Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation

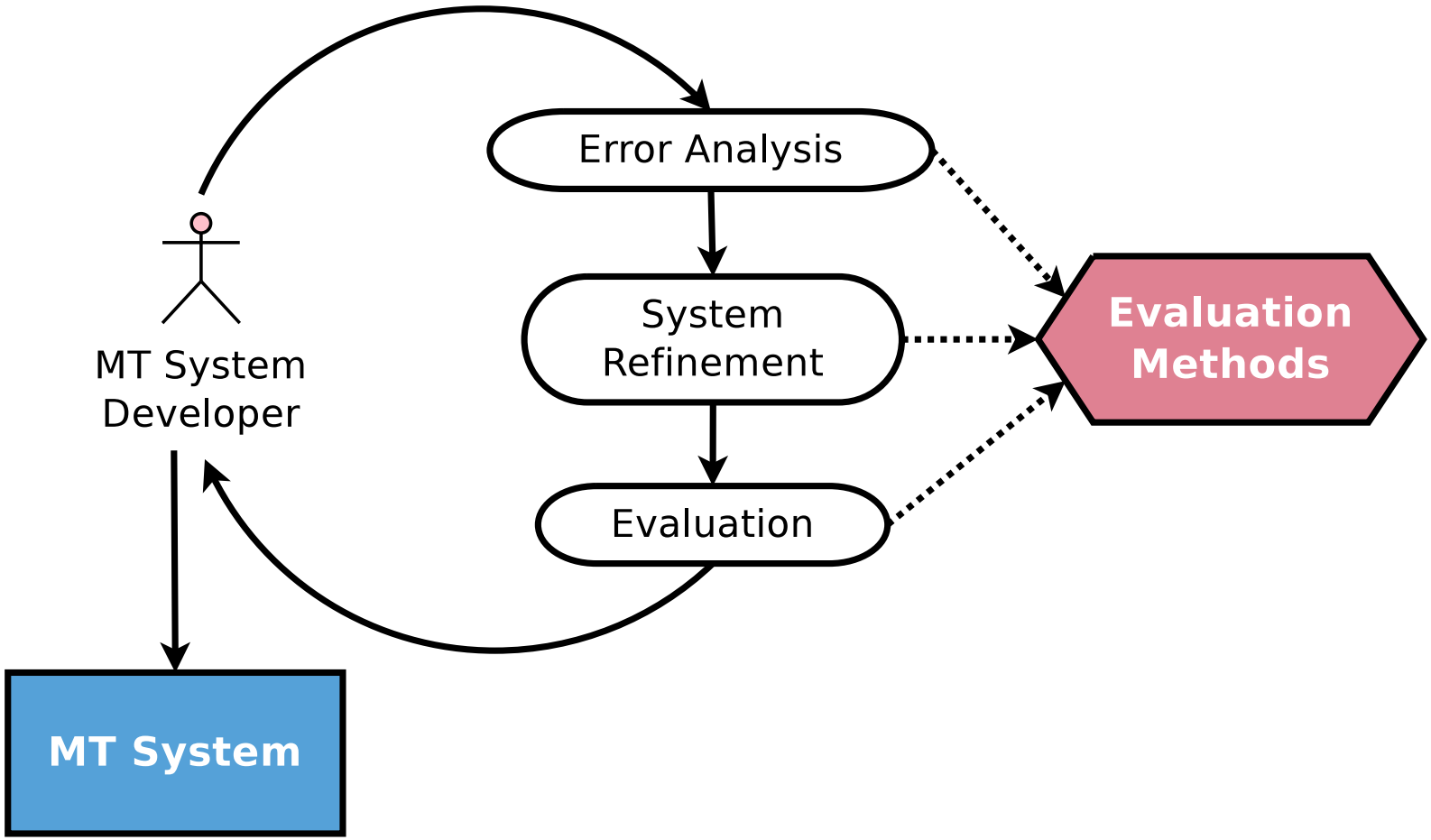
Jesús Giménez and Lluís Màrquez

Universitat Politècnica de Catalunya

Fifth MT Marathon
Le Mans, September 13-18, 2010

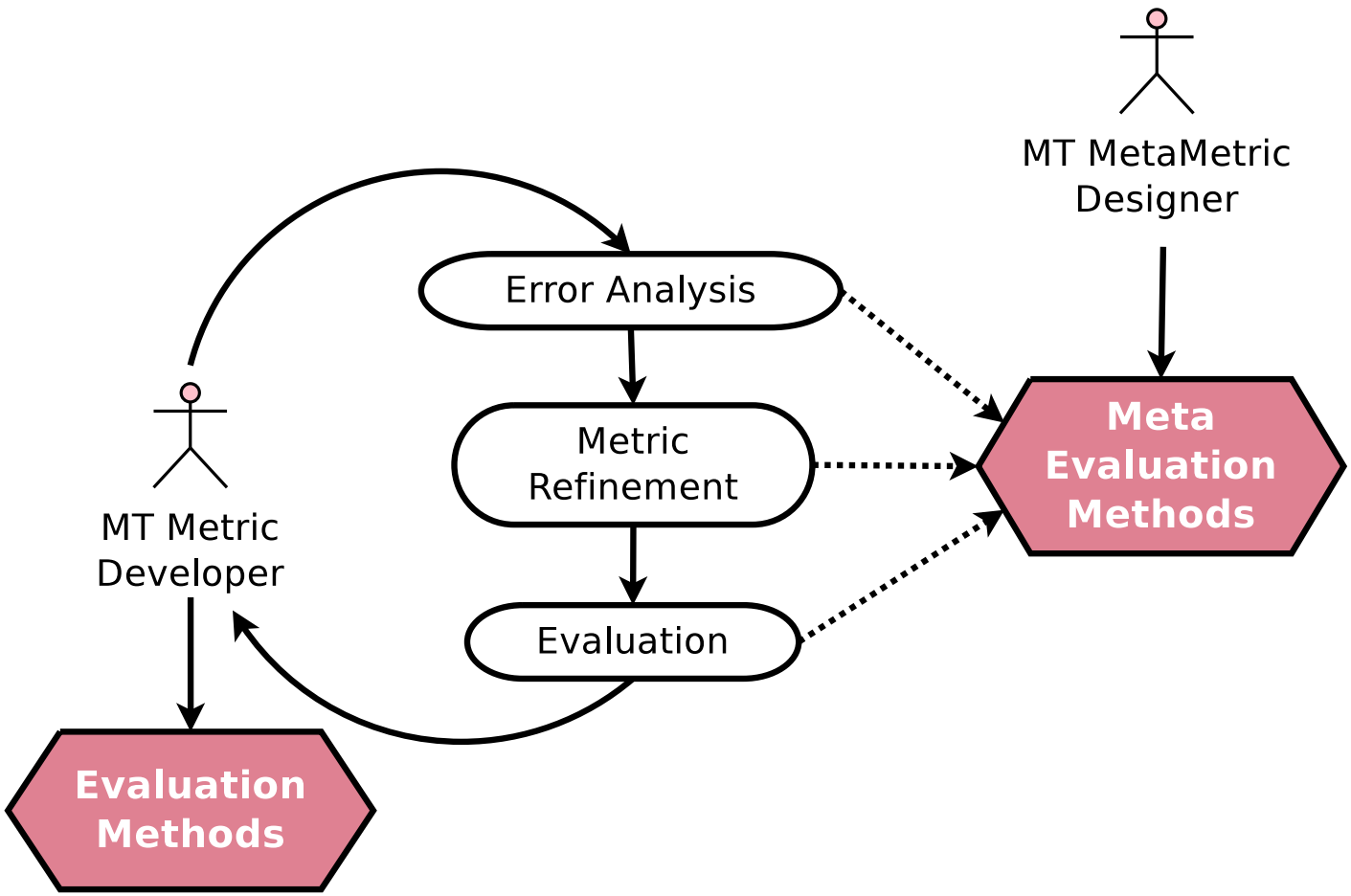
MT System Development Cycle

The Role of Evaluation Methods



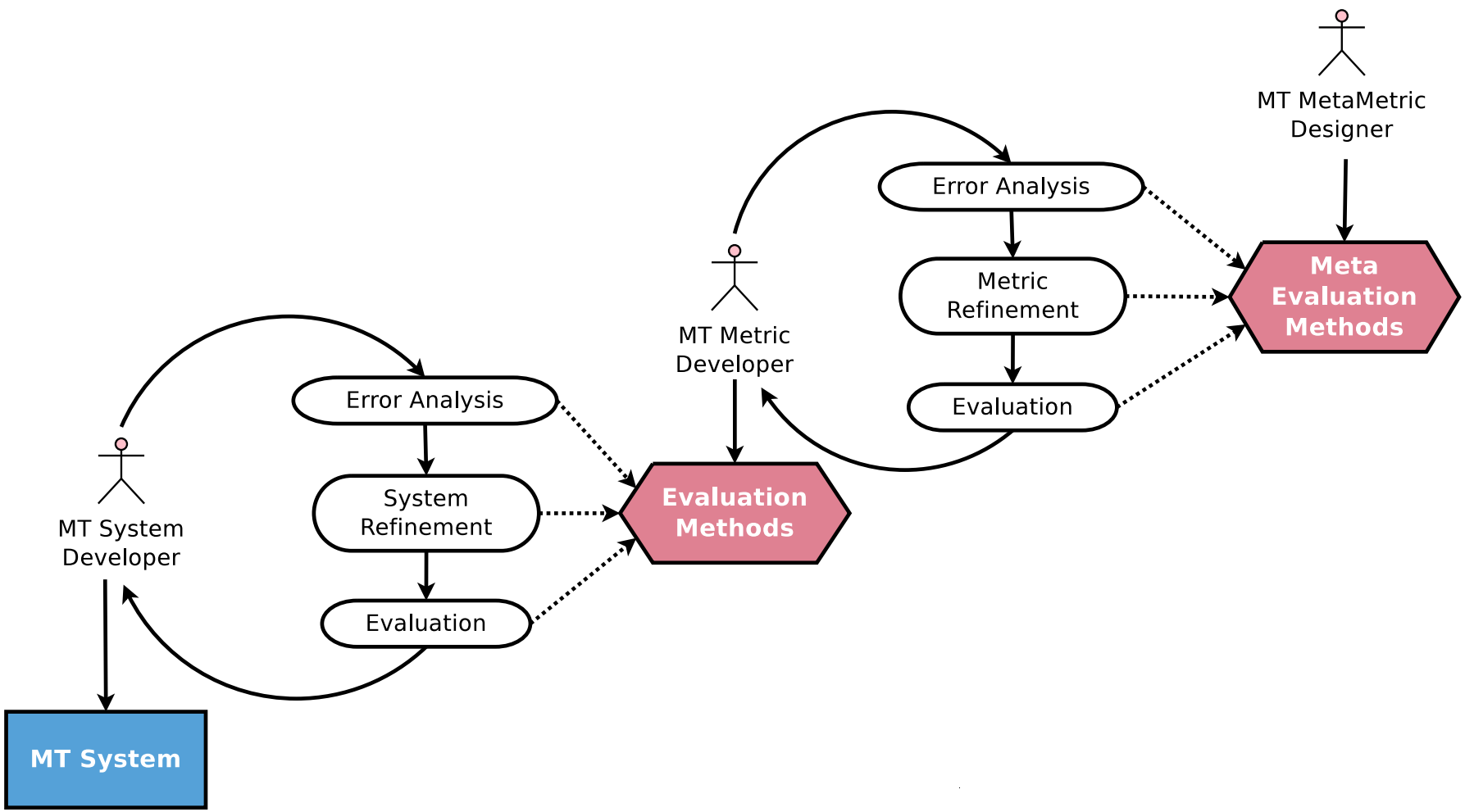
MT System Development Cycle

The Role of Evaluation Methods



MT System Development Cycle

The Role of Evaluation Methods



Tool Description

Test Suite Definition

Asiya operates over test suites (or test beds).

$\rightarrow$ a *test suite* is a collection of test cases:

- Source segment
- Candidate translation(s)
- Reference translation(s)

Tool Description

Test Suite Definition

```
Asiya.pl Asiya.config
```

src=source.xml
sys=candidates.xml
ref=references.xml

Table: Sample config file (NIST XML input format)

Tool Description

Test Suite Definition

src=source.sgm
sys=system_01.sgm
sys=system_02.sgm
sys=system_03.sgm
sys=system_04.sgm
sys=system_05.sgm
ref=reference_A.sgm
ref=reference_B.sgm
ref=reference_C.sgm

Table: Sample config file (NIST SGML input format)

Tool Description

Test Suite Definition

src=source.txt
sys=system_01.txt
sys=system_02.txt
sys=system_03.txt
sys=system_04.txt
sys=system_05.txt
ref=reference_A.txt
ref=reference_B.txt
ref=reference_C.txt

