Meritxell Gonzàlez

TALP Research Center – Universitat Politècnica de Catlaunya

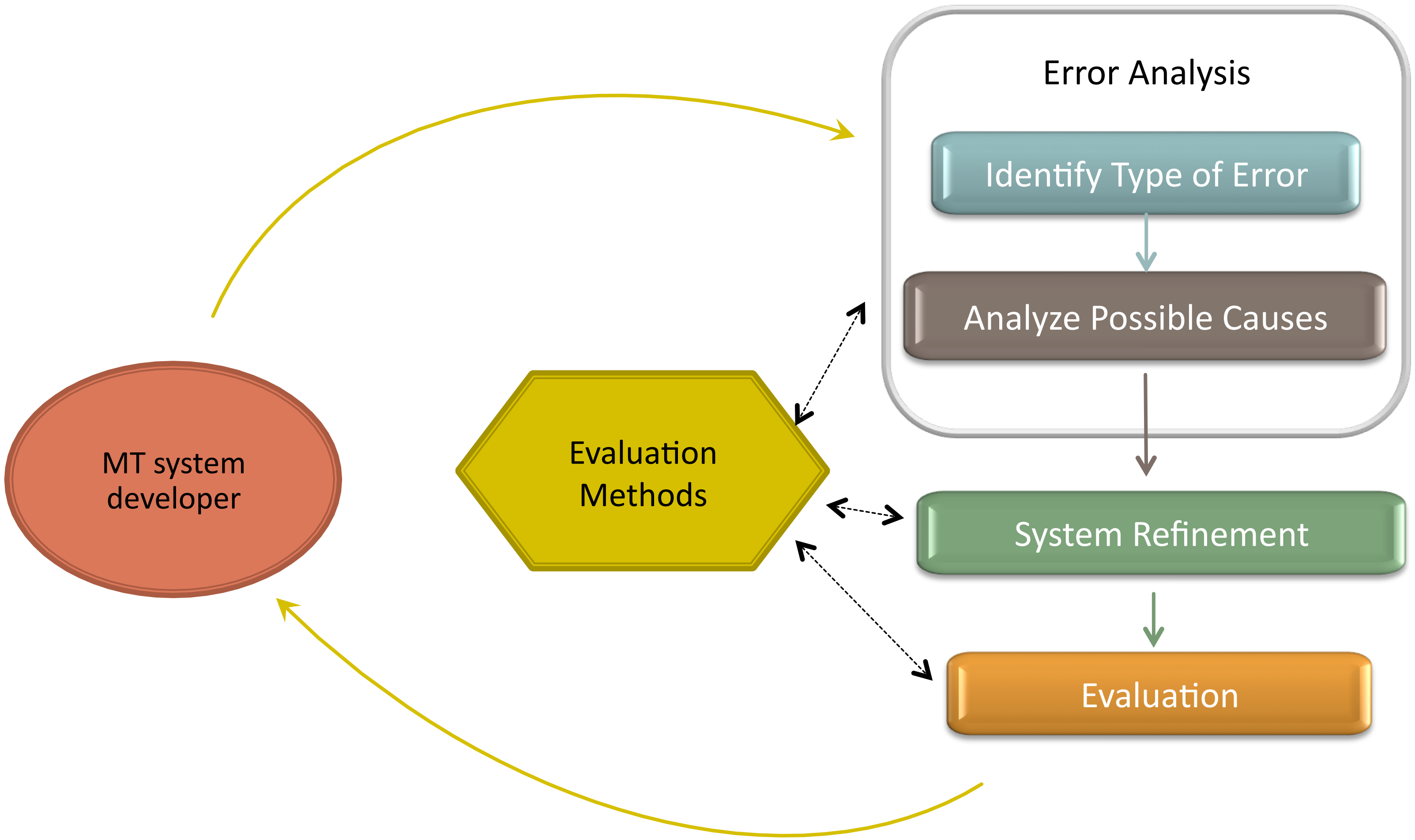
LREC 2014 – May 31st, Reykjavik, Iceland

Automatic MT Evaluation

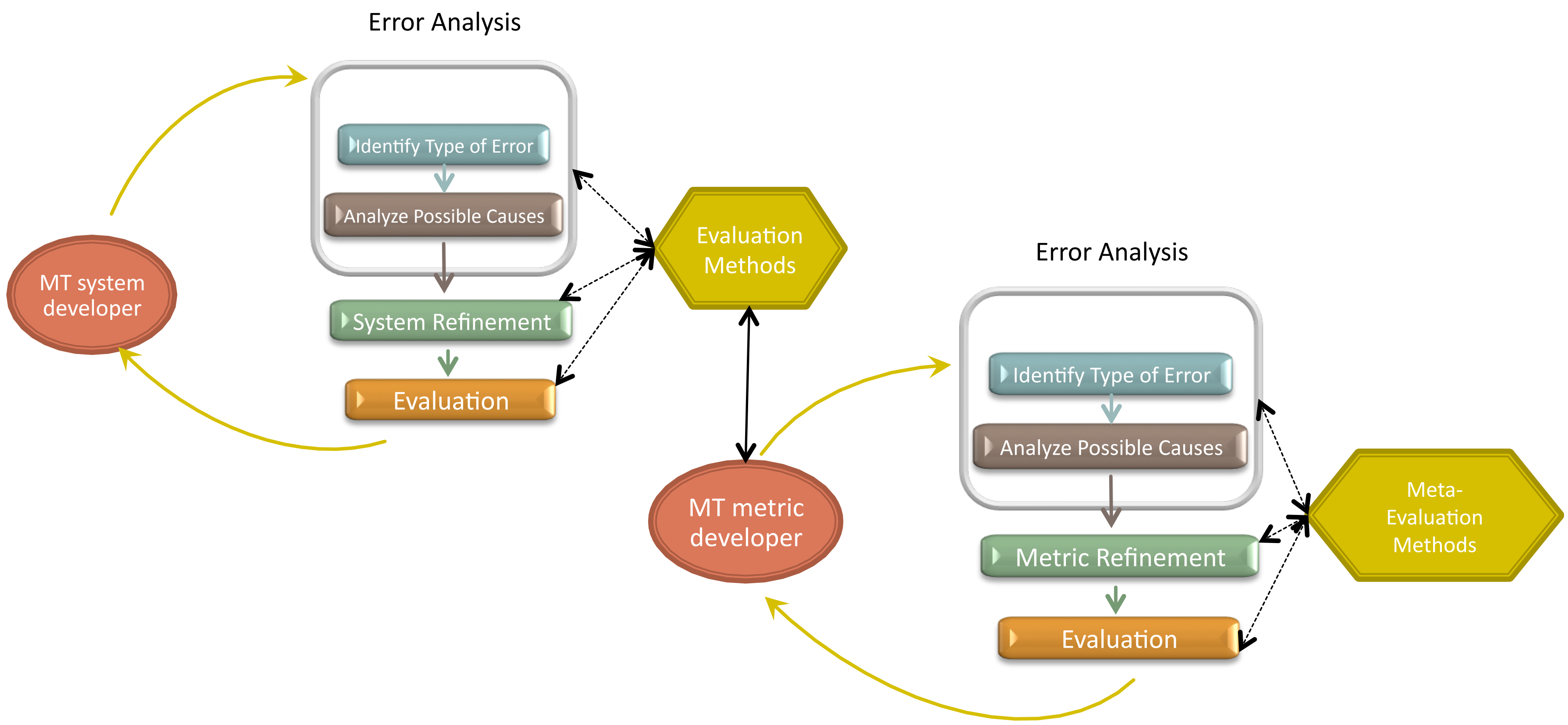
Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- Quality estimation
- The Asiya toolkit

MT Development cycle (1)



MT Development cycle (2)



Difficulties of the MT evaluation (1)

- Machine Translation is an open NLP problem
  - The correct translation is not unique.
  - The set of valid translations is not small.
  - The quality of a translation is a fuzzy concep.

