Automatic Evaluation in Machine Translation

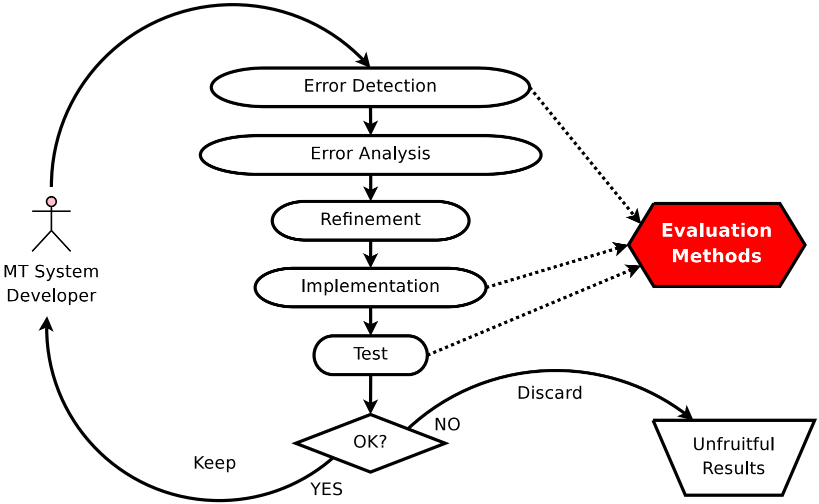
Towards Combined Linguistically-motivated Measures

Lluís Màrquez and **Jesús Giménez**
TALP Research Center
Tecnhical University of Catalonia

Machine Translation and Morphologically-rich Languages
Research Workshop of the Israel Science Foundation
University of Haifa, January 24, 2010

Talk Overview

MT System Development Cycle



Difficulties of MT Evaluation

- Machine Translation is an *open* NLP task
  - → the *correct translation* is not unique
  - → the set of valid translations is not small
  - → the *quality* of a translation is a fuzzy concept
- Quality aspects are *heterogeneous*
  - → Adequacy (or Fidelity)
  - → Fluency (or Intelligibility)
  - → Post-editing effort (time, key strokes, ...)
  - → ...
- Manual vs. automatic evaluation

MT Automatic Evaluation

Setting:

→ Compute similarity between system's output and one or several reference translations

→ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)

MT Automatic Evaluation

Setting:

→ Compute similarity between system's output and one or several reference translations

Challenge:

→ The similarity measure should be able to discriminate whether the two sentences convey the same meaning (semantic equivalence)

MT Automatic Evaluation

First Approaches:

$\longrightarrow$ Lexical similarity as a measure of quality

MT Automatic Evaluation

First Approaches:

$\longrightarrow$ Lexical similarity as a measure of quality

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

MT Automatic Evaluation

First Approaches:

→ Lexical similarity as a measure of quality

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

- **BLEU** has been widely accepted as a *'de facto'* standard

IBM BLEU metric

BLEU: a Method for Automatic Evaluation of Machine Translation

Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
IBM Research Division

"The main idea is to use a **weighted average of variable length phrase matches** against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family."

IBM BLEU metric

Conclusions of the paper (Papineni et al., 2001)

- BLEU correlates with human judgements
- It can distinguish among similar systems
- Need for multiple references or a big test with heterogeneous references
- More parametrisation in the future

Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. Cheap (vs. costly)
2. Objective (vs. subjective)
3. Reusable (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems

1. Error analysis
2. System optimization
3. System comparison

Benefits of Automatic Evaluation

Compared to manual evaluation, automatic measures are:

1. Cheap (vs. costly)
2. Objective (vs. subjective)
3. Reusable (vs. not-reusable)

Automatic evaluation metrics have notably accelerated the development cycle of MT systems

1. Error analysis
2. System optimization
3. System comparison

Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric
2. **Blind system development** → when metrics are unable to capture system improvements
3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. **System overtuning** → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture system improvements
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. System overtuning → when system parameters are adjusted towards a given metric
2. **Blind system development** → when metrics are unable to capture system improvements
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic Evaluation (compared to manual evaluation)

1. System overtuning → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture system improvements
3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems

Problems of Lexical Similarity Measures

The reliability of lexical metrics depends very strongly on the heterogeneity/representativity of reference translations.

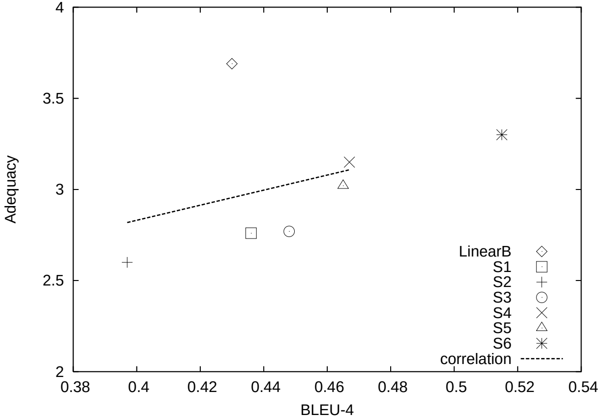
- Culy and Riehemann [CR03]
- Coughlin [Cou03]
- Callison-Burch et al. [CBOK06]

Underlying Cause

Lexical similarity is nor a *sufficient* neither a *necessary* condition so that two sentences convey the same meaning

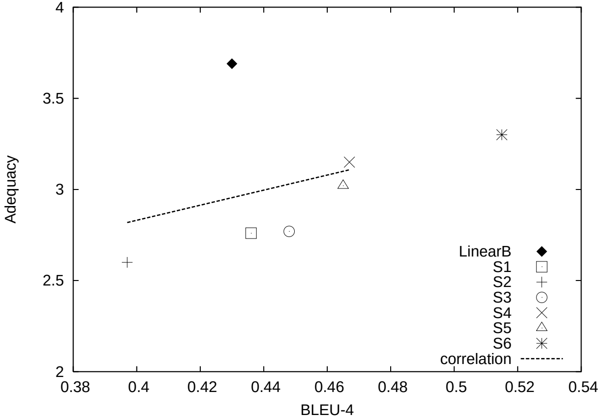
Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]



Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]



Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]

- ⟶ N-gram based metrics favor MT systems which closely replicate the lexical realization of the references
- ⟶ Test sets tend to be similar (domain, register, sublanguage) to training materials
- ⟶ Statistical MT systems heavily rely on the training data
- ⟶ Statistical MT systems tend to share the reference sublanguage and be favored by N-gram based measures

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]

→ N-gram based metrics favor MT systems which closely replicate the lexical realization of the references

→ Test sets tend to be similar (domain, register, sublanguage) to training materials

→ Statistical MT systems heavily rely on the training data

→ Statistical MT systems tend to share the reference sublanguage and be favored by N-gram based measures

Talk Overview

Can we do better?

