

MT Automatic Evaluation and Meta-evaluation with Asiya

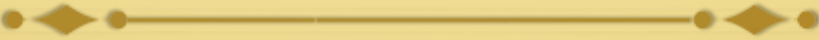
Meritxell Gonzàlez

Joint work with Jesús Giménez, Lluís Màrquez and Laura Mascarell
(some slides courtesy of Lluís Màrquez)

MT Course – Master IA

April 12th, 2013

Who am I



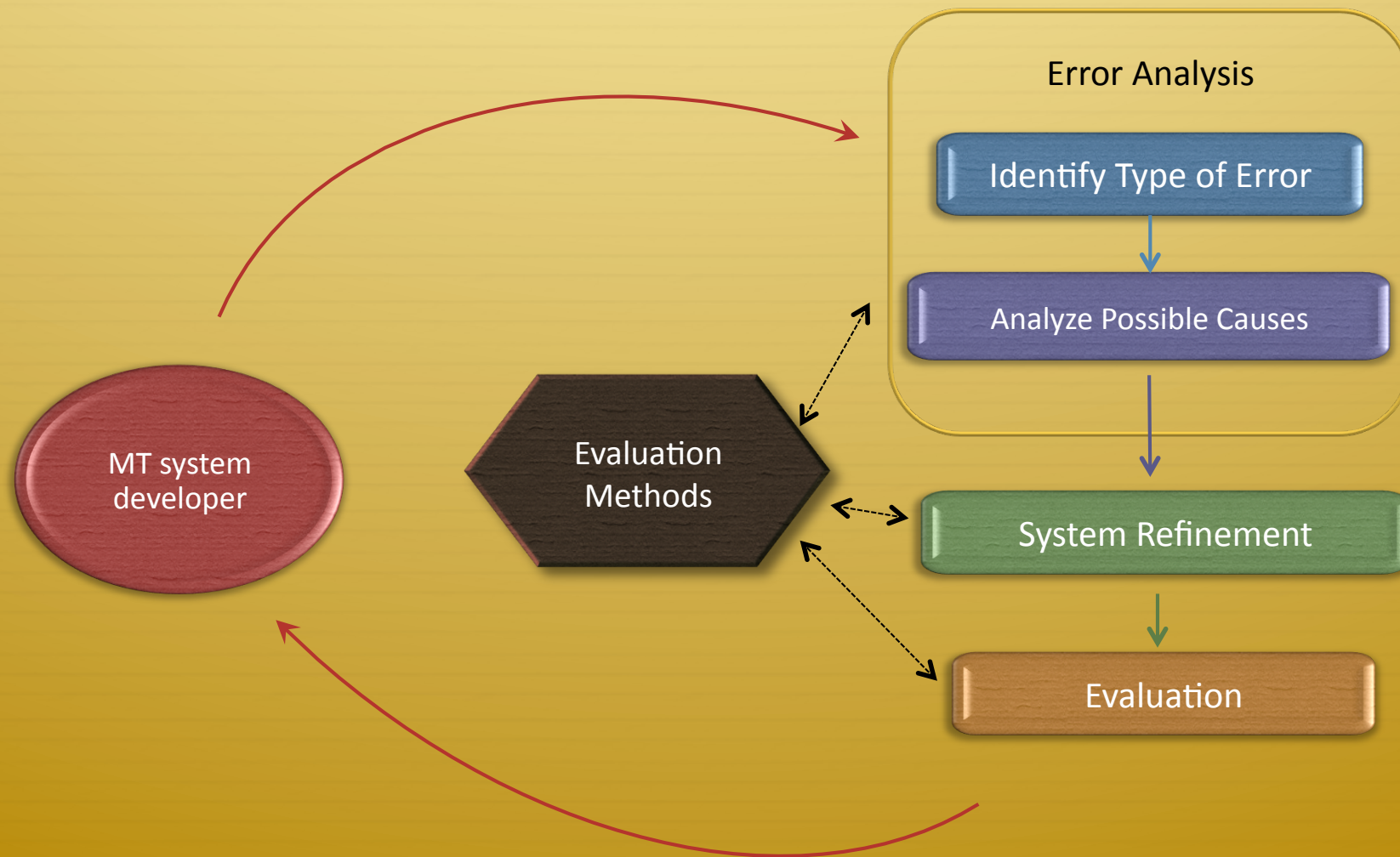
- ✦ Meritxell Gonzàlez
- ✦ Post-doc researcher at TALP Research Center
- ✦ MOLTO project: high quality and robust translation
 - ✦ Hybrid MT systems
 - ✦ Multilingual patents retrieval
- ✦ FAUST project: feedback analysis for User Adaptive statistical translation
 - ✦ Semantics for QE
 - ✦ Online tools for MT evaluation
- ✦ OpenMT-2: Traducción automática híbrida y evaluación avanzada

Overview



- ✦ Automatic MT Evaluation
- ✦ Linguistically motivated Evaluation measures
- ✦ Quality estimation
- ✦ Meta-evaluation
- ✦ The Asiya Toolkit

MT Development cycle



Difficulties of the 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 Automatic Evaluation



✦ Setting:

- ✦ Compute **similarity** between **system's output** and one or several **reference translations**.

