Empirical Machine Translation and its Evaluation

Invited Talk at the
Statistical Multilingual Analysis for
Retrieval and Translation Workshop 2009

Jesús Giménez

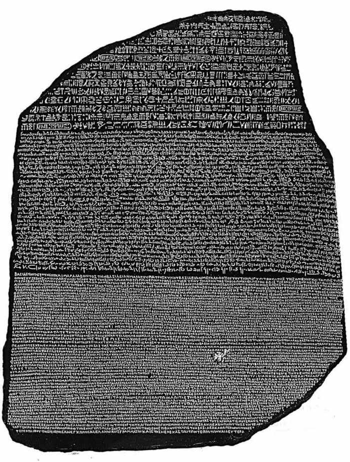
Grup de Processament del Llenguatge Natural
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

May 13, 2009

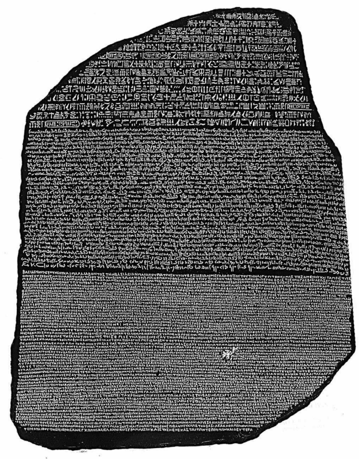
Outline

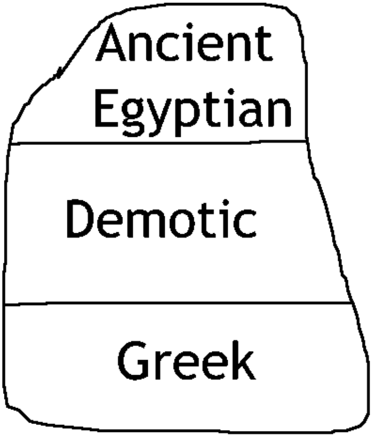
1. Empirical Machine Translation
   - Statistical Machine Translation
2. How are Empirical MT Systems Developed Today?
3. Evaluation Methods
4. Tackling the Negative Effects of Automatic Evaluation
5. Morals on This Story

Empirical Machine Translation



Empirical Machine Translation





Empirical Machine Translation

"a royal offering of Osiris, Foremost of the Westerners, the Great God, Lord of Abydos; and of Wepwawet, Lord of the Sacred Land"

Outline

Statistical Machine Translation

Translation is modeled as a decision process which may be addressed through a search over a probability space.

Statistical Machine Translation

Translation is modeled as a decision process which may be addressed through a search over a probability space.

Decision Types:

1. **Partition**
2. **Word Selection**
3. **Word Ordering**

Statistical Machine Translation

Translation is modeled as a decision process which may be addressed through a search over a probability space.

Decision Types:

1. **Partition**
   Decompose input sentence into smaller translation units
2. **Word Selection**
3. **Word Ordering**

Statistical Machine Translation

Translation is modeled as a decision process which may be addressed through a search over a probability space.

Decision Types:

1. **Partition**
   Decompose input sentence into smaller translation units
2. **Word Selection**
   Translate these units into the target language
3. **Word Ordering**

Statistical Machine Translation

Translation is modeled as a decision process which may be addressed through a search over a probability space.

Decision Types:

1. **Partition**
   Decompose input sentence into smaller translation units
2. **Word Selection**
   Translate these units into the target language
3. **Word Ordering**
   Reorder translated units

Why is SMT so Popular?

1. Theoretically well founded
2. A mighty baseline
3. Room for improvement
   - Competitive results may be attained without using any additional linguistic information further than lexical
4. Easy to build a state-of-the-art prototype system
   - Freely available components (e.g., GIZA++, SRILM, Pharaoh, MOSES, ...)

Why is SMT so Popular?

1. Theoretically well founded
2. A mighty baseline
3. Room for improvement
   - Competitive results may be attained without using any additional linguistic information further than lexical
4. Easy to build a state-of-the-art prototype system
   - Freely available components (e.g., GIZA++, SRILM, Pharaoh, MOSES, ...)

Why is SMT so Popular?

1. Theoretically well founded
2. A mighty baseline
3. Room for improvement
   - Competitive results may be attained without using any additional linguistic information further than lexical
4. Easy to build a state-of-the-art prototype system
   - Freely available components (e.g., GIZA++, SRILM, Pharaoh, MOSES, ...)

Why is SMT so Popular?

1. Theoretically well founded
2. A mighty baseline
3. Room for improvement
    - Competitive results may be attained without using any additional linguistic information further than lexical
4. Easy to build a state-of-the-art prototype system
    - Freely available components (e.g., GIZA++, SRILM, Pharaoh, MOSES, ...)

Current Trends in SMT

Linguistic Knowledge

+

Machine Learning

Current Trends in SMT

Linguistic Knowledge

$+$

Machine Learning

Current Trends in SMT

Linguistic Knowledge

$+$

Machine Learning

Current Trends in SMT

- **Word Ordering**
- **Word Selection**

Current Trends in SMT

- **Word Ordering**
  - Syntax-based translation
    - Bilingual parsing
    - Syntactic transfer
  - Dedicated discriminative models
  - A priori source reordering
  - Factored language models
- **Word Selection**

Current Trends in SMT

- **Word Ordering**
  - Syntax-based translation
    - Bilingual parsing
    - Syntactic transfer
  - Dedicated discriminative models
  - A priori source reordering
  - Factored language models
- **Word Selection**
  - Factored translation models
  - Dedicated discriminative models

Current Trends in SMT

- **Post-processing**
- **Hybridization**
- **Alternative End-to-end Architectures**

Current Trends in SMT

- **Post-processing**
    - Discriminative reranking of *n*-best lists
    - System output combination
- **Hybridization**
- **Alternative End-to-end Architectures**

Current Trends in SMT

- **Post-processing**
  - Discriminative reranking of *n*-best lists
  - System output combination
- **Hybridization**
  - RBMT and SMT (e.g., statistical post-editing)
- **Alternative End-to-end Architectures**

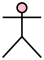
Current Trends in SMT

- **Post-processing**
  - Discriminative reranking of *n*-best lists
  - System output combination
- **Hybridization**
  - RBMT and SMT (e.g., statistical post-editing)
- **Alternative End-to-end Architectures**
  - Global on-line learning
    - Tillmann and Zhang (2006) [TZ06]
    - Liang et al. (2006) [LBCKT06]
    - Arun and Koehn (2007) [AK07]

