A Graphical Interface for MT Evaluation and Error Analysis

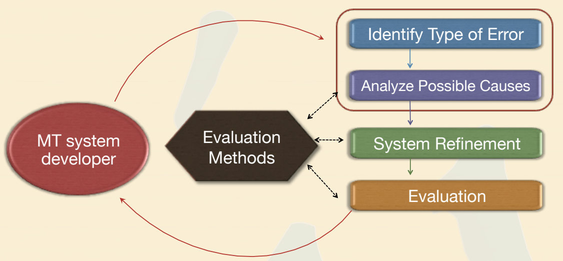
Meritxell Gonzàlez, Jesús Giménez, Lluís Màrquez
TALP Research Center, Universitat Politècnica de Catalunya

Goals

This work presents an online graphical interface to access ASIYA, an open source toolkit to evaluate automatic translations using an heterogeneous set of metrics and meta-metrics.

1. To allow MT developers to evaluate their test beds using a large set of metric scores
2. To detect and analyze the errors of the MT systems using just their Internet browsers
3. To help developers to understand the strengths and weaknesses of the evaluation measures

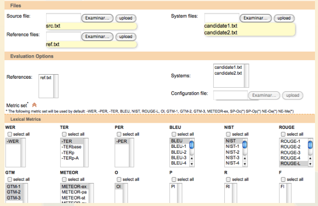
The ASIYA Toolkit



- More than 500 metric variants for MT evaluation
- Various similarity principles: precision, recall and overlap
- Different linguistic layers:
  - **Lexical similarity:** based on n-gram similarity and edit distance based on word form, *e.g., PER, TER, WER, BLEU, NIST, GTM, METEOR*
  - **Syntactic similarity:** based on part-of-speech tags, base phrase chunks, and dependency and constituency trees *e.g., SP-Overlap-POS, DP-HWCM, CP-STM*
  - **Semantic similarity:** based on named entities, semantic roles and discourse representation *e.g., NE-Overlap, SR-Overlap, DRS-Overlap*

The Online Graphical Interface (available at http://asiya.lsi.upc.edu/demo)

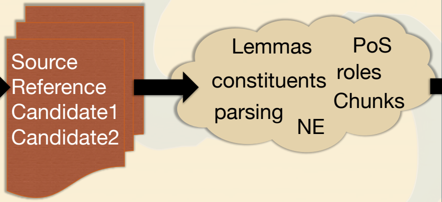
1. The online form allows to upload files and select the required options.


Files
Source file: [ ] Examinar... upload — src.txt
Reference files: [ ] Examinar... upload — ref.txt
System files: [ ] Examinar... upload — candidate1.txt, candidate2.txt

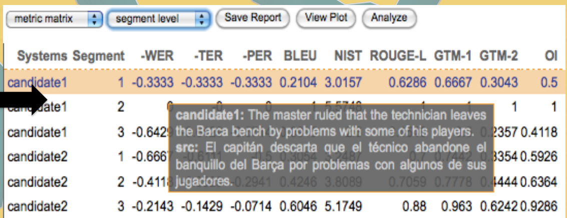
Evaluation Options
References: ref.txt
Systems: candidate1.txt, candidate2.txt
Configuration file: [ ] Examinar... upload
Metric set*
* The following metric set will be used by default: -WER, -PER, -TER, BLEU, NIST, ROUGE-L, OI, GTM-1, GTM-2, GTM-3, METEOR-ex, SP-Oc(*) SP-Op(*) NE-Oe(*) NE-Me(*)

Lexical Metrics
WER: □ select all — -WER
TER: □ select all — -TER, -TERbase, -TERp, -TERp-A
PER: □ select all — -PER
BLEU: □ select all — BLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4
NIST: □ select all — NIST, NIST-1, NIST-2, NIST-3, NIST-4
ROUGE: □ select all — ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L
GTM: □ select all — GTM-1, GTM-2, GTM-3
METEOR: □ select all — METEOR-ex, METEOR-pa, METEOR-st
O: □ select all — OI
P: □ select all — PI
R: □ select all — RI
F: □ select all — FI

2. Asiya generates a number of metric-dependent information…

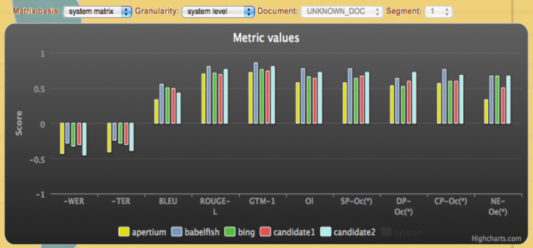


3. ...that produce an interactive evaluation report.



