

# Projects related to MT evaluation

# The dilemma of MT evaluation

- human evaluation tedious, not always effective, but no obvious bias
  - automatic evaluation obviously biased, but repeatable, "objective"
- ➔ We should make human evaluation more effective or automatic evaluation less biased

# Three basic approaches:

1. make human evaluation more effective
  - focus on effective tasks and units of a good size
  - help with automatic pre-processing
2. remove bias from automatic evaluation
  - check for violation of non-local linguistic constraints (beyond n-grams)
  - exploit linguistic knowledge about semantic equivalence (synonyms, alternative constructions)
  - give different weights to different kinds of deviations (a missing "not" is typically more severe than a missing "the")
3. switch to a task-based evaluation methodology
  - Review and extend Hervé's experiments on translation of comprehension tests

# Project 1

Review the infrastructure behind the current WMT08 evaluation campaign and see how it could be improved. We can obtain all data and software behind this, so this is only limited by our imagination and technical capabilities.

- identify some major issues with the current setup (ranking long sentences too tedious, short snippets not interesting)
- find and specify potential solutions
- provide prototypical implementations

# Project 2

If all MT output were good, one could run alignment algorithms between MT output and reference translations and obtain (modulo alignment errors) a monolingual collection of equivalent phrases (synonyms, equivalent constructions etc). For real-world MT output and alignment algorithms, one will instead find a combination of three types of phrase pairs:

- phrase pairs that are equivalent
- phrase pairs that show translations errors
- phrase pairs that show alignment errors

# Project 2 cnt'd

Try to work on this three-way classification by

- generating data and trying to annotate it manually
- using machine learning algorithms to learn the classification from a small set of examples
- building a (Web-based) GUI that makes the human classification of these cases as efficient as possible

As a side effect of this one might hope to find new ways to reduce the number of cases in class 3, but that might be a different story...

# Project 3

N-gram similarities like BLEU and NIST scores have known limitations, especially with respect to syntactic well-formedness. Existing NLP tools might be better suited for judging these aspects. We can try several approaches here, the following list is just meant as a starter:

- Send MT output through a PoS tagger and try to find "suspicious" sequences of PoS tags in the result or other properties that might contain hints for non-wellformedness
- Send MT output through a (robust) shallow parser and look for e.g. agreement violation in non-local dependencies or other problems in the resulting structures
- Send MT output through a commercial tool for grammar- and style checking, such as the products of the DFKI spinoff acrolinx in Berlin. Measure how many interesting problems are caught by their tools and how many of the problems caught are interesting
- Implement a PoS-enriched variant of BLEU score that allows for a more detailed comparison between MT systems, e.g. in the form that one systems makes more errors for nouns, whereas another makes more errors for verbs etc.

# Project 4

Elaborate the task-based evaluation approach for which Hervé built a first prototype.