$\mathbf{IQ}_{\mathrm{MT}}$: A Framework for Automatic MT Evaluation based on 'Human Likeness'

A Tutorial

Jesús Giménez (TALP-UPC)

October, 2006

Lexical Level

- Most metrics are developed on the basis of 'Human Acceptability'
- Most metrics consider only lexical similarities:
  - WER (Nießen et al., 2000)
  - BLEU (Papinenit et al., 2001)
  - NIST (Doddington, 2002)
  - PER (Leusch et al., 2003)
  - GTM (Melamed et al., 2003)
- Little effort has been devoted to:
  - introducing additional linguistic information
  - combining different metrics

Remarkable Efforts

- Use of additional knowledge at the lexical level:
  - ROUGE (stemming)
  - METEOR (stemming + WordNet lookup)
  - WNM (word frequency weighting)
- Use of syntactic knowledge:
  - HWCM, STM, DSTM (Liu and Gildea, 2005)
- Combining different metrics:
  - Kulesza and Shieber (2004)

(1) A new SuperMetric? or... (2) divide and conquer?

1. Design a new metric XXX which considers information at different linguistic levels (lexical, syntactic, semantic)
2. Design a set of *specialized metrics* which work at different levels and *combine* their scores into *a single measure of MT quality*?

The QARLA assumption

"All manual references are equally optimal and the best similarity metric is the one that identifies and uses features that are common to all references, grouping them and separating them from the automatic candidates."

- QARLA provides 3 measures:

  QUEEN determines the quality of a set of systems
  KING determines the quality of a set of metrics
  JACK determines the quality of a test set

(Amigo et al., ACL 2005)

The QUEEN assumption

"A good candidate should be similar to all models according to all metrics."

Given a set of metrics X, a set of references R, and a candidate a:

$$\text{QUEEN}_{X,R}(a) = Prob(\forall x \in X : x(a, r) \geq x(r', r''))$$

The QUEEN properties

(i) it is able to combine different similarity metrics into a single evaluation measure.

(ii) it is not affected by the scale properties of individual metrics, i.e. it does not require metric normalisation and it is not affected by metric weighting.

(iii) Candidates which are very far from the set of models all receive QUEEN=0.

(iv) The value of QUEEN is maximised for candidates that "merge" with the references according to all metrics in X.

(v) The universal quantifier on the metric parameter x implies that adding redundant metrics does not bias the result of QUEEN.

The KING measure

"A good metric should score human references higher than MT systems."

Given sets of metrics X, references R, and systems A:

$$\text{KING}_{A,R}(X) =$$

$$Prob(\forall a \in A : \text{QUEEN}_{X,R-\{r\}}(r) \geq \text{QUEEN}_{X,R-\{r\})}(a)$$

KING measures the ability of a set of metrics to discern between automatic and human translations.

The JACK measure

"References and systems should not be biased."

