MT Automatic Evaluation and Meta-evaluation

Meritxell Gonzàlez
Joint work with Jesús Giménez and Lluís Màrquez
(some slides courtesy of Lluís Màrquez)
Grial Seminar
March 7th, 2013

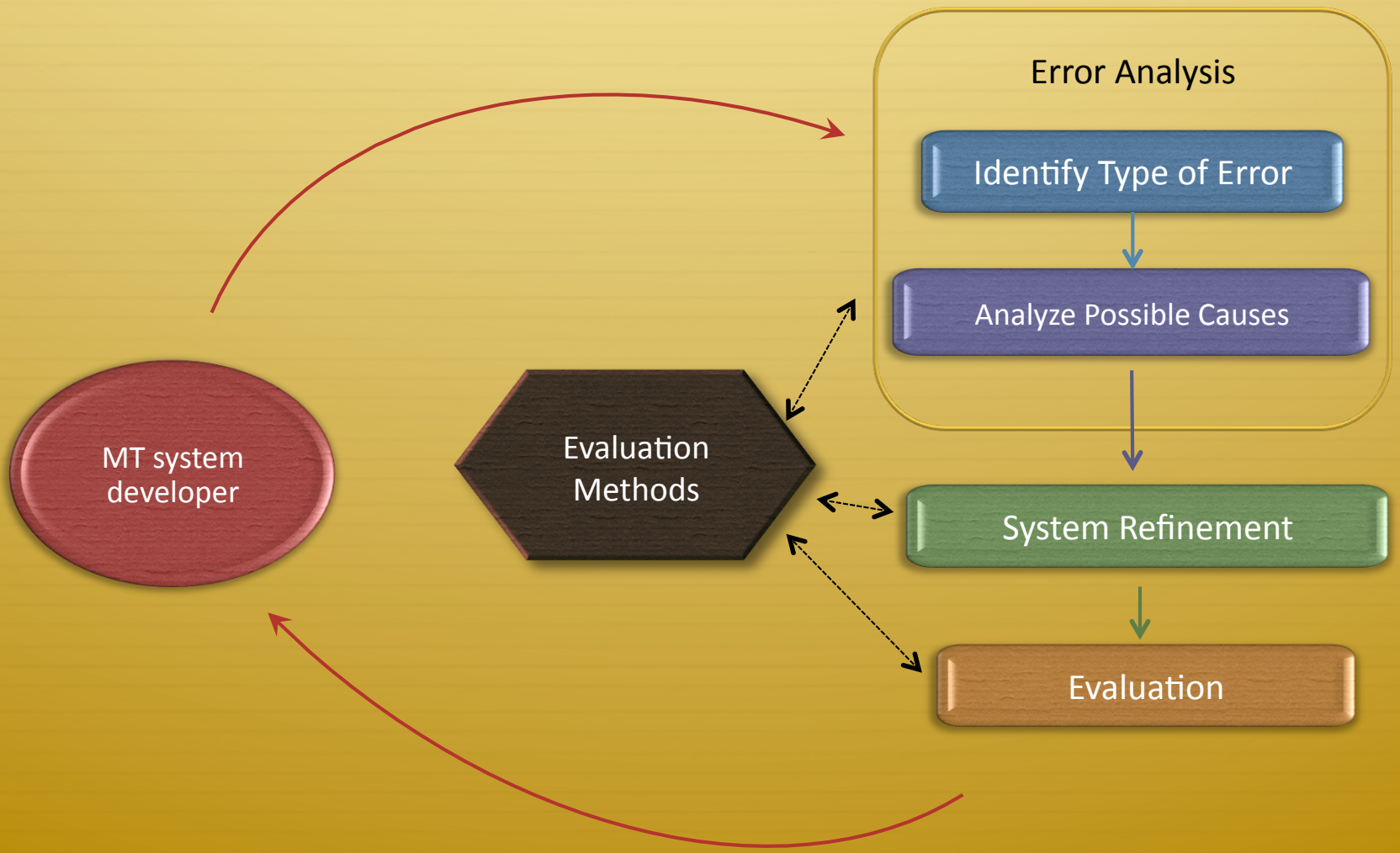
Who am I

- Meritxell Gonzàlez
- Post-doc researcher at UPC
- 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

Benefits of Automatic Evaluation

- Compared to manual evaluation, automatic measures are:
  - Cheap (vs. costly)
  - Objective (vs. subjective)
  - Reusable (vs. not-reusable)
- Automatic evaluation metrics have notably accelerated the development cycle of MT systems
  - Error analysis
  - System optimization
  - System comparison

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

IBM BLEU metric

- **BLEU: a Method for Automatic Evaluation of Machine Translation**
  Kishore Papineni, Salim Roukos, Todd Ward, Wei-Jing Zhu
  IBM Research Division (Papineni et al., 2001)

- "The main idea is to use a weighted average of variable length phrase matches against the reference translations. This view gives rise to a family of metrics using various weighting schemes. We have selected a promising baseline metric from this family."

