Evaluating Evaluation
Lessons from the WMT 2007 Shared Task

Philipp Koehn
(with a lot of help from Chris Callison-Burch)

11 September 2007

WMT Evaluation Campaign

- Annual event since 2005
  - shared task of an ACL Workshop on Machine Translation (WMT)
  - now part of EuroMatrix project
  - about 10–15 groups participate each year
  - next year: ACL 2008 (Ohio), Marathon meeting in May 2008 (Berlin)
- Goals
  - promote MT performance for **European languages**
  - **large-scale** (30 million word training), relatively **wide domain** (politics)
  - **low barrier of entry**: baseline system provided (Moses)
  - well-defined set of **homogenous** training data (opposed to NIST eval)
  - allows to for focus on **specific** problems (e.g. morphology, unknown words)
  - also an opportunity to **improve evaluation** of MT

Participants

- Some big players missing...
  (Google, IBM, RWTH Aachen, USC/ISI)
- ... but **wide variety** of systems
  - statistical phrase-based (most)
  - statistical tree-based (CMU)
  - dependency treelet system (Microsoft)
  - rule-based (Systran)
  - rule-based + statistical post-editing (Systran+NRC/Edinburgh)
  - hybrid (system combination) (Saarbrücken)
- Not a toy task: relatively **high translation quality**

Evaluation of Evaluation

- Organizer's effort mostly focused on questions of evaluation
- **Manual** evaluation: participants volunteer 8 hours worth of work
  - what metric?
  - how many judgments do we need?
  - are judges consistent?
- **Automatic** evaluation
  - evaluation of evaluation campaign
  - what automatic metrics correlate best with human judgment?

WMT Evaluation 2007

- Tasks
  - English to/from French, German Spanish, and Czech
  - test sets drawn from **Europarl** and **news commentary**
- 88 'primary' system submissions were **manually evaluated**
- Recruited 100+ judges, who contributed 330 hours for **81,000+ judgments**
  - participants in evaluation
  - students of a course on MT
  - paid students at U Edinburgh

Evaluation of Evaluation

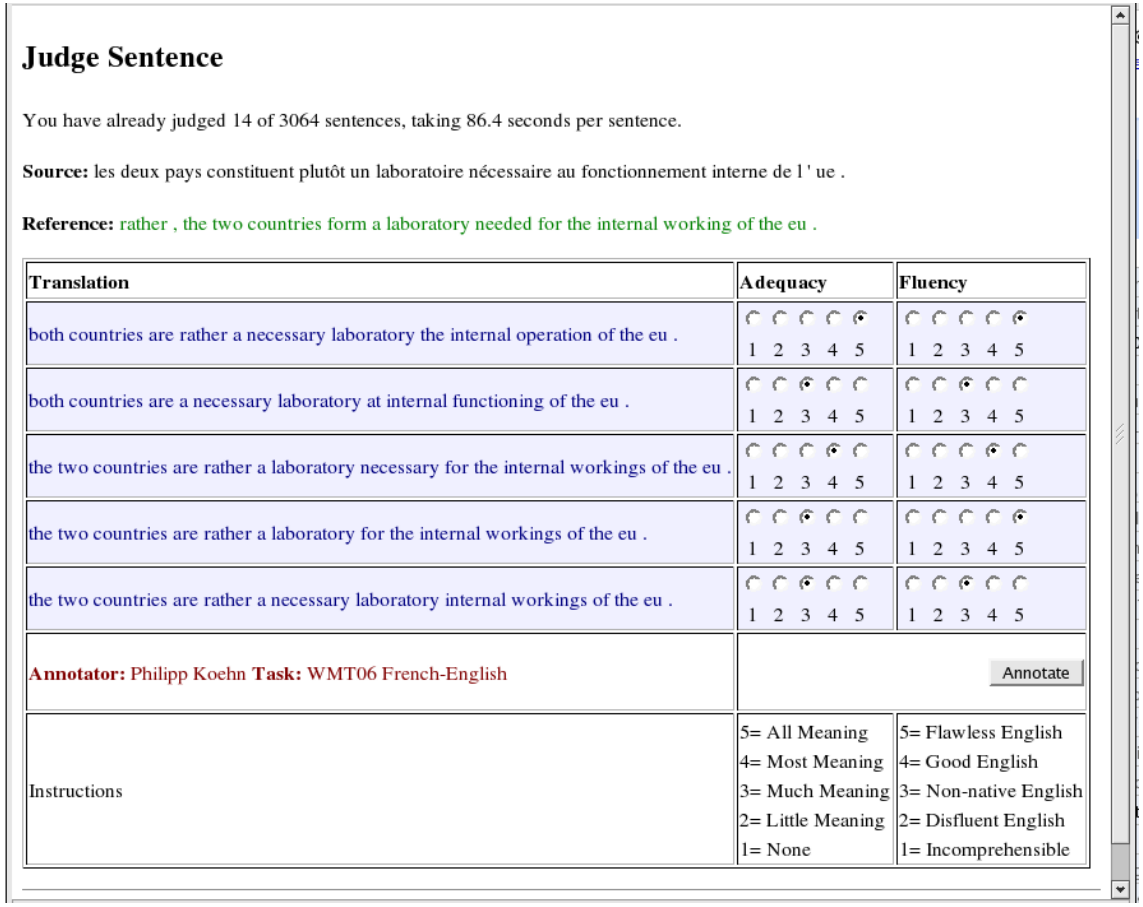
- We wanted to analyze evaluation measures and establish **best practices**
- Questions to investigate:
  - which **automatic evaluation metrics correlate** most strongly with human judgments of translation quality?
  - how **consistent** are people when they judge translation quality?
  - to what extent do they **agree** with other annotators?
  - can we **improve human evaluation**?

Fluency and Adequacy

- Traditional metric (used by NIST eval, WMT 2006, IWSLT)
- Two 5-point scales:

How much of the meaning of the reference is preserved?	How do you judge the fluency of the translation?
5 = All	5 = Flawless English
4 = Most	4 = Good English
3 = Much	3 = Non-native English
2 = Little	2 = Disfluent English
1 = None	1 = Incomprehensible

Web Tool



Judge Sentence

You have already judged 14 of 3064 sentences, taking 86.4 seconds per sentence.

Source: les deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l ' ue .

Reference: rather , the two countries form a laboratory needed for the internal working of the eu .