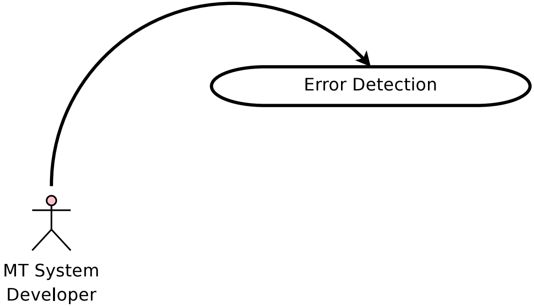
Outline

The Current System Development Cycle

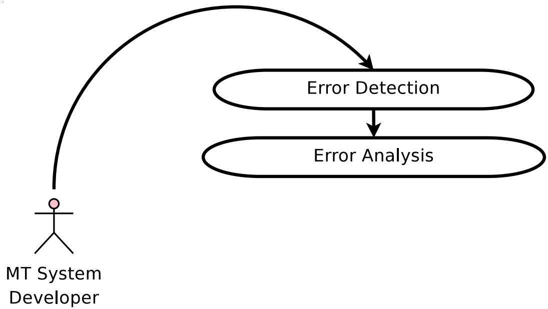
The Current System Development Cycle



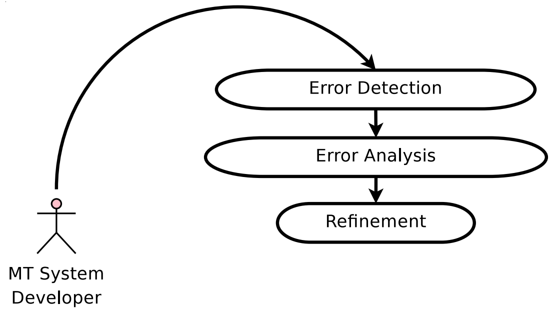
The Current System Development Cycle



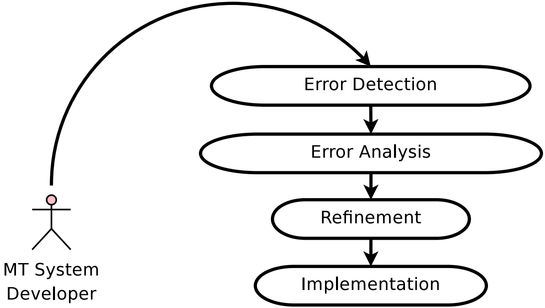
The Current System Development Cycle



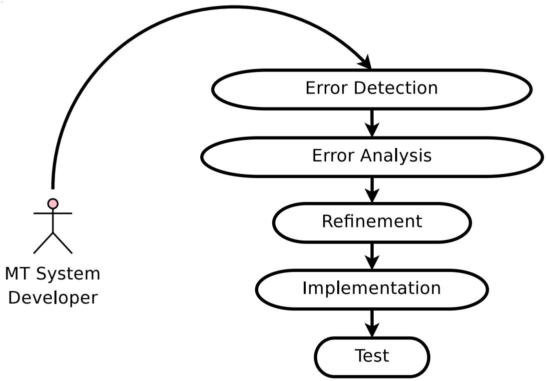
The Current System Development Cycle



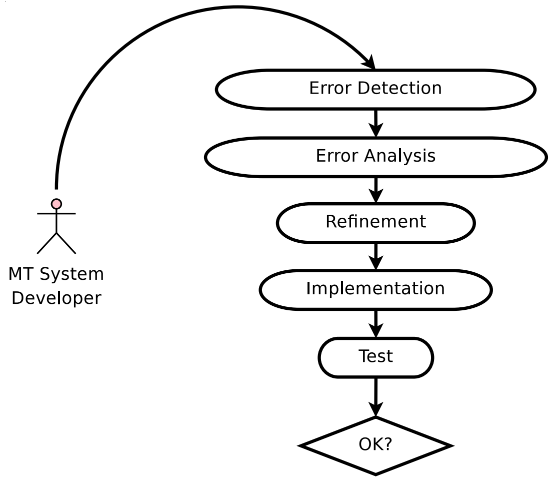
The Current System Development Cycle



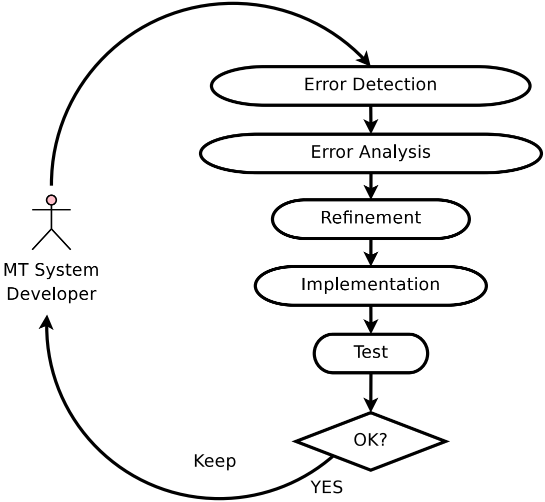
The Current System Development Cycle



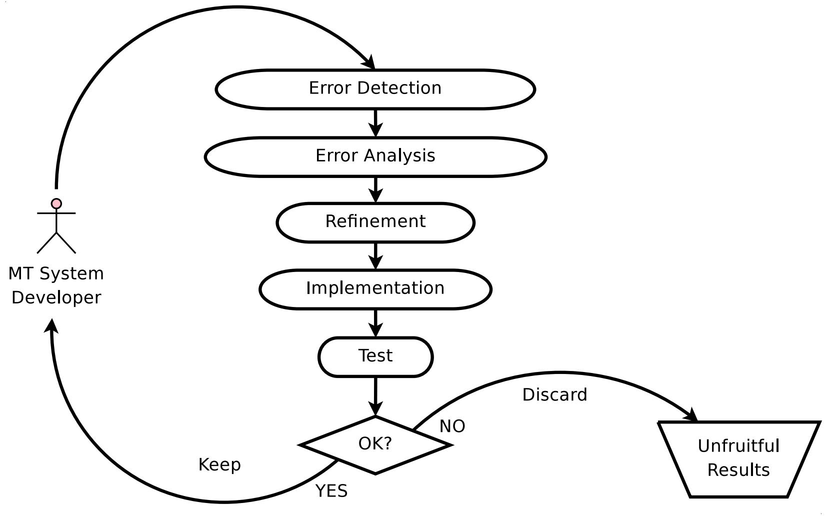
The Current System Development Cycle



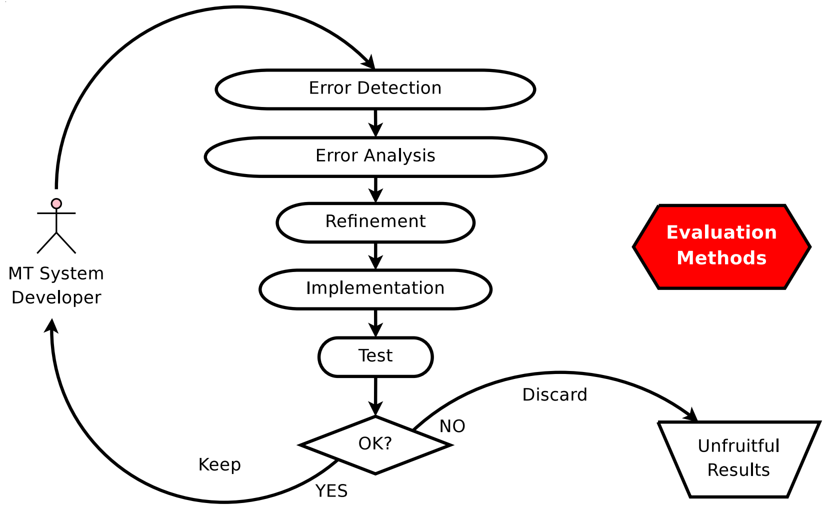
The Current System Development Cycle



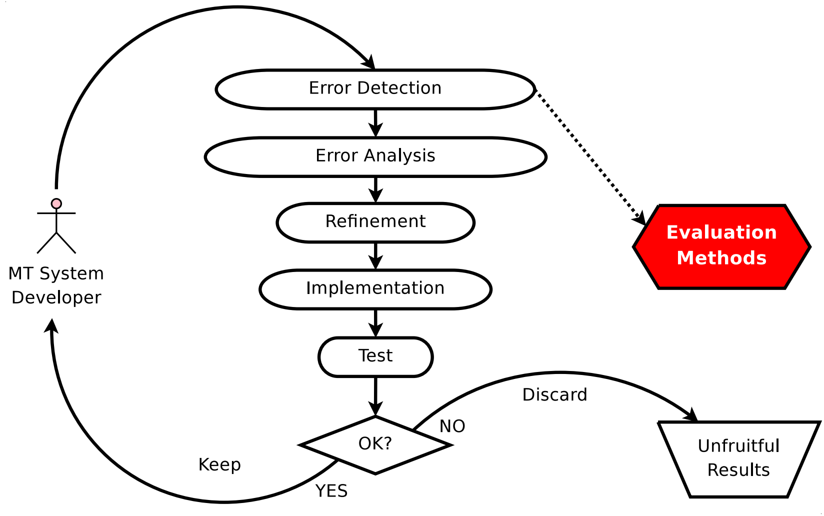
The Current System Development Cycle



