Meta-Evaluation

- Metrics are automatically evaluated on the basis of **human likeness**, i.e., in terms of their ability to distinguish manual from automatic translations.
  - ORANGE, Lin and Och [LO04]
  - KING, Amigó et al. [AGPV05]
- We use the KING measure
  - *"A metric should never rank* any *reference translation lower in quality than* any *automatic translation."*
- KING(x) serves as an estimate of the impact on system performance of the quality aspects captured by metric x

Lexical Features

Feature	Metric	KING	LinearB	Best SMT
Edit Distance	1-PER	0.63	0.65	**0.70**
	1-TER	**0.70**	0.53	**0.58**
	1-WER	0.67	0.49	**0.54**
Precision	BLEU	0.65	0.47	**0.51**
	NIST	0.69	10.63	**11.27**
Recall	ROUGE_W	0.68	0.31	**0.33**
F-measure	GTM ($e = 1$)	0.64	0.80	**0.85**
	GTM ($e = 2$)	0.66	0.31	**0.32**
	METEOR_{exact}	0.68	0.60	**0.64**
	METEOR_{wnsyn}	0.68	0.64	**0.68**

Shallow Syntactic Features

Feature	Metric	KING	LinearB	Best SMT
PoS Overlapping	SP-O_p-$\star$	0.64	0.52	**0.55**
	SP-O_p-J	0.26	0.53	**0.59**
	SP-O_p-N	0.53	0.57	**0.63**
	SP-O_p-V	0.43	0.39	**0.41**
Chunk Overlapping	SP-O_c-$\star$	0.63	0.54	**0.57**
	SP-O_c-NP	0.60	0.59	**0.63**
	SP-O_c-PP	0.38	0.63	**0.66**
	SP-O_c-VP	0.41	0.49	**0.51**
$NIST_x$	SP-$NIST_l$-5	0.69	10.78	**11.44**
	SP-$NIST_p$-5	**0.71**	8.74	**9.04**
	SP-$NIST_{iob}$-5	0.65	6.81	**6.91**
	SP-$NIST_c$-5	0.57	6.13	**6.18**

Syntactic Features (i)

Feature	Metric	KING	LinearB	Best SMT
Dependency Parsing	DP-HWC$_w$-4	0.59	0.14	0.14
	DP-HWC$_c$-4	0.48	**0.42**	0.41
	DP-HWC$_r$-4	0.52	**0.33**	0.31
	DP-O$_l$-$\star$	0.58	0.41	**0.43**
	DP-O$_c$-$\star$	0.60	0.50	**0.51**
	DP-O$_c$-a	0.30	0.51	**0.57**
	DP-O$_c$-aux	0.14	**0.56**	0.54
	DP-O$_c$-det	0.35	**0.75**	0.73
	DP-O$_c$-n	0.57	0.57	**0.59**
	DP-O$_c$-v	0.37	0.43	**0.45**

Syntactic Features (ii)

Feature	Metric	KING	LinearB	Best SMT
Dependency Parsing	DP-O_r-$\star$	**0.66**	0.36	0.36
	DP-O_r-aux	0.14	**0.56**	0.54
	DP-O_r-det	0.35	**0.75**	0.73
	DP-O_r-fc	0.21	**0.26**	0.24
	DP-O_r-i	**0.60**	**0.44**	0.43
	DP-O_r-obj	0.43	**0.36**	0.35
	DP-O_r-s	0.47	**0.52**	0,45
Constituency Parsing	CP-O_p-$\star$	**0.64**	0.52	**0.55**
	CP-O_c-$\star$	0.63	0.50	**0.53**
	CP-O_c-NP	0.61	0.55	**0.58**
	CP-O_c-PP	0.51	0.50	**0.53**
	CP-O_c-SBAR	0.36	0.36	**0.38**
	CP-STM-9	0.58	0.35	0.35

Shallow Semantic Features

Feature	Metric	KING	LinearB	Best SMT
Named Entities	NE-M_e-$\star$	0.32	0.53	**0.56**
	NE-M_e-ORG	0.11	0.27	**0.29**
	NE-M_e-PER	0.13	0.34	0.34
Semantic Roles	SR-M_r-$\star$	0.50	**0.19**	0.18
	SR-M_r-A0	0.33	**0.31**	0.30
	SR-M_r-A1	0.28	0.14	0.14
	SR-O_r	0.41	**0.64**	0.63
	SR-O_r-$\star$	**0.53**	0.36	**0.37**
	SR-O_r-AM-TMP	0.13	**0.39**	0.38