Translation	Adequacy	Fluency
both countries are rather a necessary laboratory the internal operation of the eu .	1 2 3 4 5	1 2 3 4 5
both countries are a necessary laboratory at internal functioning of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory necessary for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a laboratory for the internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
the two countries are rather a necessary laboratory internal workings of the eu .	1 2 3 4 5	1 2 3 4 5
Annotator: Philipp Koehn **Task:** WMT06 French-English		Annotate
Instructions	5= All Meaning 4= Most Meaning 3= Much Meaning 2= Little Meaning 1= None	5= Flawless English 4= Good English 3= Non-native English 2= Disfluent English 1= Incomprehensible

Example

In Le deux pays constituent plutôt un laboratoire nécessaire au fonctionnement interne de l'ue.

Ref Rather, the two countries form a laboratory needed for the internal working of the EU.

MT 1 Both countries are rather a necessary laboratory the internal operation of the EU.

MT 2 Both countries are a necessary laboratory at internal functioning of the EU.

MT 3 The two countries are rather a laboratory necessary for the internal workings of the EU.

MT 4 The two countries are rather a laboratory for the internal workings of the EU.

MT 5 The two countries are rather a necessary laboratory internal workings of the EU.

Judge each sentence in terms of **adequacy** and **fluency** on the scale of 1–5!

Judgments

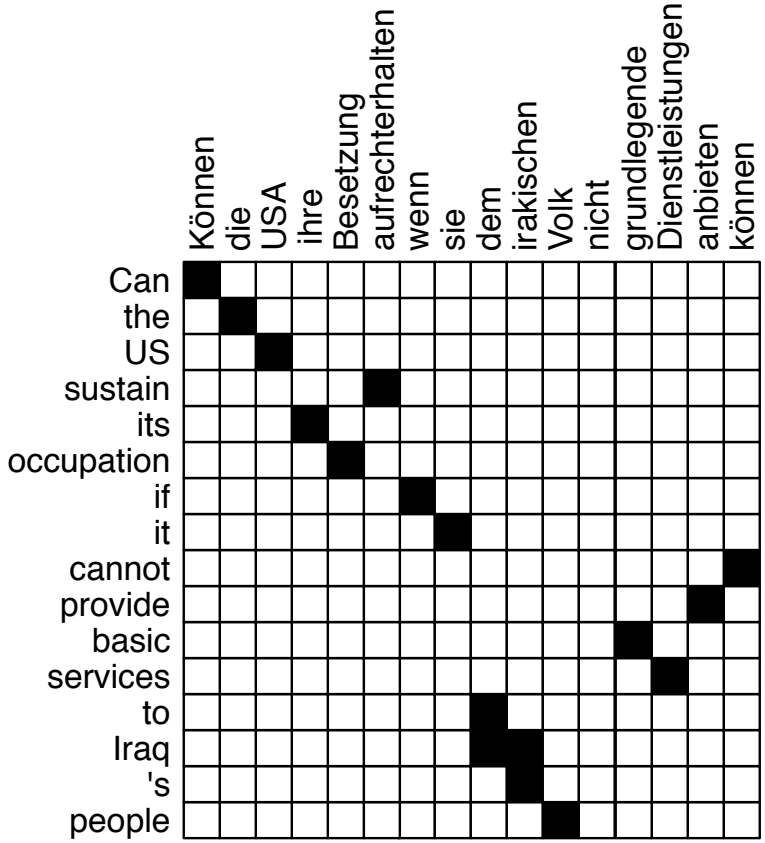
	Adequacy					Fluency				
	1	2	3	4	5	1	2	3	4	5
System 1										
System 2										
System 3										
System 4										
System 5										