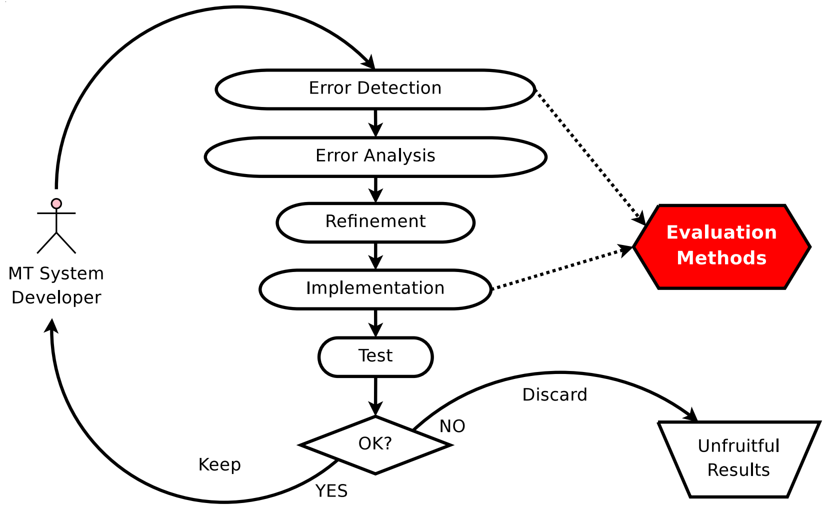
The Current System Development Cycle



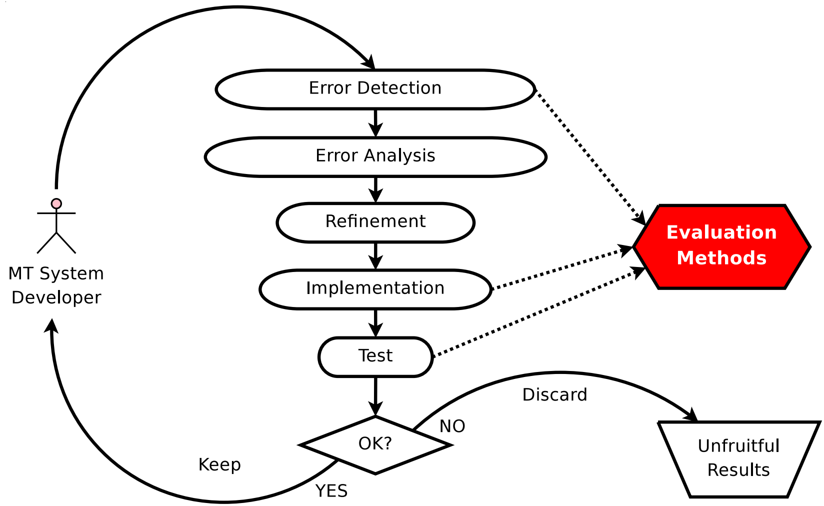
The Current System Development Cycle



The Current System Development Cycle



The Current System Development Cycle



Outline

1. Empirical Machine Translation
2. How are Empirical MT Systems Developed Today?
3. Evaluation Methods
   - Manual Evaluation
   - Automatic Evaluation
   - The Apple Collection Metaphore
4. Tackling the Negative Effects of Automatic Evaluation
5. Morals on This Story

Outline

ALPAC Approach (1966)

- **Fidelity** (or Accuracy) — (measured on a 0-9 scale)
  how much information is retained by the translated sentence compared to the original?
- **Intelligibility** — (measured on a 1-9 scale)
  how 'understandable' is the automatic translation?