Systems	Segment	-WER	-TER	-PER	BLEU	NIST	ROUGE-L	GTM-1	GTM-2	OI
candidate1	1	-0.3333	-0.3333	-0.3333	0.2104	3.0157	0.6286	0.6667	0.3043	0.5
candidate1	2	0	-0	0	1	5.5748	1	1	1	1
candidate1	3	-0.6429	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.2357	0.4118
candidate2	1	-0.6667	-0.6111	-0.5	0.3054	3.2487	0.7	0.7442	0.3354	0.5926
candidate2	2	-0.4118	[illegible]	-0.2941	0.4246	3.8089	0.7059	0.7778	0.4444	0.6364
candidate2	3	-0.2143	-0.1429	-0.0714	0.6046	5.1749	0.88	0.963	0.6242	0.9286

candidate1: The master ruled that the technician leaves the Barca bench by problems with some of his players.
src: El capitán descarta que el técnico abandone el banquillo del Barça por problemas con algunos de sus jugadores.


Matrix basis: system matrix · Granularity: system level · Document: UNKNOWN_DOC · Segment: 1

4. The visualization of the linguistic information is useful for MT developers, such as interactive annotations and parse trees:

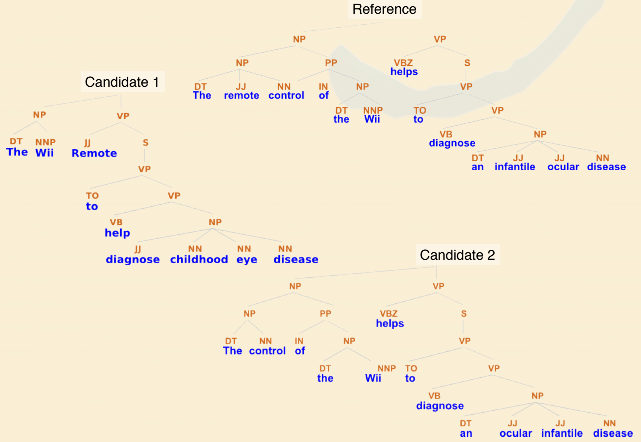
Source	El	mando	de	la	Wii	ayuda	a	diagnosticar	una	enfermedad	ocular	infantil	.

Ref	The	remote	control	of	the	Wii	helps	to	diagnose	an	infantile	ocular	disease	.
	DT	JJ	NN	IN	DT	NNP	VBZ	TO	VB	DT	JJ	J	NN	.
	B-NP	I-NP	I-NP	B-PP	B-NP	I-NP	B-VP	I-VP	I-VP	B-NP	I-NP	I-NP	I-NP	O
	O	O	O	O	O	B-ORG	O	O	O	O	O	O	O	O

Cand. 1	The	Wii	Remote	to	help	diagnose	childhood	eye	disease	.
	DT	NPP	NPP	TO	VB	VB	NN	NN	NN	.
	B-NP	I-NP	I-NP	B-VP	I-VP	I-VP	B-NP	I-NP	-NP	O
	O	B-ORG	I-ORG	O	O	O	O	O	O	O

Cand. 2	The	control	of	the	Wii	helps	to	diagnose	an	ocular	infantile	disease	.
	DT	NN	IN	DT	NNP	VBZ	TO	VB	DT	JJ	JJ	NN	.
	B-NP	I-NP	B-PP	B-NP	I-NP	B-BP	I-VP	I-VP	B-NP	I-NP	I-NP	I-NP	O
	O	O	O	O	B-ORG	O	O	O	O	O	O	O	O

Constituency parse trees:



Future Work

- Improve usability of the interface, e.g., allow input texts, dote the parse trees with more interactions.
- Create a database to save test sets and results.
- Show word alignments and use them to calculate metrics under a new principle.
- Detect and classify errors automatically.
- Create a search engine to filter results and obtain specific good/bad examples from the test set.

This research has been partially funded by the Spanish Ministry of Education and Science (OpenMT-2, TIN2009-14675-C03) and the European Community's Seventh Framework Programme under grant agreement numbers 247762 (FAUST project, FP7-ICT-2009- 4-247762) and 247914 (MOLTO project, FP7-ICT-2009-4- 247914).