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

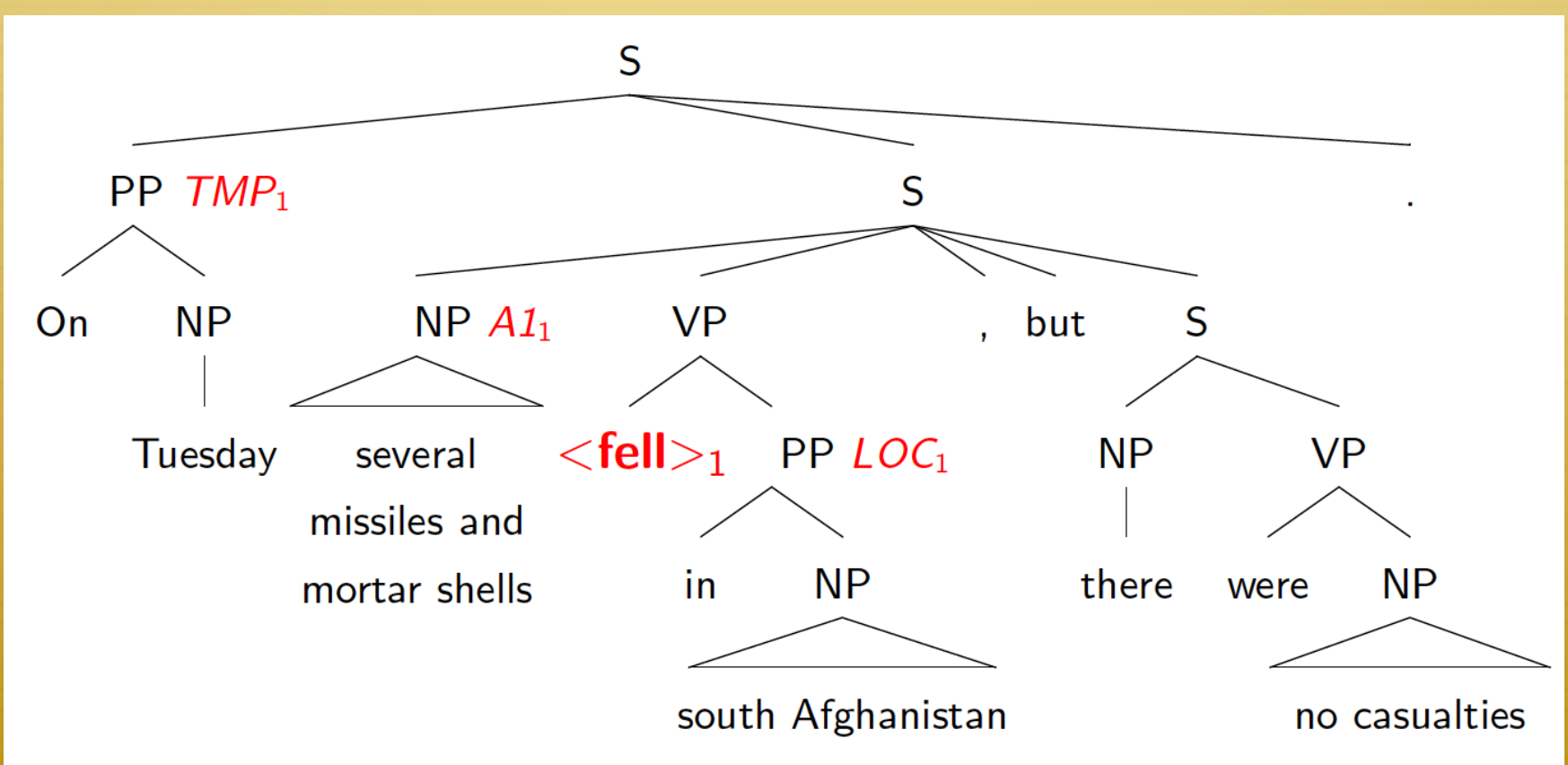
Linguistically motivated measures

- 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)
  - Rather than comparing sentences at lexical level: Compare the linguistic structures and the words within them.

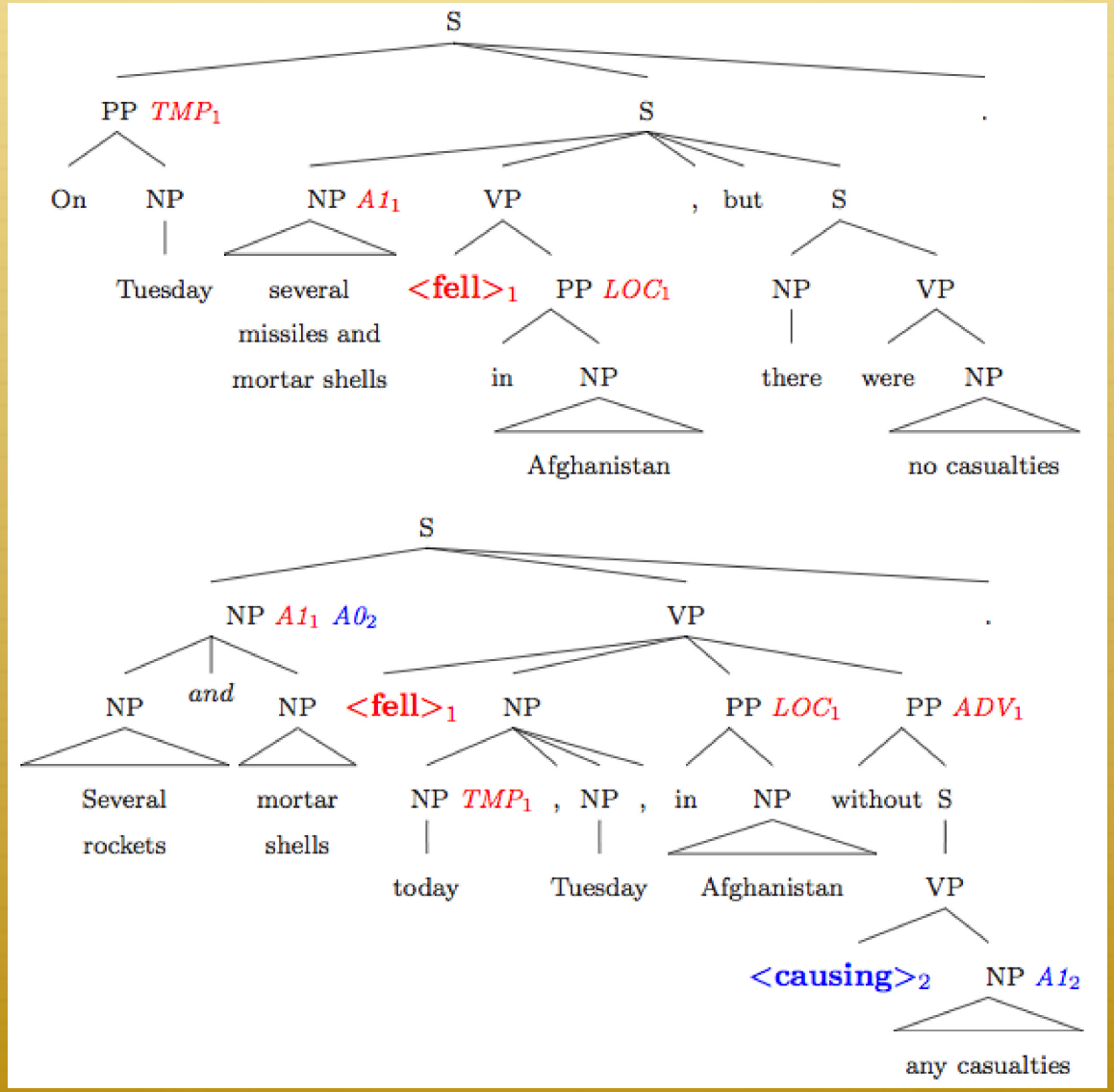
Example: Giménez and Màrquez, 2010

- 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 coecient
- 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

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 diferent 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)

Evaluation of syntactic measures

- NIST 2005 Arabic-to-English Exercise

Level	Metric	ρ all	ρ SMT
Lexical	BLEU	0.06	0.83
	METEOR	0.05	**0.90**
Syntactic	POS	0.42	0.89
	DP	**0.88**	0.86
	CP	0.74	**0.95**
Semantic	SR	0.72	**0.96**
	DR	0.92	0.92
	DR-POS	**0.97**	0.90

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

Metric Combination

- Diferent measures capture diferent 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
  - $M = \{ROUGE_W, METEOR, DP\text{-}HWC_r, DP\text{-}O_c(*), DP\text{-}O_l(*), DP\text{-}O_r(*), CP\text{-}STM_4, SR\text{-}O_r(*), SR\text{-}O_{rv}, DR\text{-}O_{rp}(*)\}$

