Different Architectures of Hybrid Machine Translation

Chrysanthi Melanou
SoSe 2018
Different approaches to Machine Translation

Knowledge-Driven MT

1. Example-Based
2. Statistics-Based

Data-Driven MT

1. Rule-Based
2. Knowledge-Based
Knowledge-driven Machine Translation

1. Rule-based MT:
   - based on linguistic information about source and target language
   - RBMT system generates input to output on the basis of morphological, syntactic and semantic analysis of both analysis.

2. Knowledge-based MT:
Data-driven Machine Translation

1. **Example-based MT:**
   - Characterised by its use of bilingual corpus with parallel texts as its main knowledge base at run-time
   - Consists of: a matching module, recombination, and smoothing

2. **Statistics-based MT:**
   - Based on statistical models whose parameters are derived from the analysis of bilingual text corpora
## Summary of Different Approaches to MT Systems

<table>
<thead>
<tr>
<th></th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Rule-Based</strong></td>
<td>1. Easy to build an initial system</td>
<td>1. Rules are formulated by experts</td>
</tr>
<tr>
<td></td>
<td>2. Based on linguistic theories</td>
<td>2. Difficult to maintain and extend</td>
</tr>
<tr>
<td></td>
<td>3. Effective for core phenomena</td>
<td>3. Ineffective for marginal phenomena</td>
</tr>
<tr>
<td><strong>Knowledge-Based</strong></td>
<td>1. Based on taxonomy of knowledge</td>
<td>1. Hard to build knowledge hierarchy</td>
</tr>
<tr>
<td></td>
<td>2. Contains an inference engine</td>
<td>2. Hard to define granularity of knowledge</td>
</tr>
<tr>
<td></td>
<td>3. Interlingual representation</td>
<td>3. Hard to represent knowledge</td>
</tr>
<tr>
<td><strong>Example-Based</strong></td>
<td>1. Extracts knowledge from corpus</td>
<td>1. Similarity measure is sensitive to system</td>
</tr>
<tr>
<td></td>
<td>2. Based on translation patterns in corpus</td>
<td>2. Search cost is expensive</td>
</tr>
<tr>
<td></td>
<td>3. Reduces the human cost</td>
<td>3. Knowledge acquisition is still problematic</td>
</tr>
<tr>
<td><strong>Statistics-Based</strong></td>
<td>1. Numerical knowledge</td>
<td>1. No linguistic background</td>
</tr>
<tr>
<td></td>
<td>2. Extracts knowledge from corpus</td>
<td>2. Search cost is expensive</td>
</tr>
<tr>
<td></td>
<td>3. Reduces the human cost</td>
<td>3. Hard to capture long distance phenomena</td>
</tr>
<tr>
<td></td>
<td>4. Model is mathematically grounded</td>
<td></td>
</tr>
</tbody>
</table>
The 2 basic areas of Machine Translation

1. Knowledge-driven MT
2. Data-driven MT

😊 Both types of systems reach comparable translation quality

😊 The level of output acceptance (in terms of understandability) in the best language directions is about 50%
Comparison of RBMT and SMT systems

<table>
<thead>
<tr>
<th>RBMT</th>
<th>SMT</th>
</tr>
</thead>
<tbody>
<tr>
<td>✅ Accurate translation</td>
<td>❌ Lose adequacy due to missing or spurious translations</td>
</tr>
<tr>
<td>❌ Weakness in lexical selection in transfer</td>
<td>❌ Difficulties to cope with linguistic phenomena</td>
</tr>
<tr>
<td>❌ Lack robustness in analysis failures sentences</td>
<td>✅ Read more fluent and are better in lexical selection</td>
</tr>
<tr>
<td></td>
<td>✅ More robust and always produce output</td>
</tr>
</tbody>
</table>
Strengths and Weaknesses of SMT vs. RBMT

(RBMT: translate pro ↔ SMT: Koehn 2005, examples from EuroParl)

EN: *I wish the negotiators continued success with their work in this important area.*