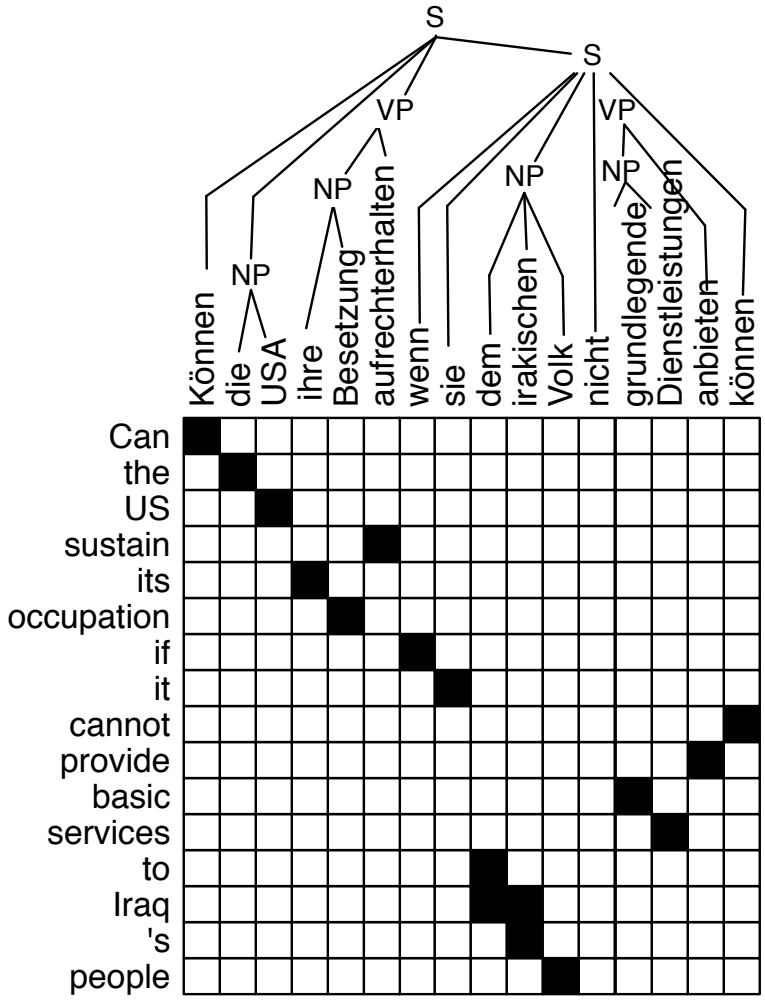
Manual Evaluation

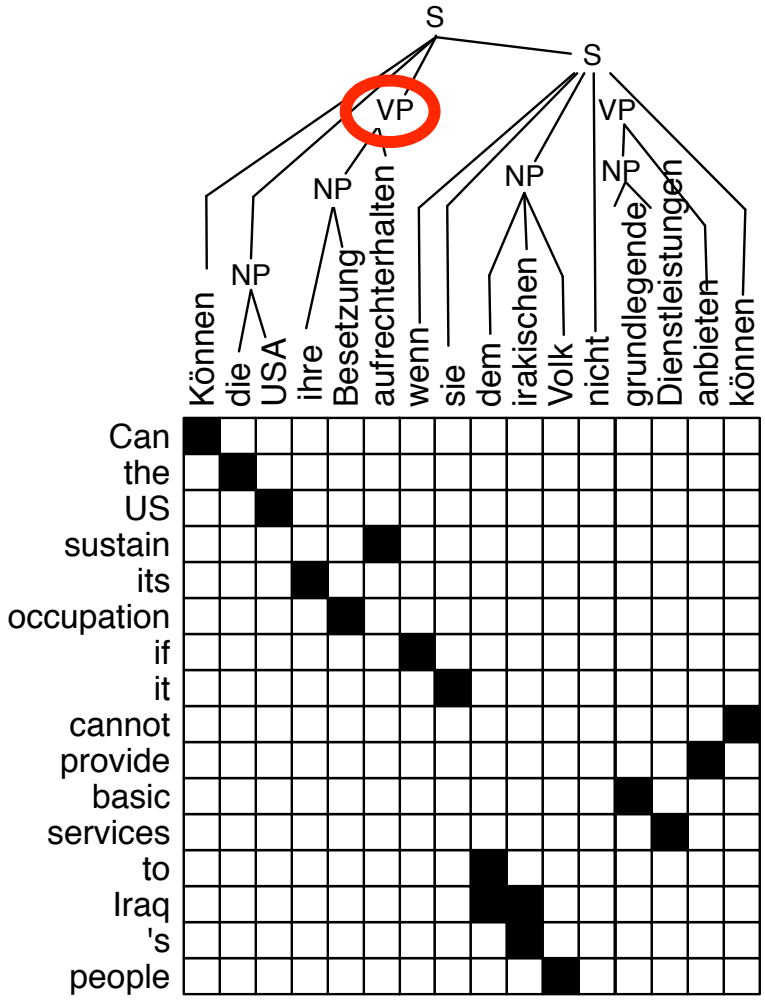
- Three different **types of evaluation**:
  - score each translation along *fluency* and *adequacy* scales
  - rank translations of sentences relative to each other
  - rank translations of sub-sentential units
- Metrics evaluated by
  - inter-annotator **agreement** (agreement with others)
  - intra-annotator agreement (self consistency)
  - average **time** to make one judgement

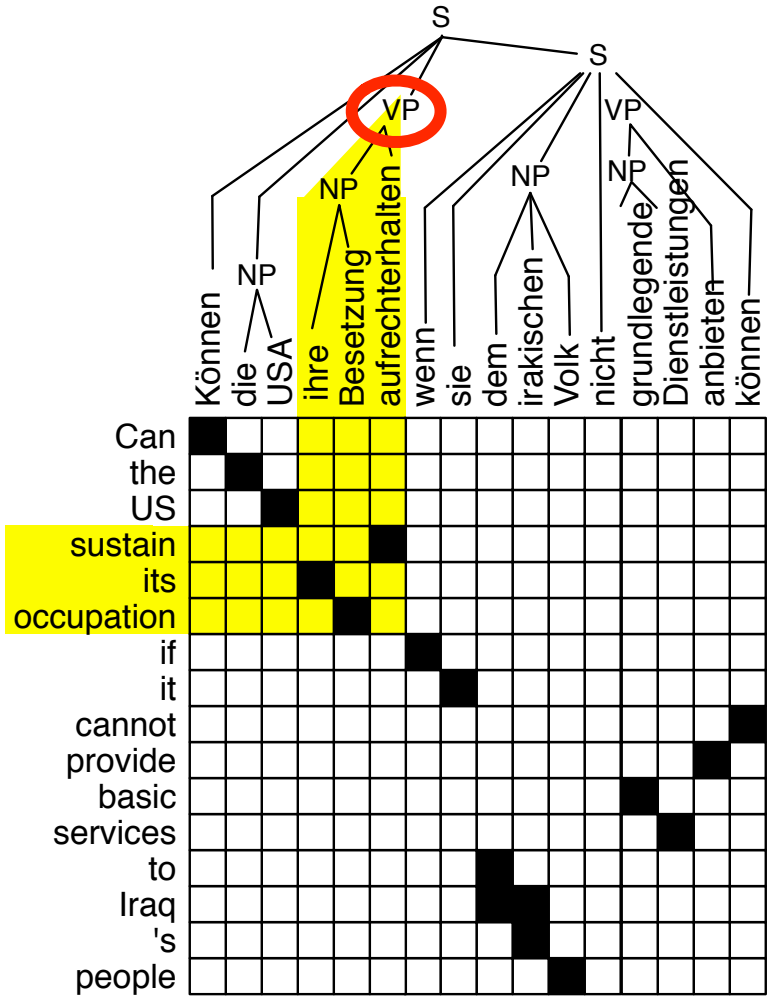
Ranking Translations of Constituents

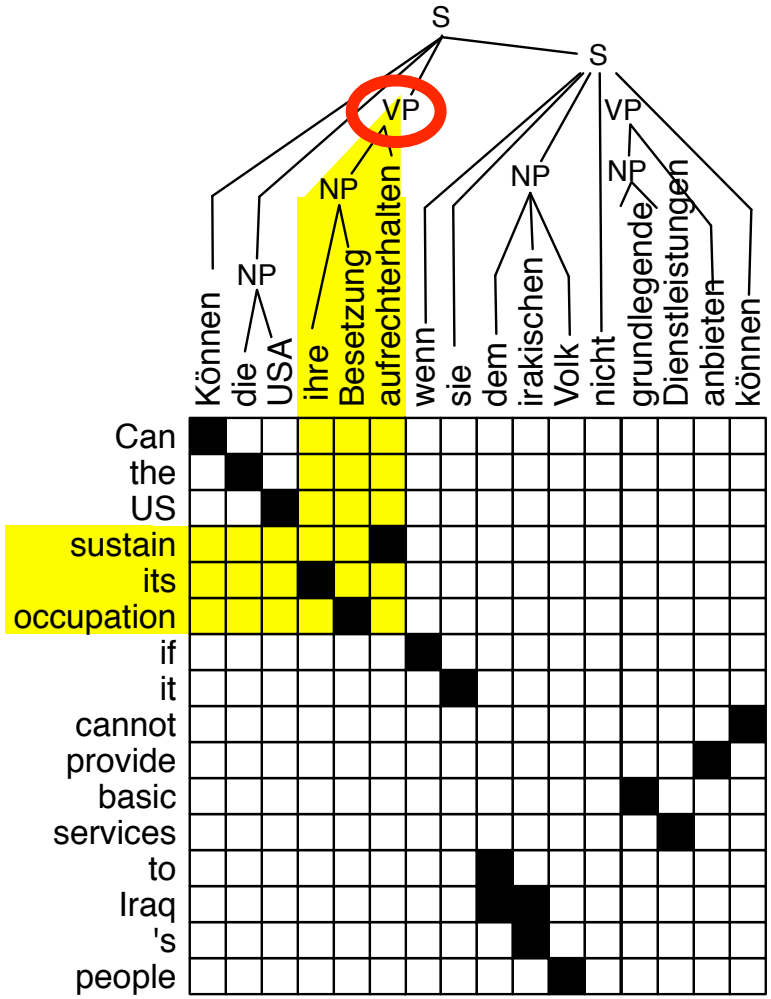
- Intuition: Ranking translations of **long sentences** is **difficult**, because systems produces errors in different parts of them
- Goal: focus attention on particular **parts of the translation** to make the task **easier**
- Method:
  1. automatically word-align source with reference and system translations
  2. **parse source** sentence
  3. select **constituents** to be judged
  4. highlight source phrase and corresponding target phrases
  5. rank those

