Automatic Evaluation in Machine Translation

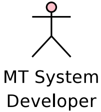
Towards Similarity Measures Based on Multiple Linguistic Layers

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

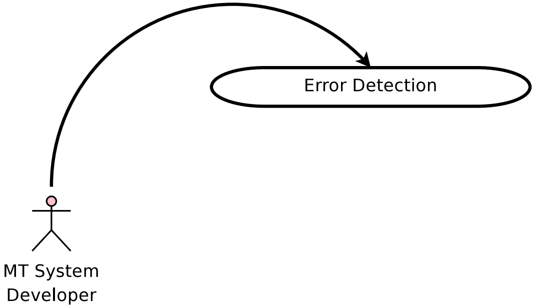
MOLTO workshop – **GF meets SMT**
Göteborg, November 5, 2010

Talk Overview

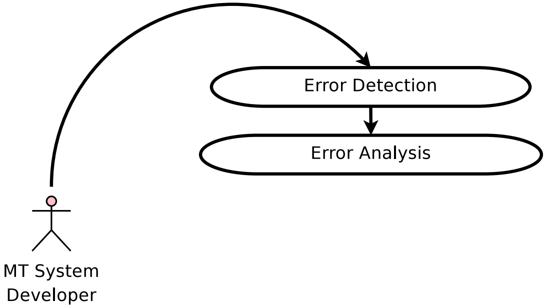
The Current System Development Cycle



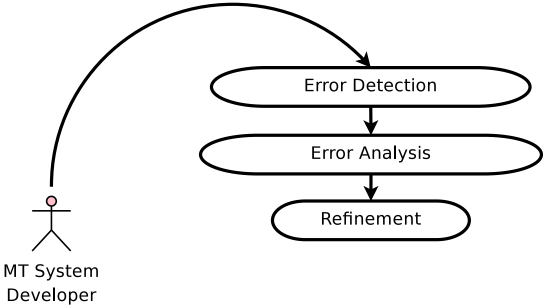
The Current System Development Cycle



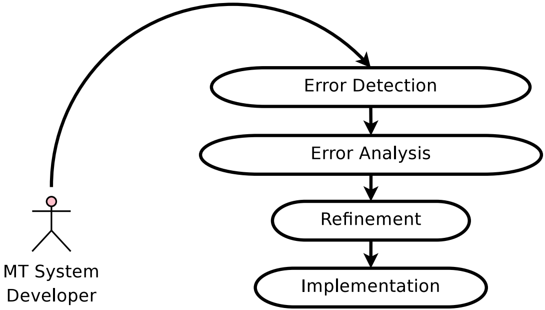
The Current System Development Cycle



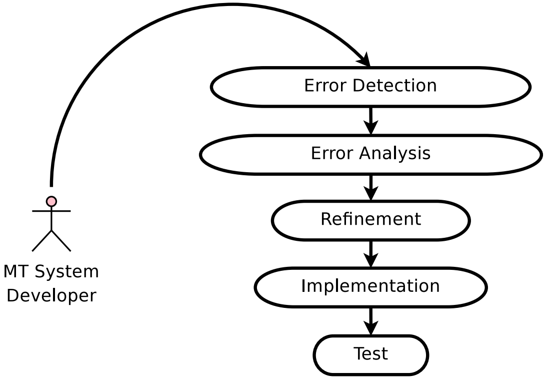
The Current System Development Cycle



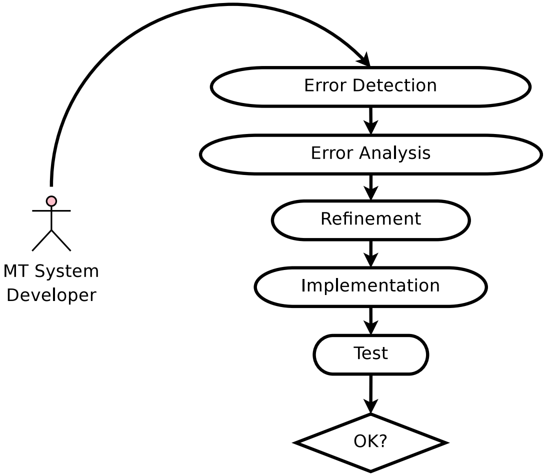
The Current System Development Cycle



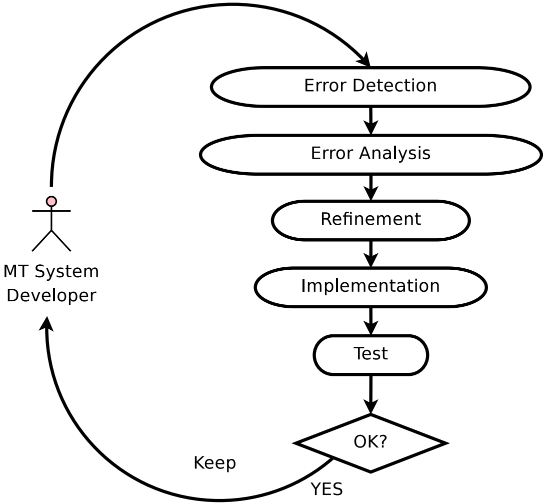
The Current System Development Cycle



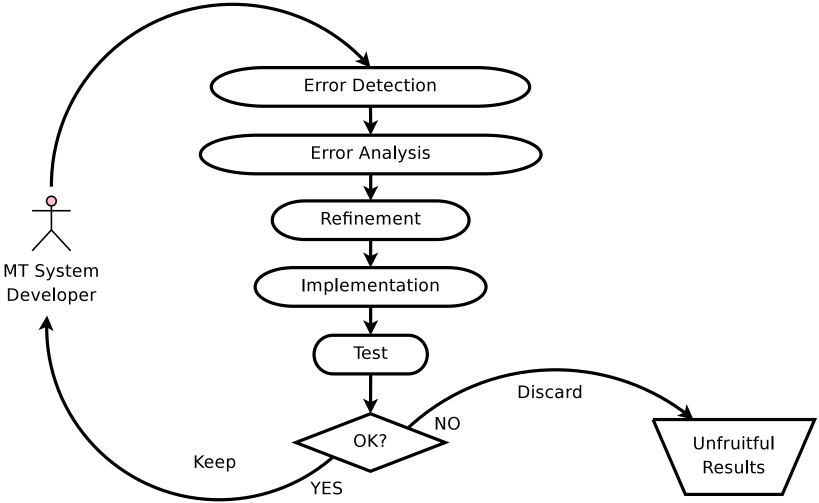
The Current System Development Cycle



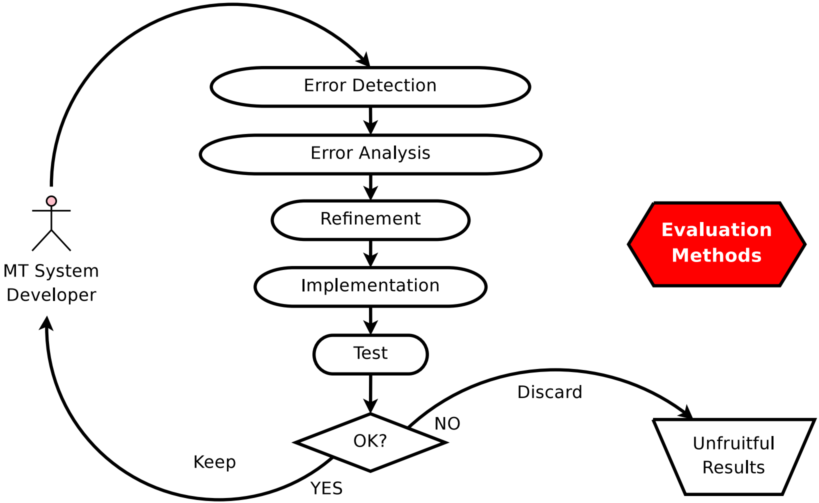
The Current System Development Cycle



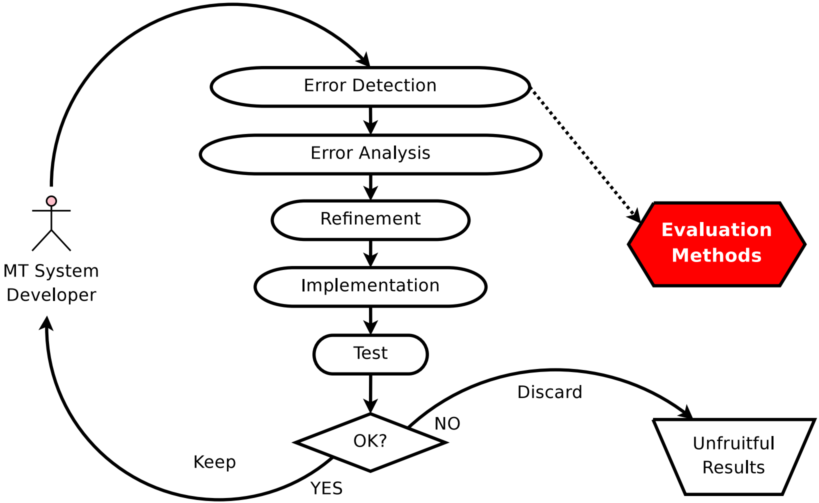
The Current System Development Cycle



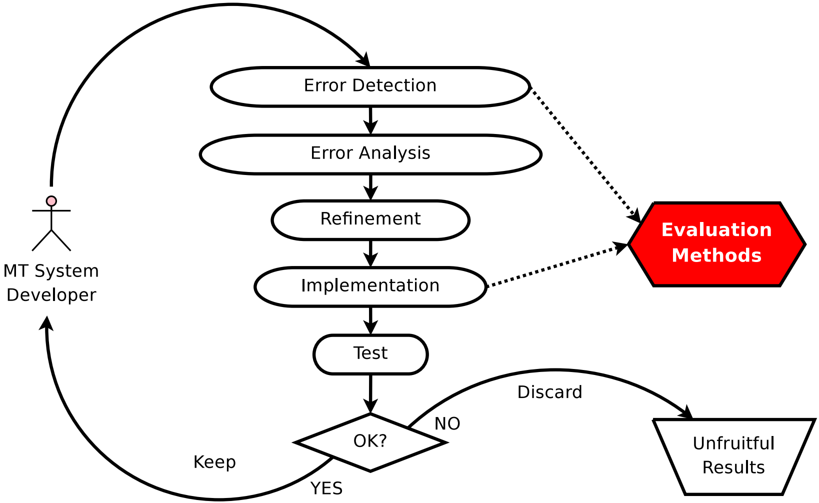
The Current System Development Cycle



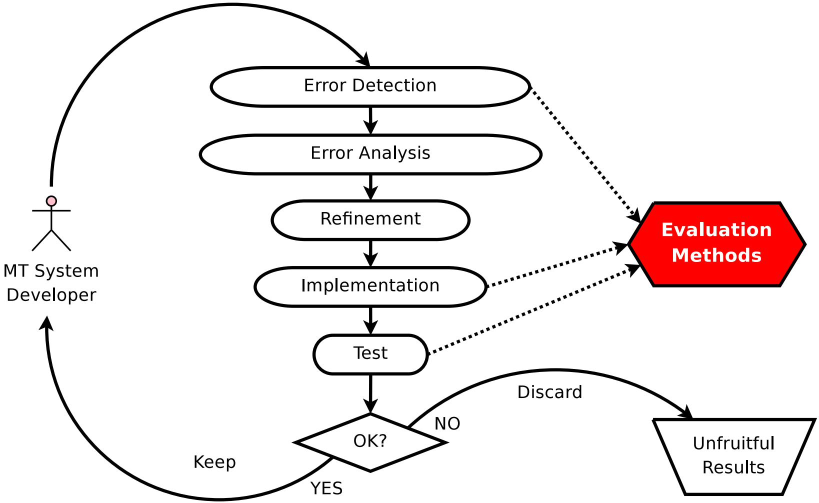
The Current System Development Cycle



The Current System Development Cycle



The Current 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, ...)
  - → ...

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 Manual Evaluation

- Many protocols for manual evaluation exist
- ARPA's Approach (since 90's):
- Adequacy (fidelity) and Fluency (intelligibility).

Score	Adequacy	Fluency
5	All information	Flawless English
4	Most	Good
3	Much	Non-native
2	Little	Disfluent
1	None	Incomprehensible

Pros and Cons of Manual Evaluation

Advantages	Disadvantages
Direct interpretation	

Pros and Cons of Manual Evaluation

Advantages	Disadvantages
Direct interpretation	Time cost
	Money cost

Pros and Cons of Manual Evaluation

Advantages	**Disadvantages**
Direct interpretation	Time cost
	Money cost
	Subjectivity
	Non-reusability

MT Automatic Evaluation

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