Difficulties of the MT evaluation (2)

- Quality aspects are heterogeneous:
  - **Adequacy (or Fidelity)**: Does the output convey the same meaning as the input sentence? Is part of the message lost, added, or distorted?
  - **Fluency (or Intelligibility)**: Is the output fluent? This involves both grammatical correctness and idiomatic word choices.
  - **Post-edition** effort: time required to repair the translation, number of key strokes, etc.

Example

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

- **The remote control of the Wii helps to diagnose an infantile ocular disease.**

- The control of the Wii **help** to diagnose an ocular **illness childish.**
- The **control of the Wii** helps to diagnose an infantile ocular disease.
- The Wii remote **helps diagnose** a childhood eye disease.
- **The Wii Remote** to help diagnose childhood eye disease.
- The control of the Wii helps to diagnose an **ocular infantile** disease.
- The **mando** of the Wii helps to diagnose an infantile ocular disease.

Manual vs. automatic evaluation

- Categorisation problem for human annotations.
  - 5-point likert scale [LDC05]
  - 4-point likert scale [TAUS13]

Adequacy	
4	All
3	Most
2	Little
1	None

Fluency	
4	Flawless
3	Good
2	Disfluent
1	Incomprehensible

- Ranking problem for human annotations [Cal12]
- Regression problem for automatic metrics.

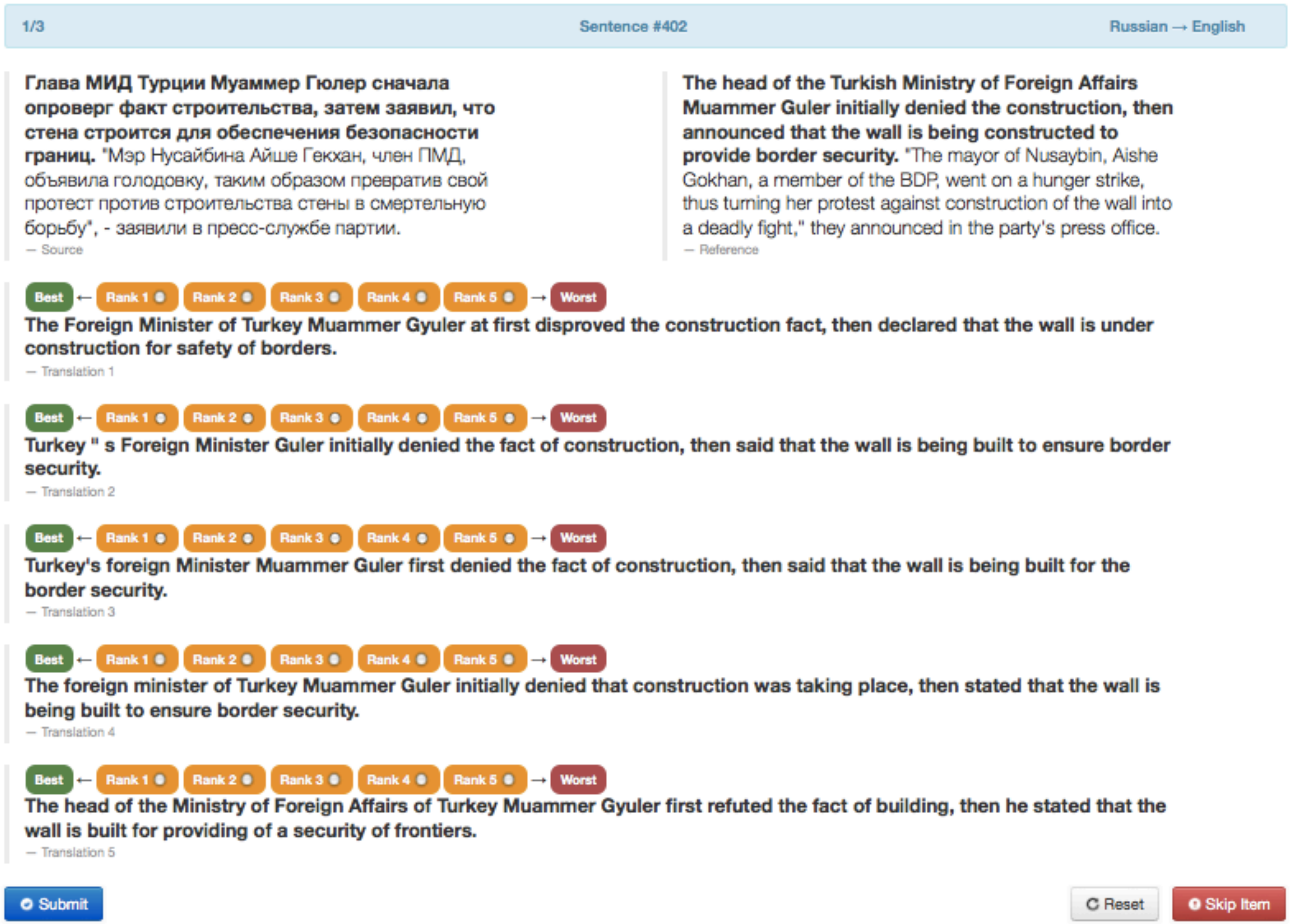
Meta-Evaluation

- Correlation with human assessments
  - **Pearson (system level)**
  - Spearman
  - **Kendall's tau (segment level)**
- Consistency (ranking)
- AvgDelta [Cal12]

Human Annotation Tools

- BLAST [Sty11] - annotate
- Appraise [Fed12] - rank
- DQF [Tau12] – best practices
- Costa MT Evaluation Tool [Chat13] – error classification

Appraise [Fed12]



Interanotator Agreement

- Cohen's kappa coefficient [Coh60]
- WMT13 [Boj13]

Pair	Inter-κ	Intra-κ
CZ-EN	0.244	0.479
EN-CZ	0.168	0.290
DE-EN	0.299	0.535
EN-DE	0.267	0.498
ES-EN	0.277	0.575
EN-ES	0.206	0.492
FR-EN	0.275	0.578
EN-FR	0.231	0.495
RU-EN	0.278	0.450
EN-RU	0.243	0.513

- Kappa interpretation [Lan77]
  - 0.0–0.2 slight
  - 0.2–0.4 fair
  - 0.4–0.6 moderate
  - 0.6–0.8 substantial
  - 0.8–1.0 almost perfect

Benefits of Automatic Evaluation (1)

- Compared to manual evaluation, automatic measures are:
  - Cheap (vs. costly)
  - Objective (vs. subjective)
  - Reusable (vs. not-reusable)

Benefits of Automatic Evaluation (2)

- Automatic evaluation metrics have notably accelerated the development cycle of MT systems
  - Error analysis
    - Identify and analyze weak points
  - System optimization
    - Ranking of N-best list and parameter estimation
  - System comparison
    - Phrase- or system-based combination

Active Topic of Research

- Annual metrics competition organized by the WMT workshop series and supported by the EC
  - http://www.statmt.org/wmt14/
  - Both Evaluation Measure and Confidence Estimation
- Biannual OpenMT metric competition organized by NIST and supported by DARPA
  - http://www.nist.gov/itl/iad/mig/openmt.cfm
  - Evaluation Measures for informal data genres and speech translations
- 1st Workshop on Asian Translation, Tokyo, October 2014
  - http://orchid.kuee.kyoto-u.ac.jp/WAT/
  - Japanese-Chinese, test data is prepared using paragraph as a unit

Overview

- Introduction
- **Automatic MT evaluation**
- Linguistically motivated evaluation measures
- Quality estimation
- The Asiya toolkit

MT Automatic Evaluation (1)

- Setting:
  - Compute the similarity between a 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 (2)

- Goals:
  - **Low cost**
  - **Tunable**
  - **Meaningful**
  - **Coherent**
  - **Consistent**