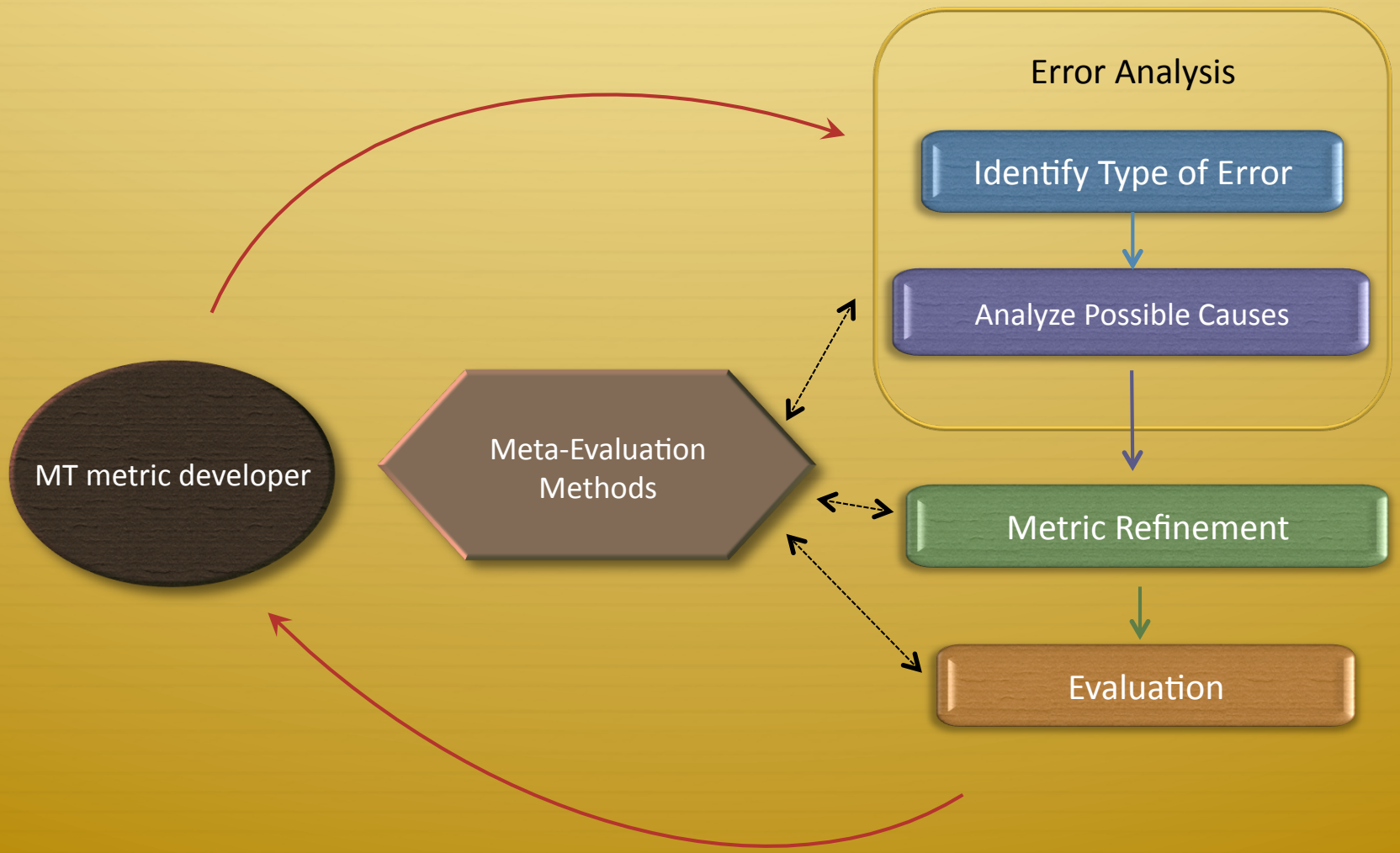
Learn new models

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

Meta evaluation

Metric-wise system development

MT Development cycle



Meta-evaluation

- Correlation with assessments
  - Pearson
  - Spearman
  - Kendall tau
- Orange [LO04]
- King [AGPV05]
- Consistency (ranking)

Conclusions

- 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)

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 error analysis

Overview

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

Metrics and Meta-metrics

- 813 metrics are available for language 'es' -> 'en'