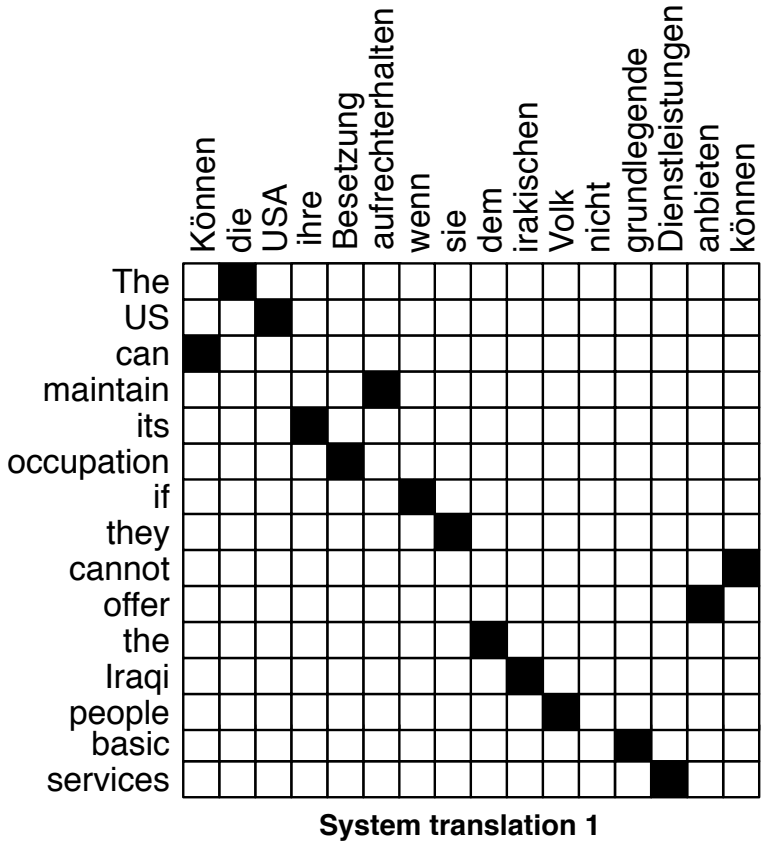

Können
die
USA
ihre
Besetzung
aufrechterhalten
wenn
sie
dem
irakischen
Volk
nicht
grundlegende
Dienstleistungen
anbieten
können
Can
the
US
sustain
its
occupation
if
it
cannot
provide
basic
services
to
Iraq
's
people


S
S
VP
NP
NP
VP
NP
NP
Können
die
USA
ihre
Besetzung
aufrechterhalten
wenn
sie
dem
irakischen
Volk
nicht
grundlegende
Dienstleistungen
anbieten
können
Can
the
US
sustain
its
occupation
if
it
cannot
provide
basic
services
to
Iraq
's
people

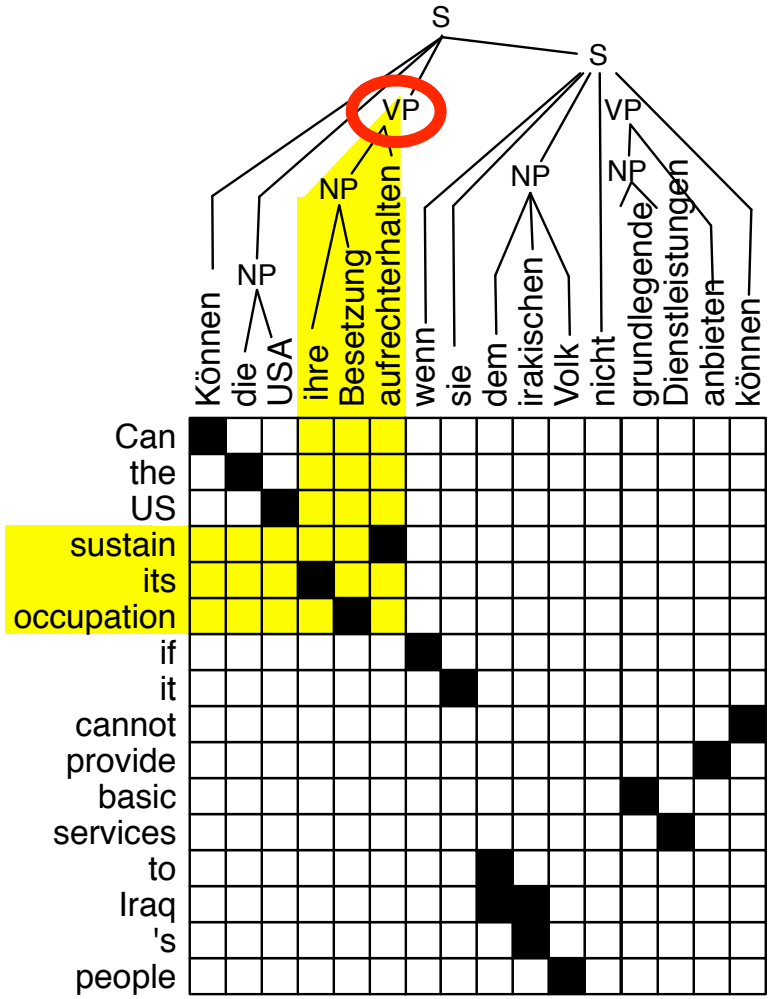

S
S
VP
NP
aufrechterhalten
NP
Besetzung
NP
USA
VP
NP
Dienstleistungen
können
Können
die
ihre
wenn
sie
dem
irakischen
Volk
nicht
grundlegende
anbieten
Can
the
US
sustain
its
occupation
if
it
cannot
provide
basic
services
to
Iraq
's
people