→ Lexical similarity as a measure of quality

MT Automatic Evaluation

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

→ 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

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

→ 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: Papineni, Roukos, Ward and Zhu [2001]

Candidate 1:
`It is a guide to action which ensures that the military always obeys the commands of the party.`

Candidate 2:
`It is to insure the troops forever hearing the activity guidebook that party direct.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Candidate 1:

`It is a guide to action which ensures that the military always obeys the commands of the party.`

Reference 1:

`It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2:

`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3:

`It is the practical guide for the army always to heed the directions of the party.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Candidate 1:
`It is a guide to action which ensures that the military always obeys the commands of the party.`

Reference 1:
`It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2:
`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3:
`It is the practical guide for the army always to heed the directions of the party.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Candidate 2:

`It is to insure the troops forever hearing the activity guidebook that party direct.`

Reference 1:

`It is a guide to action that ensures that the military will forever heed Party commands.`

Reference 2:

`It is the guiding principle which guarantees the military forces always being under the command of the Party.`

Reference 3:

`It is the practical guide for the army always to heed the directions of the party.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but:

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{1+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{2+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{3+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{4+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{5 +}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{6+}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Precision-based measure, but: $\text{Prec.} = \frac{7}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

A reference word should only be matched once.

Algorithm:

1. Count number of times w_i occurs in each reference.
2. Keep the minimun between the maximum of (1) and the number of times w_i appears in the candidate (*clipping*).
3. Add these values and divide by candidate's number of words.

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Modified 1-gram precision:

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}=2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 =$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \rightarrow \text{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\text{Max}_{(1)}=2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 = \frac{2}{}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}{=}2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision (1-gram)

Modified 1-gram precision: $P_1 = \frac{2}{7}$

Candidate:

`The the the the the the the.`

Reference 1:

`The cat is on the mat.`

Reference 2:

`There is a cat on the mat.`

1. $w_i \rightarrow \mathrm{The}$
   $\#_{w_i,R1} = 2$
   $\#_{w_i,R2} = 1$
2. $\mathrm{Max}_{(1)}=2$, $\#_{w_i,C} = 7$
   $\Rightarrow$ Min=2
3. No more distinct words

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Modified n-gram precision

- Straightforward generalisation to n-grams, P_n.
- Generalisation to multiple sentences:

$$\mathrm{P}_n = \frac{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} \mathit{Count}_{\text{clipped}}(n\text{gram})}{\sum_{C \in \{\text{candidates}\}} \sum_{n\text{gram} \in C} \mathit{Count}(n\text{gram})}$$

low n	high n
adequacy	fluency

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

BiLingual Evaluation Understudy, BLEU

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log \mathrm{P}_n\right)$$

- Geometric average of P_n (empirical suggestion).
- w_n positive weights summing to one.
- Brevity penalty.