First Approaches

- Lexical similarity as a measure of quality
  - Edit Distance: WER [Nie00], PER [Til97], TER [Sno06]
  - Precision: BLEU [Pap01], NIST [Dod02]
  - Recall: ROUGE [Lin04a]
  - Precision/Recall: GTM [Mel03], METEOR [Ban05,Den10]

Precision and Recall of Words (1)

- The remote control of the Wii helps to diagnose an infantile ocular disease.
- **The Wii remote helps diagnose a childhood eye disease.**

Precision and Recall of Words (2)

- <u>**The remote**</u> control of the <u>**Wii helps**</u> to <u>**diagnose**</u> an infantile ocular <u>**disease .**</u>

- <u>**The Wii remote helps diagnose**</u> a childhood eye <u>**disease .**</u>
  - **Precision**: $\frac{correct}{output\_length} = \frac{7}{10} = 0.7$
  - **Recall**: $\frac{correct}{reference\_length} = \frac{7}{14} = 0.5$
  - **F-measure**: $\frac{precision * recall}{(precision + recall)/2} = \frac{0.35}{0.6} = 0.583$

Precision and Recall of Words (3)

- The remote control of the Wii helps to diagnose an infantile ocular disease .
- Wii the control of the remote to diagnose disease helps an ocular infantile .
  - **Precision:** $\frac{correct}{output\_length} = \frac{14}{14} = 1.00$
  - **Recall:** $\frac{correct}{reference\_length} = \frac{14}{14} = 1.00$
  - **F-measure:** $\frac{precision * recall}{(precision + recall)/2} = \frac{1.00}{1.00} = 1.00$

No Penalty for reordering!

IBM BLEU (1)

- "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." [Pap01]

IBM BLEU (2)

- **Modified** *N*-gram precision between machine translation output and reference translation.
  - Usually with *n*-grams of size 1 to 4

- Modified *n*-gram precision on the entire corpus

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} count_{clip}(ngram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C} count_{clip}(ngram')}$$

- Brevity penalty for too short translations.

$$BP = \begin{cases} 1 & if \quad c > r \\ e^{1-\frac{r}{c}} & \end{cases}$$

  - Typically computed over the entire corpus, not single sentences.

IBM BLEU (3)

- The remote control of the Wii helps to diagnose an infantile ocular disease .
- The control of the Wii helps to diagnose an ocular infantile disease .
- $w_n = 1/4$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n \right)$$

	p_n	BP * p_n	$w_n \log P_n$
1-gram precision	13/13 = 1.0	0.926	0
2-gram precision	8/12 = 0.667	0.617	-0.405
3-gram precision	6/11 = 0.545	0.505	-0.606
4-gram precision	5/10 = 0.5	0.463	-0.693
Brevity penalty	0.926		
P_n	Same, only one sentence		
BLEU score	**0.6046**		

Problems of lexical similarities (1)

- The reliability of lexical metrics depends strongly on the heterogeneity/representativity of reference translations.

- Actually, human translations tend to score low on BLEU.

- Underlying Cause:
  - Lexical similarity is neither a sufficient nor a necessary condition so that two sentences convey the same meaning.

Problems of lexical similarities (2)

- Statistical MT systems heavily rely on the training data.
- Testsets tend to be similar (domain, register, sublanguage) to training materials.
- *N*-gram based metrics favour MT systems which closely replicate the lexical realization of the references.
- Statistical MT systems tend to share the reference sublanguage and be favoured by *n*-gram-based measures.

Overview

- Introduction
- Automatic MT evaluation
- **Linguistically motivated evaluation measures**
- Quality estimation
- The *Asiya* toolkit

Linguistically motivated measures (1)

- Extending Lexical Similarity Measures to increase robustness [Gim09]

  - Lexical variants:
    - Morphological information (i.e., stemming) **ROUGE** and **METEOR**
    - Synonymy lookup : **METEOR** (based on WordNet)

  - Paraphrasing support:
    - Extended versions of **METEOR**, **TER**

  - Equivalent reference translation graph:
    - **HyTER** [Dre12]

METEOR (1) [Ban05]

- Parameterized harmonic mean of word P and R

$$F_{mean} = \frac{P \cdot R}{\alpha \cdot P + (1-\alpha) \cdot R}$$

- Matching algorithm
  - Exact matching
  - Partial credit for matching stems
  - Partial credit for matching synonyms

$$Pen = \gamma \cdot \left(\frac{ch}{m}\right)^{\beta}$$

- *N*-gram **penalty** based on the number of **chunks** with longer length of adjacent words **matched** in both strings

- Final score:

$$METEOR = (1 - Pen) \cdot F_{mean}$$

METEOR (2) [Den10]

- Extensions
  - METEOR-NEXT
    - Weighted matches depending on the type
    - Phrase-level matches
    - new matching algorithm accounting for start-positions distance

  - Paraphrasing
    - Paraphrase tables from parallel corpora
    - Used by the paraphrase matcher

  - δ parameter: content vs. function words discrimination

More linguistically-motivated measures

- Features capturing syntactic and semantic information
  - Shallow parsing, constituency and dependency parsing, named entities, semantic roles, textual entailment, discourse representation, error categories, …

- Some linguistically-motivated measures:
  - IQmt [Gim09] – syntactic and semantics
  - MaxSim [Cha08] - syntactic
  - RTE [Pad09] – textual entailment
  - VERTa [Com14] – syntactic and semantics

Example 1: Structural Similarity (1)

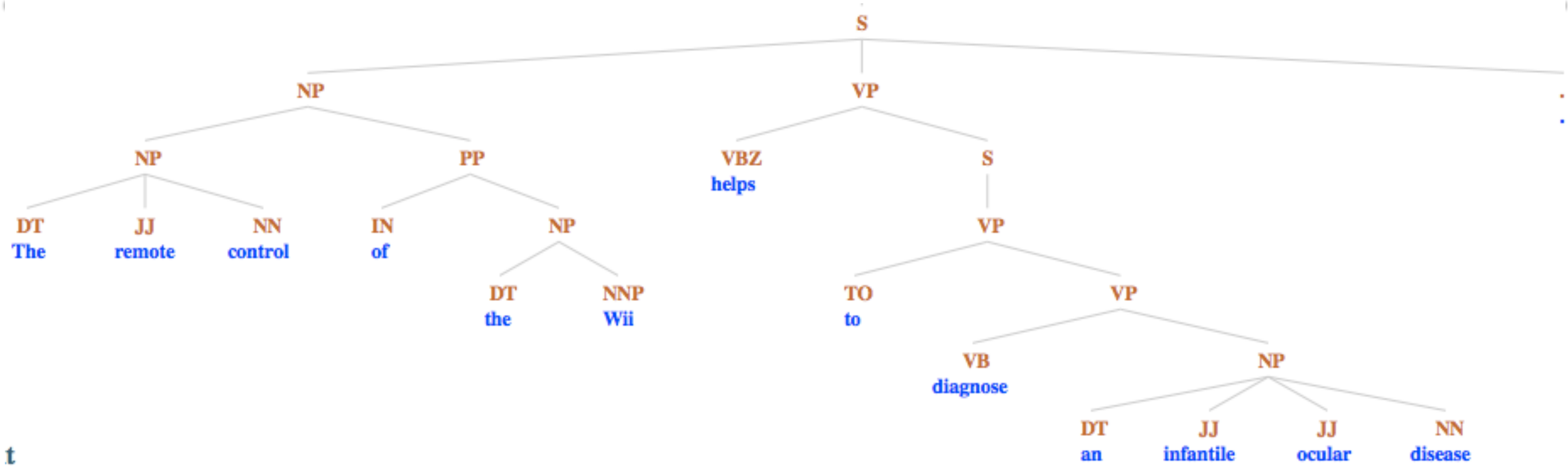
- Rather than comparing sentences at lexical level: Compare the linguistic structures and the words within them [Gim10]

- Compare different linguistic-level elements
  - Words, lemmas, POS, Chunks
  - Parsing Trees
  - Named entities and semantic roles
  - Discourse representation (logical forms)