✦ Challenge:

- ✦ The **similarity** measure should be able to discriminate whether the two sentences convey the same meaning (**semantic equivalence**)

MT Automatic Evaluation



✦ First Approaches:

✦ Lexical similarity as a measure of quality

✦ Edit Distance: WER, PER, TER

✦ Precision: **BLEU**, NIST

✦ Recall: ROUGE

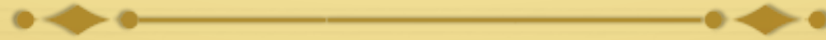
✦ Precision/Recall: GTM, METEOR

Problems of lexical similarity measures



- ✦ The **reliability** of lexical metrics depends very strongly on **the heterogeneity/representativity** of reference translations.
- ✦ Underlying Cause
 - ✦ Lexical similarity is nor a **sufficient** neither a **necessary** condition so that two sentences convey the same meaning

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

Linguistically motivated Evaluation measures



Linguistically motivated measures



- ✦ Extending Lexical Similarity Measures to increase robustness
 - ✦ Lexical variants
 - ✦ Morphological information (i.e., stemming)
ROUGE and METEOR
 - ✦ Synonymy lookup : METEOR (based on WordNet)
 - ✦ Paraphrasing support:
 - ✦ Extended versions of METEOR, TER

Linguistically motivated measures



- ✦ More linguistically-motivated measures:
 - ✦ Features capturing **syntactic** and **semantic** information
 - ✦ Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation
- ✦ Work at UPC (Jesús Giménez and Lluís Màrquez, 2007-2008)
 - ✦ Rather than comparing sentences at lexical level:
Compare the linguistic structures and the words within them.

Example: Giménez and Márquez, 2010



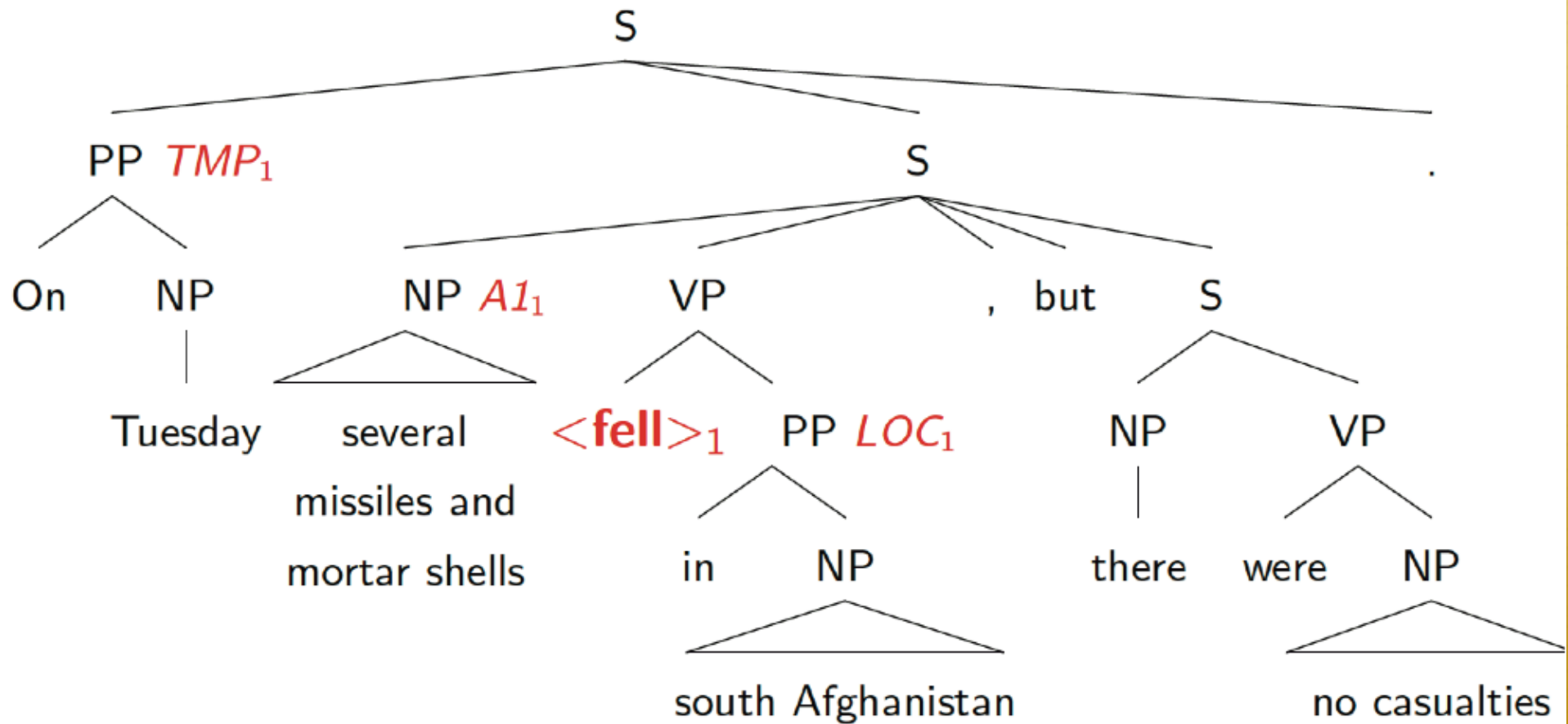
✦ Hypothesis:

- ✦ On Tuesday several missiles and mortar shells fell in south Afghanistan , but there were no casualties .