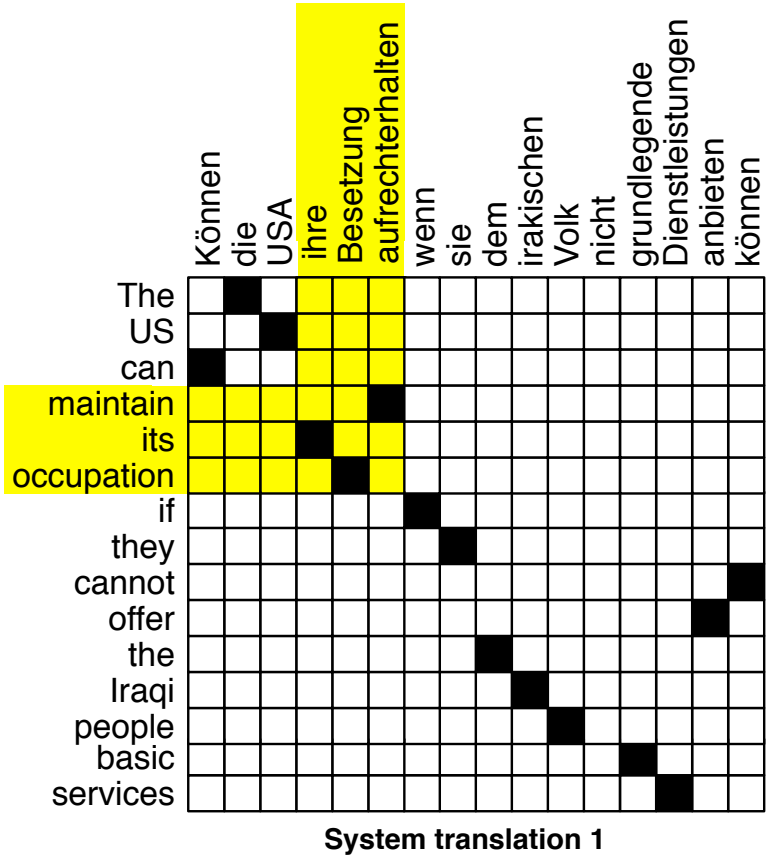

S
S
VP
NP
NP
NP
VP
NP
Können die USA ihre Besetzung aufrechterhalten wenn sie dem irakischen Volk nicht grundlegende Dienstleistungen anbieten können
Can the US sustain its occupation if it cannot provide basic services to Iraq 's people