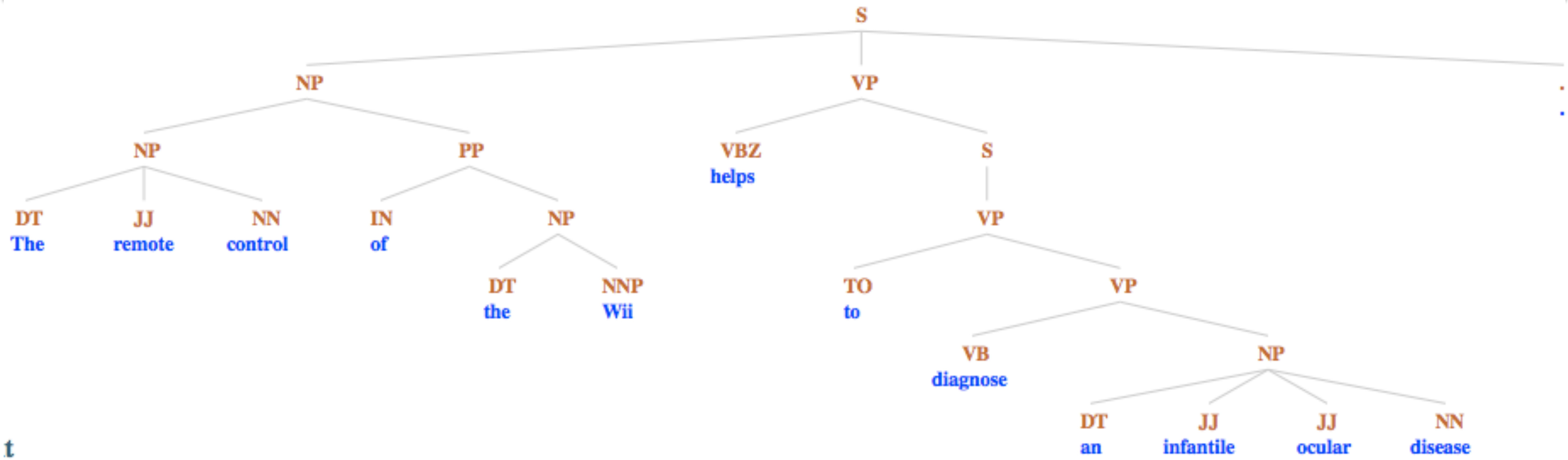
Example 1: Structural Similarity (2)

- The remote control of the Wii helps to diagnose an infantile ocular disease.
- The Wii remote **helps diagnose** a childhood eye disease.

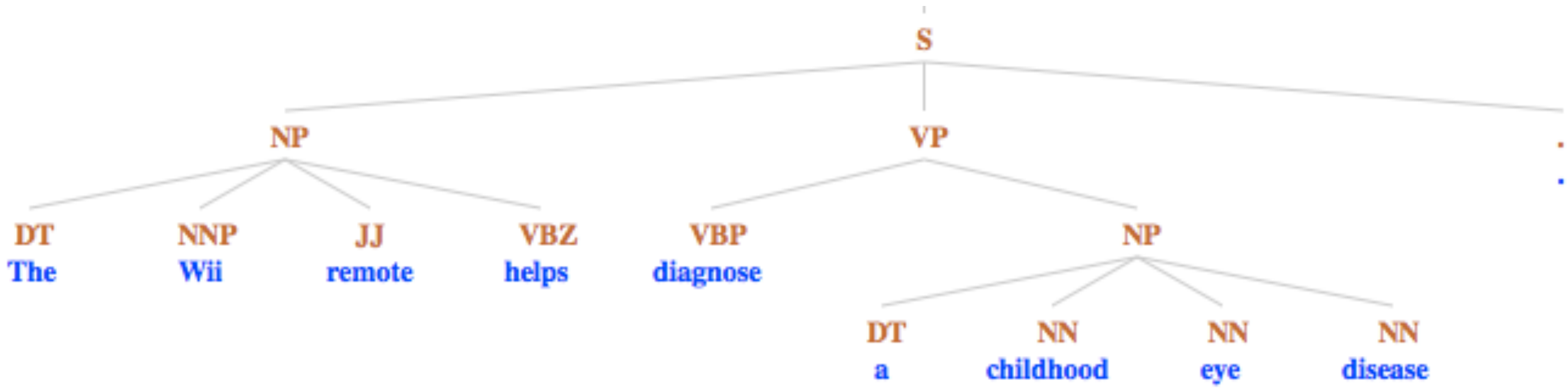
Example 1: Structural Similarity (3)



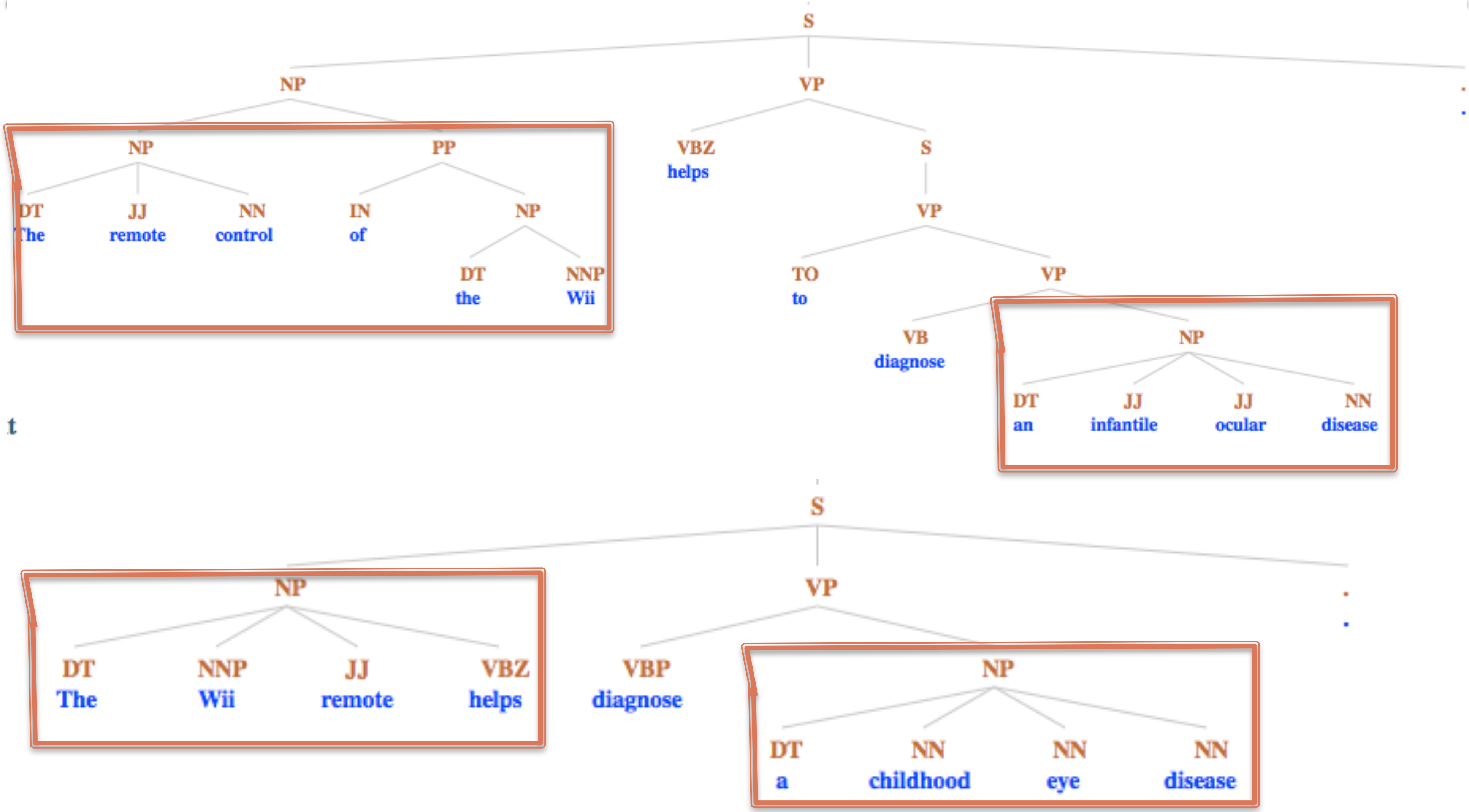
Example 1: Structural Similarity (4)



t



Example 1: Structural Similarity (4)



Measuring structural similarity (1)

- Linguistic Element (LE): abstract reference to any possible type of linguistic unit, structure, or relationship among them.
  - For instance: POS tags, word lemmas, NPs, semantic roles, dependency relations, etc.