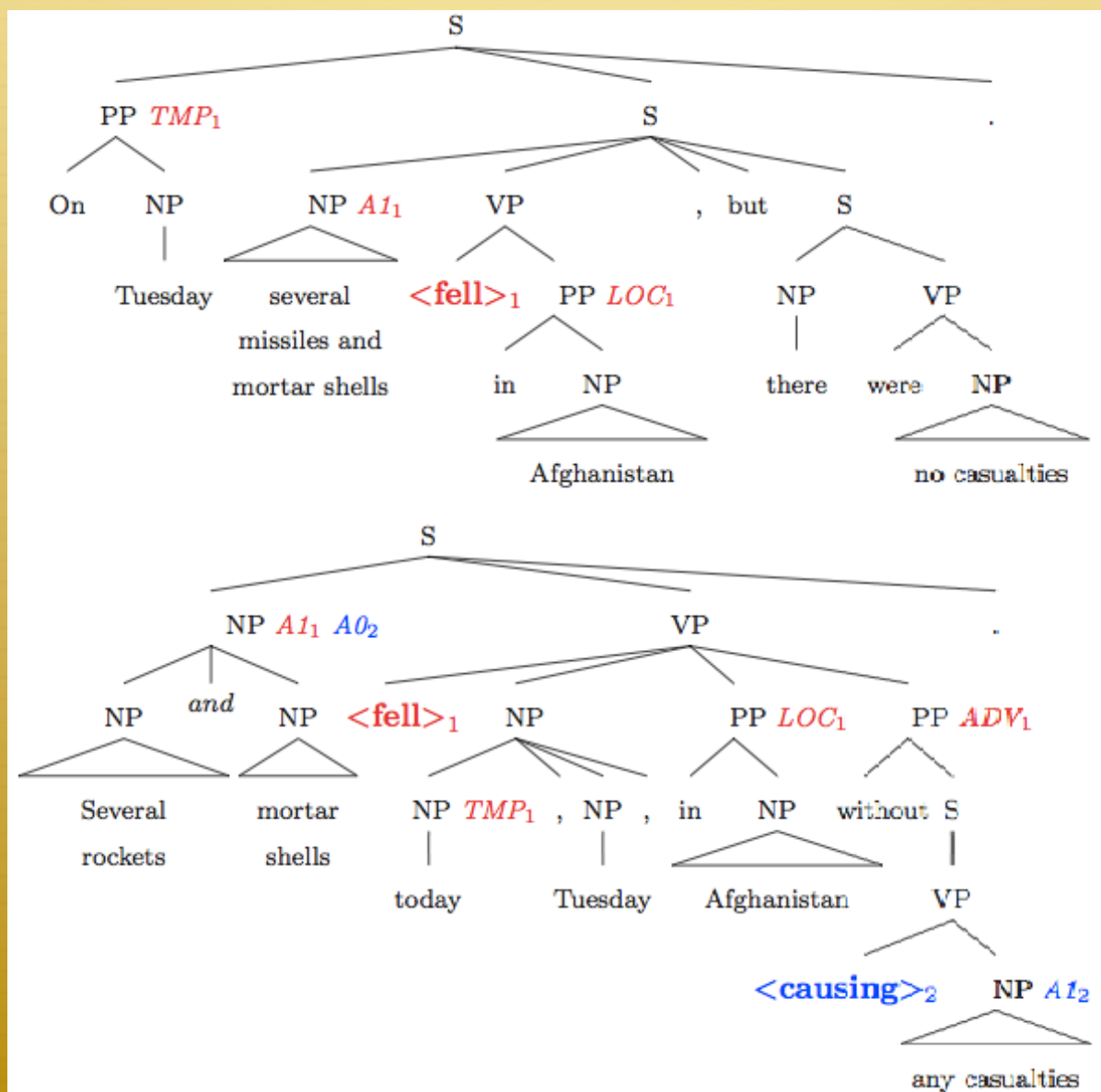
✦ Reference

- ✦ Several rockets and mortar shells fell today , Tuesday , in south Afghanistan without causing any casualties .

Example: Giménez & Màrquez, 2010



Examples: Giménez & Màrquez, 2010



Measuring structural similarity



- ✦ OVERLAP: generic similarity measure among Linguistic Elements. Inspired by the Jaccard similarity coefficient
- ✦ 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

Overlap

$$O(t) = \frac{\sum_{i \in (\text{items}_t(\text{hyp}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{hyp}}(i, t)}{\sum_{i \in (\text{items}_t(\text{hyp}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{hyp}}(i, t), \text{count}_{\text{ref}}(i, t))}$$

$$O(\star) = \frac{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \cap \text{items}_t(\text{ref}))} \text{count}_{\text{hyp}}(i, t)}{\sum_{t \in T} \sum_{i \in (\text{items}_t(\text{hyp}) \cup \text{items}_t(\text{ref}))} \max(\text{count}_{\text{hyp}}(i, t), \text{count}_{\text{ref}}(i, t))}$$

Example – Lexical Overlap



✦ Reference:

- ✦ The Spanish affiliate of the Disney Channel will debut the first totally Spanish fiction on March 4.

✦ Candidate:

- ✦ The Spanish branch of Disney Channel will wear for the first time next the 4 of March the first totally Spanish fiction product.

Example – Lexical Overlap

- ✦ hits: 15 (the min-intersection, marked as *), total 27 (the union, taking the max for each item instead of the sum, marked as @).

Candidate		reference	
'the' => 3,	@	'the' => 2,	*
'next' => 1,	@	'of' => 1,	*
'time' => 1,	@	'fiction' => 1,	* @
'of' => 2,	@	'debut' => 1,	@
'fiction' => 1,		'on' => 1,	@
'will' => 1,		'will' => 1,	* @
'.' => 1,		'.' => 1,	* @
'first' => 2,	@	'first' => 1,	*
'for' => 1,	@	'Channel' => 1,	* @
'Channel' => 1,		'affiliate' => 1,	@
'branch' => 1,	@	'4' => 1,	* @
'4' => 1,		'March' => 1,	* @
'wear' => 1,	@	'Disney' => 1,	* @
'March' => 1,		'totally' => 1,	* @
'Disney' => 1,		'Spanish' => 2,	* @
'product' => 1,	@	'The' => 1	* @
'totally' => 1,			
'Spanish' => 2,			
'The' => 1			