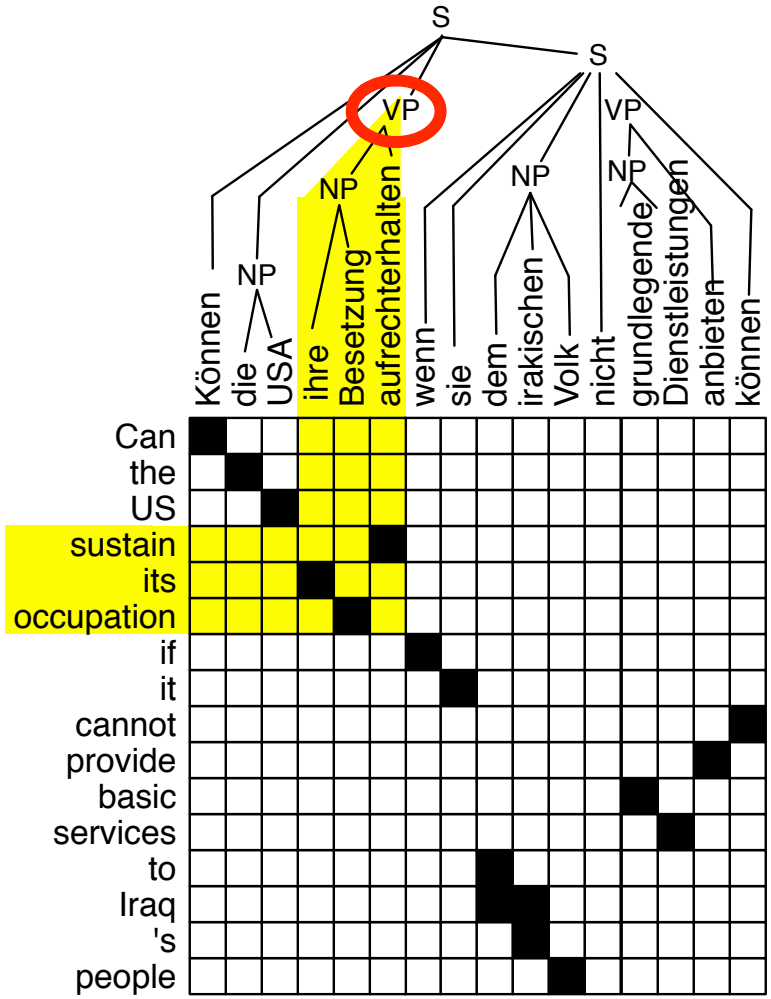


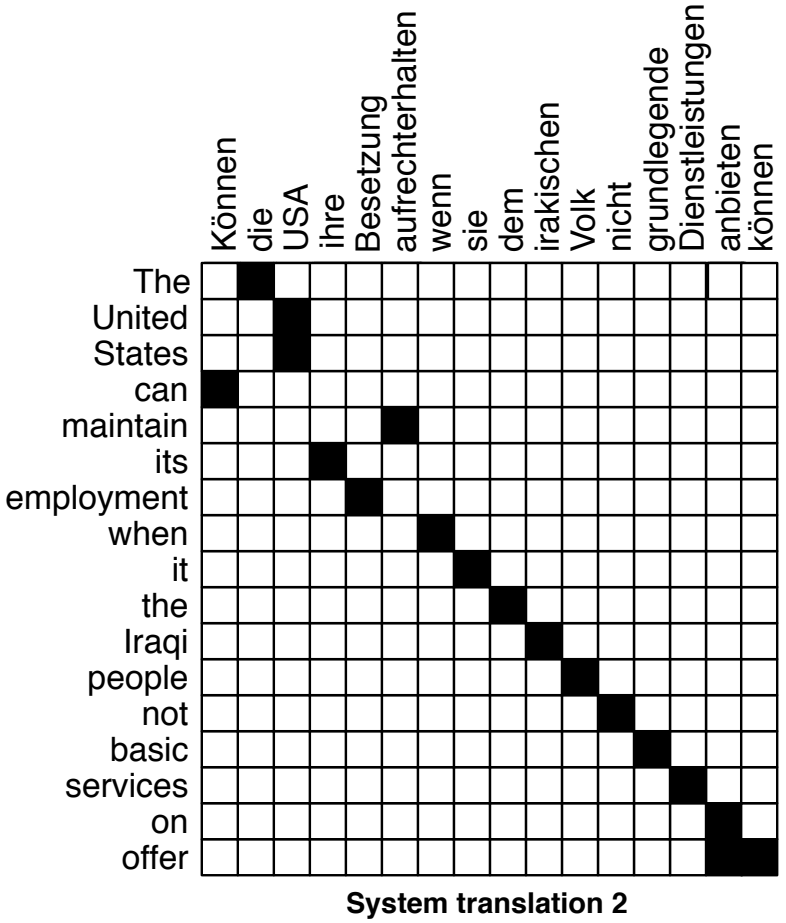
System translation 1



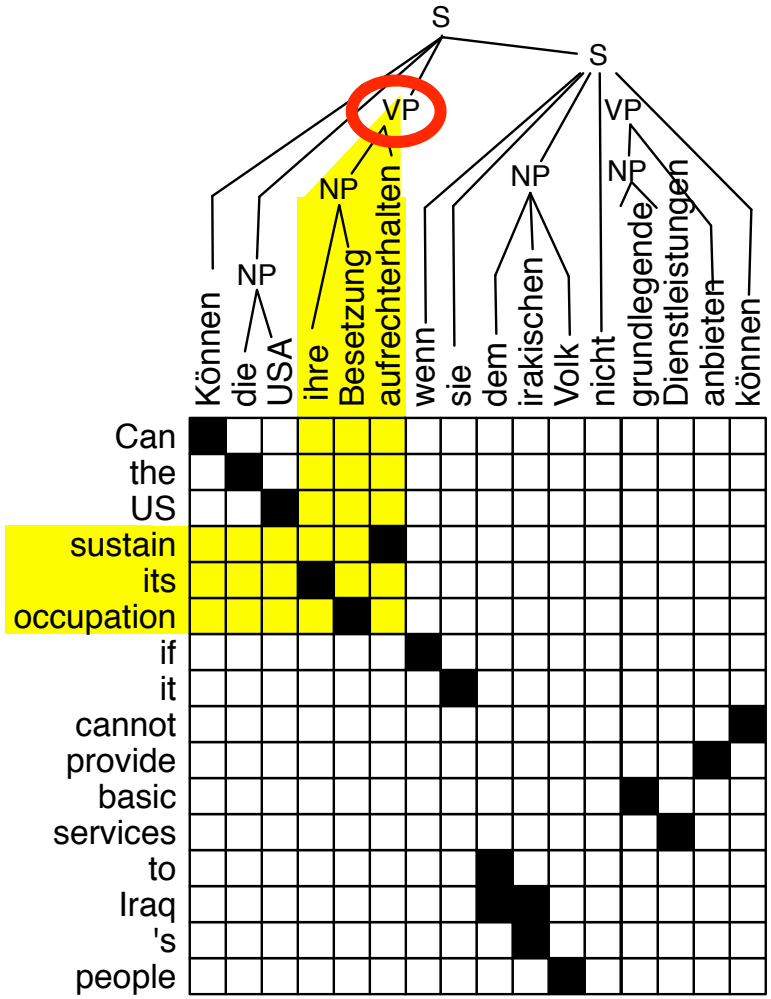


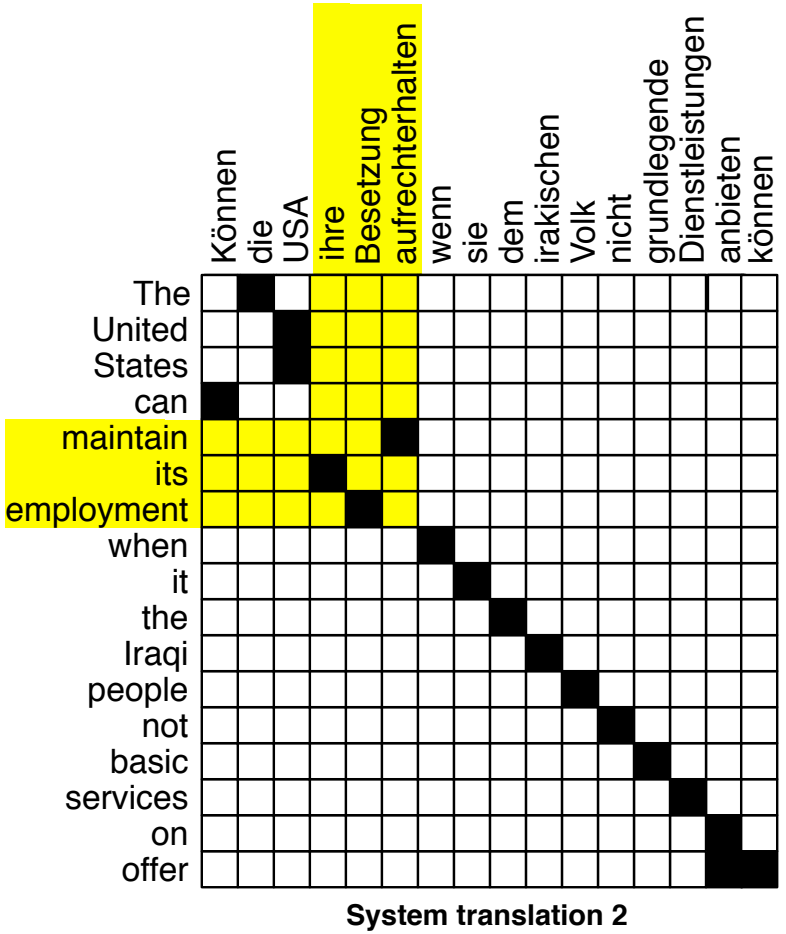
System translation 1





System translation 2





System translation 2

Results of the Meta-Evaluation

- We measured agreement among annotators using the **kappa coefficient**:

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

  where
  - $P(A)$ is the proportion of times that the annotators agree
  - $P(E)$ is the proportion of time that they would agree by chance.
- **Interpretation** of K scores varies, but:
  - .6 – .8 is good agreement
  - .4 – .6 is moderate agreement
  - $<$ .4 and we should start to worry