- A sentence can be seen as a bag (or a sequence) of LEs of a certain type

Measuring structural similarity (2)

- **OVERLAP** [Gim07]: generic similarity measure among linguistic elements inspired by the Jaccard coefficient [Jac1901]
- SEMPOS [Mac08] is a MT evaluation measure that considers several **overlapping variations**
- **MATCHING** is a more strict variant [Gim10]
  - All items inside an element are considered the same unit.
  - Computes the proportion of fully translated LEs according to their types.

Overlap (1)

$$O(t) = \frac{\sum_{i \in (items_{t}(cand) \bigcap items_t(ref))} count_{cand}(i,t)}{\sum_{i \in (items_t(cand) \bigcup items_t(ref))} \max(count_{cand}(i,t), count_{ref}(i,t))}$$

$$O(*) = \frac{\sum_{t \in T} \sum_{i \in (items_{t}(cand) \bigcap items_t(ref))} count_{cand}(i,t)}{\sum_{t \in T} \sum_{i \in (items_t(cand) \bigcup items_t(ref))} \max(count_{cand}(i,t), count_{ref}(i,t))}$$

Overlap (2)

- The remote control of the Wii helps to diagnose an infantile ocular disease.

- The Wii remote helps diagnose a childhood eye disease.

- Overlap:
  - Intersection: 13
  - Union: 25
  - Ol = 13/25 = 0.52

Words	Reference	Candidate
the	2	1
remote	1	1
control	1	
of	1	
wii	1	1
helps	1	1
to	1	
diagnose	1	1
an	1	
a		1
infantile	1	
childhood		1
ocular	1	
eye		1
disease	1	1
.	1	1

Overlap (3)

- The remote control of the Wii helps to diagnose an infantile ocular disease.
- DT JJ NN IN DT NNP VBZ TO VB DT JJ JJ NN .

- The Wii remote helps diagnose a childhood eye disease.
- DT NNP JJ VBZ VB DT NN NN NN .

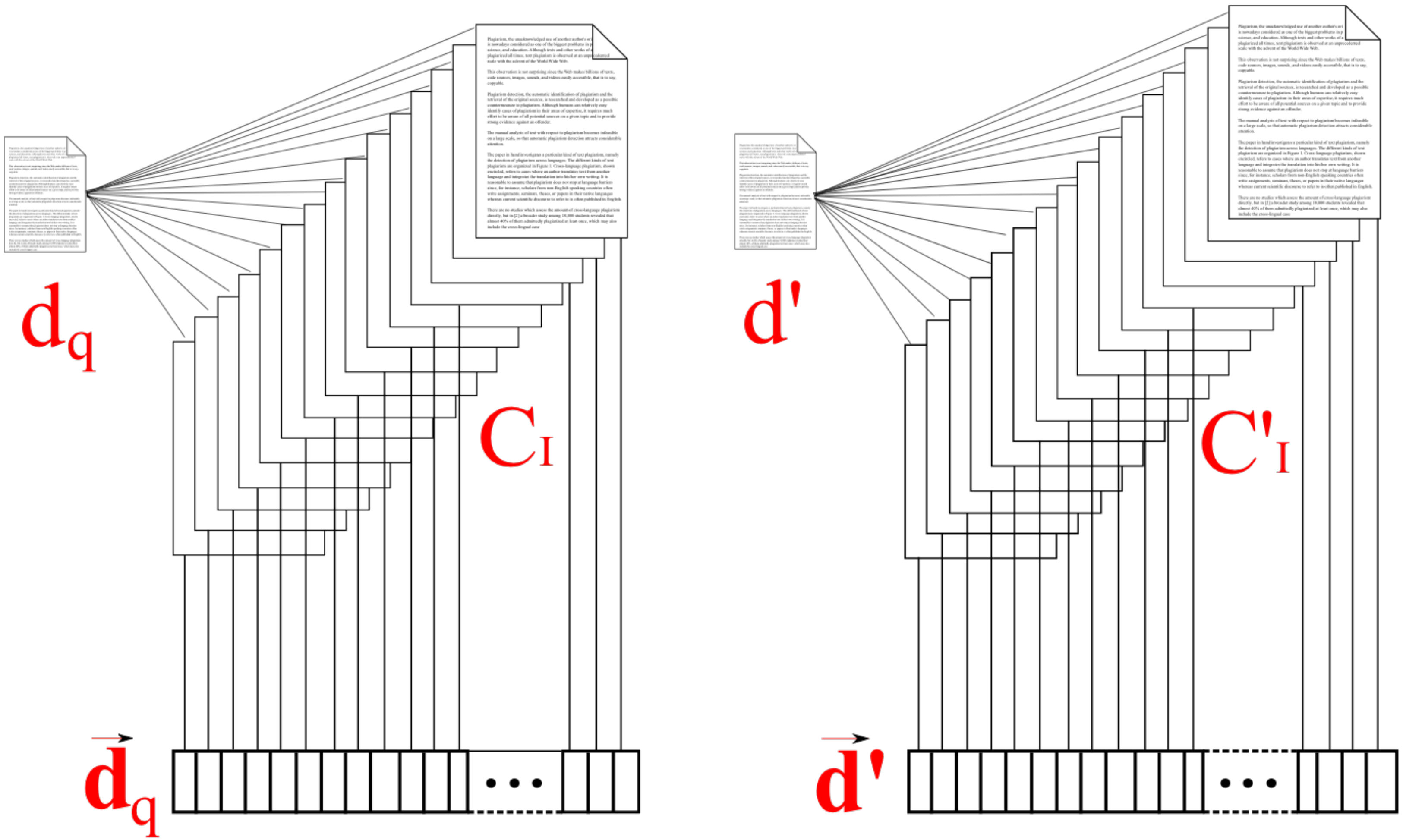
- Overlap:
  - Intersection: 9
  - Union: 15
  - Ol = 9/15 = 0.6

Words	Reference	Candidate
DT	3	2
JJ	3	1
NN	2	3
IN	1	
NNP	1	1
VBZ	1	1
TO	1	
VB	1	1
.	1	1