Measuring structural similarity



- ✦ MATCHING is a similar but more strict variant
 - ✦ All items inside an element are considered the same unit
 - ✦ Computes the proportion of fully translated LEs, according to their types

Measuring structural similarity



- ✦ Overlap and Matching have been instantiated over different linguistic level elements (for English, Spanish, Catalan, French and German)
- ✦ Words, lemmas, POS, Chunks
- ✦ Shallow, dependency and constituency parsing
- ✦ Named entities and semantic roles (es, ca, en)
- ✦ Discourse representation (logical forms) (en)

Measuring structural similarity

- ✦ The Spanish branch of Disney Channel **will** wear **for the** first time **next the 4 of March the** first totally Spanish fiction product .
- ✦ DT JJ NN IN NNP NNP MD VB IN DT JJ NN IN DT CD IN NNP DT JJ RB JJ NN NN .
- ✦ B-NP I-NP I-NP B-PP B-NP I-NP **B-VP** I-VP **B-PP** **B-NP** I-NP I-NP **B-PP** **B-NP** I-NP **B-PP** **B-NP** I-NP **B-PP** **B-NP** B-NP I-NP I-NP I-NP I-NP I-NP I-NP O
- ✦ (S1 (S (NP (NP (DT The) (NNP Spanish) (NN branch)) (PP (IN of) (NP (NNP Disney) (NNP Channel)))) (VP (MD will) (VP (VB wear) (PP (IN for) (NP (NP (DT the) (JJ first) (NN time)) (PP (IN next) (NP (NP (NP (DT the) (CD 4)) (PP (IN of) (NP (NNP March)))) (NP (DT the) (JJ first) (ADJP (RB totally) (JJ Spanish)) (NN fiction) (NN product))))))))) (. .)))

Measuring structural similarity

✦ (S
 (NP
 (NP
 (DT The) (NNP Spanish) (NN branch))
 (P
 (IN of)
 (NP (NNP Disney) (NNP Channel))))
 (VP
 (MD will)
 (VP
 (VB wear)
 (P
 (IN for)
 (NP
 (NP
 (DT the) (JJ first) (NN time))
 (P
 (IN next)
 (NP
 (NP
 (NP (DT the) (CD 4))
 (P
 (IN of) (NP (NNP March))))
 (NP
 (DT the) (JJ first) (ADJP (RB totally) (JJ Spanish))
 (NN fiction) (NN product))))))
 (. .))

Quality / Confidence Estimation



Quality Estimation



- ✦ Setting:
 - ✦ Quality assessment without reference translations
- ✦ Information available:
 - ✦ Source sentence, candidate translation(s) and, possibly, MT system information
- ✦ Motivation - usefulness:
 - ✦ System ranking
 - ✦ Hypotheses re-ranking
 - ✦ User feedback filtering
 - ✦ Measuring improvement
 - ✦ Post-edition effort

Quality Estimation Features



- ✦ System-dependent
 - ✦ internal system probabilities/scores
 - ✦ features over **n-best translation hypotheses**
 - ✦ language modeling
 - ✦ hypothesis rank
 - ✦ score ratio
 - ✦ average hypothesis length
 - ✦ length ratio
 - ✦ center hypothesis

Quality Estimation Features



- ✦ System-independent

- ✦ source (translation difficulty)

- ✦ sentence length

- ✦ Ambiguity - dictionary/alignment/WordNet-based

- ✦ (number of candidate translations per word or phrase)

- ✦ target (fluency)

- ✦ sentence length

- ✦ language modeling

- ✦ source-target (adequacy)

- ✦ length ratio

- ✦ punctuation issues

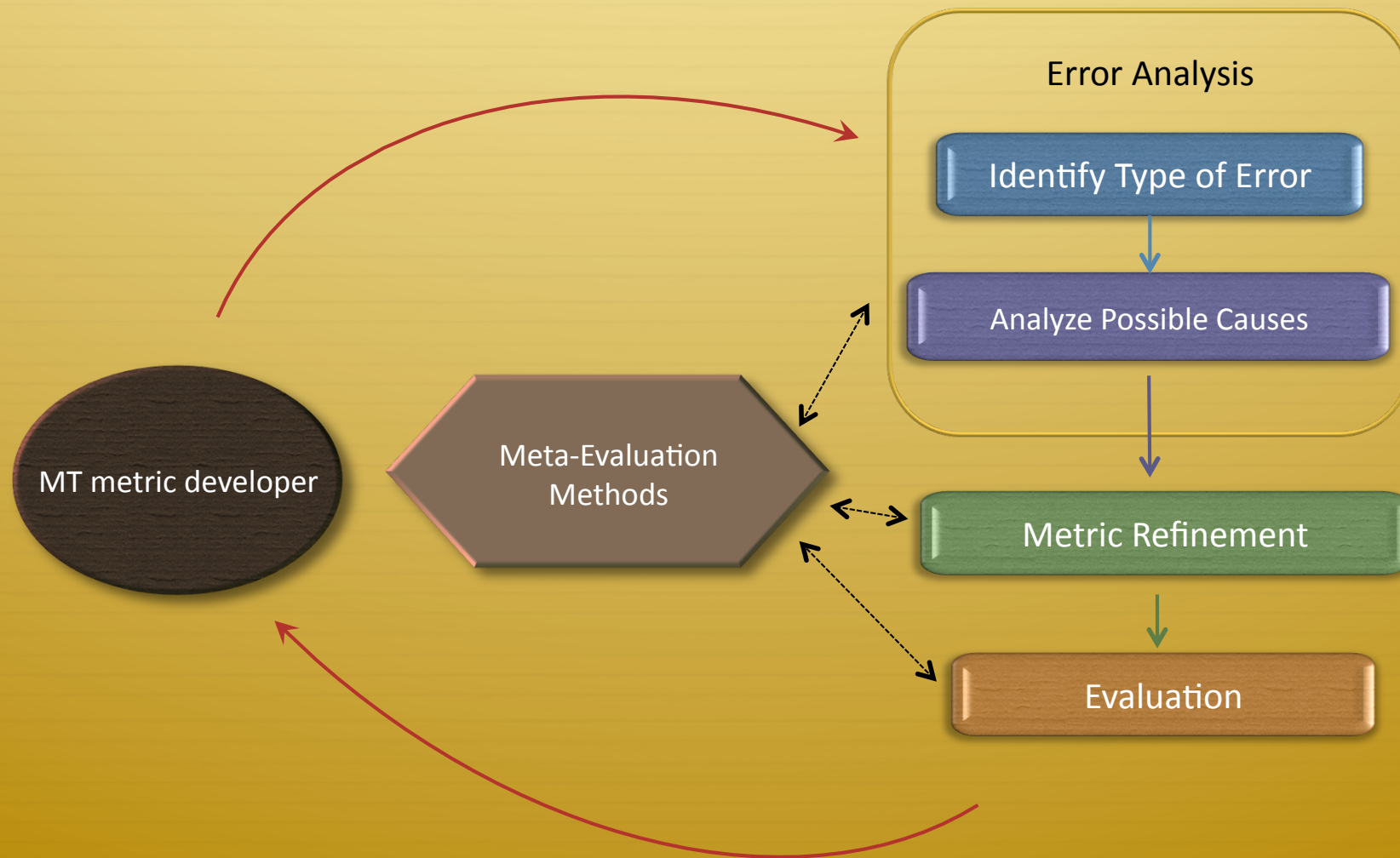
- ✦ candidate matching ! dictionary-/alignment-based