Extending Lexical Similarity Measures to increase robustness (avoid sparsity):

- Lexical variants
  - → Morphological information (i.e., stemming) ROUGE and METEOR
  - → Synonymy lookup: METEOR (based on WordNet)
- Paraphrasing support:
  - → Zhou et al. [ZLH06], Kauchak and Barzilay [KB06], Owczarzak et al. [OGGW06]
  - → New versions of METEOR, TER

Can we do better?

Extending Lexical Similarity Measures to increase robustness (avoid sparsity):

- Lexical variants
  - → Morphological information (i.e., stemming)
    ROUGE and METEOR
  - → Synonymy lookup: METEOR (based on WordNet)
- Paraphrasing support:
  - → Zhou et al. [ZLH06], Kauchak and Barzilay [KB06], Owczarzak et al. [OGGW06]
  - → New versions of METEOR, TER

Similarity Measures Based on Linguistic Features

More linguistically-motivated measures:

- Features capturing syntactic and semantic information
- Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation
- Extense bibliography in the last years: [PN07], [LG05], [AGGM06], [MB07] [OvGW07a, OvGW07b], [KSO09], [CN08], [RMDW01], [GM07, GM09], [GMGM10], [PCGJM09], etc.

Some Examples of Linguistically Motivated Measures

- Expected Dependency Pair Match (Kahn, Snover and Ostendorf; 2009)
  - ⟶ dependency parsing (PCFG + head-finding rules)
  - ⟶ precision and recall scores of various tree decompositions
  - ⟶ +synonymy +paraphrasing
- MaxSim(Chen and Ng; 2008)
  - ⟶ a general framework for arbitrary similarity functions
  - ⟶ dependency relations, lemma, parts of speech, synonymy
  - ⟶ bipartite graph to obtain an optimal matching between items
- RTE (Padó, Galley, Jurafsky and Manning, 2009)
  - ⟶ semantic equivalence based on textual entailment features
  - ⟶ alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment

Some Examples of Linguistically Motivated Measures

- Expected Dependency Pair Match (Kahn, Snover and Ostendorf; 2009)
  - ⟶ dependency parsing (PCFG + head-finding rules)
  - ⟶ precision and recall scores of various tree decompositions
  - ⟶ +synonymy +paraphrasing
- MaxSim(Chen and Ng; 2008)
  - ⟶ a general framework for arbitrary similarity functions
  - ⟶ dependency relations, lemma, parts of speech, synonymy
  - ⟶ bipartite graph to obtain an optimal matching between items
- RTE (Padó, Galley, Jurafsky and Manning, 2009)
  - ⟶ semantic equivalence based on textual entailment features
  - ⟶ alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment

Some Examples of Linguistically Motivated Measures

- Expected Dependency Pair Match (Kahn, Snover and Ostendorf; 2009)
  - ⟶ dependency parsing (PCFG + head-finding rules)
  - ⟶ precision and recall scores of various tree decompositions
  - ⟶ +synonymy +paraphrasing
- MaxSim(Chen and Ng; 2008)
  - ⟶ a general framework for arbitrary similarity functions
  - ⟶ dependency relations, lemma, parts of speech, synonymy
  - ⟶ bipartite graph to obtain an optimal matching between items
- RTE (Padó, Galley, Jurafsky and Manning, 2009)
  - ⟶ semantic equivalence based on textual entailment features
  - ⟶ alignment, semantic compatibility, insertion/deletion, preservation of reference and structural alignment

Our Approach (Giménez & Màrquez, 2010)

Work at UPC with Jesús Giménez

Rather than comparing sentences at lexical level:

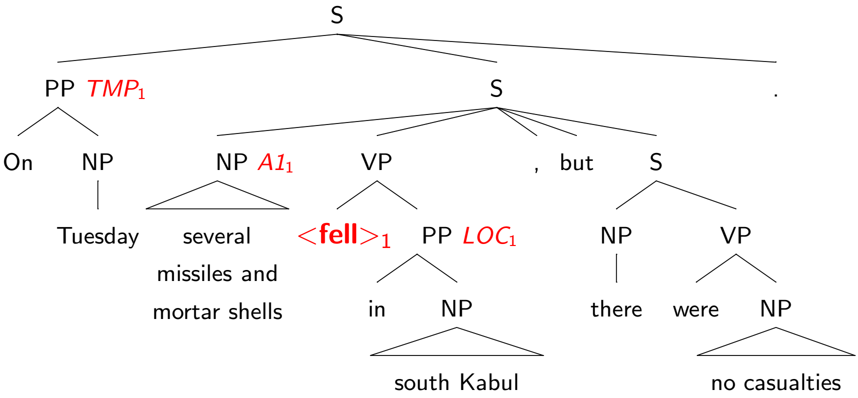
Compare the linguistic structures and the words within them