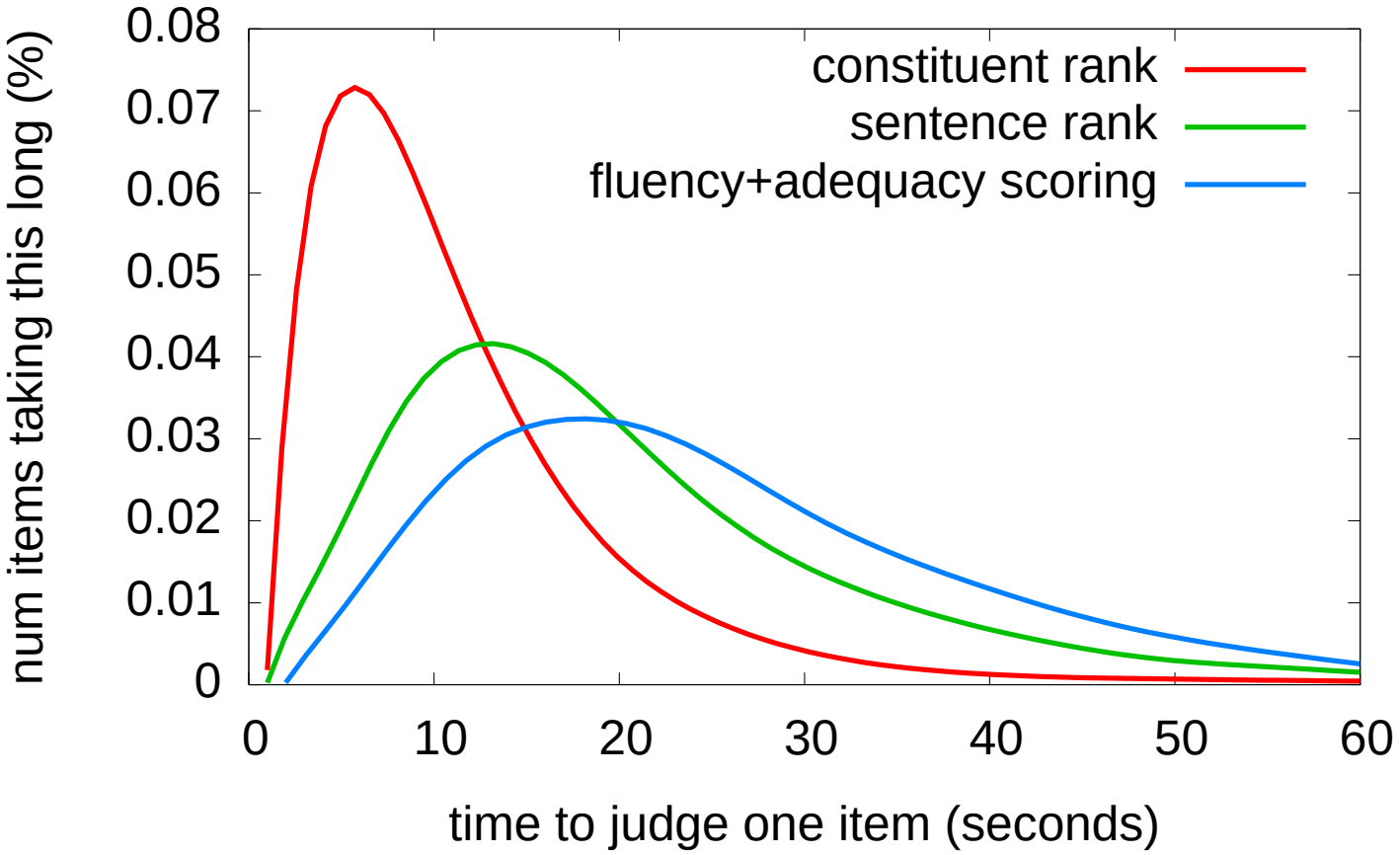
Inter-Annotator Agreement

Evaluation type	$P(A)$	$P(E)$	K
Fluency (absolute)	.400	.2	.250
Adequacy (absolute)	.380	.2	.226
Fluency (relative)	.520	.333	.281
Adequacy (relative)	.538	.333	.307
Sentence ranking	.582	.333	.373
Constituent ranking	.712	.333	.566

Intra-Annotator Agreement

Evaluation type	$P(A)$	$P(E)$	K
Fluency (absolute)	.630	.2	.537
Adequacy (absolute)	.574	.2	.468
Fluency (relative)	.690	.333	.535
Adequacy (relative)	.696	.333	.544
Sentence ranking	.749	.333	.623
Constituent ranking	.842	.333	.762

Time to judge one item



Automatic evaluation metrics

- Ranked system outputs using 11 different automatic metrics

  N-gram matching:

  Bleu, GTM, Translation Error Rate

  Flexible matching:

  Meteor, ParaEval precision, ParaEval recall

  Linguistic info:

  Dependency overlap, Semantic role overlap, WER over verbs

  Correlation-centric:

  Maximum correlation training on adequacy, and on fluency

- Meta-evaluation: Spearman's rank correlation with human judgments

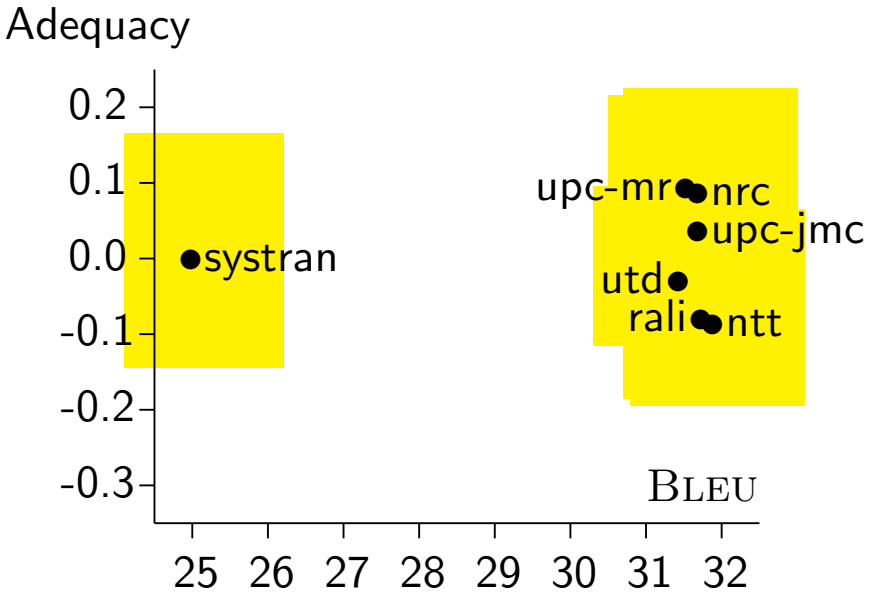
Proportion of time entries were top-ranked in manual evaluation

SYSTRAN	**32%**
University of Edinburgh	**20%**
University of Catalonia	15%
LIMSI-CNRS	13%
University of Maryland	5%
National Research Council + SYSTRAN	5%
Commercial Czech-English system	5%
University of Valencia	2%
Charles University	2%