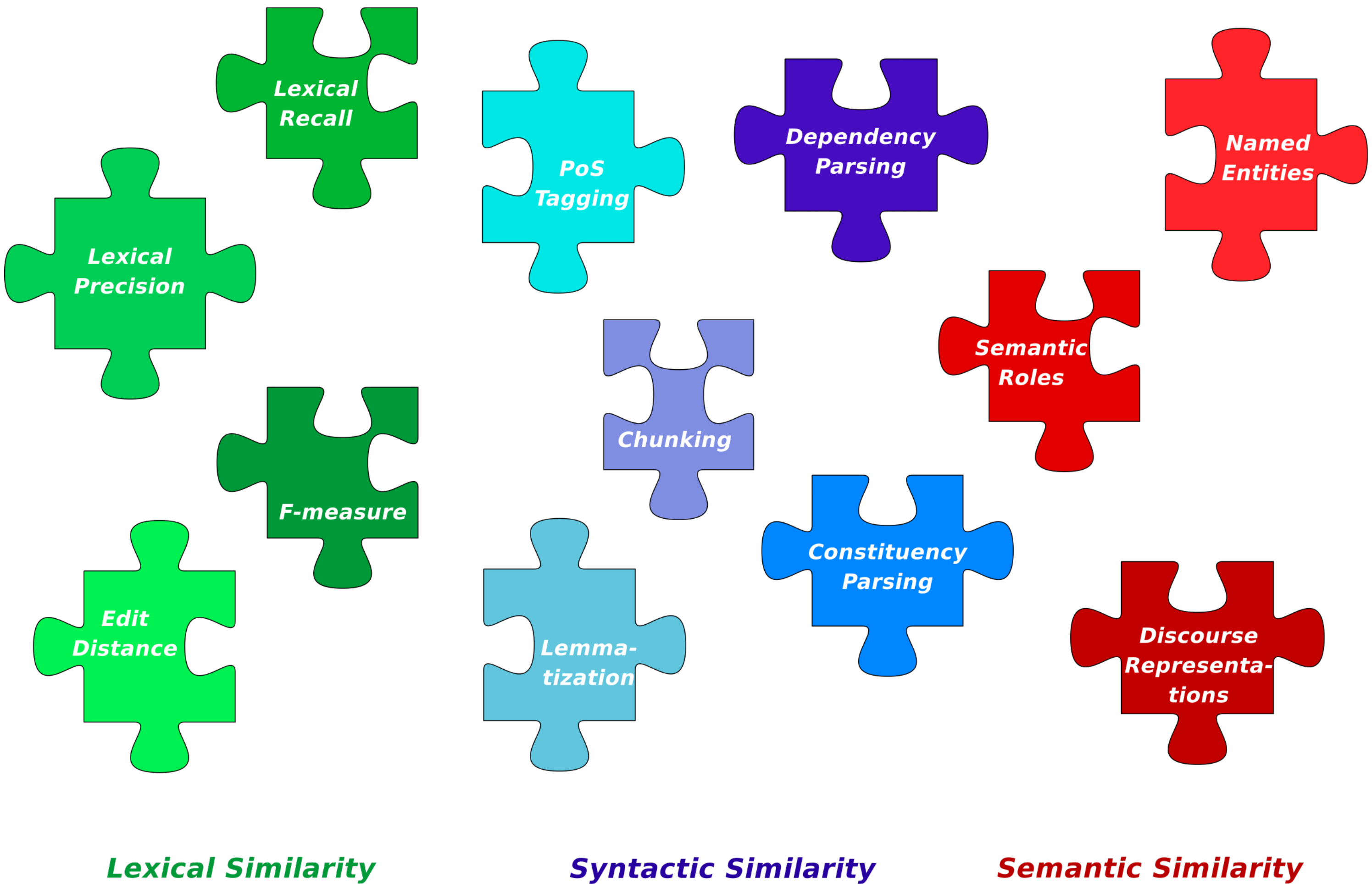
More linguistically-motivated measure (2)

- CL-Explicit Semantic Analysis
  - CL-ESA requires a significant comparable corpus C_I
  - $d_q \in L \; (d' \in L')$ is represented as a vector of relations to the index collection C_I (C_I')
  - Monolingual similarities are computed over the VSM (e.g., the cosine of the vocabulary) [Pot08]

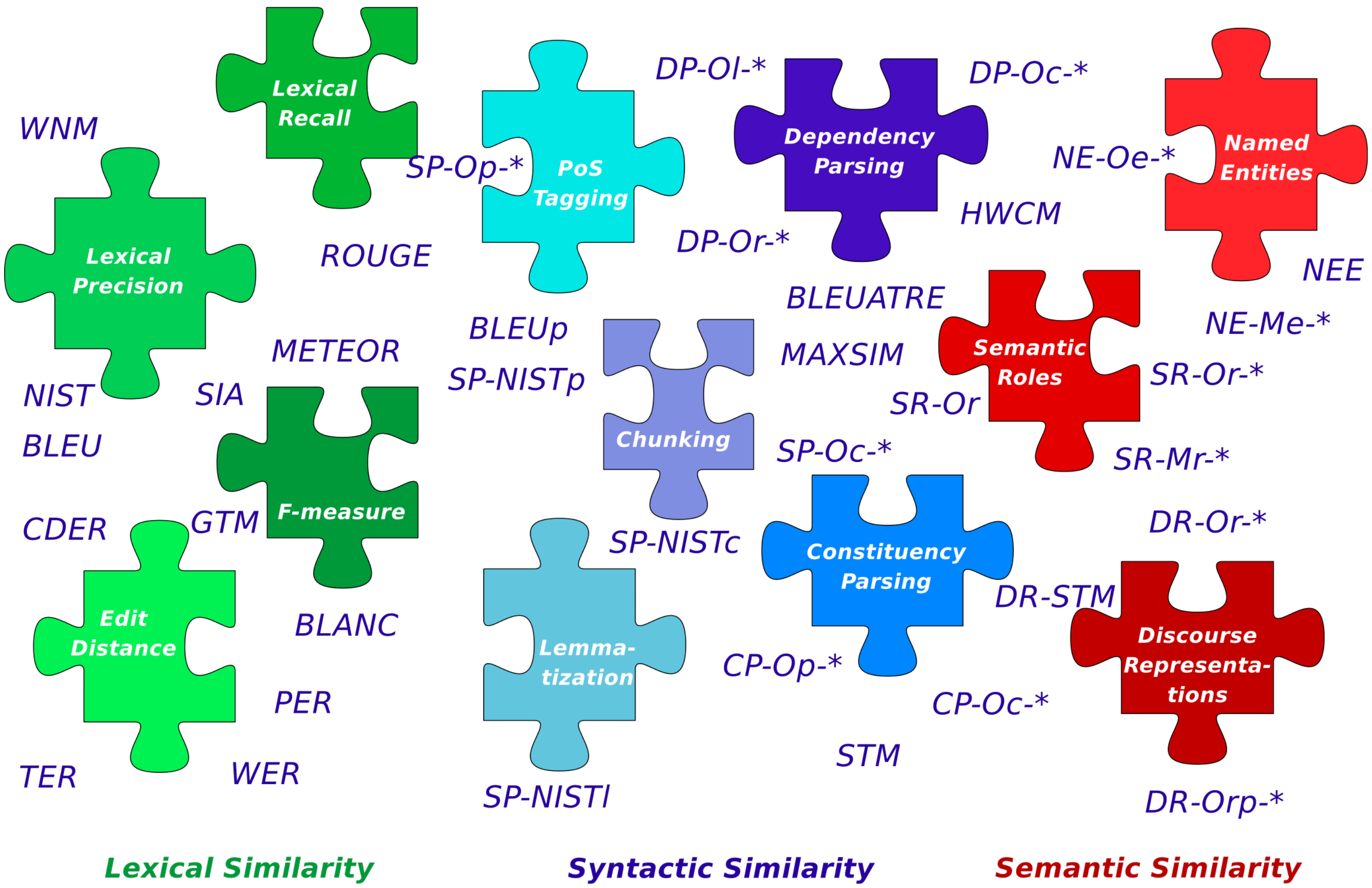
Example 2: Semantic Analysis (2)



Towards Heterogeneous MT Evaluation



Towards Heterogeneous MT Evaluation



Metric Combination

- Different measures capture different aspects of similarity

- Simple Approach
  - ULC: Uniformly-averaged linear combination of measures

- But, which ones?
  - 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$_{rp}$(*) }*

Estimate 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 on the basis of human feedback (i.e. from human assessments).