Our Approach

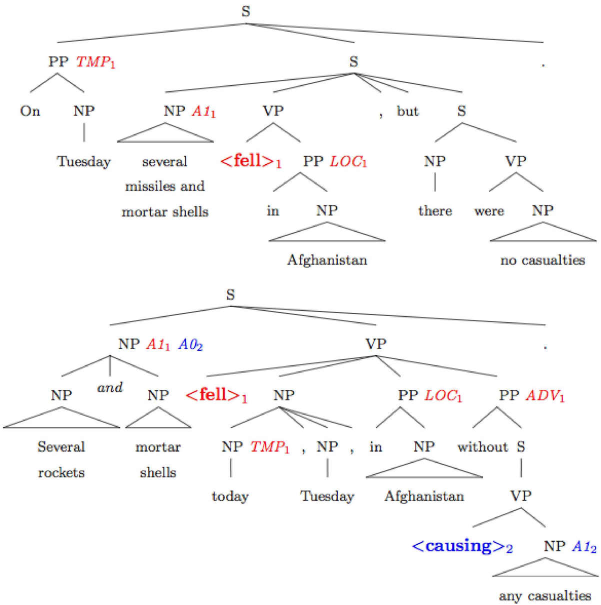
(Giménez & Màrquez, 2010)

Automatic Translation	On Tuesday several missiles and mortar shells fell in south Kabul , but there were no casualties .
Reference Translation	Several rockets and mortar shells fell today , Tuesday , in south Kabul without causing any casualties .

Our Approach (Giménez & Màrquez, 2010)



Our Approach (Giménez & Màrquez, 2010)



Measuring Structural Similarity

- OVERLAP: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient
- Linguistic element (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - → For instance: POS tags, word lemmas, NPs, syntactic phrases
  - → A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - → LEs may embed

Measuring Structural Similarity

- OVERLAP: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient
- Linguistic element (LE) = abstract reference to any possible type of linguistic unit, structure, or relationship among them
  - $\longrightarrow$ For instance: POS tags, word lemmas, NPs, syntactic phrases
  - $\longrightarrow$ A sentence can be seen as a bag (or a sequence) of LEs of a certain type
  - $\longrightarrow$ LEs may embed

Overlap among Linguistic Elements

$$\mathrm{O}(t) = \frac{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{hyp}) \,\cap\, \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{hyp}}(i, t)}{\displaystyle\sum_{i \in (\mathrm{items}_t(\mathrm{hyp}) \,\cup\, \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{hyp}}(i, t), \mathrm{count}_{\mathrm{ref}}(i, t))}$$

t is the LE type
'hyp': hypothesized translation
'ref': reference translation
$\mathrm{items}_t(s)$: set of items occurring inside LEs of type t
$\mathrm{count}_s(i, t)$: occurrences of item i in s inside a LE of type t

Overlap among Linguistic Elements

Coarser variant: micro-averaged overlap over all types

$$\mathrm{O}(\star) = \frac{\displaystyle\sum_{t \in T} \sum_{i \in (\mathrm{items}_t(\mathrm{hyp}) \,\cap\, \mathrm{items}_t(\mathrm{ref}))} \mathrm{count}_{\mathrm{hyp}}(i, t)}{\displaystyle\sum_{t \in T} \sum_{i \in (\mathrm{items}_t(\mathrm{hyp}) \,\cup\, \mathrm{items}_t(\mathrm{ref}))} \max(\mathrm{count}_{\mathrm{hyp}}(i, t), \mathrm{count}_{\mathrm{ref}}(i, t))}$$

T: set of all LE types associated to the given LE class

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict variant
  - → All items inside an element are considered the same unit
  - → Computes the proportion of fully translated LEs, according to their types
- Other possible extensions:
  - → *n*-gram matching within LEs
  - → Synonymy lookup

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict variant
  - → All items inside an element are considered the same unit
  - → Computes the proportion of fully translated LEs, according to their types
- Other possible extensions:
  - → *n*-gram matching within LEs
  - → Synonymy lookup

Overlap/Matching among Linguistic Elements

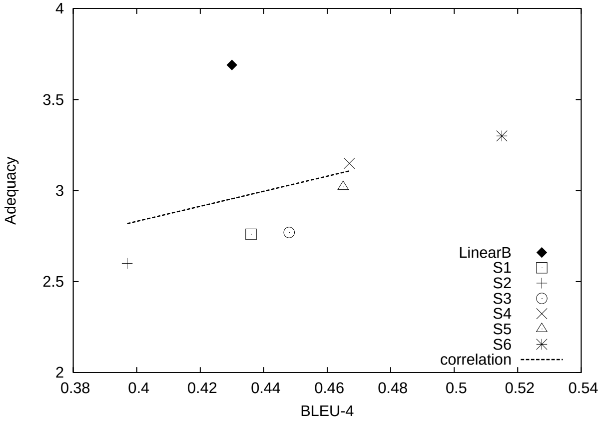
- Overlap and Matching have been instantiated over different linguistic level elements (for English)
  - → Words, lemmas, POS
  - → Shallow, dependency and constituency parsing
  - → Named entities and semantic roles
  - → Discourse representation (logical forms)
- Open source software: ASIYA, Open Toolkit for Automatic MT (Meta-)Evaluation (formerly IQMT)
  http://www.lsi.upc.es/∼nlp/Asiya/

Overlap/Matching among Linguistic Elements

- Overlap and Matching have been instantiated over different linguistic level elements (for English)
  - $\longrightarrow$ Words, lemmas, POS
  - $\longrightarrow$ Shallow, dependency and constituency parsing
  - $\longrightarrow$ Named entities and semantic roles
  - $\longrightarrow$ Discourse representation (logical forms)
- Open source software: ASIYA, Open Toolkit for Automatic MT (Meta-)Evaluation (formerly IQMT)
  http://www.lsi.upc.es/∼nlp/Asiya/

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise [CBOK06, KM06]



Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

NIST 2005 Arabic-to-English Exercise

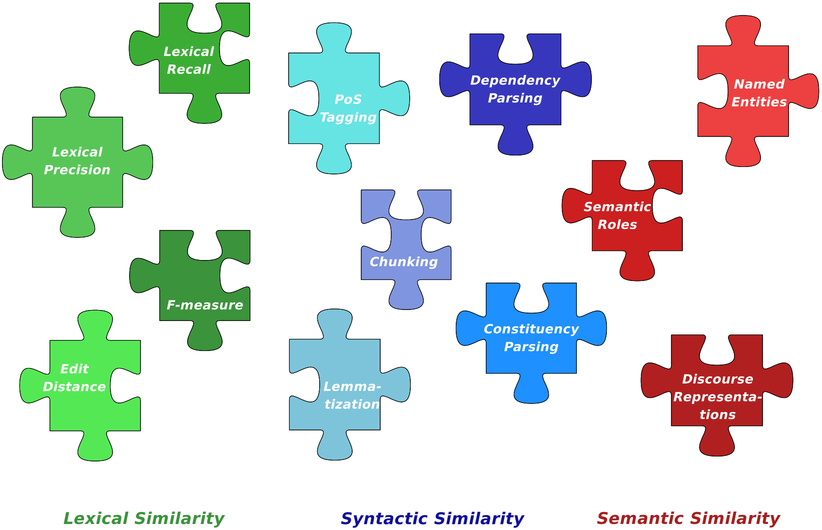
Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Evaluating Heterogeneous Features

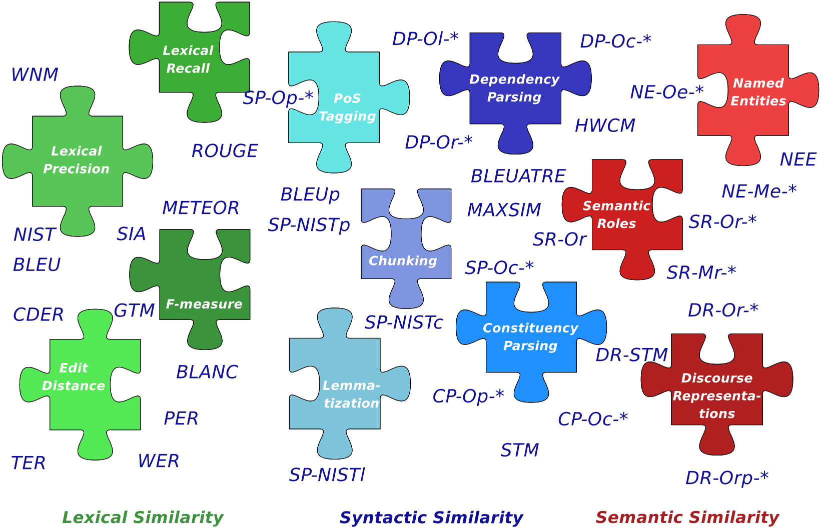
NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ_{all}	ρ_{SMT}
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	Parts-of-speech	0.42	0.89
	Dependencies (HWC)	**0.88**	0.86
	Constituents (STM)	0.74	**0.95**
Semantic	Semantic Roles	0.72	**0.96**
	Discourse Repr.	0.92	0.92
	Discourse Repr. (PoS)	**0.97**	0.90

Towards Heterogeneous Automatic MT Evaluation



Towards Heterogeneous Automatic MT Evaluation



Recent Works on Metric Combination

Different metrics capture different aspects of similarity
Suitable for combination

- Corston-Oliver et al. [COGB01]
- Kulesza and Shieber [KS04]
- Gamon et al. [GAS05]
- Akiba et al. [AIS01]
- Quirk [Qui04]
- Liu and Gildea [LG07]
- Albrecht and Hwa [AH07a]
- Paul et al. [PFS07]
- Ye et al. [YZL07]
- Giménez and Màrquez [GM08]

Recent Works on Metric Combination

Different metrics capture different aspects of similarity
Suitable for combination

- Corston-Oliver et al. [COGB01]
- Kulesza and Shieber [KS04]
- Gamon et al. [GAS05]
- Akiba et al. [AIS01]
- Quirk [Qui04]
- Liu and Gildea [LG07]
- Albrecht and Hwa [AH07a]
- Paul et al. [PFS07]
- Ye et al. [YZL07]
- Giménez and Màrquez [GM08]

The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

$$\mathrm{ULC}_M(hyp, ref) = \frac{1}{|M|} \sum_{m \in M} m(hyp, ref)$$

- Simple hill climbing approach to find the best subset of measures M on a development corpus

$M = \{$ 'ROUGE_W', 'METEOR', '$\mathit{DP\text{-}HWC}_r$', '$\mathit{DP\text{-}O}_c(\star)$', '$\mathit{DP\text{-}O}_l(\star)$', '$\mathit{DP\text{-}O}_r(\star)$', '$\mathit{CP\text{-}STM}_4$', '$\mathit{SR\text{-}O}_r(\star)$', '$\mathit{SR\text{-}O}_{rv}$', '$\mathit{DR\text{-}O}_{rp}(\star)$' $\}$

The Most Simple Approach: ULC

- Uniformly averaged linear combination of measures (ULC):

$$\mathrm{ULC}_M(hyp, ref) = \frac{1}{|M|} \sum_{m \in M} m(hyp, ref)$$

- Simple hill climbing approach to find the best subset of measures M on a development corpus

$M = \{$ '$ROUGE_W$', '$METEOR$', '$DP\text{-}HWC_r$', '$DP\text{-}O_c(\star)$', '$DP\text{-}O_l(\star)$', '$DP\text{-}O_r(\star)$', '$CP\text{-}STM_4$', '$SR\text{-}O_r(\star)$', '$SR\text{-}O_{rv}$', '$DR\text{-}O_{rp}(\star)$' $\}$

Evaluation of ULC

WMT 2008 meta-evaluation results (into-English)

Measure	ρ_{sys}	consistency$_{snt}$
ULC	**0.83**	**0.56**
DP-O$_r$($\star$)	**0.83**	0.51
DR-O$_r$($\star$)	0.80	0.50
METEOR$_{ranking}$	0.78	0.51
SR-O$_r$($\star$)	0.77	0.50
METEOR$_{baseline}$	0.75	0.51
PoS-BLEU	0.75	0.44
PoS-4gram-F	0.74	0.50
BLEU	0.52	—
BLEU$_{stem+wnsyn}$	0.50	0.51
...		