IBM BLEU: Papineni, Roukos, Ward and Zhu [2001]

Paper's Conclusions

- 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

Automatic evaluations 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

Automatic evaluations 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

1. System overtuning → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture system improvements (e.g., JHU'03)
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic 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 (e.g., JHU'03)
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

Risks of Automatic 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 (e.g., JHU'03)
3. Unfair system comparisons → when metrics are unable to reflect difference in quality between MT systems

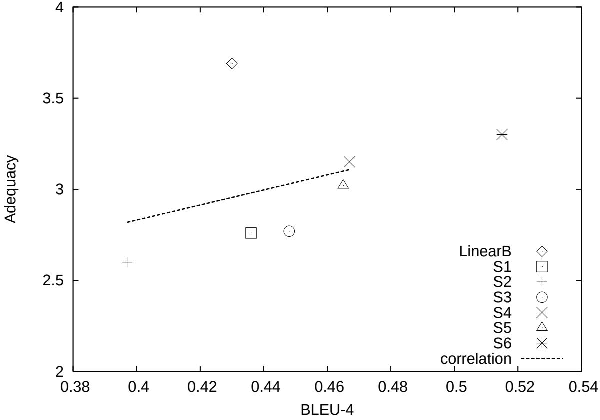
Risks of Automatic Evaluation

1. System overtuning → when system parameters are adjusted towards a given metric
2. Blind system development → when metrics are unable to capture system improvements (e.g., JHU'03)
3. **Unfair system comparisons** → when metrics are unable to reflect difference in quality between MT systems

Talk Overview

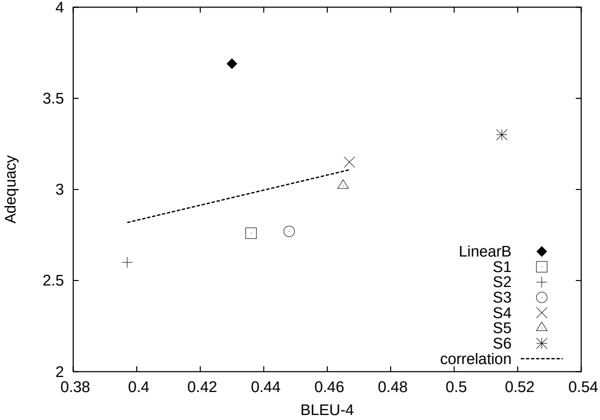
Problems of Lexical Similarity Measures

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

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
Sentence #498

Automatic Translation (LinearB)	On Tuesday several missiles and mortar shells fell in southern Israel , but there were no casualties .
Reference Translation	Several Qassam rockets and mortar shells fell today, Tuesday , in southern Israel without causing any casualties .

Only one 4-gram in common!

Problems of Lexical Similarity Measures

NIST 2005 Arabic-to-English Exercise
Sentence #498

Automatic Translation (LinearB)	On Tuesday several missiles **and mortar shells fell** in southern Israel , but there were no casualties .
Reference Translation	Several Qassam rockets **and mortar shells fell** today, Tuesday , in southern Israel without causing any casualties .

Only one 4-gram in common!

The Limits of Lexical Similarity

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

- Culy and Riehemann [CR03]
- Coughlin [Cou03]

Underlying Cause

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

Talk Overview

Extending Lexical Similarity Measures

Increase robustness (avoid sparsity):

- Lexical variants
  - → Morphological variations (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]

Similarity Measures Based on Linguistic Features

- Syntactic Similarity
  - → Shallow Parsing
    - Popovic and Ney [PN07]
    - Giménez and Màrquez [GM07]
  - → Constituency Parsing
    - Liu and Gildea [LG05]
    - Giménez and Màrquez [GM07]
  - → Dependency Parsing
    - Liu and Gildea[LG05]
    - Amigó et al. [AGGM06]
    - Mehay and Brew [MB07]
    - Owczarzak et al. [OvGW07a, OvGW07b]
    - Kahn et al. [KSO09]
    - Chan and Ng [CN08]

Similarity Measures Based on Linguistic Features

- Semantic Similarity
  - → Named Entities

    Reeder et al. [RMDW01]
    Giménez and Màrquez [GM07]

  - → Semantic Roles

    Giménez and Màrquez [GM07]

  - → Textual Entailment

    Padó et al. [PCGJM09]

  - → Discourse Representations

    Giménez and Màrquez [GM09]

Our Approach (Giménez & Màrquez, 2010)

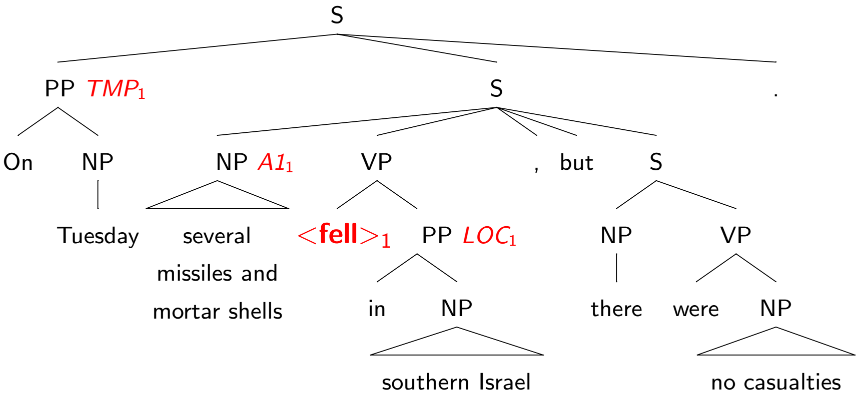
- Rather than comparing sentences at lexical level:

> Compare the linguistic structures and the words within them

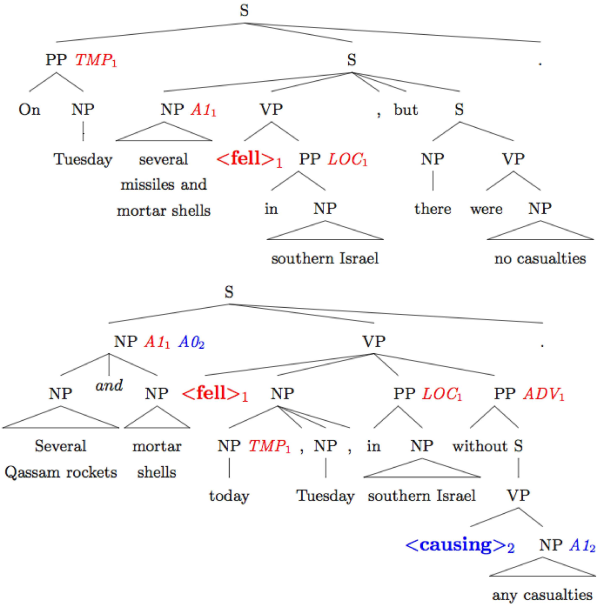
Our Approach

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

Our Approach



Our Approach



Measuring Structural Similarity

- 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

- Generic Similarity measure among LEs: OVERLAP
  Inspired by the Jaccard similarity coefficient

  Precision/Recall/F_1 can also be used

Measuring Structural Similarity

- 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

- Generic Similarity measure among LEs: OVERLAP
  Inspired by the Jaccard similarity coefficient
  Precision/Recall/F_1 can also be used

Measuring Structural Similarity

- 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
- Generic Similarity measure among LEs: OVERLAP

  Inspired by the Jaccard similarity coefficient

  Precision/Recall/F_1 can also be used

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 among Linguistic Elements

- The overlap measures can be instantiated at all levels of linguistic information to provide concrete similarity measures
- Lexical overlap over word forms O_l
- Average lexical overlap among semantic roles: SR-$O_r - (*)$

Example: Lexical Overlaping

hyp on **tuesday several** missiles **and mortar shells fell in southern israel ,** but there were no **casualties .**

ref **several** qassam rockets **and mortar shells fell** today , **tuesday , in southern israel** without causing any **casualties .**

$$hyp \cap ref = \{ \text{'tuesday', 'several', 'and', 'mortar', 'shells', 'fell', 'in', 'southern', 'israel', ',', 'casualties', '.'} \}$$

$$hyp \cup ref = \{ \text{'on', 'tuesday', 'several', 'missiles', 'and', 'mortar', 'shells', 'fell', 'in', 'southern', 'israel', ',', 'but', 'there', 'were', 'no', 'casualties', '.', 'qassam', 'rockets', 'today', ',', 'without', 'causing', 'any'} \}$$

$$O_l = \frac{|hyp \cap ref|}{|hyp \cup ref|} = \frac{12}{25} \qquad P = \frac{|hyp \cap ref|}{|hyp|} = \frac{12}{18} \qquad R = \frac{|hyp \cap ref|}{|ref|} = \frac{12}{19}$$

Example: Average lexical overlaping among semantic roles

hyp_{A1} = { **‘several’**, ‘missiles’, **‘and’**, **‘mortar’**, **‘shells’** }
ref_{A1} = { ‘several’, ‘qassam’, ‘rockets, ‘and’, ‘mortar’, ‘shells’, ‘any’, ‘casualties’ }

$\text{hyp}_{\text{A0}} = \emptyset$
ref_{A0} = { ‘several’, ‘qassam’, ‘rockets, ‘and’, ‘mortar’, ‘shells’ }
hyp_{TMP} = { ‘on’, ‘tuesday’ }
ref_{TMP} = { ‘today’ }
hyp_{LOC} = { **‘in’**, **‘southern’**, **‘israel’** }
ref_{LOC} = { ‘in’, ‘southern’, ‘israel’ }
$\text{hyp}_{\text{ADV}} = \emptyset$
ref_{ADV} = { ‘without’, ‘causing’, ‘any’, ‘casualties’ }

$\text{SR-}O_r(\text{A1}) = \frac{4}{9}$ $\text{SR-}O_r(\text{TMP}) = \frac{0}{3}$ $\text{SR-}O_r(\text{ADV}) = \frac{0}{4}$
$\text{SR-}O_r(\text{A0}) = \frac{0}{6}$ $\text{SR-}O_r(\text{LOC}) = \frac{3}{3}$

$$\text{SR-}O_r(\star) = \frac{4+0+0+3+0}{9+6+3+3+4} = \frac{7}{25} = 0.28$$

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict measure
  - → All items inside an element are considered the same unit
  - → Computes the proportion of fully translated LEs, according to their types
- Overlap and Matching have been instantiated over different linguistic level elements (for Englsih)
  - → Words, lemmas, POS
  - → Shallow, dependency and constituency parsing
  - → Named entities and semantic roles
  - → Discourse representation (logical forms)
- Freely available software: IQ$_{MT}$ framework
  http://www.lsi.upc.es/∼nlp/IQMT/

