Will there be winners?

Gr. Thurmair,
Linguatec
Prague, 2009-05-14

Baseline

Outcome of the MT Workshop Athens, 2009:
1. Knowledge-driven systems are still not beaten in quality by data-driven systems in the majority of language directions researched

2. MT output acceptability is between 50% and 20%, depending on language direction

Winners look different!

Error analysis
- knowledge-driven systems:
  - Parse failures / robustness, lexical selection, fluency
- data-driven systems:
  - Non-local phenomena / word order, output grammaticality, accuracy, domain dependency
Hybrid MT systems

• Progress by exploiting all available resources
  – Dictionaries and grammars (knowledge-driven systems)
  – Phrase tables and language models (data-driven systems)
  – Identify and use the 'knowledge' encoded in them

• Build flexible system architectures
  – Support the use of all types of resources
  – Scale the systems according to their availability
    • in case of "less resourced" translation directions
  – types of architectures:
    • Coupling (linear, parallel)
    • Extensions of RMT / SMT skeletons
    • Hybrid combinations of components

Analysis

• Identify source content: what needs to be translated!
  Don’t start target sentence without knowing what you want to say!
  • Source language is not just a bag of words (word semantics);
    You can only generate what you have analysed

  – Strengthen analysis capacity!
    • This is a knowledge-driven task!
      • E.g.: syntactic functions / cases; intonation (Kuo/Ramsay 08)
        tense & aspect, pronouns, gapping, ...
    • Must be supported by data-driven resources
      • Tree banks, collocations, etc.
    • Go beyond sentence boundary
      • NE coreference; text grammars, discourse referents
    • Develop robust fallbacks for analysis failures
Transfer

- Maintain translation **accuracy**
  - No missing words, no spurious add-ons

- Use all existing knowledge sources
  - (All) possible translations are coded in dictionaries!
    - Phrase table contents should be subsets of those ...
      => use dictionaries for cleanup / control
      => add dictionary information to phrase tables
  - Phrase tables provide **probabilities** to translations
    - RMT Transfer components should use them
  - Structural (and complex lexical) transfer
    - Source and target trees are not isomorphic ...

- *Don’t decide* on transfer selection with limited knowledge
  - Give generation a **lattice** of translation options

Generation

- Improve **fluency / grammaticality** of translation result

- Build a knowledge-driven skeleton of target text
  - Constituent ordering, based on syntactic functions / cases
    - (SOV>SVO; do-insertion; complex VP split, ...)
  - Ensure grammaticality of the output
    - This is the most significant human evaluation criterion

- Use language model information for fine tuning
  - for selection of lexical units
  - for handling ’language-use’ dependent phenomena, e.g.:
    - Adverb placement
    - Preposition selection
Usage / Context

- MT is not a playground for machine learning (or linguistic) approaches

- Worry about the (real) user problems in translation
  - deficient input
  - customer specific terminology
  - translation memory integration
  - domain adaptation
  - embedded solutions

- Winners will be systems with \textit{user/market acceptance}
  - Improve \textit{MT quality}, by integrating all available resources
  - Improve domain \textit{integration} / application orientation

\textbf{Thank you for your attention}

g.thurmair@linguatec.de