Semantic Features

Feature	Metric	KING	LinearB	Best SMT
Discourse Representations	DR-O$_r$-$\star$	**0.59**	**0.36**	0.34
	DR-O$_r$-card	0.12	**0.49**	0.45
	DR-O$_r$-dr	0.56	**0.43**	0.40
	DR-O$_r$-eq	0.12	**0.17**	0.16
	DR-O$_r$-named	0.38	**0.48**	0.45
	DR-O$_r$-pred	0.55	**0.38**	0.36
	DR-O$_r$-prop	0.39	**0.27**	0.24
	DR-O$_r$-rel	0.56	**0.38**	0.34
	DR-STM-9	0.40	0.26	0.26

Outline

Ex: Thousand Monks

Ref 1:	Over 1000 monks and nuns , observers and scientists from over 30 countries and the host country attended the religious summit held for the first time in Myanmar which started today , Thursday .
2:	More than 1000 monks , nuns , observers and scholars from more than 30 countries , including the host country , participated in the religious summit which Myanmar hosted for the first time and which began on Thursday .
3:	The religious summit , staged by Myanmar for the first time and began on Thursday , was attended by over 1,000 monks an nuns , observers and scholars from more than 30 countries and host Myanmar .
4:	More than 1,000 monks , nuns , observers and scholars from more than 30 countries and the host country Myanmar participated in the religious summit , which is hosted by Myanmar for the first time and which began on Thursday .
5:	The religious summit , which started on Thursday and was hosted for the first time by Myanmar , was attended by over 1,000 monks and nuns , observers and scholars from more than 30 countries and the host country Myanmar .

Ex: Thousand Monks

Info:	**(1)** → **subject:** over/more_than 1,000 monks and nuns, observers and scientists/scholars from over/more_than 30 countries , and/including the host country **action:** attended/participated_in **(2)** → **subject:** the religious summit **action:** began/started **temporal:** on Thursday **(3)** → **object:** the religious summit **action:** hosted **subject:** by Myanmar **mode:** for the first time
LinearB:	1000 monks from more than 30 States and the host State Myanmar attended the Summit , which began on Thursday , hosted by Myanmar for the first time .
Best SMT:	Religious participated in the summit , hosted by Myanmar for the first time began on Thursday , as an observer and the world of the 1000 monk nun from more than 30 countries and the host state Myanmar .

Ex: Thousand Monks - Lexical Features

Feature	Metric	LinearB	Best SMT
Human	Adequacy	**3**	2
	Fluency	**3.5**	2
Edit Distance	1-PER	0.64	**0.69**
	1-TER	**0.53**	0.51
	1-WER	0.40	**0.48**
Precision	BLEU	0.44	**0.45**
	NIST	9.04	**9.96**
Recall	ROUGE_W	0.22	**0.23**
F-measure	GTM ($e = 2$)	0.30	**0.32**
	METEOR_{wnsyn}	0.59	**0.64**

Ex: Thousand Monks - Shallow Syntactic Features

Feature	Metric	LinearB	Best SMT
PoS Overlapping	SP-O_p-$\star$	**0.52**	0.51
	SP-O_p-IN	**0.71**	0.67
	SP-O_p-NN	**0.67**	0.38
	SP-O_p-NNP	0.60	**0.75**
	SP-O_p-V	0.40	**0.75**
Chunk Overlapping	SP-O_c-$\star$	0.56	**0.60**
	SP-O_c-NP	0.56	**0.60**
	SP-O_c-PP	**0.80**	0.71
NIST$_x$	SP-NIST$_p$	6.21	**8.36**
	SP-NIST$_c$	**6.43**	6.25
	SP-NIST$_{iob}$	5.78	**6.41**

Ex: Thousand Monks - Syntactic Features

Feature	Metric	LinearB	Best SMT
Dependency Parsing	DP-HWC$_w$-4	**0.17**	0.16
	DP-O_r-$\star$	**0.46**	0.44
	DP-O_r-mod	**0.62**	0.41
	DP-O_r-obj	**0.29**	0.00
	DP-O_r-pcomp-n	**0.71**	0.39
	DP-O_r-rel	**0.33**	0.00
Constituency Parsing	CP-O_c-$\star$	**0.59**	0.48
	CP-O_c-NP	**0.59**	0.55
	CP-O_c-PP	**0.57**	0.54
	CP-O_c-SB	**0.73**	0.00
	CP-O_c-VP	**0.64**	0.42
	CP-STM-9	**0.34**	0.23