- METRICS = { -PER, -TER, -TERbase, -TERp, -TERp-A, -WER, ALGNp, ALGNr, ALGNs, BLEU, BLEU-1, BLEU-2, BLEU-3, BLEU-4, BLEUi-2, BLEUi-3, BLEUi-4, CE-BiDictA, CE-BiDictO, CE-Nc, CE-Ne, CE-Oc, CE-Oe, CE-Op, CE-ippl, CE-ipplC, CE-ipplP, CE-length, CE-logp, CE-logpC, CE-logpP, CE-long, CE-oov, CE-short, CE-scrippl, CE-scripplC, CE-scripplP, CE-srclen, CE-srclogp, CE-srclogpC, CE-srclogpP, CE-srcoov, CE-symbols, CP-Oc(*), CP-Oc(ADJP), CP-Oc(ADVP), CP-Oc(CONJP), CP-Oc(FRAG), CP-Oc(INTJ), CP-Oc(LST), CP-Oc(NAC), CP-Oc(NP), CP-Oc(NX), CP-Oc(O), CP-Oc(PP), CP-Oc(PRN), CP-Oc(PRT), CP-Oc(QP), CP-Oc(RRC), CP-Oc(S), CP-Oc(SBAR), CP-Oc(SINV), CP-Oc(SQ), CP-Oc(UCP), CP-Oc(VP), CP-Oc(WHADJP), CP-Oc(WHADVP), CP-Oc(WHNP), CP-Oc(WHPP), CP-Oc(X), CP-Op(#), CP-Op($), CP-Op('), CP-Op((), CP-Op()), CP-Op(*), CP-Op(,), CP-Op(.), CP-Op(:), CP-Op(CC), CP-Op(CD), CP-Op(DT), CP-Op(EX), CP-Op(F), CP-Op(FW), CP-Op(IN), CP-Op(J), CP-Op(JJ), CP-Op(JJR), CP-Op(JJS), CP-Op(LS), CP-Op(MD), CP-Op(N), CP-Op(NN), CP-Op(NNP), CP-Op(NNPS), CP-Op(NNS), CP-Op(P), CP-Op(PDT), CP-Op(POS), CP-Op(PRP), CP-Op(PRP$), CP-Op(R), CP-Op(RB), CP-Op(RBR), CP-Op(RBS), CP-Op(RP), CP-Op(SYM), CP-Op(TO), CP-Op(UH), CP-Op(V), CP-Op(VB), CP-Op(VBD), CP-Op(VBG), CP-Op(VBN), CP-Op(VBP), CP-Op(VBZ), CP-Op(W), CP-Op(WDT), CP-Op(WP$), CP-Op(WP), CP-Op(WRB), CP-Op(``), CP-STM-1, CP-STM-2, CP-STM-3, CP-STM-4, CP-STM-5, CP-STM-6, CP-STM-7, CP-STM-8, CP-STM-9, CP-STMi-2, CP-STMi-3, CP-STMi-4, CP-STMi-5, CP-STMi-6, CP-STMi-7, CP-STMi-8, CP-STMi-9, DP-HWCM_c-1, DP-HWCM_c-2, DP-HWCM_c-3, DP-HWCM_c-4, DP-HWCM_r-1, DP-HWCM_r-2, DP-HWCM_r-3, DP-HWCM_r-4, DP-HWCM_w-1, DP-HWCM_w-2, DP-HWCM_w-3, DP-HWCM_w-4, DP-HWCMi_c-2, DP-HWCMi_c-3, DP-HWCMi_c-4, DP-HWCMi_r-2, DP-HWCMi_r-3, DP-HWCMi_r-4, DP-HWCMi_w-2, DP-HWCMi_w-3, DP-HWCMi_w-4, DP-Oc(*), DP-Oc(a), DP-Oc(as), DP-Oc(aux), DP-Oc(be), DP-Oc(c), DP-Oc(comp), DP-Oc(det), DP-Oc(have), DP-Oc(n), DP-Oc(postdet), DP-Oc(ppspec), DP-Oc(predet), DP-Oc(prep), DP-Oc(saidx), DP-Oc(sentadjunct), DP-Oc(subj), DP-Oc(that), DP-Oc(u), DP-Oc(v), DP-Oc(vbe), DP-Oc(xsaid), DP-Ol(*), DP-Ol(1), DP-Ol(2), DP-Ol(3), DP-Ol(4), DP-Ol(5), DP-Ol(6), DP-Ol(7), DP-Ol(8), DP-Ol(9), DP-Or(*), DP-Or(amod), DP-Or(amount-value), DP-Or(appo), DP-Or(appo-mod), DP-Or(as-arg), DP-Or(as1), DP-Or(as2), DP-Or(aux), DP-Or(be), DP-Or(being), DP-Or(by-subj), DP-Or(c), DP-Or(cn), DP-Or(comp1), DP-Or(conj), DP-Or(desc), DP-Or(dest), DP-Or(det), DP-Or(else), DP-Or(fc), DP-Or(gen), DP-Or(guest), DP-Or(have), DP-Or(head), DP-Or(i), DP-Or(inv-aux), DP-Or(inv-have), DP-Or(lex-dep), DP-Or(lex-mod), DP-Or(mod), DP-Or(mod-before), DP-Or(neg), DP-Or(nn), DP-Or(num), DP-Or(num-mod), DP-Or(obj), DP-Or(obj1), DP-Or(obj2), DP-Or(p), DP-Or(p-spec), DP-Or(pcomp-c), DP-Or(pcomp-n), DP-Or(person), DP-Or(pnmod), DP-Or(poss), DP-Or(post), DP-Or(pre), DP-Or(pred), DP-Or(punc), DP-Or(rel), DP-Or(s), DP-Or(sc), DP-Or(subcat), DP-Or(subclass), DP-Or(subj), DP-Or(title), DP-Or(vrel), DP-Or(wha), DP-Or(whn), DP-Or(whp), DPm-HWCM_c-1, DPm-HWCM_c-2, DPm-HWCM_c-3, DPm-HWCM_c-4, DPm-HWCM_r-1, DPm-HWCM_r-2, DPm-HWCM_r-3, DPm-HWCM_r-4, DPm-HWCM_w-1, DPm-HWCM_w-2, DPm-HWCM_w-3, DPm-HWCM_w-4, DPm-HWCMi_c-2, DPm-HWCMi_c-3, DPm-HWCMi_c-4, DPm-HWCMi_r-2, DPm-HWCMi_r-3, DPm-HWCMi_r-4, DPm-HWCMi_w-2, DPm-HWCMi_w-3, DPm-HWCMi_w-4, DPm-Oc(abbrev), DPm-Oc(acomp), DPm-Oc(advcl), DPm-Oc(advmod), DPm-Oc(agent), DPm-Oc(amod), DPm-Oc(appos), DPm-Oc(arg), DPm-Oc(attr), DPm-Oc(aux), DPm-Oc(auxpass), DPm-Oc(cc), DPm