- Examples:
  - AMBER [Che12] – downhill simplex
  - SIMBLEU (ROSE) [Son11] – SVM
  - SPEDE [Wan12] – pFSM for regression
  - TERRORCAT [Fis12] - SVM on error categories

Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- **Quality estimation**
- The *Asiya* toolkit

Quality Estimation (1)

- Setting:
  - Quality assessment without reference translations
- Information available:
  - Source sentence, candidate translation(s) and, possibly, MT system information
- Motivation:
  - System ranking (system selection)
  - Hypotheses re-ranking (parameter optimization)
  - Feedback filtering (especially end-users)
  - Post-edition effort (industry pricing)

Quality Estimation (2)

- Relevant Work:
  - Johns Hopkins University Summer Workshop, 2003. “Confidence Estimation for Machine Translation”. [Bla04]
- Recent work:
  - (Specia et al., 2009;2010), (Soricut and Echihabi, 2010), (Giménez and Specia 2010), (Pighin et al., 2011), (Avramidis, 2012)
- WMT shared task on Quality Estimation
  - [Cal12] WMT12 – 11 participants
  - [Boj13] WMT13 – 14 participants
  - (3d edition at WMT 2014)

Quality Estimation Features (1)

- System-dependent
  - internal system probabilities/scores (automatic score)
  - features over **n-best translation hypotheses**
    - language modelling
    - candidates rank
    - score ratio
    - average candidates length
    - length ratio
    - …

Quality Estimation Features (2)

- System-independent
  - source (translation difficulty)
    - Source sentence length
    - Ambiguity → dictionary/alignment/WordNet-based
      - e.g, number of candidate translations per word or phrase
  - target (fluency)
    - OOV
    - Language models: perplexity, log probability

Quality Estimation Features (3)

- System-independent
  - source-target (adequacy)
    - length factor
    - punctuation and symbols concurrency
    - candidate matching → dictionary-/alignment-based
    - character *n*-grams [McN04]
    - pseudo-cognates [Sim92]
    - word alignments [Gon14]

QE Challenges

- QE is a difficult task
  - Few corpus available
  - Too domain-oriented

DE-EN, task 1.2, QE2013	Kendall's τ ties ignored
DFKI-logregFss33	0.31
DFKI-logregFss24	0.28
UPC-1	0.27
UPC-2	0.24
DCU-CCG	0.18
CNGL-SVRPLSF1	0.17
CNGL-SVRF1	0.17
Baseline	0.08
Oracle BLEU	0.22
Oracla METEOR-ex	0.20

Overview

- Introduction
- Automatic MT evaluation
- Linguistically motivated evaluation measures
- Quality estimation
- **The Asiya toolkit**

Asiya

- Asiya is an Open Toolkit for Automatic Machine Translation and (Meta-)Evaluation
  http://asiya.lsi.upc.edu

- Asiya provides:
  - Automatic evaluation measures using several linguistic layers for a variety of languages
  - Quality Estimation measures
  - Meta-evaluation metrics
  - Learning schemes

Asiya

- Languages:
  - English, Spanish, Catalan
  - Czech, French, German and Russian with limited resources
- Similarity principles
  - Precision, Recall, Overlap, Matching, …
- Linguistic layers:
  - Lexical, Syntactic, Semantic, Discourse

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-srcippl, CE-srcipplC, CE-srcipplP, 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, Fl, 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, Rl, 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 }

Asiya how to (1)

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

Asiya how to (2)

- Asiya.pl 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
#------------------------------------------------------------
```

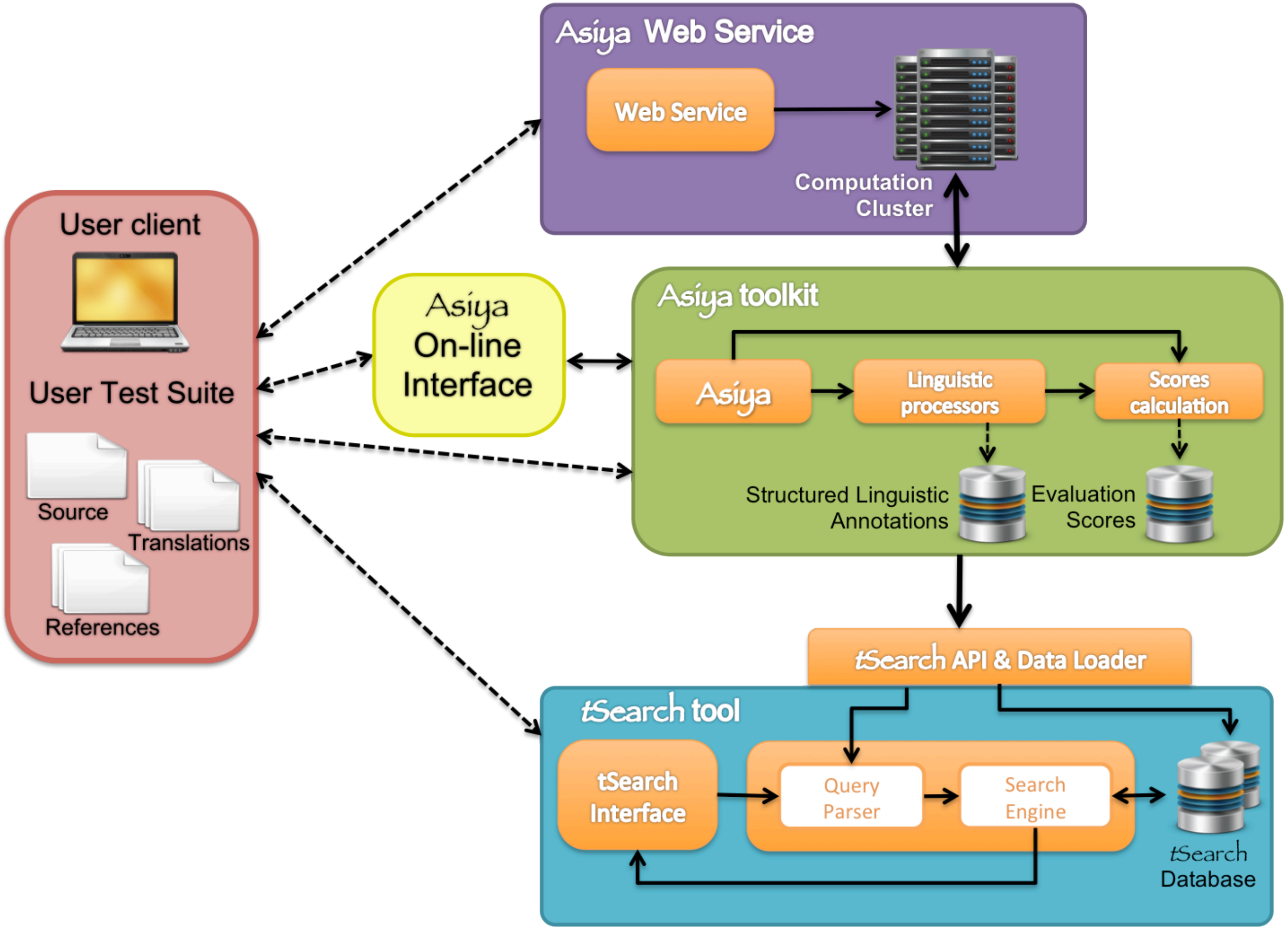
Asiya how to (3)

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

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
#-----------------------------------------------------------
```

Asiya Interfaces



Hands-on

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(*), $DP\text{-}HWC_r$, $DP\text{-}O_r$(*), $CP\text{-}STM_4$*
- 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

References

- [LDC05] NIST Multimodal Information Group. NIST 2005 Open Machine Translation (OpenMT)
- [TAUS13] TAUS. Quality Evaluation using Adequacy and/or Fluency Approaches. https://evaluation.taus.net/resources/adequacy-fluency-guidelines
- **[Tau12]** *Nora Aranberri and Rahzeb Choudhury*. **Advancing Best Practices in Machine Translation Quality Evaluation**. TAUS 2012.
- [Sty11] Sara Stymne. BLAST: A Tool for Error Analysis of Machine Translation Output. 2011.
- [Fed12] Christian Federmann. Appraise: An Open-Source Toolkit for Manual Phrase-Based Evaluation of Translations. LREC 2012.
- [Cha13] Konstantinos Chatzitheodorou, Stamatis Chatzistamatis. COSTA MT Evaluation Tool: An Open Toolkit for Human Machine Translation Evaluation. The Prague Bulletin of Mathematical Linguistics No. 100, 2013, pp. 83–89.