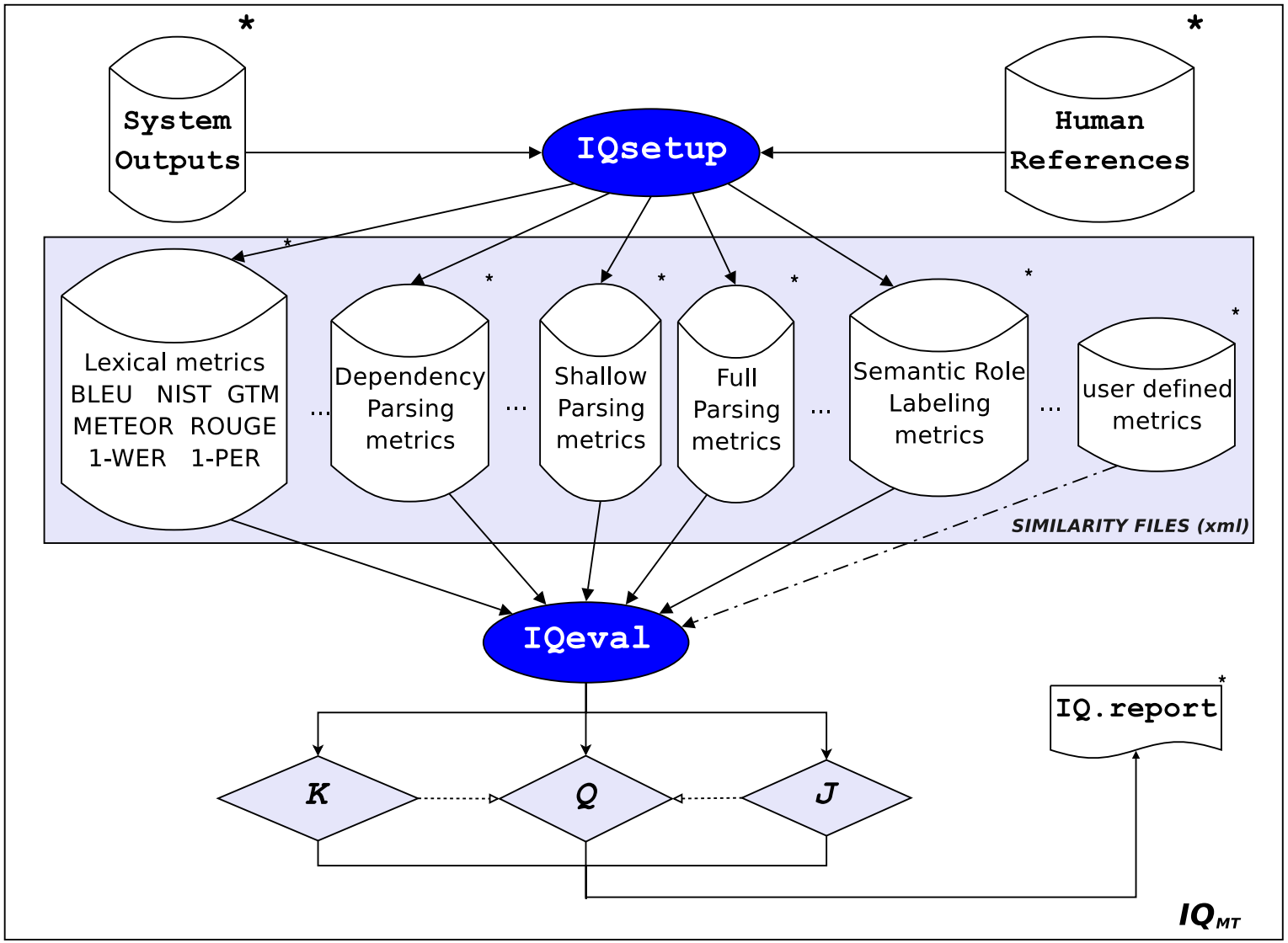
Given sets of metrics X, references R, and systems A:

$$\text{JACK}(A, R, X) =$$

$$Prob(\exists a, a' \in A : \text{QUEEN}_{X,R}(a) > 0$$
$$\wedge\ \text{QUEEN}_{X,R}(a') > 0 \wedge\ \ \forall x \in X : x(a, a') \leq x(a, r)$$

A high JACK value means that most references are closely and heterogeneously surrounded by automatic translations.

System Architecture



IQsetup <config-file>

'a priori' computing of similarities for given sets of metrics X, references R, and systems A:

```
#SRC ============================================
source=./data/dev.src.spa.iso
#REF ============================================
ref=./data/dev.src.eng.iso
ref=./data/dev.ref1.eng.iso
ref=./data/dev.ref2.eng.iso
#OUT ============================================
sys=./data/sys_1.out
...
sys=./data/sys_n.out
#----------------------------------------------
```

IQsetup <config-file>

```
NAME=NIST_mt05_AE                /* testbed name (id) */
IQMT=/usr/local/IQMT-2.0         /* IQMT path */
LANG=ENG                         /* target language */
CASE=all                         /* case [lc|all]

doBLEU                           /* metric set selection*/
doNIST
doGTM
doMETEOR
doWER
doPER
doROUGE
```

