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
Asiya operates over test suites (or test beds).

→ 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)
<table>
<thead>
<tr>
<th>src</th>
<th>source.sgm</th>
</tr>
</thead>
<tbody>
<tr>
<td>sys</td>
<td>system_01.sgm</td>
</tr>
<tr>
<td>sys</td>
<td>system_02.sgm</td>
</tr>
<tr>
<td>sys</td>
<td>system_03.sgm</td>
</tr>
<tr>
<td>sys</td>
<td>system_04.sgm</td>
</tr>
<tr>
<td>sys</td>
<td>system_05.sgm</td>
</tr>
<tr>
<td>ref</td>
<td>reference_A.sgm</td>
</tr>
<tr>
<td>ref</td>
<td>reference_B.sgm</td>
</tr>
<tr>
<td>ref</td>
<td>reference_C.sgm</td>
</tr>
</tbody>
</table>

**Table:** Sample config file (NIST SGML input format)
<table>
<thead>
<tr>
<th>src</th>
<th>source.txt</th>
</tr>
</thead>
<tbody>
<tr>
<td>sys</td>
<td>system_01.txt</td>
</tr>
<tr>
<td>sys</td>
<td>system_02.txt</td>
</tr>
<tr>
<td>sys</td>
<td>system_03.txt</td>
</tr>
<tr>
<td>sys</td>
<td>system_04.txt</td>
</tr>
<tr>
<td>sys</td>
<td>system_05.txt</td>
</tr>
<tr>
<td>ref</td>
<td>reference_A.txt</td>
</tr>
<tr>
<td>ref</td>
<td>reference_B.txt</td>
</tr>
<tr>
<td>ref</td>
<td>reference_C.txt</td>
</tr>
</tbody>
</table>

**Table:** Sample config file (RAW input format)
• **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]
Asiya.pl -v -eval single Asiya.config
Asiya.pl -v -eval single Asiya.config

Metric Repository:

- **Lexical** (Precision, Recall, F₁, 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, F₁, 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]
Asiya.pl  -v  -eval <schemes>  Asiya.config
Asiya.pl -v -eval <schemes> Asiya.config

Schemes:

- **Single** metric scores
- **Ulc** normalized arithmetic mean of metric scores [GM10]
- **Queen** scores [AGPV05]
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]
Asiya.pl -v -eval <schemes> Asiya.config
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...
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 $\rho$ [Spe04]
  - Kendall $\tau$ [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 $\rho$ [Spe04]
  - Kendall $\tau$ [Ken55]
- ORANGE [LO04]
- KING [AGPV05]
- Consistency [CBKMS09]
Asiya.pl -v -metaeval <schemes> <criteria>
Asiya.config

Criteria:

- Correlation with human assessments
  - Pearson $r$ [Pea14]
  - Spearman $\rho$ [Spe04]
  - Kendall $\tau$ [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:

- $\alpha$ 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:
  • $\alpha$ significance level
  • number of resamplings

[GO TO DEMO]
Tool Description

Meta-Evaluation Options

Asiya.pl -v -optimize <schemes> <criteria>
Asiya.config
Asiya.pl -v -optimize <schemes> <criteria>
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
Asiya.pl -v -optimize <schemes> <criteria>
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 → 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 → 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
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.

Chris Callison-Burch, Philipp Koehn, Christof Monz, and Josh Schroeder.
Findings of the 2009 Workshop on Statistical Machine Translation.

Bradley Efron and Robert Tibshirani.
Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy.
R. A. Fisher. 
On a Distribution Yielding the Error Functions of Several Well Known Statistics. 

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

Maurice Kendall. 
*Rank Correlation Methods*. 

Philipp Koehn. 
Chin-Yew Lin and Franz Josef Och.  

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.  