Ex: Thousand Monks - Semantic Features

Feature	Metric	LinearB	Best SMT
Semantic Roles	SR-O_r	**0.84**	0.25
	SR-O_r-$\star$	**0.56**	0.18
	SR-O_r-A0	**0.44**	0.10
	SR-O_r-A1	**0.57**	0.28
Discourse Representations	DR-O_r-$\star$	**0.45**	0.34
	DR-O_r-dr	**0.57**	0.40
	DR-O_r-nam	**0.75**	0.24
	DR-O_r-pred	0.44	**0.45**
	DR-O_r-rel	**0.51**	0.32
	DR-STM-9	**0.32**	0.29

Outline

Outline

Heterogeneous Automatic MT Error Analysis

- We have presented a valid path towards *heterogeneous automatic MT error analysis*:
  - Our approach allows developers to rapidly obtain detailed qualitative linguistic reports on their system's capabilities.
  - Human efforts may concentrate on high-level analysis.

Hey! Linguistic Metrics are Not the Panacea[1]

- Linguistic metrics rely on:
  1. the representativity of the set of human references
     - lexicon
     - grammar
     - style...
  2. automatic linguistic processors are
     - domain-dependent
     - language-dependent
     - prone to error
     - slow

Sentence level performance must be improved!

[1] Panacea: a remedy for all ills or difficulties (see cure-all).

Hey! Linguistic Metrics are Not the Panacea[1]

- Linguistic metrics rely on:
  1. the representativity of the set of human references
     - lexicon
     - grammar
     - style...
  2. automatic linguistic processors are
     - domain-dependent
     - language-dependent
     - prone to error
     - slow

Sentence level performance must be improved!

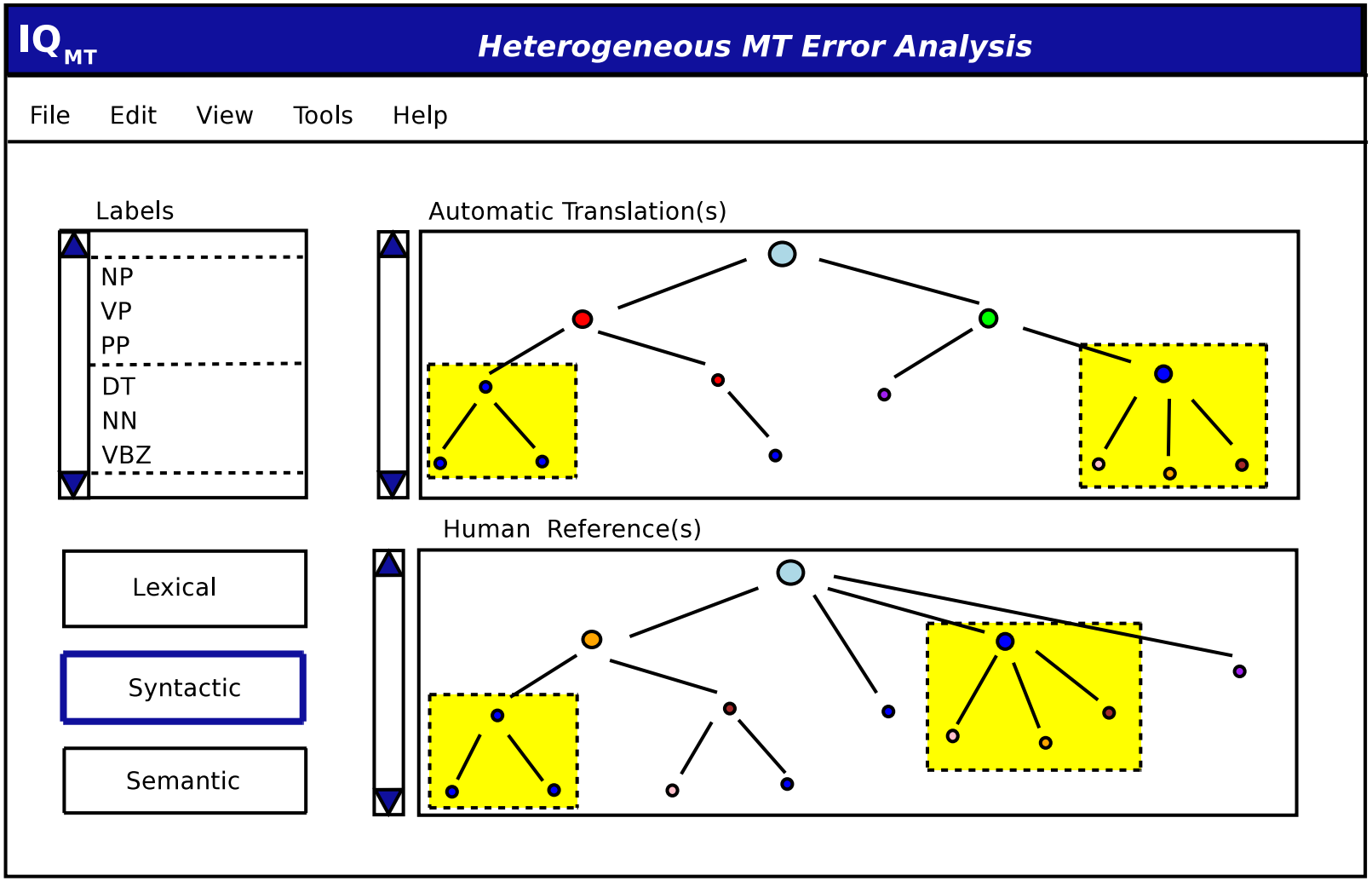
[1] Panacea: a remedy for all ills or difficulties (see cure-all).

