**Machine Translation**

Rule-based MT & MT evaluation

Jörg Tiedemann

jorg.tiedemann@lingfil.uu.se
Department of Linguistics and Philology
Uppsala University

---

**What do we need?**

- **Direct translation**: a huge dictionary
- **Transfer-based translation**: grammars & rules
  - rules for source language analysis (syntactic/semantic)
  - rules for source-to-target transfer
  - rules for target language generation
- **Interlingua-based translation**: the same but no transfer

---

**What are the problems?**

- **Direct translation**: dictionary has to cover all cross-lingual phenomena
  - need to include contextual information in dictionary (long phrases)
  - problems with non-compositionality and ambiguity
  - inflectional agreement, shifts in word order & structure
  → direct translation systems include simplistic rules
Direct Translation

- **simplistic approach:** only low-level pre/post-processing (tokenization, etc ...)
- **advanced approach:** handle some specific phenomena
  - identification & handling of syntactic ambiguity
  - morphological processing/synthesis
  - word re-ordering rules
  - rules for prepositions
  - handling of compounds and idioms, ...

(Advanced) Direct Translation

Is it feasible?

- a lot of compositionality in natural language
- many similarities between languages (especially between related languages)
- example: Systran (in daily use by the European Commission)
  - > 1.6 million dictionary units
  - dictionaries for different domains
  - more-and-more transfer based

→ many data-driven MT systems ~ direct translation systems

Transfer-based Translation

Motivation:

- complete analysis of source language sentences
- transfer step covers divergences between languages
- handle lexical & structural ambiguity in one formalism

→ What kind of information/tools do we need?

What kind of information/tools do we need?

- source language parser (morpho-syntactic analysis)
- transfer engine (e.g. unification based grammar)
- target language generator

→ modular design
Transfer-based MT

Syntactic Transfer rules (systematic structural differences)

- English to Spanish:
  - $\text{NP} \rightarrow \text{Adjective1 Noun2} \Rightarrow \text{NP} \rightarrow \text{Noun2 Adjective1}$

- Chinese to English:
  - $\text{VP} \rightarrow \text{PP[+Goal]} \ V \Rightarrow \text{VP} \rightarrow \ V \ PP[+Goal]$

- English to Japanese:
  - $\text{VP} \rightarrow \text{V NP} \Rightarrow \text{VP} \rightarrow \text{NP V}$
  - $\text{NP} \rightarrow \text{NP1 RelClause2} \Rightarrow \text{NP} \rightarrow \text{RelClause2 NP1}$

Need preference mechanism for rule selection!

- on $\rightarrow$ på
- come.vb $\rightarrow$ kom.vb
- come on $\rightarrow$ kom igen
- sit.vb on NP $\rightarrow$ sitta.vb på NP
- sit.vb on the couch $\rightarrow$ sitta.vb i soffan

Common: preference for more specific rules
(the same applies to structural transfer rules)

What are the problems?

- lots of grammar engineering (writing rules ...)
- language-pair specific rules
- exponential ambiguity vs. early commitment
- handle variation & define preference
- coverage & robustness

→ Good quality can be achieved but low coverage!
Interlingua-based Translation

Advantages:
▶ no language-pair specific transfer
▶ simple (?) to add new languages
   (add new analysis/generation component)

Disadvantages:
▶ need to design interlingua that covers all language phenomena
▶ need semantic representation (and that's hard!)
▶ may even fail for simple examples (that can be translated literally)

Classical Rule-based Translation

→ Too much manual work involved!

Is there no hope for transfer-based systems?
▶ Domain-specific tasks
▶ Rule-induction
▶ Hybrid systems

Rule-based Translation

Domain-specific MT
▶ high quality translation for specific domains
▶ controlled languages:
   ▶ complete coverage of source language (lexicon, grammar)
     & terminology in that domain
   ▶ reduced (no?) ambiguity
   ▶ requires language checker tools
     (for source language documents)

→ high quality & high consistency

Second Part: MT evaluation

▶ How can we measure MT quality?
▶ How can we compare MT engines?
▶ How can we measure progress in MT development?
What do we expect from MT?

- adequacy & informativeness (preserve meaning)
- fluency & grammaticality (translation needs to be natural)
- acceptance (for its task), cost-efficiency

Evaluation is difficult!

- What is the best translation? (language variation!)
- Subjective aspects (What is “fluent”? Clarity? Style?)
- What is “grammatical”? 
- What is “adequate”? (Is it possible to be adequate?)