Evaluation of ULC

WMT 2009 meta-evaluation results (into-English)

Measure	ρ_{sys}	**consistency$_{\text{snt}}$**
ULC	**0.83**	**0.54**
maxsim	0.80	0.52
rte(absolute)	0.79	0.53
meteor-rank	0.75	0.49
rte(pairwise)	0.75	0.51
terp	-0.72	0.50
meteor-0.6	0.72	0.49
meteor-0.7	0.66	0.49
bleu-ter/2	0.58	—
nist	0.56	—
wpF	0.56	0.52
ter	-0.54	0.45
...		

Portability Across Corpora

NIST 2004/2005 MT Evaluation Campaigns

	$\mathbf{AE}_{2004}$	$\mathbf{CE}_{2004}$	$\mathbf{AE}_{2005}$	$\mathbf{CE}_{2005}$
#references	5	5	5	4
#outputs$_{\text{ass.}}$	5/5	10/10	6/7	5/10
#sentences$_{\text{ass.}}$	347/1,353	447/1,788	266/1,056	272/1,082
Avg. Adequacy	2.81/5	2.60/5	3.00/5	2.58/5
Avg. Fluency	2.56/5	2.41/5	2.70/5	2.47/5

Portability Across Corpora

Meta-evaluation of ULC across test beds
(Pearson Correlation)

	$\mathbf{AE_{04}}$	$\mathbf{CE_{04}}$	$\mathbf{AE_{05}}$	$\mathbf{CE_{05}}$
ULC $(_{AE04})$	0.6392	0.6294	0.5327	0.5695
ULC $(_{CE04})$	0.6306	0.6333	0.5115	0.5692
ULC $(_{AE05})$	0.6175	0.6029	0.5450	0.5706
ULC $(_{CE05})$	0.6218	0.6208	0.5270	0.6047
Max Indiv.	0.5877	0.5955	0.4960	0.5348

Linguistic Measures at International Campaigns

- NIST 2004/2005
  - → Arabic-to-English / Chinese-to-English
  - → Broadcast news / weblogs / dialogues
- WMT 2007-2010
  - → Translation between several European languages
  - → European Parliament Proceedings / Out-of-domain News
- IWSLT 2005-2008
  - → Spoken language translation
  - → Chinese-to-English

Linguistic Measures at International Campaigns

- NIST 2004/2005
  - $\longrightarrow$ Arabic-to-English / Chinese-to-English
  - $\longrightarrow$ Broadcast news / weblogs / dialogues
- WMT 2007-2010
  - $\longrightarrow$ Translation between several European languages
  - $\longrightarrow$ European Parliament Proceedings / Out-of-domain News
- IWSLT 2005-2008
  - $\longrightarrow$ Spoken language translation
  - $\longrightarrow$ Chinese-to-English

Controversial results at NIST Metrics MATR08/09 Challenges!

Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures
2. Work on the combination of measures
   → Learning combined similarity measures
3. Porting measures to languages other than English
   → Need of linguistic analyzers
4. Use measures for semi–automatic error analysis
   → (Web) Graphical interface

Ongoing and Future Work

1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures
2. Work on the combination of measures
   → Learning combined similarity measures
3. Porting measures to languages other than English
   → Need of linguistic analyzers
4. Use measures for semi–automatic error analysis
   → (Web) Graphical interface

Ongoing and Future Work

1. Metaevaluation of measures
   $\longrightarrow$ Better understand differences between lexical and higher level measures

2. Work on the combination of measures
   $\longrightarrow$ Learning combined similarity measures

3. Porting measures to languages other than English
   $\longrightarrow$ Need of linguistic analyzers

4. Use measures for semi–automatic error analysis
   $\longrightarrow$ (Web) Graphical interface

Ongoing and Future Work

1. Metaevaluation of measures
   $\longrightarrow$ Better understand differences between lexical and higher level measures
2. Work on the combination of measures
   $\longrightarrow$ Learning combined similarity measures
3. Porting measures to languages other than English
   $\longrightarrow$ Need of linguistic analyzers
4. Use measures for semi–automatic error analysis
   $\longrightarrow$ (Web) Graphical interface

Talk Overview

Confidence Estimation

New setting:

→ Quality evaluation without reference translations

Motivation:

→ Ranking of several candidate translations when translating new examples

Information available:

→ Source sentence, candidate translation(s), and (possibly) system information

Confidence Estimation

New setting:

→ Quality evaluation without reference translations

Motivation:

→ Ranking of several candidate translations when translating new examples

Information available:

→ Source sentence, candidate translation(s), and (possibly) system information

Johns Hopkins University Summer Workshop, 2003
"Confidence Estimation for Machine Translation" [BFF+03]

Confidence Estimation

→ **Classification according to the target function**

- *Human likeness*
  → discern between human and automatic translations
  - Classification
- *Human acceptability*
  → emulate the behavior of human assessors
  - Classification [GAS05]
  - Linear Regression [Qui04, AH07b, SG10]
  - Ranking [SE10]

Confidence Estimation

Features to train the quality measures:

- System-dependent
- System-independent

Confidence Estimation

Features to train the quality measures:

- System-dependent
  - → internal system probabilities/scores
  - → features over *n*-best translation hypotheses
    - language modeling
    - hypothesis rank
    - score ratio
    - average hypothesis length
    - length ratio
    - center hypothesis
- System-independent

Confidence Estimation

Features to train the quality measures:

- System-dependent
- System-independent
  - → source (translation difficulty)
    - sentence length
    - ambiguity → dictionary/alignment/WordNet-based (number of candidate translations per word or phrase)
  - → target (fluency)
    - sentence length
    - language modeling
  - → source-target (adequacy)
    - length ratio
    - punctuation issues
    - candidate matching → dictionary-/alignment-based