Meta evaluation



Metric-wise system development

MT Development cycle



Meta-evaluation



- ✦ Correlation with assessments

 - ✦ Pearson

 - ✦ Spearman

 - ✦ Kendall tau

- ✦ Orange [LO04]

- ✦ King [AGPV05]

- ✦ Consistency (ranking)


Metric Combination

- ✦ Different measures capture different aspects of similarity
 - ✦ Suitable for combination
- ✦ Simple Approach: ULC
 - ✦ Uniformly averaged linear combination of measures (ULC):
- ✦ Simple hill climbing approach to find the best subset of measures M on a development corpus
 - ✦ $M = \{ROUGE_w, METEOR, DP-HWC_r, DP-O_c(*), DP-O_l(*), DP-O_r(*), CP-STM_4, SR-O_r(*), SR-O_{rv}, DR-O_{rp}(*)\}$

Learn new models



- ✦ The goal is to combine the scores conferred by different evaluation measures into a single measure of quality such that their relative contribution is adjusted based on human feedback (i.e., from human assessments).
- ✦ Asiya integrates a Perceptron scheme.
- ✦ Do you want to develop other schema?



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

Asiya



- ✦ Asiya provides:
 - ✦ Automatic evaluation measures using several linguistic layers for a variety of languages
 - ✦ Quality Estimation measures
- ✦ Meta-evaluation metrics
- ✦ Learning schemes
- ✦ Web graphical interface for semi-automatic error analysis
 - ✦ (video demo: <http://nlp.lsi.upc.edu/asiya/asiya-demo.mov>)
- ✦ Remote Web Service
- ✦ Translation Search (tSearch) application for **qualitative** error analysis