MT evaluation

Manual evaluation

- ask actual users to rate translations
- statistics over user responses
- separate evaluations of adequacy & fluency
- requires guidelines
- task-specific evaluation

Automatic evaluation

- compare to reference translations
- approximations by measuring overlaps
- strong bias but useful for rapid development

Typical setup:

<table>
<thead>
<tr>
<th>Adequacy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 = All</td>
<td>5 = Flawless English</td>
</tr>
<tr>
<td>4 = Most</td>
<td>4 = Good English</td>
</tr>
<tr>
<td>3 = Much</td>
<td>3 = Non-native English</td>
</tr>
<tr>
<td>2 = Little</td>
<td>2 = Disfluent English</td>
</tr>
<tr>
<td>1 = None</td>
<td>1 = Incomprehensible</td>
</tr>
</tbody>
</table>

Strong correlations if evaluated together
→ Separate evaluation on different examples?
Manual MT evaluation

Compare MT engines:

▶ rank proposed translations
▶ measure relative quality
▶ could include manual translation
▶ could rank selected segments only

→ simpler task, better agreement, less guidelines

Task-specific evaluation

Different tasks require different types of quality!

- **browsing quality**: Is the translation understandable in its context? (main contents is clear)
- **post-editing quality**: How many edit operations are required to turn it into a good translation?
- **publishing quality**: How many human interventions are necessary to make the entire document ready for printing?

→ Difficult to have a general framework!

Automatic Evaluation

- constant evaluation is necessary for system development
- ... but manual evaluation is too expensive!

→ Automatic evaluation is required!

Comparison of MT output with reference translations:
BLEU, NIST, METEOR, WER, PER, TER, hTER, ROUGE ...
Automatic Evaluation

Why are there so many automatic evaluation measures?

- only approximations of adequacy & fluency
- different types of correlations with human evaluation
- possible bias towards certain approaches
- tuning on automatic measures makes them inappropriate

The “BLEU-score Revolution”

Basic idea:

- translation is better if it is closer to given (correct) reference translations
- “closeness” can be measured in terms of N-gram overlaps → modified form of precision
- add “brevity penalty” to account for sentence length

→ High correlation with human judgments (0.99 & 0.96 in original paper)!

Candidate (1) shares more words and word-N-grams with the reference translations than candidate (2)!
→ Compute precision scores (proportion of correct N-grams)
The “BLEU-score Revolution”

Modified N-gram precision (for each N-gram):

\[
\text{count}_{\text{clip}} = \min (\text{count}_{\text{candidate}}, \max \text{count}_{\text{reference}})
\]

→ Avoid to count correct N-grams more often than they appear in any reference translation!

Example

Candidate: the the the the the the the the the the.
Reference 1: The cat is on the mat.
Reference 2: There is a cat on the mat.

\[
\text{count}_{\text{clip}}(\text{the}) = 2
\]

\[
\rho_{\text{unigram}} = 2/7 \text{ (unigram precision)}
\]

Brevity penalty (BP) for short candidates (c):

\[
BP \begin{cases} 
1 & \text{if } c > r \\
\exp(1 - r/c) & \text{if } c \leq r
\end{cases}
\]

Putting it all together:

\[
\text{BLEU} = BP \times \exp \left( \sum_{n=1}^{N} w_n \log p_n \right)
\]

Usually \( w_n = 1/N \) and \( N = 4 \)

What’s good about BLEU:

▶ easy to compute
▶ gives scores from 0 to 100%
▶ can be used to measure system development
▶ can quickly test different system parameters
The “BLEU-score Revolution”

What’s risky with BLEU:

▶ systems are tuned for optimizing BLEU scores → strong bias, less correlation with human judgments
▶ often only one reference translation (not good!)
▶ difficult to compare systems with generally different approaches
▶ difficult to compare performance on different language pairs
▶ even more difficult to compare results on different domains & text types

Alternative Measures

After BLEU many evaluation measures have been proposed:

Other evaluation metrics

- **NIST**: BLEU + n-gram weights according to informativeness (rare → more informative)
- **METEOR**: harmonic mean of unigram precision and recall + synonym expansion & stemming
- **WER, PER, TER, hTER**: based on edit distance (insertion, deletion, substitution, moving)
- **Dependency overlap**: overlap in grammatical relations
- **Semantic role overlap**: lexical overlap between semantic roles

Metrics can be combined to better correlate with human judgments! → Automatically train combination weights!

Summary on MT Evaluation

▶ automatic evaluation is (very) popular but risky
▶ human evaluation is safe but expensive
▶ automatic measures are great for system development
▶ lots of discussion about MT evaluation

→ Don’t forget to look at actual MT output!

Next

Lab:

1. evaluation experiment: play a little game
   ▶ try to guess the type of translations (automatic or manual)
   ▶ test if automatic translations are understandable or not
   ▶ challenge the system and find out MT weaknesses

2. evaluate on-line translation services
   ▶ run Google Translate and evaluate results

Next lecture:

▶ The amazing utility of parallel corpora