Confidence Estimation

Features to train the quality measures:

- System-dependent
- System-independent

Remark: most valuable features

- System-dependent
- Based on *n*-best lists
- Capturing target text properties

The FAUST Project (2010-2013)

- Feedback Analysis for User Adaptive Statistical Translation
- Theme FP7-ICT-2009-4
- Objective 2.2: Language-based interaction
- Coordinator: University of Cambridge (Bill Byrne)
- http://divf.eng.cam.ac.uk/faust

Goal Develop interactive machine translation systems which adapt rapidly and intelligently to user feedback

FAUST: On-line Confidence Estimation

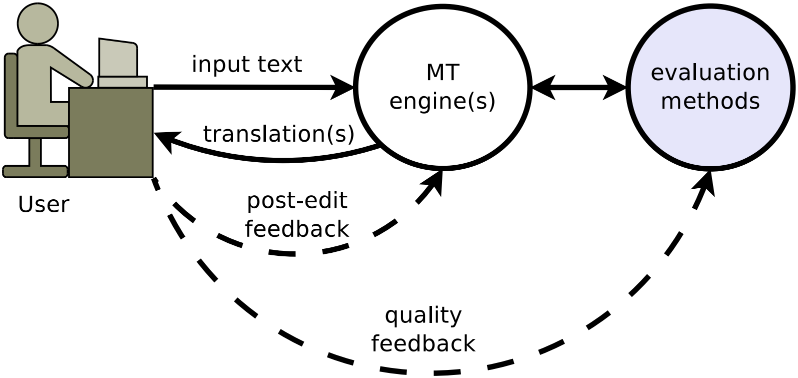
CE-related challenge

⟶ Create novel automatic metrics of translation quality which reflect preferences learned from user feedback

- State of the art: MT relies on metrics which do not reflect user interest
- FAUST: MT metrics as models of user feedback

⟶ Keywords: on-line, adaptive

FAUST: On-line Confidence Estimation



FAUST: On-line Confidence Estimation

source Eric es alto

Ta Eric is high

Tb Eric is tall

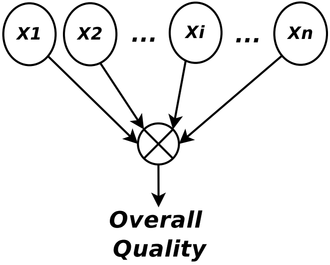
○ Ta is better than Tb

● Tb is better than Ta

○ Ta and Tb are equally good (or bad)

quality(Tb) > quality(Ta) ?

FAUST: On-line Confidence Estimation

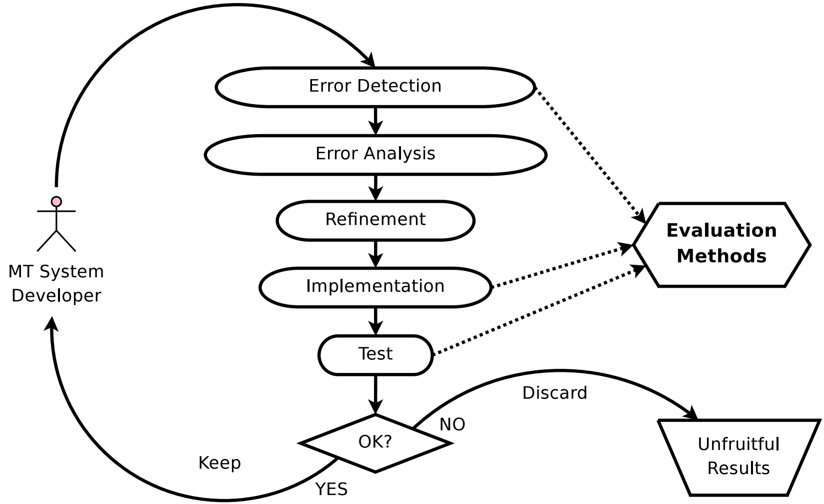


Ongoing work:

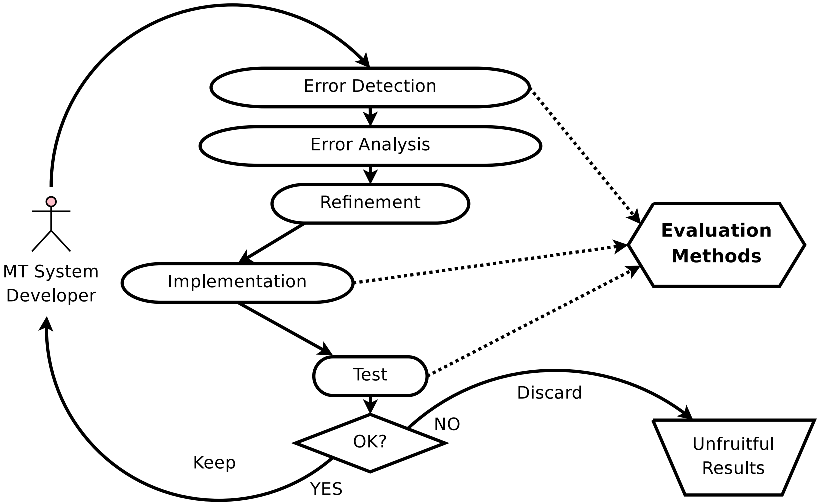
- Preliminary set of 14 CE measures (= features)
- Learn to rank pairwise comparisons
- Ranking perceptron (with linear and polynomial kernels)
- Promising results on an initial batch setting