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

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Other Manual Measures

- Comprehension Evaluation
- Cloze Test (blank-filling)
- Read Time
- Required Post-Editing (measured on key strokes)
- Post-Edit Time
- Meaning Maintenance (measured on a 1-5 scale)
- Clarity (measured on a 0-3 scale)
- Preferred Translation
- Quality Panel Evaluation

Pros and Cons of Manual Evaluation

Advantages	**Disadvantages**

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

Pros and Cons of Manual Evaluation

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

Outline

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

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

Benefits of Automatic Evaluation

- Automatic evaluations are:
  - Costless (vs. costly)
  - Objective (vs. subjective)
  - 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:
  - Costless (vs. costly)
  - Objective (vs. subjective)
  - 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

Negative Consequences of Automatic Evaluation

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

Negative Consequences of Automatic Evaluation

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

Negative Consequences of Automatic Evaluation

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

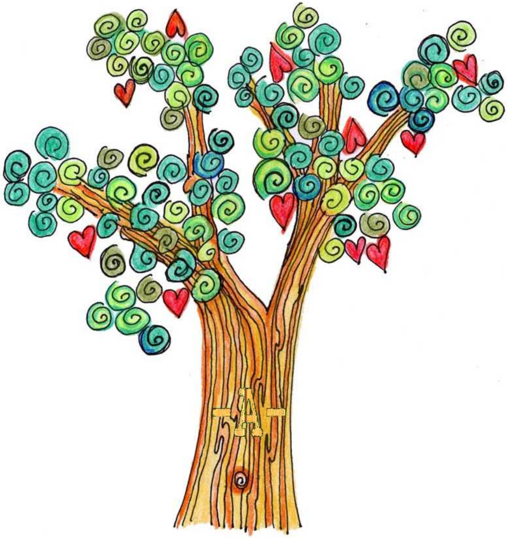
Negative Consequences of Automatic Evaluation

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

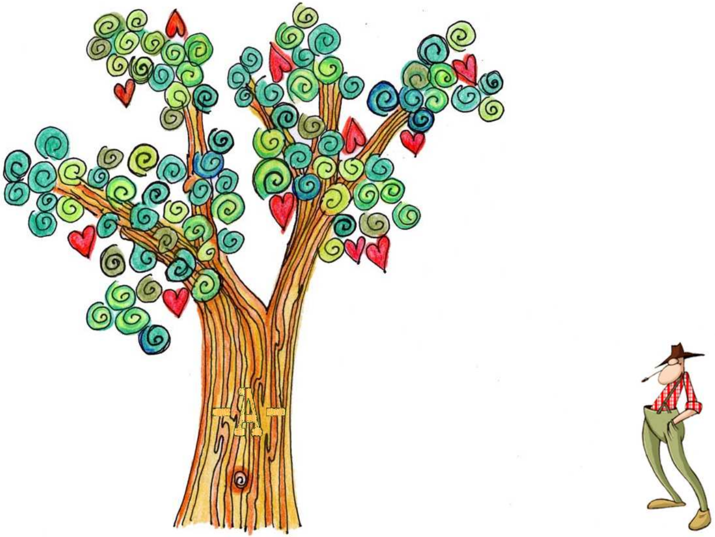
Outline

1. Empirical Machine Translation
2. How are Empirical MT Systems Developed Today?
3. Evaluation Methods
   - Manual Evaluation
   - Automatic Evaluation
   - The Apple Collection Metaphore
4. Tackling the Negative Effects of Automatic Evaluation
5. Morals on This Story

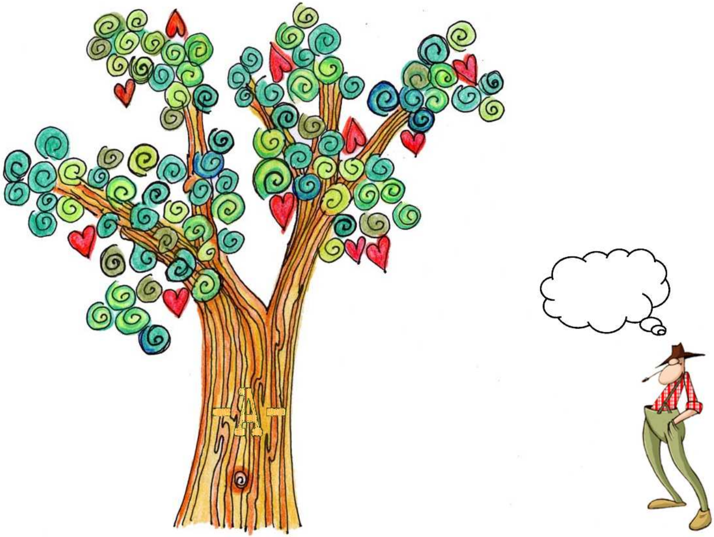
The Problem of Apple Collection (AC)



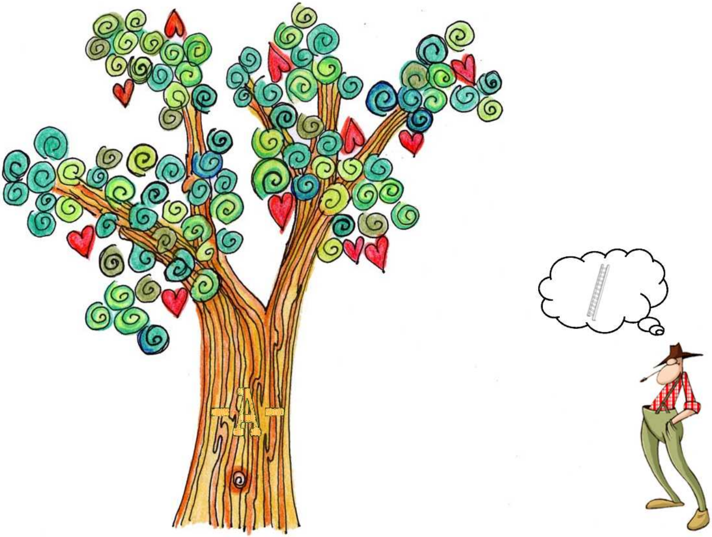
The Problem of Apple Collection (AC)