References

- [Coh60] Cohen, Jacob (1960). "A coefficient of agreement for nominal scales". Educational and Psychological Measurement 20 (1): 37–46.

- [Boj13] Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. *Findings of the 2013 Workshop on Statistical Machine Translation.* WMT13, 2013.

- [Lan77] Landis, J.R.; Koch, G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics* **33** (1): 159–174.

- [Cal12] Chris Callison-Burch, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut and Lucia Specia. *Findings of the 2012 Workshop on Statistical Machine Translation. WMT13*

References

- [Nie00] Nie en, S., Och, F. J., Leusch, G., & Ney, H. An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research. Proceedings of the 2nd International Conference on Language Resources and Evaluation. LREC, 2000.
- [Til97] Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. Accelerated DP based Search for Statistical Translation. Proceedings of European Conference on Speech Communication and Technology. 1997.
- [Sno06] Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. A Study of Translation Edit Rate with Targeted Human Annotation. Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (AMTA) pp. 223-231, 2006.
- [Pap01] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. Bleu: a method for automatic evaluation of machine translation, RC22176 (Technical Report). IBM T.J. Watson Research Center. 2001.

References

- [Dodo2] Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. Proceedings of the 2nd International Conference on Human Language Technology(pp. 138-145). 2002.
- [Lin04a] Lin, C.-Y., & Och, F. J. Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statics. Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL). 2004.
- [Mel03] Melamed, I. D., Green, R., & Turian, J. P. Precision and Recall of Machine Translation. Proceedings of the Joint Conference on Human Language Technology and the North American Chapter of the Association for Computational Linguistics (HLT-NAACL). 2003.
- [Ban05] Satanjeev Banerjee and Alon Lavie, "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments", *Proceedings of the ACL 2005 Workshop on Intrinsic and Extrinsic Evaluation Measures for MT and/or Summarization, 2005.*

References

- *[Den10]* Michael Denkowski and Alon Lavie, "METEOR-NEXT and the METEOR Paraphrase Tables: Improved Evaluation Support For Five Target Languages", *Proceedings of the ACL 2010 Joint Workshop on Statistical Machine Translation and Metrics MATR, 2010.*
- [Dre12] Dreyer, Markus and Marcu, Daniel. HyTER: Meaning-equivalent Semantics for Translation Evaluation. In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. NAACL HLT, 2012
- [Gim09] Jesús Giménez and Lluís Màrquez. On the Robustness of Syntactic and Semantic Features for Automatic MT Evaluation. In *Proceedings of the Fourth Workshop on Statistical Machine Translation, WMT 2009*, Athens, Greece, 2009.
- [Cha08] Chan YS, Ng HT. MAXSIM: a maximum similarity metric for machine translation evaluation. In Proceedings of ACL-08/HLT, pp 55-62. 2008.

References

- [Pad09] S. Pado and M. Galley and D. Jurafsky and C. Manning. Robust Machine Translation Evaluation with Entailment Features. Proceedings of ACL, 2009.
- [Com14] Elisabet Comelles, Jordi Atserias, Victoria Arranz, Irene Castellon and Jordi Sesé. VERTa: Facing a Multilingual Experience of a Linguistic MT Evaluation. LREC, 2014.
- [Gim10] Jesús Giménez, Lluís Màrquez. Linguistic Measures for Automatic Machine Translation Evaluation. *Machine Translation*, Springer Netherlands, 2010.
- [Gim07] Jesús Giménez and Lluís Màrquez. Linguistic Features for Automatic Evaluation of Heterogeneous MT Systems. In *Proceedings of WMT 2007 (ACL'07)*, June 2007.
- [Jac01] Jaccard, Paul. Étude comparative de la distribution florale dans une portion des Alpes et des Jura", *Bulletin de la Société Vaudoise des Sciences Naturelles* 37: 547–579. 1901.

References

- [Mac08] Machácek, Matous and Bojar, Ondrej. Approximating a Deep-syntactic Metric for MT Evaluation and Tuning. Proceedings of the Sixth Workshop on Statistical Machine Translation. WMT '11, 2011.
- [Pot08] Martin Potthast, Benno Stein, Maik Anderka. A Wikipedia-Based Multilingual Retrieval Model. In Advances in Information Retrieval, Vol. 4956, pp. 522-530. 2008.
- [Che12] Boxing Chen, Roland Kuhn, and George Foster. Improving amber, an MT evaluation metric. In Proceedings of the Seventh Workshop on Statistical Machine Translation. ACL 2012.
- [Son11] Xingyi Song and Trevor Cohn. Regression and ranking based optimisation for sentence level MT evaluation. In Proceedings of the Sixth Workshop on Statistical Machine Translation . 2011.
- [Wan12] Mengqiu Wang and Christopher Manning. SPEDE: Probabilistic edit distance metrics for MT evaluation. In Proceedings of the Seventh Workshop on Statistical Machine Translation. ACL 2012.

References

- [Fis12] Mark Fishel, Rico Sennrich, Maja Popovic, and Ondrej Bojar. 2012. TerrorCat: a translation error categorization-based MT quality metric. In Proceedings of the Seventh Workshop on Statistical Machine Translation. ACL 2012.
- [Bla04] Blatz, John and Fitzgerald, Erin and Foster, George and Gandrabur, Simona and Goutte, Cyril and Kulesza, Alex and Sanchis, Alberto and Ueffing, Nicola. Confidence Estimation for Machine Translation. Proceedings of the 20th International Conference on Computational Linguistics, COLING 2004.
- Giménez and Specia, 2010. Lucia Specia and Jesús Giménez. Combining Confidence Estimation and Reference-based Metrics for Segment-level MT Evaluation. In *Ninth Conference of the Association for Machine Translation in the Americas*, AMTA 2010.
- [Specia et al., 2010] Lucia Specia, Dhwaj Raj, and Marco Turchi. Machine translation evaluation versus quality estimation. *Machine Translation*, 24(1): 39–50, Springer Netherlands, 2010.

References

- Special et al., 2009. Lucia Specia, Marco Turchi, Zhuoran Wang, John Shawe-Taylor, and Craig Saunders. Improving the Confidence of Machine Translation Quality Estimates. In *Machine Translation Summit XII*, 2009.
- Soricut and Echihabi, 2010. Radu Soricut and Abdessamad Echihabi. TrustRank: Inducing Trust in Automatic Translations via Ranking. *Proceedings of the Association for Computational Linguistics Conference (ACL-2010). 2010.*
- Pighin et al., 2012. Daniele Pighin and Meritxell González and Lluís Màrquez. The UPC Submission to the WMT 2012 Shared Task on Quality Estimation *Proceedings of the 7th Workshop on Statistical Machine Translation* pg. 127--132. ACL 2012.
- Avramidis 2012. Quality Estimation for Machine Translation output using linguistic analysis and decoding features
  E Avramidis. Seventh Workshop on Statistical Machine Translation, 2012.
- [McN04] Paul McNamee and James Mayfield. Character N-Gram Tokenization for European Language Text Retrieval. Information Retrieval , 7(1-2):73–97. 2004.
- [Sim92] Michel Simard, George F. Foster, and Pierre Isabelle. Using Cognates to Align Sentences in Bilingual Corpora. In Proceedings of the Fourth International Conference on Theoretical and Methodological Issues in Machine Translation. 2012.
- [Gon14] Meritxell Gonzàlez and Alberto Barrón-Cedeño and Lluís Màrquez. IPA and STOUT: Leveraging Linguistic and Source-based Features for Machine Translation Evaluation. Ninth Workshop on Statistical Machine Translation (WMT2014).

Additional Slides

Evaluation of syntactic measures

- NIST 2005 Arabic-to-English Exercise

Level	Metric	ρall	ρSMT
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