Talk Overview

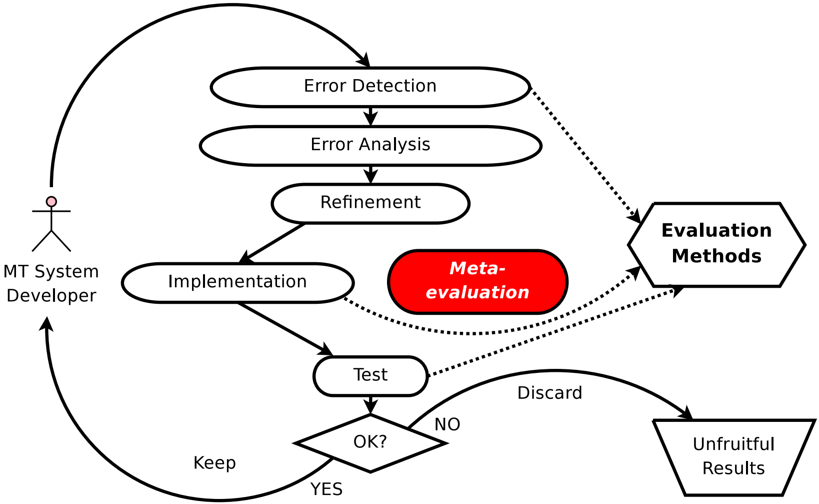
Metricwise System Development



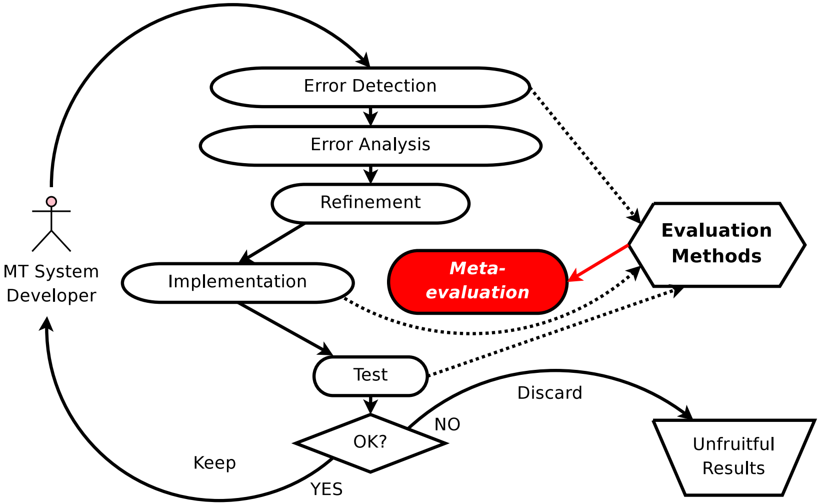
Metricwise System Development



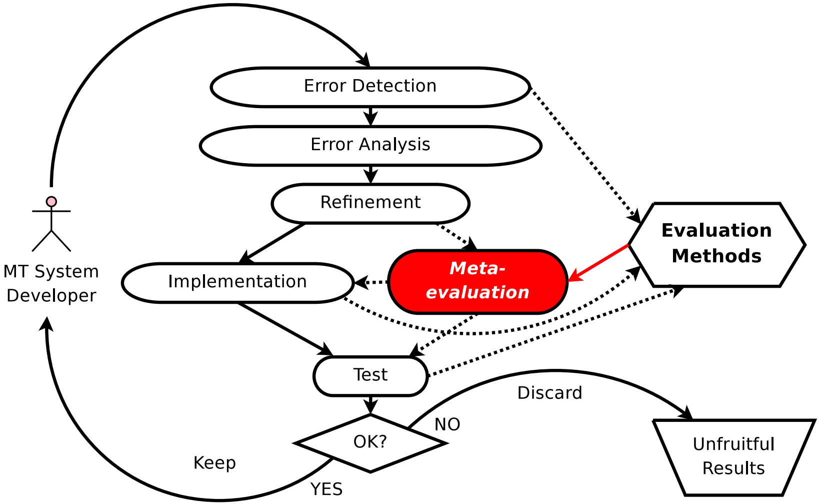
Metricwise System Development



Metricwise System Development



Metricwise System Development



Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   - → Quality aspects are heterogeneous and diverse
4. What can we do?
   - → Advance towards heterogeneous evaluation methods
   - → Metricwise system development
     - Always meta-evaluate
       (make sure your metric fits your purpose)
   - → Resort to manual evaluation
     - Always conduct manual evaluations
       (contrast your automatic evaluations)
     - Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. What can we do?
   → Advance towards heterogeneous evaluation methods
   → Metricwise system development
      Always meta-evaluate
      (make sure your metric fits your purpose)
   → Resort to manual evaluation
      Always conduct manual evaluations
      (contrast your automatic evaluations)
      Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   - → Quality aspects are heterogeneous and diverse
4. What can we do?
   - → Advance towards heterogeneous evaluation methods
   - → Metricwise system development
     - Always meta-evaluate (make sure your metric fits your purpose)
   - → Resort to manual evaluation
     - Always conduct manual evaluations (contrast your automatic evaluations)
     - Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. What can we do?
   → Advance towards heterogeneous evaluation methods
   → Metricwise system development
      Always meta-evaluate
      (make sure your metric fits your purpose)
   → Resort to manual evaluation
      Always conduct manual evaluations
      (contrast your automatic evaluations)
      Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. What can we do?
   → Advance towards heterogeneous evaluation methods
   → Metricwise system development
      Always meta-evaluate
      (make sure your metric fits your purpose)
   → Resort to manual evaluation
      Always conduct manual evaluations
      (contrast your automatic evaluations)
      Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   → Quality aspects are heterogeneous and diverse
4. What can we do?
   → Advance towards heterogeneous evaluation methods
   → Metricwise system development

      Always meta-evaluate
      (make sure your metric fits your purpose)

   → Resort to manual evaluation

      Always conduct manual evaluations
      (contrast your automatic evaluations)
      Always do error analysis (semi-automatic)

Summary

1. Empirical MT is a very active research field
2. Evaluation methods play a crucial role
3. Measuring overall translation quality is hard
   - → Quality aspects are heterogeneous and diverse
4. What can we do?
   - → Advance towards heterogeneous evaluation methods
   - → Metricwise system development

     Always meta-evaluate
     (make sure your metric fits your purpose)
   - → Resort to manual evaluation

     Always conduct manual evaluations
     (contrast your automatic evaluations)
     Always do error analysis (semi-automatic)