A State-of-the-Art Empirical AC System



A State-of-the-Art Empirical AC System



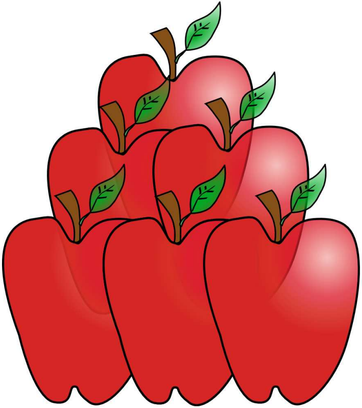
A State-of-the-Art Empirical AC System

The Apple Store

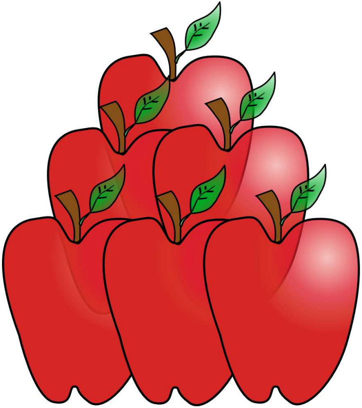
The Apple Store

The Apple Store

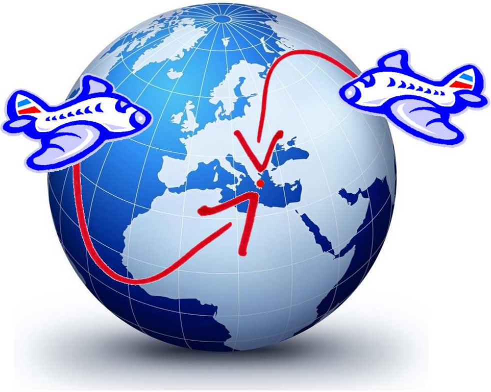
AC Evaluation



AC Evaluation



International AC Evaluation Campaign

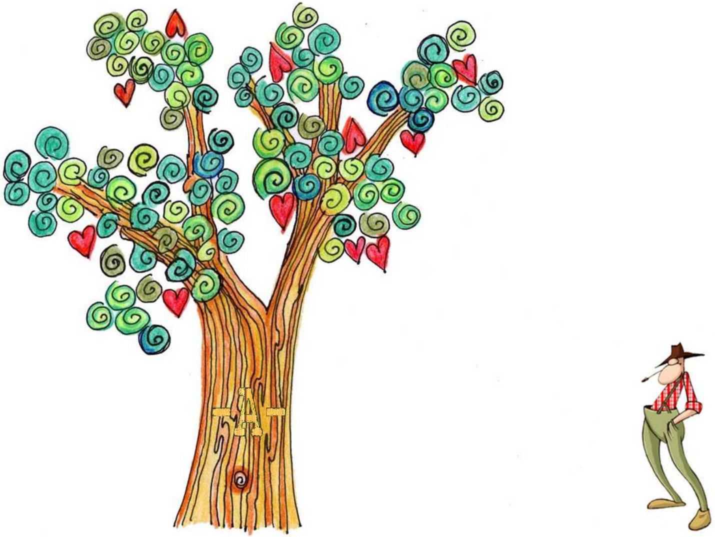


Ladder-based AC Systems

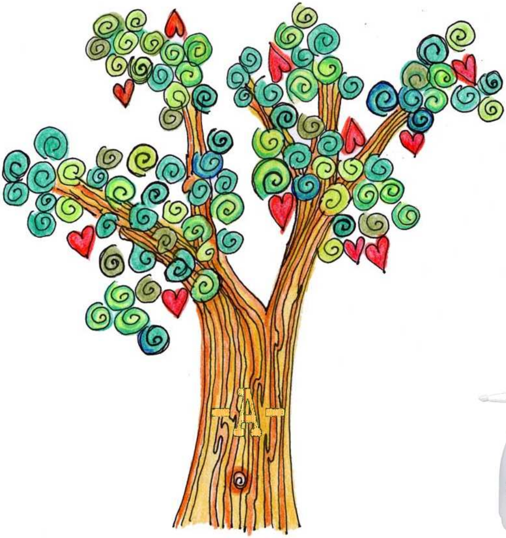
Ladder/Basket-based Hybrid AC

Ladder/Basket-based Hybrid AC

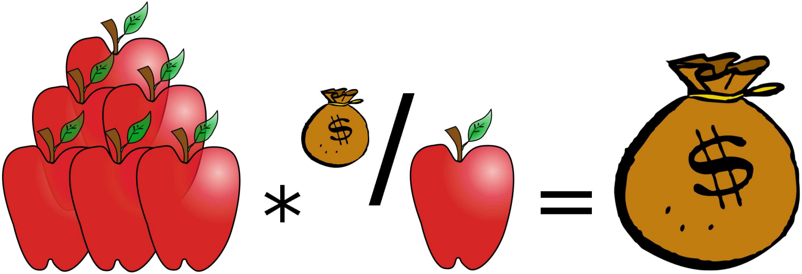
Fertilization Techniques for AC



Fertilization Techniques for AC



AC Evaluation (at the Farm)

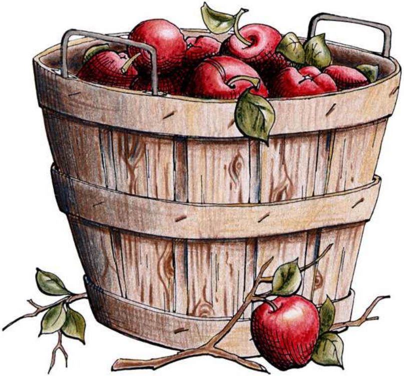


AC Evaluation (at the Apple Store)

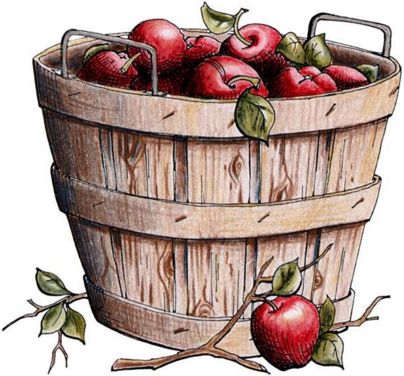
AC Evaluation (at the Apple Store)

AC Evaluation (at the Apple Store)

AC Evaluation (at the Apple Store)

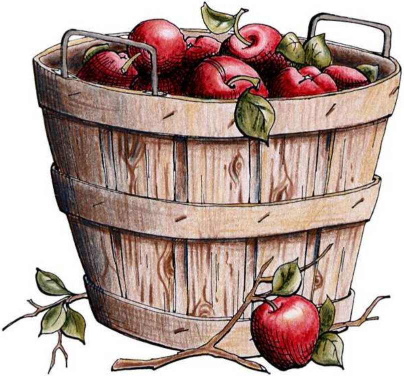


AC Evaluation (at the Apple Store)



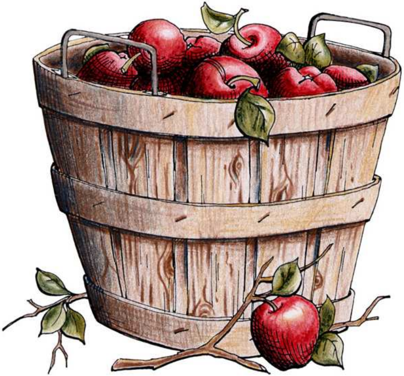
size

AC Evaluation (at the Apple Store)



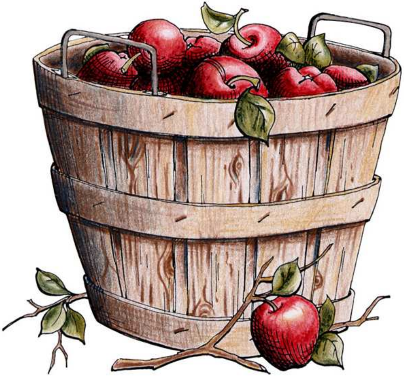
size
color

AC Evaluation (at the Apple Store)