'*IQXML*' similarity files

```
<IQ metric="BLEU-4" ref="R0"
          score="0.3945" target="S0">
  <S n="1">0.3033</S>
  <S n="2">0.5833</S>
  ...
  <S n="1007">0.6852</S>
  <S n="1008">0.8333</S>
</IQ>
```

IQeval

IQ methodology in 4 steps:

1. IQsetup
2. IQeval [-doKING | -optimizeKING]
3. IQeval -doQUEEN -M optimal_set_of_metrics
4. IQeval -doJACK

Experimental Setting

- Openlab 2006 data (TC-STAR Consortium)
- Spanish-to-English
- 1,281,427 sentences for training
- 1,008 sentences for test
  - 3 human references
  - Outputs by 4 MT systems (WB, SYSTRAN, PB, PB++)
- 26 metrics from 7 different families
  (BLEU, NIST, GTM, WER, PER, ROUGE, METEOR)

Standard Metrics Outside QARLA

METRIC	WB	SYSTRAN	PB	PB++
1-PER	0.66	0.70	**0.74**	**0.74**
1-WER	0.58	0.60	**0.64**	0.63
BLEU-3	0.50	0.56	**0.66**	**0.66**
GTM-2	0.33	0.36	**0.41**	**0.41**
MTR-exact	0.57	0.65	0.69	**0.70**
NIST-3	8.79	9.59	10.66	**10.72**
RG-L	0.56	0.63	0.66	**0.67**

Evaluating with IQ_{MT}. KING

Evaluation metric	KING
1-PER	0.30
1-WER	0.34
BLEU-3	0.32
GTM-2	0.32
MTR-exact	0.29
NIST-3	**0.37**
RG-L	0.33

$\mathbf{X} = \{\text{NIST-2, NIST-3, NIST-4, and 1-WER}\}$

$\text{KING}_{A,R}(X) = \mathbf{0.38}$ (descriptive power)

Evaluating with IQ_{MT}. QUEEN (+JACK)

MT System	**QUEEN**$_{X,R}$
WB	0.31
SYSTRAN	0.39
PB	0.45
PB++	**0.46**

$\text{JACK}(A, R, X) = \mathbf{0.77}$ (test set reliability)

Error Analysis based on QUEEN

source	los ciudadanos esperan de nosotros algo más que la simple **gestión de las crisis ; esperan señales** y una política sostenible en estos ámbitos .

systems

WB	the citizens expect of us something more than the simple **management of the crisis** and a sustainable policy in these areas . **expectantly signals**
SYS-TRAN	the citizens wait for of us something more than the simple **management of the crises; they wait for signals** and a sustainable policy in these scopes.
PB	the citizens expect us any more than simply **managing crises ; they hope signals** and a sustainable policy in these areas .
PB++	the citizens expect us something more than simply **crisis management ; they expect signs** and a sustainable policy in these areas .

references

R0	the public expect more than just **crisis management ; they expect signs** , and a sustainable policy in these fields .
R1	citizens expect something more of us than just simple **crisis management ; they expect signs** and sustainable policies in these areas .
R2	the citizens expect from us something more than a simple **crisis management ; they expect signs** and a sustainable policy in these matters .

Conclusions and Further Work

- MT Evaluation based on 'Human Likeness'
- MT Evaluation over Metric Combinations
- Applications
  - MT evaluation of heterogenous MT systems (SMT, rule-based)
  - MT Development (tuning of parameters)
  - MT Error Analysis
- New metrics
  - Dependency Parsing
  - Shallow Parsing (PoS-tagging, lemma, chunking)
  - Full Parsing
  - Shallow Semantics (NERC, SRL)

Thanks

$\mathrm{IQ}_{\mathrm{MT}}$ v1.3 is freely available at:

`http://www.lsi.upc.edu/~nlp/IQMT`