-Oc(ccomp), DPm-Oc(comp), DPm-Oc(complm), DPm-Oc(conj), DPm-Oc(cop), DPm-Oc(csubj), DPm-Oc(csubjpass), DPm-Oc(dep), DPm-Oc(det), DPm-Oc(dobj), DPm-Oc(expl), DPm-Oc(infmod), DPm-Oc(iobj), DPm-Oc(mark), DPm-Oc(mod), DPm-Oc(mwe), DPm-Oc(neg), DPm-Oc(nn), DPm-Oc(npadvmod), DPm-Oc(nsubj), DPm-Oc(nsubjpass), DPm-Oc(num), DPm-Oc(number), DPm-Oc(obj), DPm-Oc(parataxis), DPm-Oc(partmod), DPm-Oc(pobj), DPm-Oc(poss), DPm-Oc(possessive), DPm-Oc(preconj), DPm-Oc(predet), DPm-Oc(prep), DPm-Oc(prt), DPm-Oc(punct), DPm-Oc(purpcl), DPm-Oc(quantmod), DPm-Oc(rcmod), DPm-Oc(ref), DPm-Oc(rel), DPm-Oc(sdep), DPm-Oc(subj), DPm-Oc(tmod), DPm-Oc(xcomp), DPm-Oc(xsubj), DPm-Ol(1), DPm-Ol(2), DPm-Ol(3), DPm-Ol(4), DPm-Ol(5), DPm-Ol(6), DPm-Ol(7), DPm-Ol(8), DPm-Ol(9), DPm-Or(abbrev), DPm-Or(acomp), DPm-Or(advcl), DPm-Or(advmod), DPm-Or(agent), DPm-Or(amod), DPm-Or(appos), DPm-Or(arg), DPm-Or(attr), DPm-Or(aux), DPm-Or(auxpass), DPm-Or(cc), DPm-Or(ccomp), DPm-Or(comp), DPm-Or(complm), DPm-Or(conj), DPm-Or(cop), DPm-Or(csubj), DPm-Or(csubjpass), DPm-Or(dep), DPm-Or(det), DPm-Or(dobj), DPm-Or(expl), DPm-Or(infmod), DPm-Or(iobj), DPm-Or(mark), DPm-Or(mod), DPm-Or(mwe), DPm-Or(neg), DPm-Or(nn), DPm-Or(npadvmod), DPm-Or(nsubj), DPm-Or(nsubjpass), DPm-Or(num), DPm-Or(number), DPm-Or(obj), DPm-Or(parataxis), DPm-Or(partmod), DPm-Or(pobj), DPm-Or(poss), DPm-Or(possessive), DPm-Or(preconj), DPm-Or(predet), DPm-Or(prep), DPm-Or(prt), DPm-Or(punct), DPm-Or(purpcl), DPm-Or(quantmod), DPm-Or(rcmod), DPm-Or(ref), DPm-Or(rel), DPm-Or(sdep), DPm-Or(subj), DPm-Or(tmod), DPm-Or(xcomp), DPm-Or(xsubj), DR-Fr(*), DR-Frp(*), DR-Ol, DR-Or(*), DR-Or(*)_b, DR-Or(*)_i, DR-Or(alfa), DR-Or(card), DR-Or(drs), DR-Or(eq), DR-Or(imp), DR-Or(merge), DR-Or(named), DR-Or(not), DR-Or(or), DR-Or(pred), DR-Or(prop), DR-Or(rel), DR-Or(smerge), DR-Or(timex), DR-Or(whq), DR-Or-(dr), DR-Orp(*), DR-Orp(*)_b, DR-Orp(*)_i, DR-Orp(alfa), DR-Orp(card), DR-Orp(dr), DR-Orp(drs), DR-Orp(eq), DR-Orp(imp), DR-Orp(merge), DR-Orp(named), DR-Orp(not), DR-Orp(or), DR-Orp(pred), DR-Orp(prop), DR-Orp(rel), DR-Orp(smerge), DR-Orp(timex), DR-Orp(whq), DR-Pr(*), DR-Prp(*), DR-Rr(*), DR-Rrp(*), DR-STM-1, DR-STM-2, DR-STM-3, DR-STM-4, DR-STM-4_b, DR-STM-4_i, DR-STM-5, DR-STM-6, DR-STM-7, DR-STM-8, DR-STM-9, DR-STMi-2, DR-STMi-3, DR-STMi-4, DR-STMi-5, DR-STMi-6, DR-STMi-7, DR-STMi-8, DR-STMi-9, DRdoc-Ol, DRdoc-Or(*), DRdoc-Or(*)_b, DRdoc-Or(*)_i, DRdoc-Or(alfa), DRdoc-Or(card), DRdoc-Or(dr), DRdoc-Or(drs), DRdoc-Or(eq), DRdoc-Or(imp), DRdoc-Or(merge), DRdoc-Or(named), DRdoc-Or(not), DRdoc-Or(or), DRdoc-Or(pred), DRdoc-Or(prop), DRdoc-Or(rel), DRdoc-Or(smerge), DRdoc-Or(timex), DRdoc-Or(whq), DRdoc-Orp(*), DRdoc-Orp(*)_b, DRdoc-Orp(*)_i, DRdoc-Orp(alfa), DRdoc-Orp(card), DRdoc-Orp(dr), DRdoc-Orp(drs), DRdoc-Orp(eq), DRdoc-Orp(imp), DRdoc-Orp(merge), DRdoc-Orp(named), DRdoc-Orp(not), DRdoc-Orp(or), DRdoc-Orp(pred), DRdoc-Orp(prop), DRdoc-Orp(rel), DRdoc-Orp(smerge), DRdoc-Orp(timex), DRdoc-Orp(whq), DRdoc-STM-1, DRdoc-STM-2, DRdoc-STM-3, DRdoc-STM-4, DRdoc-STM-4_b, DRdoc-STM-4_i, DRdoc-STM-5, DRdoc-STM-6, DRdoc-STM-7, DRdoc-STM-8, DRdoc-STM-9, DRdoc-STMi-2, DRdoc-STMi-3, DRdoc-STMi-4, DRdoc-STMi-5, DRdoc-STMi-6, DRdoc-STMi-7, DRdoc-STMi-8, DRdoc-STMi-9, F1, GTM-1, GTM-2, GTM-3, METEOR-ex, METEOR-pa, METEOR-st, METEOR-sy, NE-Me(*), NE-Me(ANGLE_QUANTITY), NE-Me(DATE), NE-Me(DISTANCE_QUANTITY), NE-Me(LANGUAGE), NE-Me(LOC), NE-Me(MEASURE), NE-Me(METHOD), NE-Me(MISC), NE-Me(MONEY), NE-Me(NUM), NE-Me(ORG), NE-Me(PER), NE-Me(PERCENT), NE-Me(PROJECT), NE-Me(SIZE_QUANTITY), NE-Me(SPEED_QUANTITY), NE-Me(SYSTEM), NE-Me(TEMPERATURE_QUANTITY), NE-Me(TIME), NE-Me(WEIGHT_QUANTITY), NE-Oe(*), NE-Oe(**), NE-Oe(ANGLE_QUANTITY), NE-Oe(DATE), NE-Oe(DISTANCE_QUANTITY), NE-Oe(LANGUAGE), NE-Oe(LOC), NE-Oe(MEASURE), NE-Oe(METHOD), NE-Oe(MISC), NE-Oe(MONEY), NE-Oe(NUM), NE-Oe(O), NE-Oe(ORG), NE-Oe(PER), NE-Oe(PERCENT), NE-Oe(PROJECT), NE-Oe(SIZE_QUANTITY), NE-Oe(SPEED_QUANTITY), NE-Oe(SYSTEM), NE-Oe(TEMPERATURE_QUANTITY), NE-Oe(TIME), NE-Oe(WEIGHT_QUANTITY), NIST, NIST-1, NIST-2, NIST-3, NIST-4, NIST-5, NISTi-2, NISTi-3, NISTi-4, NISTi-5, Ol, PER, Pl, ROUGE-1, ROUGE-2, ROUGE-3, ROUGE-4, ROUGE-L, ROUGE-S*, ROUGE-SU*, ROUGE-W, RI, SP-Oc(*), SP-Oc(ADJP), SP-Oc(ADVP), SP-Oc(CONJP), SP-Oc(INTJ), SP-Oc(LST), SP-Oc(NP), SP-Oc(O), SP-Oc(PP), SP-Oc(PRT), SP-Oc(SBAR), SP-Oc(UCP), SP-Oc(VP), SP-Op(#), SP-Op($), SP-Op('), SP-Op((), SP-Op()), SP-Op(*), SP-Op(,), SP-Op(.), SP-Op(:), SP-Op(CC), SP-Op(CD), SP-Op(DT), SP-Op(EX), SP-Op(F), SP-Op(FW), SP-Op(IN), SP-Op(J), SP-Op(JJ), SP-Op(JJR), SP-Op(JJS), SP-Op(LS), SP-Op(MD), SP-Op(N), SP-Op(NN), SP-Op(NNP), SP-Op(NNPS), SP-Op(NNS), SP-Op(P), SP-Op(PDT), SP-Op(POS), SP-Op(PRP$), SP-Op(PRP), SP-Op(R), SP-Op(RB), SP-Op(RBR), SP-Op(RBS), SP-Op(RP), SP-Op(SYM), SP-Op(TO), SP-Op(UH), SP-Op(V), SP-Op(VB), SP-Op(VBD), SP-Op(VBG), SP-Op(VBN), SP-Op(VBP), SP-Op(VBZ), SP-Op(W), SP-Op(WDT), SP-Op(WP$), SP-Op(WP), SP-Op(WRB), SP-Op(``), SP-cNIST, SP-cNIST-1, SP-cNIST-2, SP-cNIST-3, SP-cNIST-4, SP-cNIST-5, SP-cNISTi-2, SP-cNISTi-3, SP-cNISTi-4, SP-cNISTi-5, SP-iobNIST, SP-iobNIST-1, SP-iobNIST-2, SP-iobNIST-3, SP-iobNIST-4, SP-iobNIST-5, SP-iobNISTi-2, SP-iobNISTi-3, SP-iobNISTi-4, SP-iobNISTi-5, SP-lNIST, SP-lNIST-1, SP-lNIST-2, SP-lNIST-3, SP-lNIST-4, SP-lNIST-5, SP-lNISTi-2, SP-lNISTi-3, SP-lNISTi-4, SP-lNISTi-5, SP-pNIST, SP-pNIST-1, SP-pNIST-2, SP-pNIST-3, SP-pNIST-4, SP-pNIST-5, SP-pNISTi-2, SP-pNISTi-3, SP-pNISTi-4, SP-pNISTi-5, SR-Fr(*), SR-MFr(*), SR-MPr(*), SR-MRr(*), SR-Mr(*), SR-Mr(*)_b, SR-Mr(*)_i, SR-Mr(A0), SR-Mr(A1), SR-Mr(A2), SR-Mr(A3), SR-Mr(A4), SR-Mr(A5), SR-Mr(AA), SR-Mr(AM-ADV), SR-Mr(AM-CAU), SR-Mr(AM-DIR), SR-Mr(AM-DIS), SR-Mr(AM-EXT), SR-Mr(AM-LOC), SR-Mr(AM-MNR), SR-Mr(AM-MOD), SR-Mr(AM-NEG), SR-Mr(AM-PNC), SR-Mr(AM-PRD), SR-Mr(AM-REC), SR-Mr(AM-TMP), SR-Mra(*), SR-Mrv(*), SR-Mrv(*)_b, SR-Mrv(*)_i, SR-Mrv(A0), SR-Mrv(A1), SR-Mrv(A2), SR-Mrv(A3), SR-Mrv(A4), SR-Mrv(A5), SR-Mrv(AA), SR-Mrv(AM-ADV), SR-Mrv(AM-CAU), SR-Mrv(AM-DIR), SR-Mrv(AM-DIS), SR-Mrv(AM-EXT), SR-Mrv(AM-LOC), SR-Mrv(AM-MNR), SR-Mrv(AM-MOD), SR-Mrv(AM-NEG), SR-Mrv(AM-PNC), SR-Mrv(AM-PRD), SR-Mrv(AM-REC), SR-Mrv(AM-TMP), SR-Nv, SR-Ol, SR-Or, SR-Or(*), SR-Or(*)_b, SR-Or(*)_i, SR-Or(A0), SR-Or(A1), SR-Or(A2), SR-Or(A3), SR-Or(A4), SR-Or(A5), SR-Or(AA), SR-Or(AM-ADV), SR-Or(AM-CAU), SR-Or(AM-DIR), SR-Or(AM-DIS), SR-Or(AM-EXT), SR-Or(AM-LOC), SR-Or(AM-MNR), SR-Or(AM-MOD), SR-Or(AM-NEG), SR-Or(AM-PNC), SR-Or(AM-PRD), SR-Or(AM-REC), SR-Or(AM-TMP), SR-Or_b, SR-Or_i, SR-Ora, SR-Ora(*), SR-Orv, SR-Orv(*), SR-Orv(*)_b, SR-Orv(*)_i, SR-Orv(A0), SR-Orv(A1), SR-Orv(A2), SR-Orv(A3), SR-Orv(A4), SR-Orv(A5), SR-Orv(AA), SR-Orv(AM-ADV), SR-Orv(AM-CAU), SR-Orv(AM-DIR), SR-Orv(AM-DIS), SR-Orv(AM-EXT), SR-Orv(AM-LOC), SR-Orv(AM-MNR), SR-Orv(AM-MOD), SR-Orv(AM-NEG), SR-Orv(AM-PNC), SR-Orv(AM-PRD), SR-Orv(AM-REC), SR-Orv(AM-TMP), SR-Orv_b, SR-Orv_i, SR-Ov, SR-Pr(*), SR-Rr(*), TER, TERbase, TERp, TERp-A, WER }

