**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**

- **simpistic 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?

**Transfer-based Translation**

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:
  - NP → Adjective1 Noun2 ⇒ NP → Noun2 Adjective1
- Chinese to English:
- English to Japanese:
  - VP → V NP ⇒ VP → V
  - NP → NP1 RelClause2 ⇒ NP → RelClause2 NP1

Need preference mechanism for rule selection!

- on → på
- come.vb → kom.vb
- sit.vb on NP → sitta.vb på NP

→ Common: preference for more specific rules

Transfer-based MT

What are the problems?

- lots of grammar engineering (writing rules ...)
- language-pair specific rules
- exponential ambiguity
- variation & 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 (direct) examples

Classical Rule-based Translation

→ Too much manual work involved!

Is there no hope for rule-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
  ▶ reduce 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)

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

Manual MT evaluation

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?

Compare MT engines:

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

→ simpler task, better agreement, less guidelines
Rule-based MT  MT evaluation

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!

Possibly: Decide quality level depending on evaluation results

Manual MT evaluation

What are the problems?

▶ need volunteers (every time we want to evaluate) → expensive evaluation!
▶ could be hard to setup
▶ subjective measures & disagreement between annotators

→ Difficult to find a better solution ....

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, ROUGE ...

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”

- introduced in 2002 by Papineni et al
- desperately needed by rapid MT development
- quickly adapted by statistical MT community
- created a boom in MT research/experiments

→ Many MT papers report only BLEU scores and don’t even look at the translations ...

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)!

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

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

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

Example

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

\[
\text{count}_{\text{clip}}(\text{the}) = 2
\]
\[
\rho_{\text{unigram}} = 2/7 \text{ (unigram precision)}
\]
The “BLEU-score Revolution”

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

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

Putting it all together:

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

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

What’s risky with BLEU:
- systems are tuned for optimizing BLEU scores
  $\rightarrow$ strong bias, less correlation with human judgments
- often only one reference translation
- 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**: 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. try to manually evaluate on-line translation services
2. 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

Next lecture:
▶ The amazing utility of parallel corpora (part I)