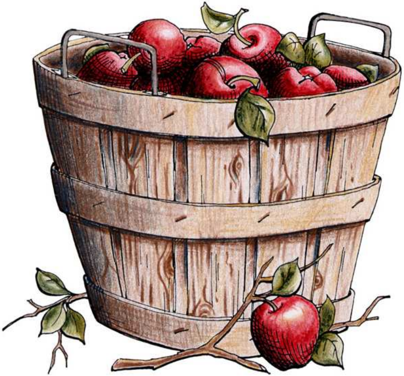
size
color
shape

AC Evaluation (at the Apple Store)



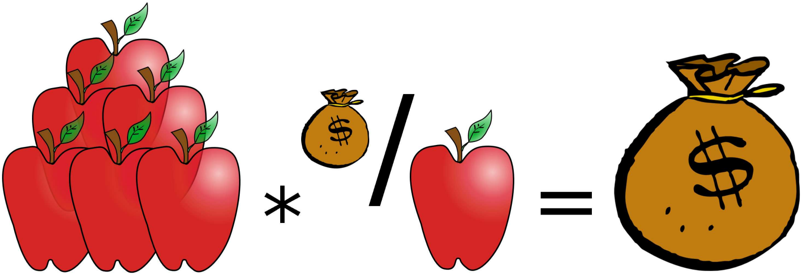
size
color
shape
taste

AC Evaluation (at the Apple Store)

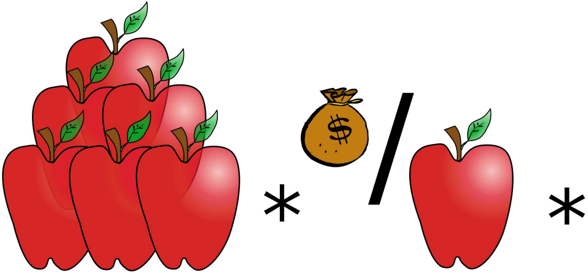


size
color
shape
taste
flavor

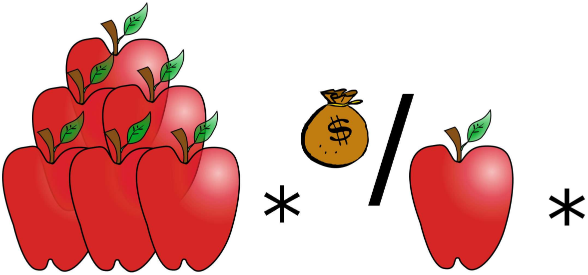
AC Evaluation (at the Apple Store)



AC Evaluation (at the Apple Store)

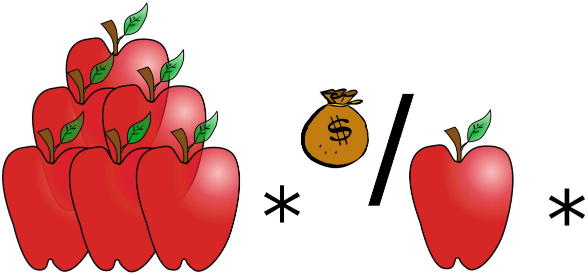


AC Evaluation (at the Apple Store)



Q(size, color, shape, test, flavor, ...)

AC Evaluation (at the Apple Store)

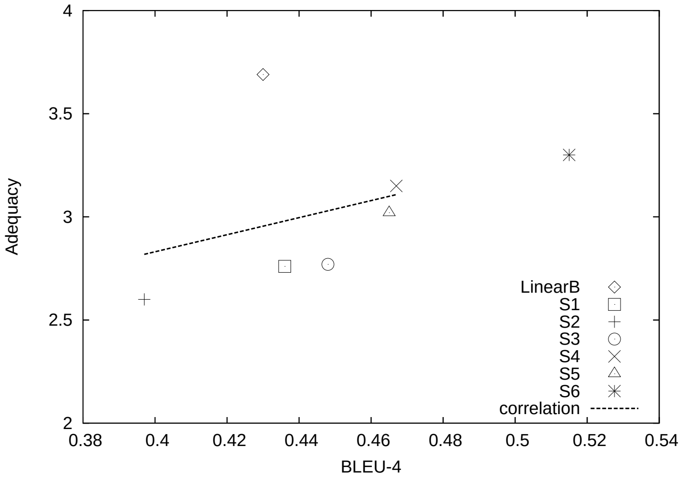


Q(size, color, shape, test, flavor, ...) =

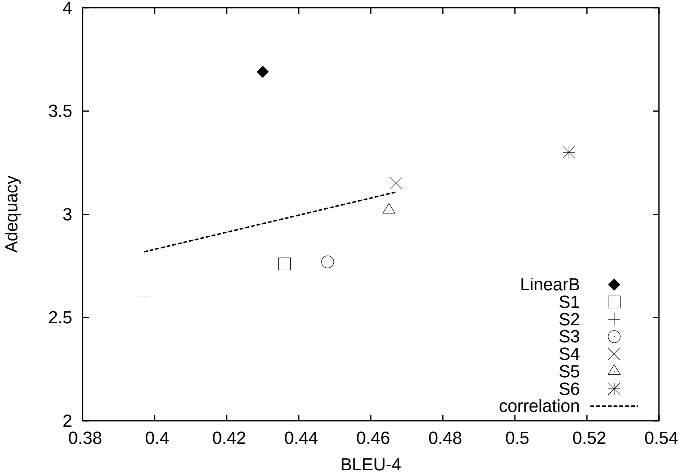
Outline

Outline

NIST 2005 Arabic-to-English Exercise



NIST 2005 Arabic-to-English Exercise



NIST 2005 Arabic-to-English Exercise

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 .

Only one 4-gram in common!

NIST 2005 Arabic-to-English Exercise

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 .

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.

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.