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict measure
  - → All items inside an element are considered the same unit
  - → Computes the proportion of fully translated LEs, according to their types
- Overlap and Matching have been instantiated over different linguistic level elements (for Englsih)
  - → Words, lemmas, POS
  - → Shallow, dependency and constituency parsing
  - → Named entities and semantic roles
  - → Discourse representation (logical forms)
- Freely available software: IQ$_{\text{MT}}$ framework
  http://www.lsi.upc.es/∼nlp/IQMT/

Overlap/Matching among Linguistic Elements

- Matching is a similar but more strict measure
  - → All items inside an element are considered the same unit
  - → Computes the proportion of fully translated LEs, according to their types
- Overlap and Matching have been instantiated over different linguistic level elements (for Englsih)
  - → Words, lemmas, POS
  - → Shallow, dependency and constituency parsing
  - → Named entities and semantic roles
  - → Discourse representation (logical forms)
- Freely available software: IQ$_{\text{MT}}$ framework
  http://www.lsi.upc.es/∼nlp/IQMT/

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

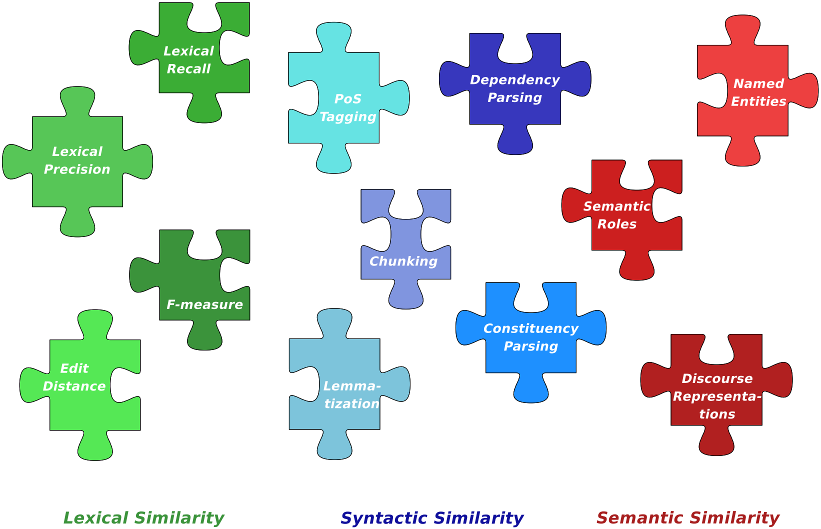
Overlap vs. F_1

NIST 2005 Arabic-to-English Exercise

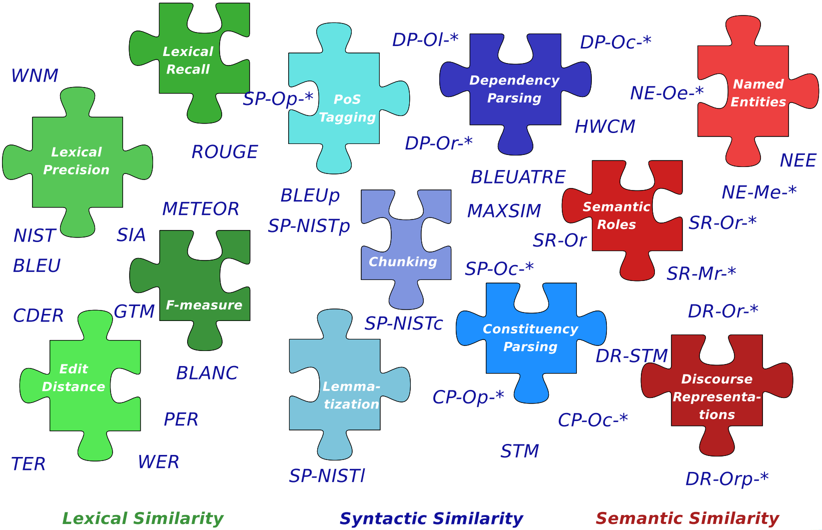
	Measure	Spearman ρ	Pearson r	SMT Pearson r
Overlap	O_l	0.3561	0.0464	0.8460
	SR-$O_r(\star)$	0.7901	0.6719	0.9087
	SR-$M_r(\star)$	0.8242	0.7887	0.8966
	DR-$O_r(\star)$	0.7901	0.6243	0.9336
	DR-$O_{rp}(\star)$	1.0000	0.8932	0.9718
F_1	O_l	0.3561	0.0283	0.8386
	SR-$O_r(\star)$	0.7901	0.6675	0.9057
	SR-$M_r(\star)$	0.7022	0.7658	0.8812
	DR-$O_r(\star)$	0.7022	0.5700	0.9082
	DR-$O_{rp}(\star)$	1.0000	0.9092	0.9751