Table: Sample config file (RAW input format)

Tool Description

General Options

- **Input Format:** raw text, NIST XML/SGML
- **Language Pair:** source/target language and case sensitivity
- **Predefined Sets** of metrics, systems and references

Tool Description

General Options

- **Input Format:** raw text, NIST XML/SGML
- **Language Pair:** source/target language and case sensitivity
- **Predefined Sets** of metrics, systems and references

[GO TO DEMO]

Tool Description

Evaluation Options

```
Asiya.pl -v -eval single Asiya.config
```

Tool Description

Evaluation Options

`Asiya.pl -v -eval single Asiya.config`

Metric Repository:

- **Lexical** (Precision, Recall, F_1, Overlap, Error Rate)
- **Shallow Syntactic** (Lemmatization, PoS Tagging, and Base Phrase Chunking)
- **Syntactic** (Constituency and Dependency Parsing)
- **Shallow Semantic** (Semantic Roles and Named Entities)
- **Semantic** (Discourse Representations)

Tool Description

Evaluation Options

`Asiya.pl -v -eval single Asiya.config`

Metric Repository:

- **Lexical** (Precision, Recall, F_1, Overlap, Error Rate)
- **Shallow Syntactic** (Lemmatization, PoS Tagging, and Base Phrase Chunking)
- **Syntactic** (Constituency and Dependency Parsing)
- **Shallow Semantic** (Semantic Roles and Named Entities)
- **Semantic** (Discourse Representations)

[GO TO DEMO]

Tool Description

Evaluation Options

`Asiya.pl -v -eval <schemes> Asiya.config`

Tool Description

Evaluation Options

`Asiya.pl -v -eval <schemes> Asiya.config`

Schemes:

- **Single** metric scores
- **Ulc** normalized arithmetic mean of metric scores [GM10]
- **Queen** scores [AGPV05]

Tool Description

Evaluation Options

`Asiya.pl -v -eval <schemes> Asiya.config`

Schemes:

- **Single** metric scores
- **Ulc** normalized arithmetic mean of metric scores [GM10]
- **Queen** scores [AGPV05]

[GO TO DEMO]

Tool Description

General Options

`Asiya.pl -v -eval <schemes> Asiya.config`

Tool Description

General Options

`Asiya.pl -v -eval <schemes> Asiya.config`

- **Output Format:** metric matrix, system matrix
  NIST/WMT score files
- **Other Options:**
  - Include reference scores
  - Granularity → system/document/segment level
  - LaTeX / PDF output
  - etc...

Tool Description

General Options

`Asiya.pl -v -eval <schemes> Asiya.config`

- **Output Format:** metric matrix, system matrix
  NIST/WMT score files
- **Other Options:**
  - Include reference scores
  - Granularity → system/document/segment level
  - LaTeX / PDF output
  - etc...

[GO TO DEMO]

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval` <*schemes*> <*criteria*>
`Asiya.config`

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria> Asiya.config`

Criteria:

- Correlation with human assessments
  - Pearson r [Pea14]
  - Spearman ρ [Spe04]
  - Kendall τ [Ken55]
- ORANGE [LO04]
- KING [AGPV05]
- Consistency [CBKMS09]

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval` *<schemes>* *<criteria>* `Asiya.config`

Criteria:

- Correlation with human assessments
  - Pearson r [Pea14]
  - Spearman ρ [Spe04]
  - Kendall τ [Ken55]
- ORANGE [LO04]
- KING [AGPV05]
- Consistency [CBKMS09]

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria> Asiya.config`

Criteria:

- Correlation with human assessments
  - Pearson r [Pea14]
  - Spearman ρ [Spe04]
  - Kendall τ [Ken55]
- ORANGE [LO04]
- KING [AGPV05]
- Consistency [CBKMS09]

[GO TO DEMO]

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria>`
`Asiya.config`

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria>`
`-ci <method> Asiya.config`

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria>`
`-ci <method> Asiya.config`

Statistical significance tests:

- Fisher [Fis24]
- Bootstrap resampling [ET86]
- Paired bootstrap resampling [Koe04]
- Options:
  - α significance level
  - number of resamplings