Extending the Reference Material

- 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]

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]

Linguistic Features

- Semantic Similarity
  - Named Entities
    - Reeder et al. [RMDW01]
    - Giménez and Màrquez [GM07]
  - Semantic Roles
    - Giménez and Màrquez [GM07]
  - Discourse Representations
    - Giménez and Màrquez [GM09]

Linguistic 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

Linguistic 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

Linguistic 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

Linguistic 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

Linguistic Features at International Campaigns

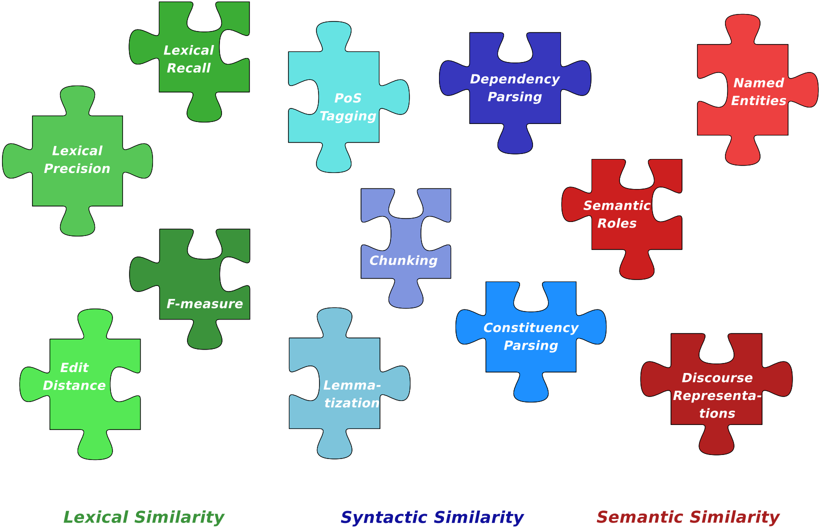
- NIST 2004/2005
  - Arabic-to-English / Chinese-to-English
  - Broadcast news / weblogs / dialogues
- WMT 2007-2009
  - Translation between several European languages
  - European Parliament Proceedings / Out-of-domain News
- IWSLT 2005-2008
  - Spoken language translation
  - Chinese-to-English

Linguistic Features at International Campaigns

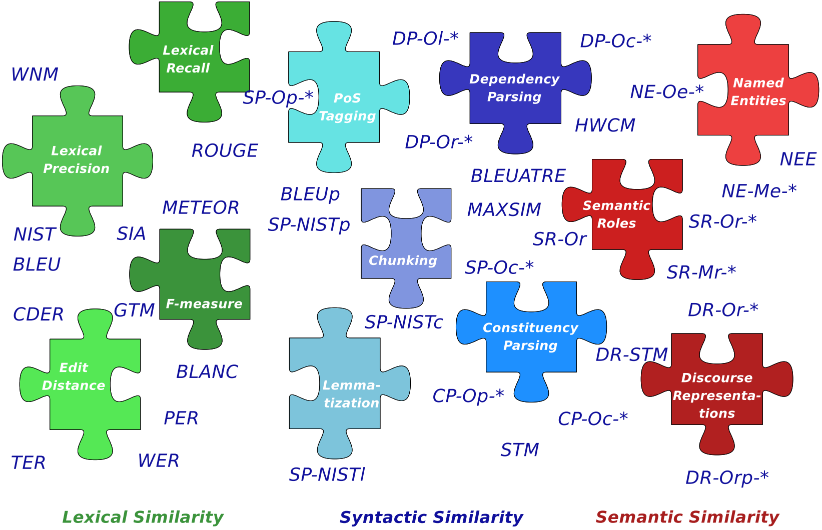
- NIST 2004/2005
  - Arabic-to-English / Chinese-to-English
  - Broadcast news / weblogs / dialogues
- WMT 2007-2009
  - Translation between several European languages
  - European Parliament Proceedings / Out-of-domain News
- IWSLT 2005-2008
  - Spoken language translation
  - Chinese-to-English

Controversial results at the NIST Metrics MATR08 Challenge!

Towards Heterogeneous Automatic MT Evaluation



Towards Heterogeneous Automatic MT Evaluation

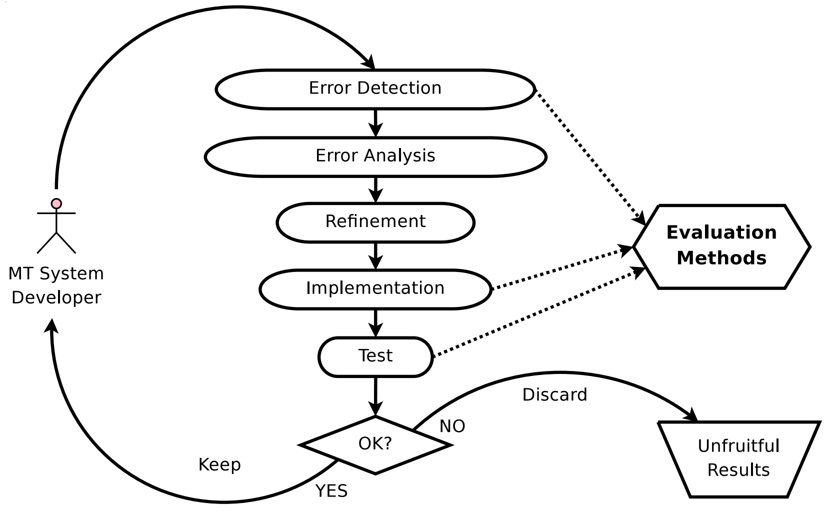


Recent Works on Metric Combination

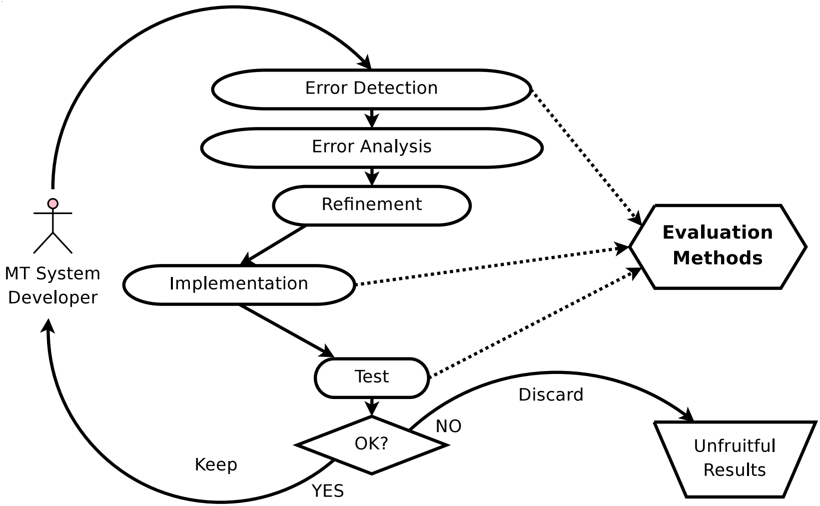
- Corston-Oliver et al. [COGB01]
- Kulesza and Shieber [KS04]
- Gamon et al. [GAS05]
- Akiba et al. [AIS01]
- Quirk [Qui04]
- Liu and Gildea [LG07]
- Albrecht and Hwa [AH07]
- Paul et al. [PFS07]
- Ye et al. [YZL07]
- Giménez and Màrquez [GM08]