Talk Overview

Towards Heterogeneous Automatic MT Evaluation



Towards Heterogeneous Automatic MT Evaluation



MOLTO

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 [AH07]
- 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 [AH07]
- 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', 'DP-HWC_r', '$\text{DP-O}_c(\star)$', '$\text{DP-O}_l(\star)$', '$\text{DP-O}_r(\star)$', 'CP-STM_4', '$\text{SR-O}_r(\star)$', 'SR-O_{rv}', '$\text{DR-O}_{rp}(\star)$' }

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$_{\mathrm{snt}}$**
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 Domains

NIST 2004/2005 MT Evaluation Campaigns

	AE$_{2004}$	**CE**$_{2004}$	**AE**$_{2005}$	**CE**$_{2005}$
#references	5	5	5	4
#outputs$_{\mathrm{ass.}}$	5/5	10/10	6/7	5/10
#sentences$_{\mathrm{ass.}}$	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 Domains

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 over Low-quality Translations

IWSLT 2006 MT Evaluation Campaign (Chinese-to-English)

	CRR	**ASR$_r$**	**ASR$_s$**
#references	7	7	7
#outputs$_{\text{ass.}}$	6/14	6/14	6/13
#sentences$_{\text{ass.}}$	400/500	400/500	400/500
Avg. Adequacy	1.40/5	1.02/5	0.93/5
Avg. Fluency	1.16/5	0.98/5	0.98/5

Linguistic Measures over Low-quality Translations

IWSLT 2006 MT Evaluation Campaign (Chinese-to-English)

Similarity	Measure	CRR	ASR_r	ASR_s
	1-WER	0.4737	0.5029	0.4814
	BLEU	0.5401	0.5337	0.5187
	NIST	0.5275	0.5348	0.5269
Lexical	O_l	0.5679	0.6166	0.5830
	GTM_2	**0.6211**	**0.6410**	**0.6117**
	$ROUGE_W$	0.5815	0.6048	0.5812
	METEOR	0.4373	0.4964	0.4798
	ULC	0.4956	0.5137	0.5270
	ULC_{opt}	0.6406	0.6688	0.6371

Linguistic Measures over Low-quality Translations

IWSLT 2006 MT Evaluation Campaign (Chinese-to-English)

Similarity	**Measure**	**CRR**	**ASR$_r$**	**ASR$_s$**
	1-WER	0.4737	0.5029	0.4814
	BLEU	0.5401	0.5337	0.5187
	NIST	0.5275	0.5348	0.5269
Lexical	O_l	0.5679	0.6166	0.5830
	GTM$_2$	**0.6211**	**0.6410**	**0.6117**
	ROUGE$_W$	0.5815	0.6048	0.5812
	METEOR	0.4373	0.4964	0.4798
	ULC	0.4956	0.5137	0.5270
	ULC$_{opt}$	0.6406	0.6688	0.6371

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
  - → 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

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

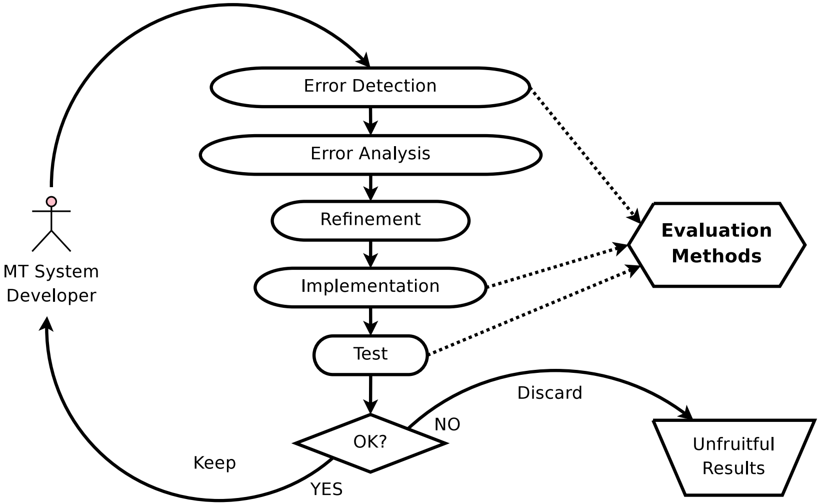
1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures
2. Work on the combination of measures
   → Learning combined similarity measures
3. Porting measures to languages other than English
   → Need of linguistic analyzers
4. Use measures for semi–automatic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

Ongoing and Future Work

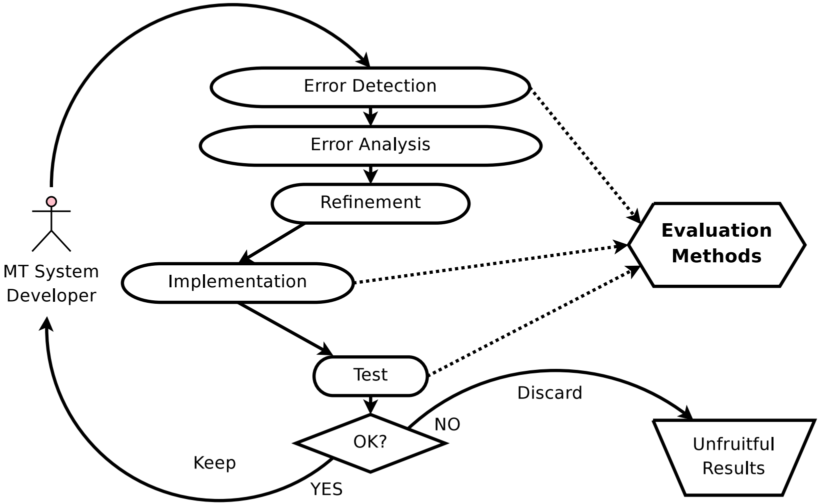
1. Metaevaluation of measures
   → Better understand differences between lexical and higher level measures
2. Work on the combination of measures
   → Learning combined similarity measures
3. Porting measures to languages other than English
   → Need of linguistic analyzers
4. Use measures for semi–automatic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