Automatic Evaluation in Machine Translation

Towards Combined Linguistically-motivated Measures

Lluís Màrquez and **Jesús Giménez**
TALP Research Center
Tecnhical University of Catalonia

Machine Translation and Morphologically-rich Languages
Research Workshop of the Israel Science Foundation
University of Haifa, January 24, 2010

On-line Confidence Estimation

Preliminary set of features

Metric	**Description**
CE-BiDictO	bilingual dictionary based overlap
CE-N_c	source/candidate phrase chunk ratio
CE-N_e	source/candidate named entity ratio
CE-O_c	source/candidate phrase chunk overlap
CE-O_e	source/candidate named entity overlap
CE-O_p	source/candidate part-of-speech overlap
CE-ippl	candidate language model inverse perplexity
CE-ippl$_C$	candidate chunk language model inverse perplexity
CE-ippl$_P$	candidate PoS language model inverse perplexity
CE-length	source/candidate length ratio
CE-long	source/candidate length ratio (penalize short candidates)
CE-oov	candidate language model out-of-vocabulary tokens ratio
CE-short	source/candidate length ratio (penalize long candidates)
CE-symbols	symbol overlap (punctuation, etc.)

Enrique Amigó, Jesús Giménez, Julio Gonzalo, and Lluís Màrquez.
MT Evaluation: Human-Like vs. Human Acceptable.
In *Proceedings of the Joint 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics (COLING-ACL)*, pages 17–24, 2006.

Joshua Albrecht and Rebecca Hwa.
A Re-examination of Machine Learning Approaches for Sentence-Level MT Evaluation.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 880–887, 2007.

Joshua Albrecht and Rebecca Hwa.
Regression for Sentence-Level MT Evaluation with Pseudo References.
In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 296–303, 2007.

Yasuhiro Akiba, Kenji Imamura, and Eiichiro Sumita.
Using Multiple Edit Distances to Automatically Rank Machine Translation Output.
In *Proceedings of Machine Translation Summit VIII*, pages 15–20, 2001.

John Blatz, Erin Fitzgerald, George Foster, Simona Gandrabur, Cyril Goutte, Alex Kulesza, Alberto Sanchis, and Nicola Ueffing.
Confidence estimation for machine translation. Final Report of Johns Hopkins 2003 Summer Workshop on Speech and Language Engineering.
Technical report, Johns Hopkins University, 2003.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn.
Re-evaluating the Role of BLEU in Machine Translation Research.
In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2006.

- Simon Corston-Oliver, Michael Gamon, and Chris Brockett.
  A Machine Learning Approach to the Automatic Evaluation of Machine Translation.
  In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 140–147, 2001.
- Deborah Coughlin.
  Correlating Automated and Human Assessments of Machine Translation Quality.
  In *Proceedings of Machine Translation Summit IX*, pages 23–27, 2003.
- Christopher Culy and Susanne Z. Riehemann.
  The Limits of N-gram Translation Evaluation Metrics.
  In *Proceedings of MT-SUMMIT IX*, pages 1–8, 2003.
- Michael Gamon, Anthony Aue, and Martine Smets.
  Sentence-Level MT evaluation without reference translations: beyond language modeling.
  In *Proceedings of EAMT*, pages 103–111, 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*, pages 256–264, 2007.

Jesús Giménez and Lluís Màrquez.
Heterogeneous Automatic MT Evaluation Through Non-Parametric Metric Combinations.
In *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, pages 319–326, 2008.

Jesús Giménez and Lluís Màrquez.
On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation.
In *Proceedings of the 4th Workshop on Statistical Machine Translation (EACL 2009)*, 2009.

David Kauchak and Regina Barzilay.
Paraphrasing for Automatic Evaluation.

In *Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL)*, pages 455–462, 2006.

Philipp Koehn and Christof Monz.
Manual and Automatic Evaluation of Machine Translation between European Languages.
In *Proceedings of the NAACL Workshop on Statistical Machine Translation*, pages 102–121, 2006.

Alex Kulesza and Stuart M. Shieber.
A learning approach to improving sentence-level MT evaluation.
In *Proceedings of the 10th International Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, pages 75–84, 2004.

Ding Liu and Daniel Gildea.
Syntactic Features for Evaluation of Machine Translation.

In *Proceedings of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization*, pages 25–32, 2005.

Ding Liu and Daniel Gildea.
Source-Language Features and Maximum Correlation Training for Machine Translation Evaluation.
In *Proceedings of the 2007 Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 41–48, 2007.

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)*, pages 148–155, 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*, pages 80–87, 2007.

Karolina Owczarzak, Josef van Genabith, and Andy Way.
Labelled Dependencies in Machine Translation Evaluation.
In *Proceedings of the ACL Workshop on Statistical Machine Translation*, pages 104–111, 2007.

Michael Paul, Andrew Finch, and Eiichiro Sumita.

Reducing Human Assessments of Machine Translation Quality to Binary Classifiers.
In *Proceedings of the 11th Conference on Theoretical and Methodological Issues in Machine Translation (TMI)*, 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.

Chris Quirk.
Training a Sentence-Level Machine Translation Confidence Metric.
In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, pages 825–828, 2004.

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 Machine Translation Summit VIII*, pages 55–59, 2001.

Radu Soricut and Abdessamad Echihabi.
Trustrank: Inducing trust in automatic translations via ranking.
In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 612–621, Uppsala, Sweden, July 2010. Association for Computational Linguistics.

Lucia Specia and Jesús Giménez.
Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation.

In *Proceedings of the 1st Conference of the Association for Machine Translation in the Americas (AMTA)*, 2010.

Yang Ye, Ming Zhou, and Chin-Yew Lin.
Sentence Level Machine Translation Evaluation as a Ranking.
In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 240–247, 2007.

Liang Zhou, Chin-Yew Lin, and Eduard Hovy.
Re-evaluating Machine Translation Results with Paraphrase Support.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 77–84, 2006.