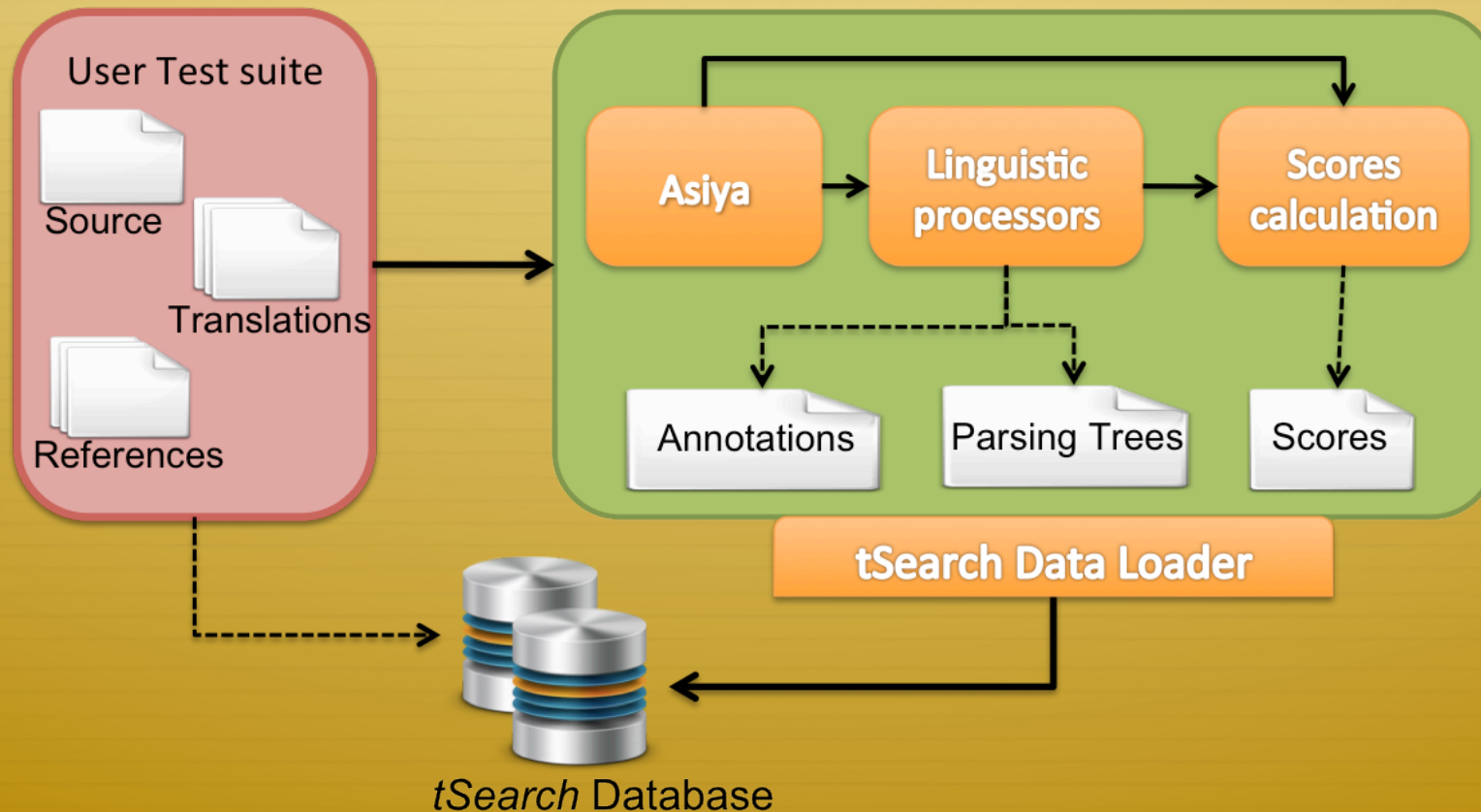
Overview



- ✦ Languages:
 - ✦ English, Spanish, Catalan
 - ✦ Also: Arabic, Czech, French, German, Romanian
- ✦ Similarity principles
 - ✦ Precision, recall, overlap, matching, ...
- ✦ Linguistic layers:
 - ✦ Lexical, Syntactic, Semantic
 - ✦ Confidence estimation

The Test Suit

- ✦ Asiya operates over test suites (or test beds).



Example

- ✦ El capitán descarta que el técnico abandone el banquillo del Barça por problemas con algunos de sus jugadores.
- ✦ The captain rejects that the coach leaves the Barça bench due problems with some of the players.
- ✦ The captain **descarta** that the technician abandon the **banquillo** of the Barça by problems with some of his players.
- ✦ The captain **discards** that the **technician** leaves **the bench of the Barça** by problems with some of his players.
- ✦ The captain **dismisses** the **technician** leaves the Barca bench due to problems with some of **their** players.
- ✦ The **master ruled** that the technician leaves the Barca bench by problems with some of his players.
- ✦ The captain rejects that the technician leaves the bench of the Barça for problems with some **of his(her,your) players.**
- ✦ The captain discards that the technician leaves the bench of the Barça **by problems** with some of his players.

The Test Suit

✦ Asiya.pl Asiya.config

✦ Asiya.config:

✦ Input format

✦ Raw

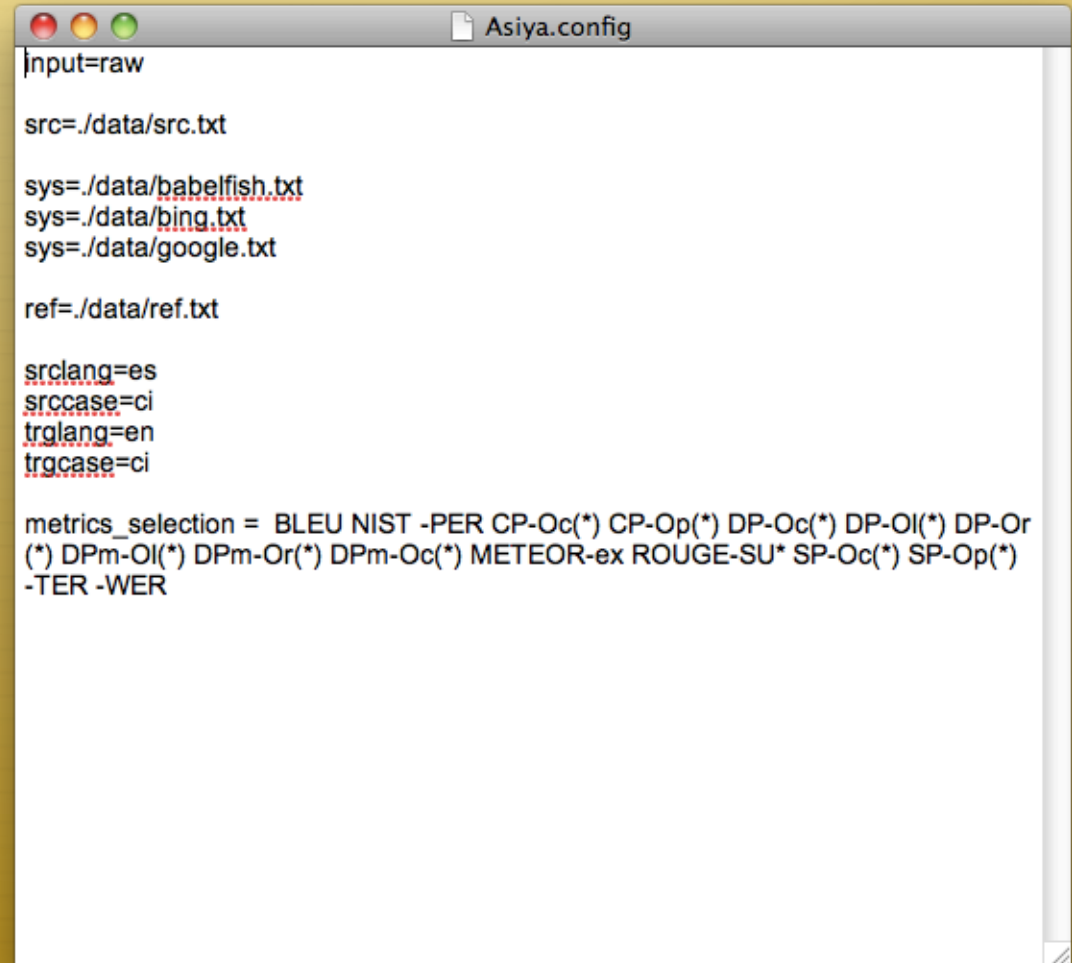
✦ Nist

✦ Language pair

✦ Srclang

✦ Trglang

✦ Predefined sets of metrics, systems and references

A screenshot of a text editor window titled "Asiya.config". The window contains the following configuration text:

```
input=raw

src=./data/src.txt

sys=./data/babelfish.txt
sys=./data/bing.txt
sys=./data/google.txt

ref=./data/ref.txt

srclang=es
srccase=ci
trglang=en
trgcase=ci

metrics_selection = BLEU NIST -PER CP-Oc(*) CP-Op(*) DP-Oc(*) DP-Ol(*) DP-Or(*)
DPm-Ol(*) DPm-Or(*) DPm-Oc(*) METEOR-ex ROUGE-SU* SP-Oc(*) SP-Op(*)
-TER -WER
```

Eval

✦ Eval <schema>

- ✦ Single
- ✦ ULC
- ✦ Queen [AGPV05]

✦ Meta-Eval

✦ Learn

✦ Output format

- ✦ Metric matrix
- ✦ System matrix
- ✦ Nist

✦ Granularity

- ✦ System, document, segment

✦ Pdf, tex

Meta-Eval

- ✦ Eval
- ✦ Meta-Eval <schemas> <criteria>
-ci <method> Asiya.config
 - ✦ Fisher [Fis24]
 - ✦ Bootstrap resampling [ET86]
 - ✦ Paired bootstrap resampling [Koe04]
Orange [LO04]
 - ✦ Options:
 - ✦ significance level
 - ✦ Asiya.pl -v -optimize <schemes>
<criteria>number of
resamplings
- ✦ Learn

Learning

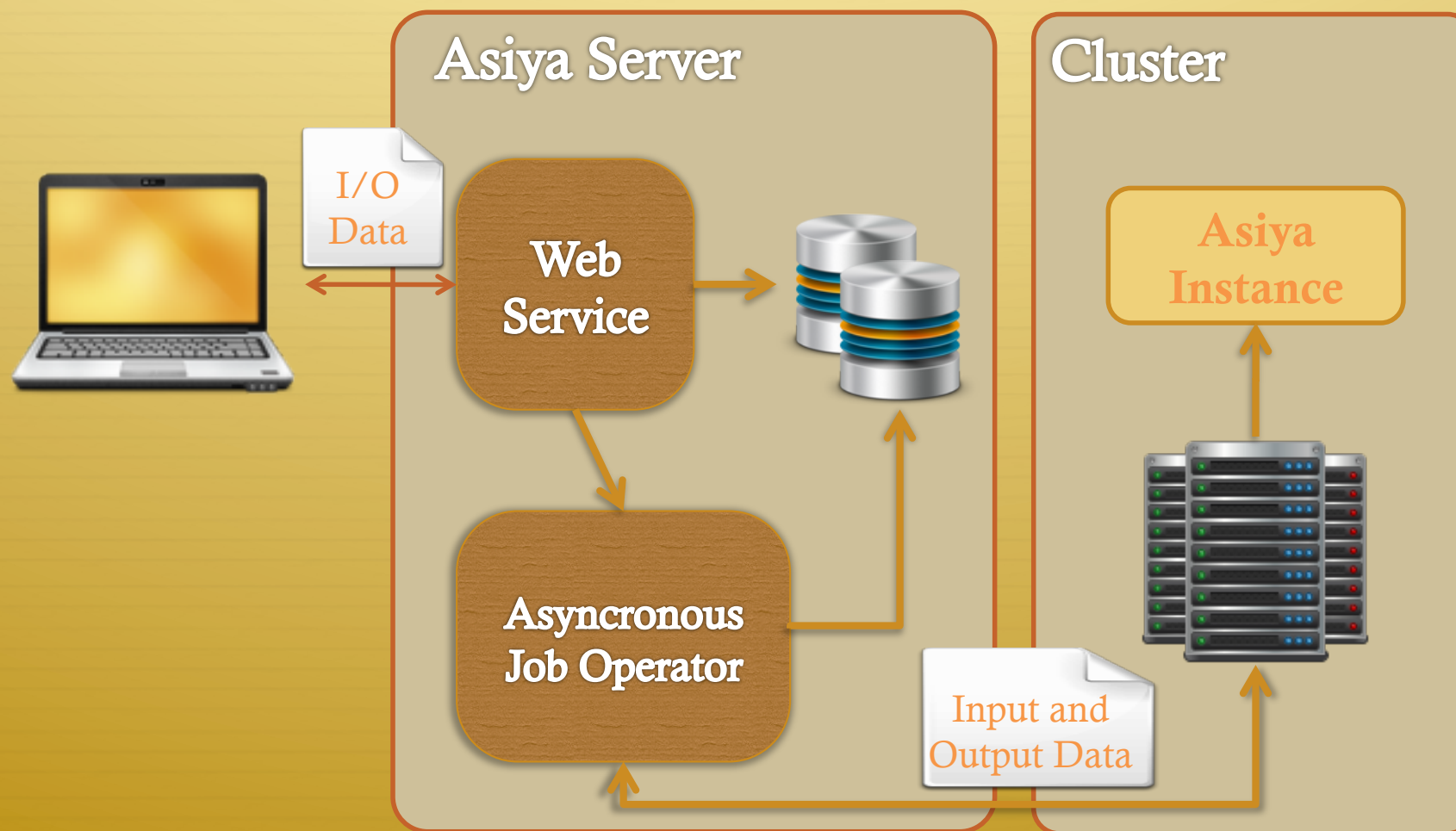
- ✦ Eval
- ✦ Meta-Eval
- ✦ Learn <scheme>
 - ✦ Perceptron
 - ✦ model <s>
- ✦ Asiya.pl -eval single -model <s>