Outline

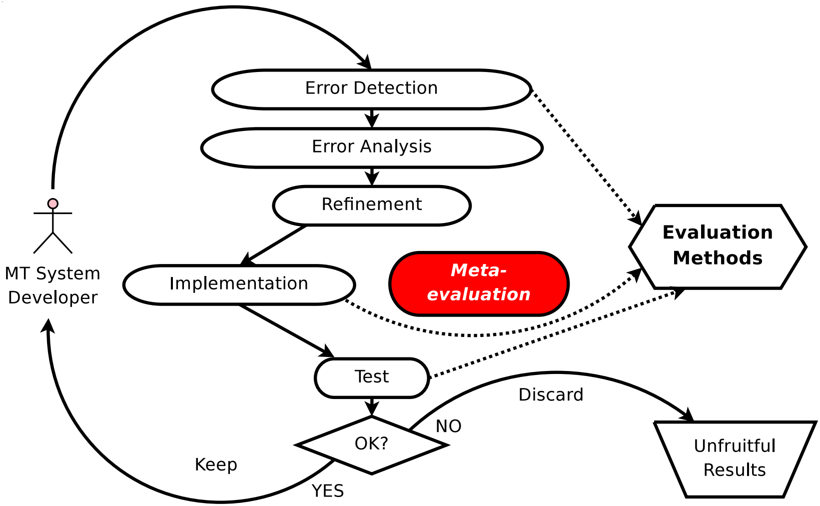
Metric Selection



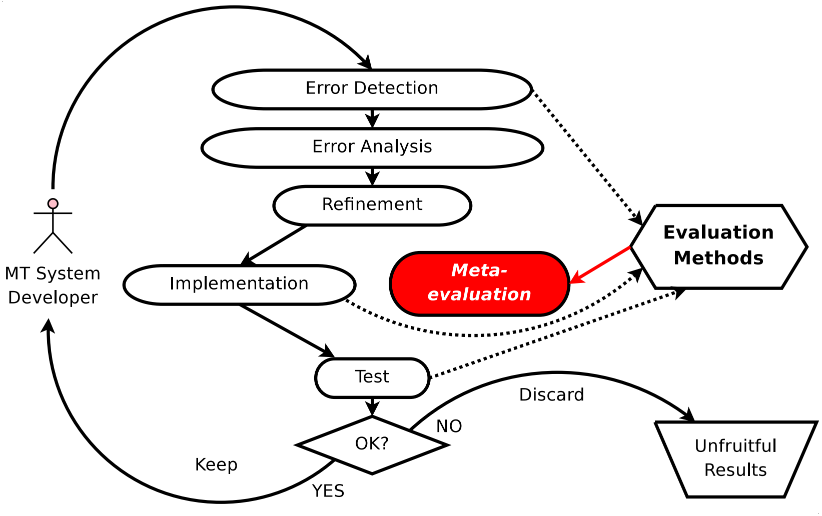
Metric Selection



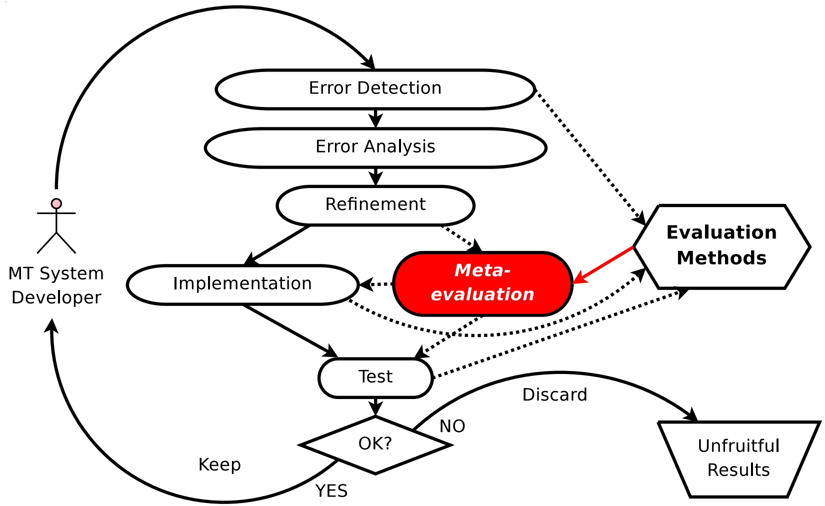
Metric Selection



Metric Selection



Metric Selection



Outline

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)

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)

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)

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)

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)

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)

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)

Thanks for your Attention

Thanks!

Empirical Machine Translation and its Evaluation

Invited Talk at the
Statistical Multilingual Analysis for
Retrieval and Translation Workshop 2009

Jesús Giménez

Grup de Processament del Llenguatge Natural
Departament de Llenguatges i Sistemes Informàtics
Universitat Politècnica de Catalunya

May 13, 2009

Dedicated Lexical Selection

Jesús Giménez and Lluís Màrquez, 2008. *Discriminative Phrase Selection for Statistical Machine Translation*. In Learning Machine Translation, NIPS Series, MIT Press.

- Related work
- Differences

Dedicated Lexical Selection

- Related work
  - Bangalore et al. (2007), Venkatapathy&Bangalore (2007)
  - Carpuat and Wu (2006, 2007, 2008)
  - Giménez and Màrquez (2007, 2008), España et al. (2008)
  - Specia et al. (2007, 2008)
  - Stroppa et al. (2007)
  - Vickrey et al. (2005)
- Differences

Dedicated Lexical Selection

- Related work
  - Bangalore et al. (2007), Venkatapathy&Bangalore (2007)
  - Carpuat and Wu (2006, 2007, 2008)
  - Giménez and Màrquez (2007, 2008), España et al. (2008)
  - Specia et al. (2007, 2008)
  - Stroppa et al. (2007)
  - Vickrey et al. (2005)
- Differences
  - Task (language-pair, domain)
  - System (learning scheme, SMT architecture)
  - Evaluation (BLEU/lexical/linguistic-based, manual)

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.

- Abhishek Arun and Philipp Koehn.
  Online Learning Methods For Discriminative Training of Phrase Based Statistical Machine Translation.
  In *Proceedings of MT SUMMIT XI*, pages 15–20, 2007.

- 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.

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.

Percy Liang, Alexandre Bouchard-CÃ´té, Dan Klein, and Ben Taskar.
An End-to-End Discriminative Approach to Machine Translation.
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 761–768, 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 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.

Christoph Tillmann and Tong Zhang.
A Discriminative Global Training Algorithm for Statistical MT.

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 721–728, 2006.

- 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.