MT Evaluation measures: Be careful what you wish for . . .

Gregor Leusch
leusch@cs.informatik.rwth-aachen.de

LREC Perspectives on Machine Translation Evaluation May 19, 2010

Human Language Technology and Pattern Recognition
Lehrstuhl für Informatik 6
Computer Science Department
RWTH Aachen University, Germany

Overview

Always keep in mind:

- **Where/When/Why/What do we evaluate?**
- **Tune the system to the measure, not vice versa!**

Where?

Where/when do we evaluate MT systems?

1. **Compare different approaches, teams (development, application)**
2. **Compare different settings for the same approach (development)**
3. **Optimize one approach (development)**
4. **Sanity checks (development, application)**
5. **Cost estimate (application)**

Why?

What is our intention when we evaluate MT systems?

- Spot problems within systems or translations → Analytical evaluation
  - MT Systems are cooperative
  - *Cf: Confidence estimation*
  - Measures may (and will) be biased
  - Measures may (and will) be easily cheated – but systems will never learn how
  - Fine-grained is good enough

- Give a (numerical) estimate on the quality of MT systems
  "On average, B is better than A" → Quantitative/comparative evaluation
  - MT Systems are your adversary
  - MT measure must not have any (unknown) bias
  - MT measure must be stable against "cheating"
  - Coarse-grained is good enough

What?

What is our focus when we evaluate MT systems?

Or: What do we consider a "good" translation?

- **A "beautiful" translation human readers can't spot from human translations?**
- **A translation that can easily be edited into a "beautiful" one?**
- **A translation which contains all the facts (more or less readably)?**
- **A translation which human readers can easily scan?**
- **A translation that can easily be used by an information retrieval system?**

Focus should even influence our decision for a human evaluation measure!

Focus?

Focus should even influence our decision for a human evaluation measure!

But: Usually, even in Evaluation campaigns, more like “you know a good translation if you see one” . . .

⇝ Hardly modelled in automatic MT evaluation at all!

⇝ At best, measure with high correlation with human judgment.

Tuning measures to systems

Recent trend in MT evaluation:

- **Find as many features correlating with human judgement as possible**
- **Or even: Find as many plausibly sounding features as possible**
- **Combine them**
- **Tune weights on old evaluation data**

Sounds good, gives good correlation on old data.

But what you are *really* doing is: You train a measure to recognize previously good systems!

Be careful what you wish for

. . .

- Tune weights on old evaluation data

Sounds good, gives good correlation on old data.

Assumption: This will also give good correlation on any new data — *whatever people do!*

But: This is not how MT research works!

- Some of these features might explain old data pretty well ⇝ get large weights
- MT systems have dozens to ten thousands of features by now
- Dozens to ten thousands of parameters are tuned to the evaluation measure and thus to these highly-weighted evaluation features
- We can safely assume that MT systems learn to cheat – whether we want or not

The question we should ask

...

- We can safely assume that MT systems learn to cheat – whether we want or not

So, the question we should ask about evaluation measures is

- not: What score will a good translation have?
- neither: Can we construct good translations with a bad score? *(We will always be able to...)*
- but: How could a translation look like (in the worst case) which achieves a better score?

Be careful what you wish for: Recall vs. Precision

Measures tuned on “old” human data weight recall much higher than precision:

Measure	weight on	
	recall / deletions	precision / insertions
TERp	85%	15%
METEOR -adq	80%	20%
-rank	95%	5%
CD6P4ER	80%	20%