Outline

Ongoing Steps...

1. Improving sentence level behavior:
   - Backing off to lexical similarity [GM08b]
   - Working on metric combinations [GM08a]
2. Porting metrics to languages other than English (e.g., Castilian Spanish, Catalan...)

A New Interface



Thanks for your Attention

IQ_{MT} v2.0 is publicly available at:

`http://www.lsi.upc.edu/~nlp/IQMT`

Towards Heterogeneous Automatic MT Error Analysis

(6th LREC)

Jesús Giménez and Lluís Màrquez

—

TALP Research Center
Technical University of Catalonia

May 29, 2008


Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez.
MT Evaluation: Human-Like vs. Human Acceptable.
In *Proceedings of COLING-ACL06*, 2006.


Enrique Amigó, Julio Gonzalo, Anselmo Peñas, and Felisa Verdejo.
QARLA: a Framework for the Evaluation of Automatic Sumarization.
In *Proceedings of the 43th Annual Meeting of the Association for Computational Linguistics*, 2005.


Jesús Giménez and Lluís Màrquez.
Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems.
In *Proceedings of the ACL Workshop on Statistical Machine Translation*, 2007.


Jesús Giménez and Lluís Màrquez.
Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations.
In *Proceedings of IJCNLP*, 2008.


Jesús Giménez and Lluís Màrquez.
On the Robustness of Linguistic Features for Automatic MT Evaluation.
In *Proceedings of the ELRA Workshop on Evaluation. Looking into the Future of Evaluation: when automatic metrics meet task-based and performance-based approaches*, 2008.


David Kauchak and Regina Barzilay.
Paraphrasing for Automatic Evaluation.
In *Proceedings of NLH-NAACL*, 2006.


Ding Liu and Daniel Gildea.

Syntactic Features for Evaluation of Machine Translation.
In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005.


Chin-Yew Lin and Franz Josef Och.
ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.
In *Proceedings of COLING*, 2004.


Audrey Le and Mark Przybocki.
NIST 2005 machine translation evaluation official results.
Technical report, NIST, August 2005.


Dennis Mehay and Chris Brew.
BLEUATRE: Flattening Syntactic Dependencies for MT Evaluation.

In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 2007.


Karolina Owczarzak, Declan Groves, Josef Van Genabith, and Andy Way.
Contextual Bitext-Derived Paraphrases in Automatic MT Evaluation.
In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA)*, 2006.


Karolina Owczarzak, Josef van Genabith, and Andy Way.
Dependency-Based Automatic Evaluation for Machine Translation.
In *Proceedings of SSST, NAACL-HLT/AMTA Workshop on Syntax and Structure in Statistical Translation*, 2007.


Maja Popovic and Hermann Ney.

Word Error Rates: Decomposition over POS classes and Applications for Error Analysis.
In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 48–55, Prague, Czech Republic, June 2007. Association for Computational Linguistics.


Florence Reeder, Keith Miller, Jennifer Doyon, and John White.
The Naming of Things and the Confusion of Tongues: an MT Metric.
In *Proceedings of the Workshop on MT Evaluation "Who did what to whom?"at MT Summit VIII*, 2001.


Liang Zhou, Chin-Yew Lin, and Eduard Hovy.
Re-evaluating Machine Translation Results with Paraphrase Support.

In *Proceedings of EMNLP*, 2006.