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 operates over test suites (or test beds).
  - a test suite is a collection of test cases:
    - Source segment
    - Candidate translation(s)
    - Reference translation(s)

The Test Suit

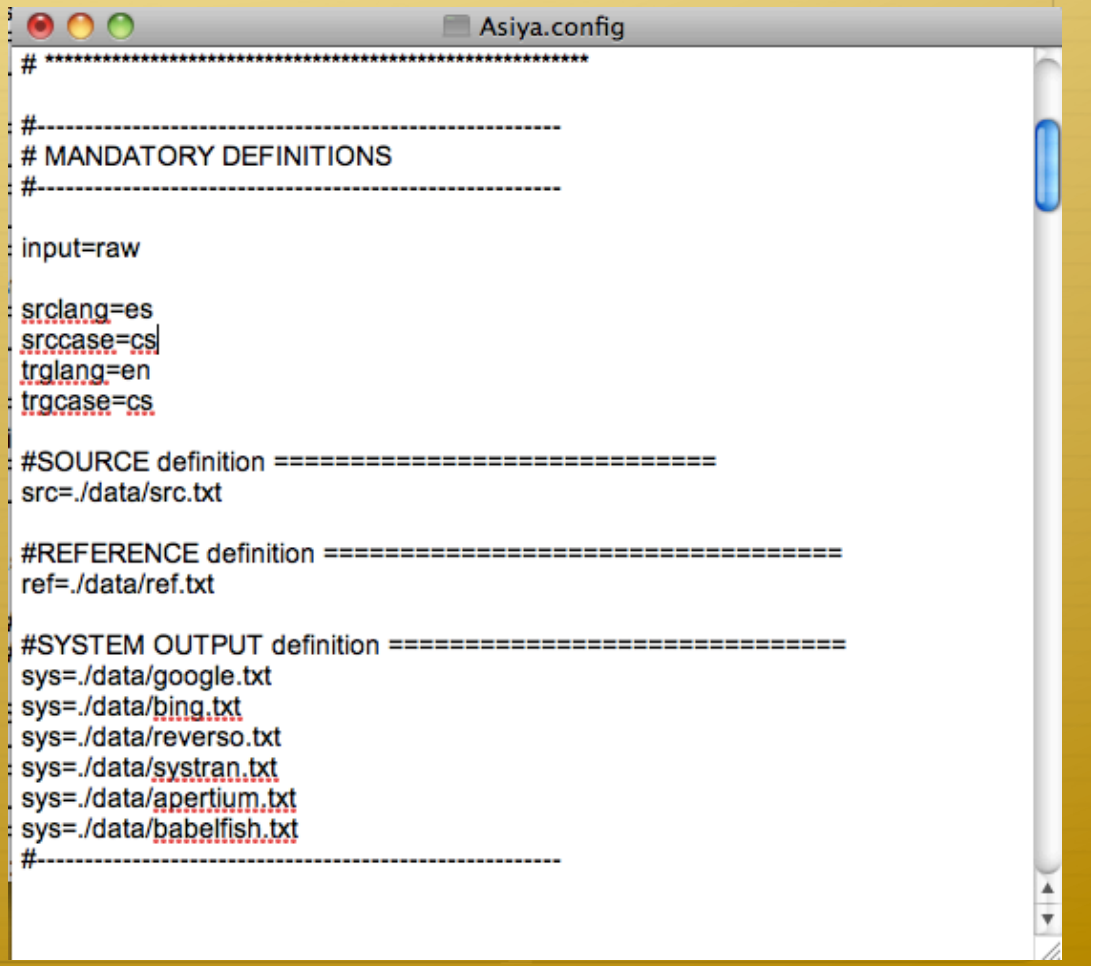
- Asiya.pl Asiya.config
- Asiya.config:

Asiya.config

```
# ******************************************************

#-------------------------------------------------------
# MANDATORY DEFINITIONS
#-------------------------------------------------------

input=raw

srclang=es
srccase=cs
trglang=en
trgcase=cs

#SOURCE definition ==============================
src=./data/src.txt

#REFERENCE definition ===================================
ref=./data/ref.txt

#SYSTEM OUTPUT definition ===============================
sys=./data/google.txt
sys=./data/bing.txt
sys=./data/reverso.txt
sys=./data/systran.txt
sys=./data/apertium.txt
sys=./data/babelfish.txt
#-------------------------------------------------------
```

General Options

- Input format
  - Raw
  - Nist
- Language pair
  - Srclang
  - Trglang
- Predefined sets of metrics, systems and references



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>
  - Correlation with assessments
    - Pearson
    - Spearman
    - Kendall tau
  - Orange [LO04]
  - King [AGPV05]
  - Consistency
- Learn

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

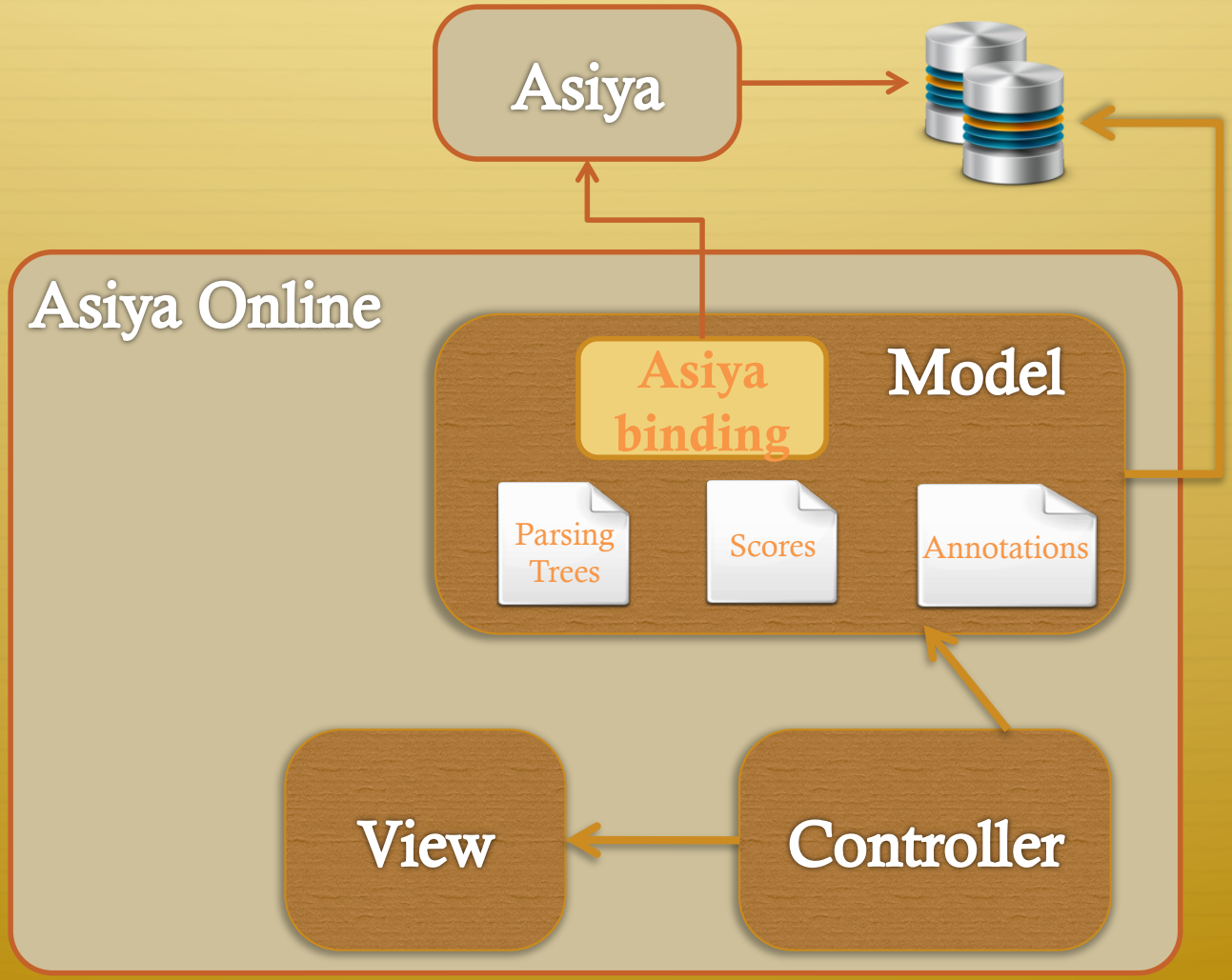
Learning

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

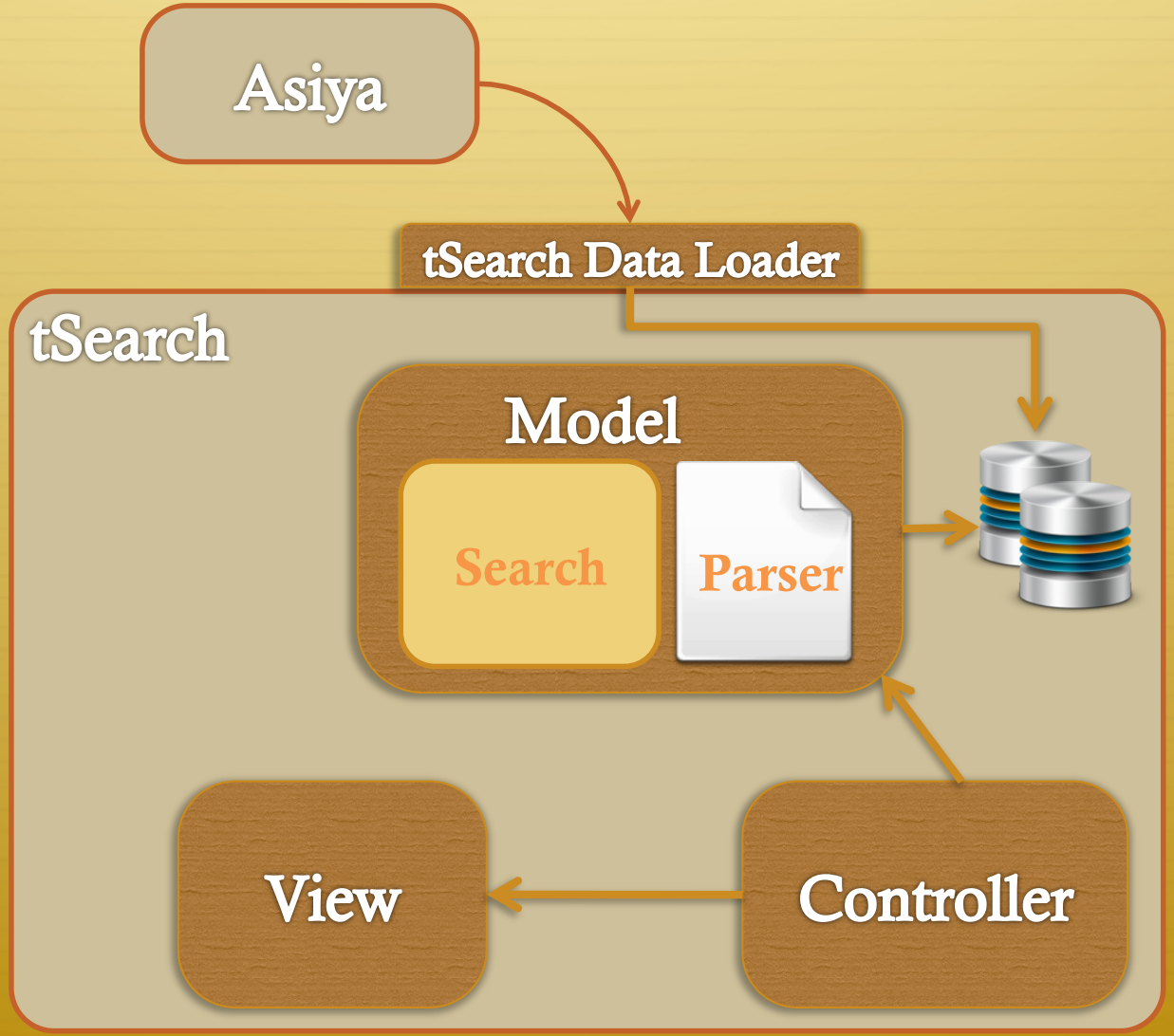
Asiya Interfaces

- Asiya 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 specic criteria related to their quality (as assessed by the automatic scores).
- Asiya Web Service
  - A RESTful web service to access the Asiya evaluation.

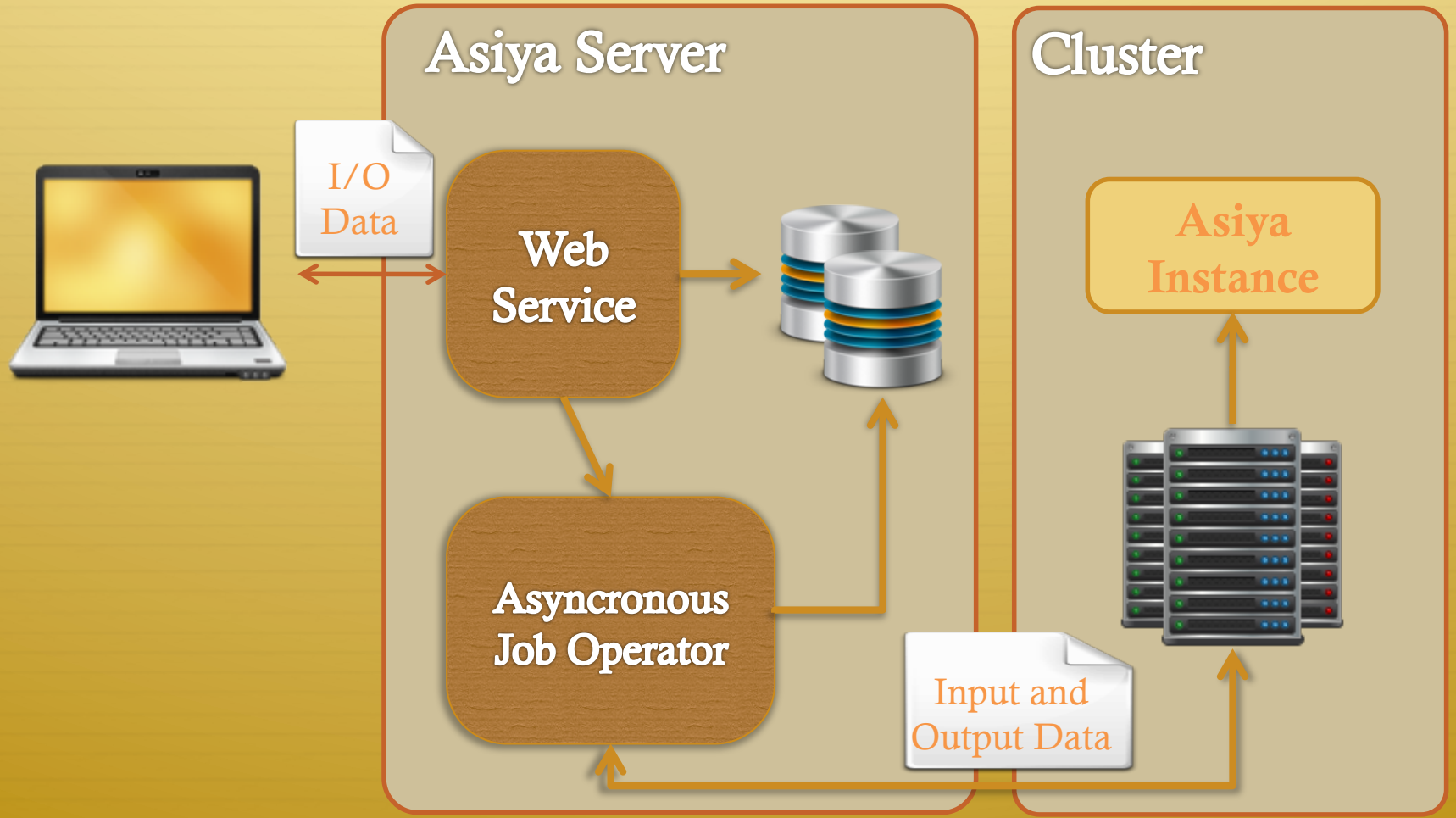
Asiya Online Interface



Asiya tSearch



Asiya Web Service



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.