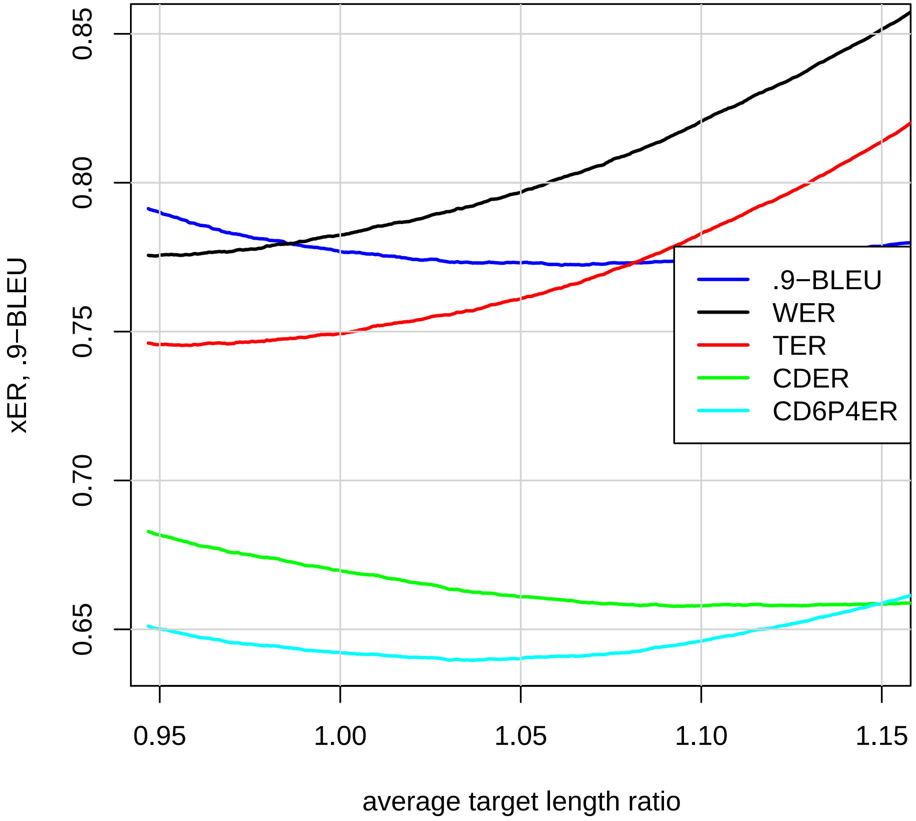
Good strategy for a system that will be scored on these measures:

- **Throw in unsure words**
- **Throw in frequent words**

But: *Is this what we want?*

Example: Length preference for some evaluation measures

Experiment: Vary the Word Penalty
(which has a direct effect on the hypothesis length)



How will a new measure behave?

Be careful what you wish for: Similarity to human translations

Just an anecdote – do not cite me here . . .

Someone proposed a measure some time ago based on the following logic:

1. **A good MT translation looks like a human translation**
2. **A feature identifying human translation is a good MT evaluation measure**
3. **Human translations are much more free in their choice of words than MT**
4. **Automatic word alignment is much more difficult with a free choice of words**
5. **⇝ Our measure: The worse the word alignment, the better the translation**

The Measure of Personal Choice

a.k.a. Reproducibility, Comparability . . .

If we give people the opportunity to tune their own measure – they will use it.

But: “Let the system with best human score win” is as good a training criterion as “Let system X win”.

**I definitely do not allege foul play – it is just hard to prove otherwise.
And the same hold also for evaluations.**

To get a feeling for a measure

- By now, we know how much +0.1 in BLEU is, or +0.5, or +2.0.
- Maybe also for TER, HTER, METEOR.
- For each new measure (and language), we need to get a feeling for it
- Maybe not even possible for measures tuned to certain conditions . . .

My conclusions

- Use the Known Evil! (At least in non-cooperative settings)
  - We know where systems can cheat (and where not)
  - We know what happens if we tune against a measure
  - We have fair baseline systems available

- Define at least a stress test set for new evaluation measures
  - Translations from a tuning run
  - Simple hacks – deletion, insertion of frequent words/phrases, . . .
  - Small mistakes with grave consequences?
    - missing V or negation,
    - transposition of S and O
    - . . .
  - Dependency on source difficulty
  - . . .

Thank you

Comments?