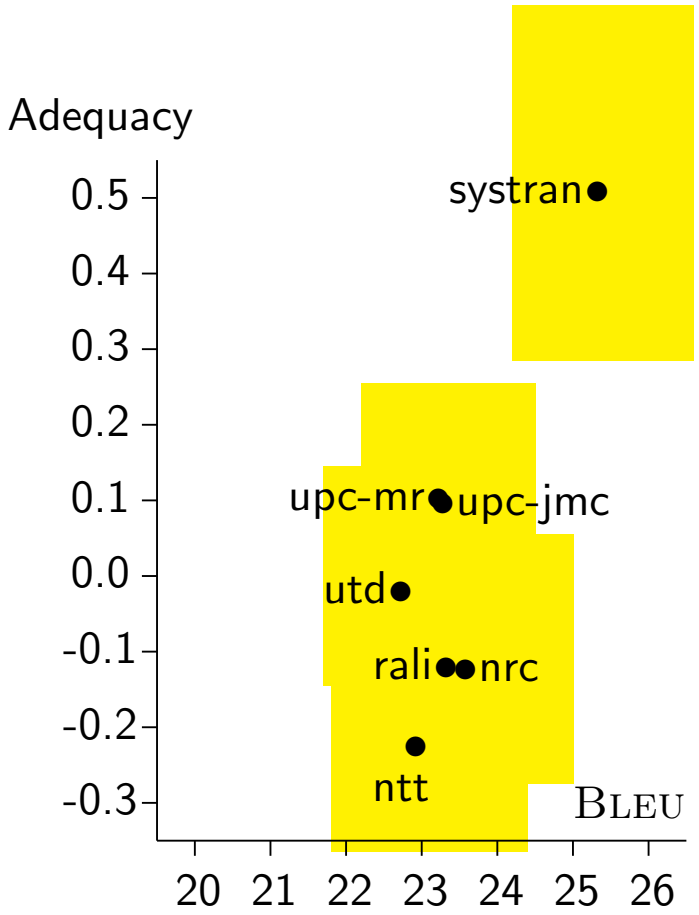
Proportion of time entries were top-ranked by automatic metrics

University of Edinburgh	**41%**
University of Catalonia	12%
LIMSI-CNRS	12%
University of Maryland	9%
Carnegie Mellon University	8%
Charles University	4%
University of California at Berkeley	3%
National Research Council + SYSTRAN	2%
SYSTRAN	**2%**
Saarland University	0.8%

Systran puzzle (WMT 2006)



- *English–French, adequacy vs.* BLEU, *in-domain*
- see also Callison-Burch et al.'s **critique of** BLEU [EACL 2006]



Mystery resolved?

- *English–French, adequacy vs.* BLEU
- **out-of-domain**
- Systran: **best** BLEU, **best** *manual*

→ lack of correlation only due to the **overly literalness** of BLEU?

Correlation

	Adequacy	Fuency	Rank	Constituent	Overall
Semantic role	**.77**	**.84**	**.80**	.74	**.79**
ParaEval-Recall	.71	.74	.77	**.80**	.76
Meteor	.71	.72	.75	.67	.71
Bleu	*.69*	*.72*	*.67*	*.60*	*.67*
Max adeq corr	.65	.66	.66	.53	.63
Max flu corr	.64	.65	.66	.51	.61
GTM	.66	.67	.62	.50	.61
Dependency overlap	.64	.64	.60	.51	.60
ParaEval-Precision	.64	.65	.61	.49	.60
1-TER	.61	.54	.52	.51	.54
1-WER of verbs	.38	.42	.43	.30	.38

Semantic Role Overlap

- Proposed by Giménez and Màrquez2007 [WMT 2007]
- Solves the Linear-B [NIST 2005] and Systran [WMT 2006] puzzle
  - NIST 2005: correlation of 0.6–0.7 vs. 0.06 for BLEU
  - WMT 2006: correlation of 0.9–0.95 vs. 0.6–0.85 for BLEU
- Checks if arguments/adjuncts to verbs overlap
- Tunable?

Lessons for Automatic Metrics

- Still an **essential tool** when building SMT systems
- Research papers should also **report manual** evaluation
- Consistent **bias** in automatic metrics when comparing different type of systems
- Improving automatic evaluation is a **well-defined task**
  - goal: better correlation with human judgments
  - impossible to *game* the metric
  - fast to compute to be usable in tuning

Lessons for Manual Metrics

- Agreement was low for fluency and adequacy scores
- We should research ways of **improving manual evaluation** so that it is
  - more consistent
  - faster / cheaper
  - easier to perform
  - re-usable
- Are we asking the right question?
  - we do not care, how **good** machine translation is
  - we do care, how **useful** machine translation is

Future Evaluations

- Euromatrix project starts an **ongoing online evaluation** later this year
- Goals:
  - provide common test sets and training data,
  - provide means for asynchronous evaluation
  - collect translations, show off best of best
- Expanded in scope to translation between **all 23 official European languages**
  - that's 253 language pairs, and 506 directions!
  - *you* could have the best Latvian-Maltese translation system in the world!
- **Continue annual evaluation**, which will focus on a subset of languages and do extensive manual evaluation
  - next year will include Hungarian
  - **ideas for manual evaluation welcome!**

Best German-English Systems

- German → English Europarl:

  SYSTRAN > liu > uedin = upc > cmu-uka > nrc > saar

- German → English News Corpus:

  SYSTRAN > uedin > upc > nrc > saar

- English → German Europarl:

  UEDIN > systran = upc > cmu-uka > nrc > saar

- English → German News Corpus:

  SYSTRAN > upc > uedin > nrc > ucb > saar

Best Spanish-English Systems

- Spanish → English Europarl:

  UPC = UEDIN > upv > cmu-syntax > cmu-uka = systran > nrc > saar

- Spanish → English News Corpus:

  UPC > uedin > systran > cmu-uka > nrc > upv > saar

- English → Spanish Europarl:

  UEDIN > upc = upv > cmu-uka > nrc = systran

- English → Spanish News Corpus:

  SYSTRAN > upc > cmu-uka > ucb > uedin > nrc = upv

Best French-English Systems

- French → English Europarl:

  LIMSI = UEDIN > systran-nrc = upc > nrc > systran > saar

- French → English News Corpus:

  LIMSI > upc = uedin > systran > systran-nrc > nrc > saar

- English → French Europarl:

  LIMSI > systran-nrc = uedin > upc > nrc = systran > saar

- English → French News Corpus:

  SYSTRAN-NRC=SYSTRAN > limsi > nrc = ucb = uedin > ucb > saar

Best Czech-English Systems

- Czech → English News Corpus:

  UMD > cu > uedin > pct

- English → Czech News Corpus:

  PCT > umd > uedin