Tool Description

Meta-Evaluation Options

`Asiya.pl -v -metaeval <schemes> <criteria>`
`-ci <method> Asiya.config`

Statistical significance tests:

- Fisher [Fis24]
- Bootstrap resampling [ET86]
- Paired bootstrap resampling [Koe04]
- Options:
  - α significance level
  - number of resamplings

[GO TO DEMO]

Tool Description

Meta-Evaluation Options

`Asiya.pl -v` $\frac{\texttt{-optimize } \langle schemes \rangle \ \langle criteria \rangle}{\texttt{Asiya.config}}$

Tool Description

Meta-Evaluation Options

`Asiya.pl -v` $\frac{\texttt{-optimize} \ \langle schemes \rangle \ \langle criteria \rangle}{\texttt{Asiya.config}}$

Metric set optimization (greedy):

1. Metrics are ranked by their individual quality
2. They are progressively added to the optimal set if and only if, when doing so, quality increases

Tool Description

Meta-Evaluation Options

`Asiya.pl -v` $\frac{\texttt{-optimize } \langle schemes \rangle \ \langle criteria \rangle}{\texttt{Asiya.config}}$

Metric set optimization (greedy):

1. Metrics are ranked by their individual quality
2. They are progressively added to the optimal set if and only if, when doing so, quality increases

[GO TO DEMO]

Ongoing and Future Work

Augment the (meta-)metric repositories

Metrics and meta-metrics:

- Port metrics to languages other than English
  (Arabic, Czech, French, German, Czech, Romanian, Spanish)
- More sophisticated metric combination schemes
- Alternative meta-evaluation criteria
- Confidence estimation metrics

Ongoing and Future Work

Augment the (meta-)metric repositories

Metrics and meta-metrics:

- Port metrics to languages other than English
  (Arabic, Czech, French, German, Czech, Romanian, Spanish)
- More sophisticated metric combination schemes
- Alternative meta-evaluation criteria
- Confidence estimation metrics

Ongoing and Future Work

A web interface for Asiya

Interface:

- Upload test suites $\rightarrow$ download results
- (Meta-)Evaluation reports
- Error analysis
  - visualization of linguistic structures
- User authentication/authorization/profile
- This week:
  1. learned Catalyst (`http://www.catalystframework.org/`)
  2. started implementation
     - test suite upload
     - system-level evaluation report

Ongoing and Future Work

A web interface for Asiya

Interface:

- Upload test suites $\rightarrow$ download results
- (Meta-)Evaluation reports
- Error analysis
  - visualization of linguistic structures
- User authentication/authorization/profile
- This week:
  1. learned Catalyst (`http://www.catalystframework.org/`)
  2. started implementation
     - test suite upload
     - system-level evaluation report

Thanks For Your Attention!

http://www.lsi.upc.edu/~nlp/Asiya/

Asiya: An Open Toolkit for Automatic Machine Translation (Meta-)Evaluation

Jesús Giménez and Lluís Màrquez

Universitat Politècnica de Catalunya

Fifth MT Marathon
Le Mans, September 13-18, 2010


Enrique Amigó, Julio Gonzalo, Anselmo Penas, and Felisa Verdejo.
QARLA: a Framework for the Evaluation of Automatic Summarization.
In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 280–289, 2005.


Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder.
Findings of the 2009 Workshop on Statistical Machine Translation.
In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 1–28, 2009.


Bradley Efron and Robert Tibshirani.
Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.
Statistical Science, 1(1):54–77, 1986.


R. A. Fisher.
On a Distribution Yielding the Error Functions of Several Well Known Statistics.
In *Proceedings of the International Congress of Mathematics*, volume 2, pages 805–813, 1924.


Jesús Giménez and Lluís Màrquez.
Linguistic Features for Automatic MT Evaluation.
To Appear in Machine Translation, 2010.


Maurice Kendall.
Rank Correlation Methods.
Hafner Publishing Co, 1955.


Philipp Koehn.
Statistical Significance Tests for Machine Translation Evaluation.
In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 388–395, 2004.


Chin-Yew Lin and Franz Josef Och.
ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation.
In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 501–507, 2004.


Karl Pearson.
The life, letters and labours of Francis Galton.
1914.
(3 volumes: 1914, 1924, 1930).


Charles Spearman.
The Proof and Measurement of Association Between Two Rings.
American Journal of Psychology, 15:72–101, 1904.