Asiya Interfaces

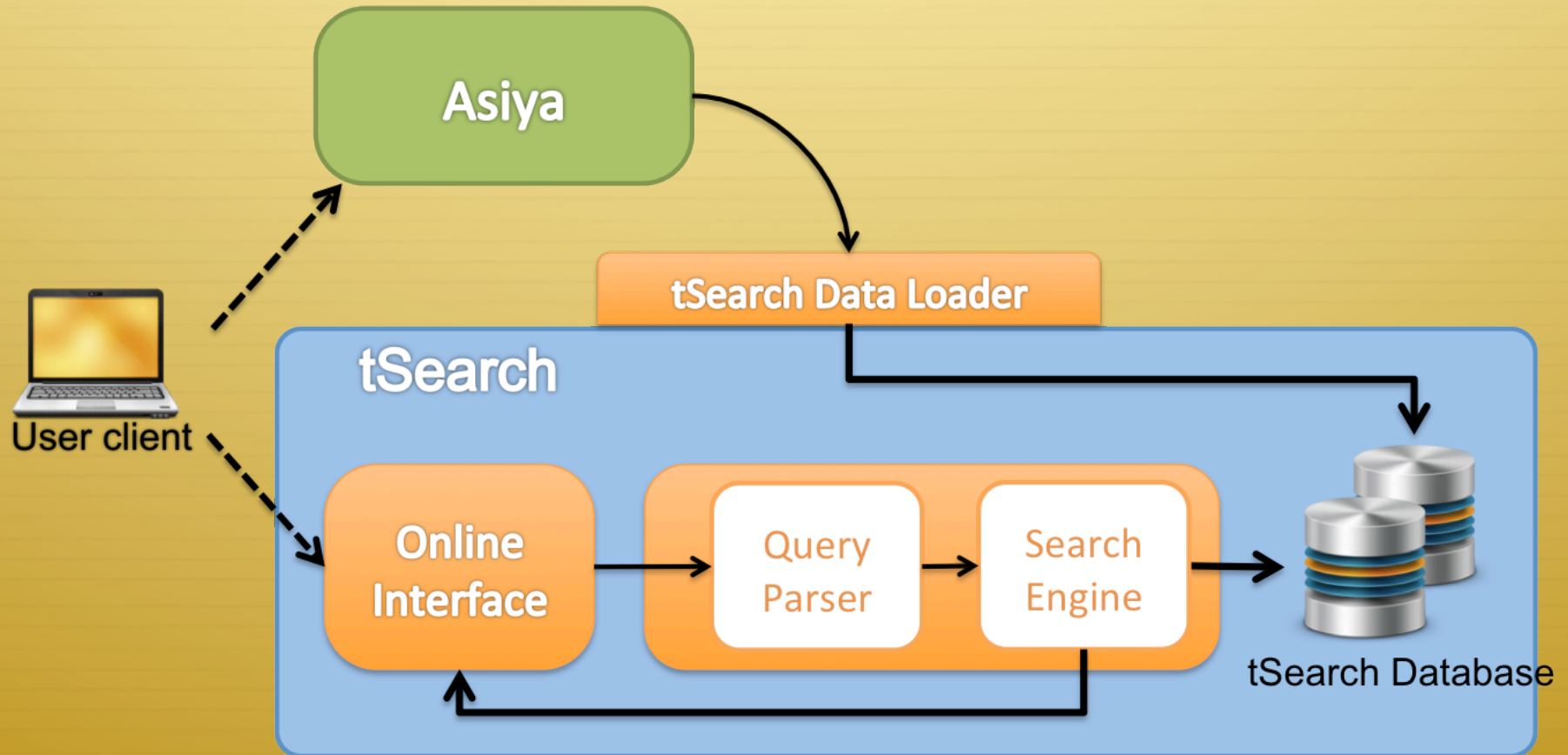


- ✦ Asiya Web Service
 - ✦ A RESTful web service to access the Asiya evaluation.
- ✦ Asiya Online Interface
 - ✦ A graphical interface to access an on-line version of Asiya.
- ✦ Asiya tSearch
 - ✦ online interface that allows to search for output translations (of a given testbed) that match some specific criteria related to their quality (as assessed by the automatic scores).

Asiya Web Service



Asiya tSearch



Try it!

✦ <http://asiya.lsi.upc.edu>

- ✦ Choose the languages
- ✦ Write some sentences or upload a SMALL file. Try to introduce several errors:
 - ✦ lexical disagreement, missing prepositions,
- ✦ Use some linguistic measures in addition to the lexical ones:
 - ✦ *BLEU, NIST, ROUGE_w, METEOR-pa*
 - ✦ *SP-Op(*), DP-HWC_r, DP-O_r(*), CP-STM₄*
- ✦ Run it and look how the segment level scores identify the errors in each sentence
- ✦ Look at the parse trees
- ✦ Use the tSearch interface to find interesting sentences according to the scores and the parse trees

References



- ✦ 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.

References



- ✦ 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.

References



- ✦ 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.