Talk Overview

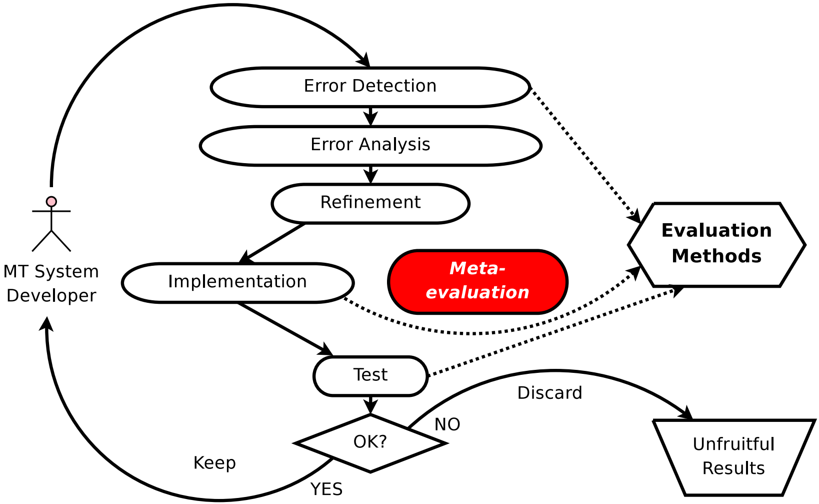
Metricwise System Development



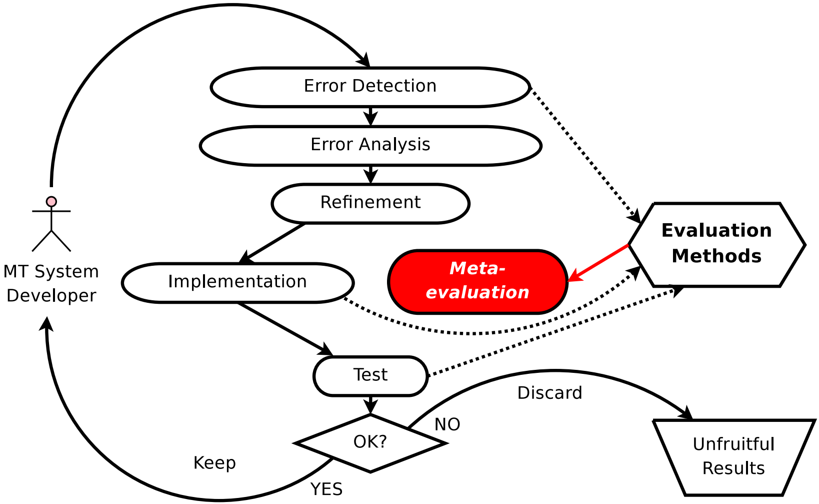
Metricwise System Development



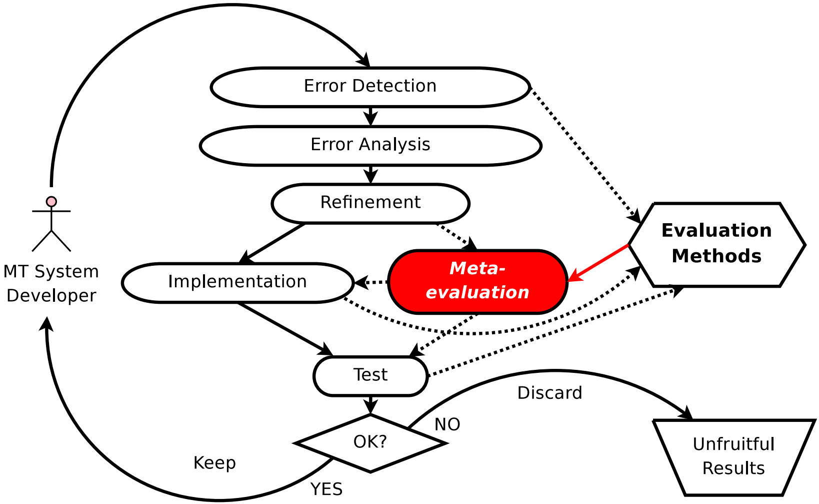
Metricwise System Development



Metricwise System Development



Metricwise System Development



Summary and Recommendations

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 and Recommendations

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 and Recommendations

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 and Recommendations

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 and Recommendations

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 and Recommendations

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 and Recommendations

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 Similarity Measures Based on Multiple Linguistic Layers

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

MOLTO workshop – **GF meets SMT**
Göteborg, November 5, 2010

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.

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.

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.

- 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.