RBMT: *Ich wünsche, dass die Unterhändler Erfolg mit ihrer Arbeit in diesem wichtigen Bereich fortsetzten.*

*continued:* Verb instead of adjective

SMT: *Ich wünsche der Verhandlungsführer fortgesetzte Erfolg bei ihrer Arbeit in diesem wichtigen Bereich.*

*three wrong inflectional endings*
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td>47</td>
<td>41</td>
<td>44</td>
<td>30</td>
<td>10</td>
<td>8</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>German</td>
<td>48</td>
<td>24</td>
<td>8</td>
<td>10</td>
<td>4</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>French</td>
<td>40</td>
<td>23</td>
<td>11</td>
<td>13</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>41</td>
<td>7</td>
<td>11</td>
<td>9</td>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italian</td>
<td>29</td>
<td>10</td>
<td>13</td>
<td>9</td>
<td>4</td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Portuguese</td>
<td>29</td>
<td>5</td>
<td>7</td>
<td>8</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dutch</td>
<td>10</td>
<td>2</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polish</td>
<td>7</td>
<td>2</td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latin</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Greek</td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Czech</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hungarian</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Swedish</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Finnish</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Slovak</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Romanian</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Danish</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
What is Hybrid Machine Translation?

Systems which try to profit from the respective other approach and avoid mistakes for which solutions already exist must be hybrid solutions, combining knowledge-driven and data-driven elements.
What is coupling?

Coupling means that two or more existing machine translation systems are used to produce improved MT output.

1. Serial coupling
2. Parallel coupling
Serial Coupling

Statistical post-editing (PSE) of RBMT:

• Consists in modifying RBMT output by means of a SMT post-editing component.

• Both RBMT and SMT technologies will be combined in order to increase the overall translation quality.
✓ The output tends to be grammatical and the main effect of the combination is an increase in lexical selection quality (one the weak points of RBMT)

✓ Care to avoid errors by the SMT postprocessor like: syntactic structure of the output and to keep the Named Entities of the output

✓ Use a RBMT system‘s syntactic structure and only local alternatives (using POS information)
Parallel Coupling

• This coupling employs several MT systems in parallel and uses some mechanism to select or produce the best output from the result set.

Example 1:

The best translations from a list of n-best translations.

They search for the best n-grams in all output hypotheses available and then select the best hypothesis from the candidate list.
Example 2:

It uses confusion networks and generates an output sentence on the basis of the variable MT outputs.

A skeleton is selected as a basis and for each position of the skeleton the best translation alternative is identified and composed to the overall output sentence.

To overcome the risk of skeleton selection, techniques have been applied to build confusion network such as to let every hypothesis be the skeleton and calculate the overall best solution.
Architecture Extensions

• By extension we mean that the system architecture basically follows the RBMT or SMT paradigm but is modified by including resources of the respective other approach.

• Modifications can occur as:

  Pre-editing

  Core modification
RBMT Extensions

• Approaches to improve RBMT systems with data-driven procedures focus on two problems:

  **Pre-editing:** is tried both on the dictionary side (by running Term-Extraction tools) and on the grammar side (by automatically extracting grammar rules from corpora)

  **System core modification:** is attempted by adding probability information (to the analysis / parsing process) and by manipulating the transfer selection process
1. RBMT – Pre-editing

a. Learning of dictionary entries

Pre-editing in RBMT systems means to apply data-driven techniques for terminology extraction from corpora either on a monolingual system (to find missing entries in the system’s dictionaries) or from bilingual corpora (to find translation candidates and to load them into the system dictionary).

The challenges are:

• Recognition of multiword terms.
• Linguistic annotation of the recognised terms.
b. Learning of rules in RBMT

The challenge for learning grammar rules seems to be that very many rule candidates are identified, even for small corpora and that is difficult to select the low-frequent „good“ rules from noise produced by the extraction technique, which causes parse failures.
2. RBMT core system modifications

Hybrid approaches focus more on translation selection in the transfer phase, which is one of the weaknesses of RBMT system

Traditional approaches to RBMT transfer selection rely on two techniques:

• Assignement of subject area codes to translation Tests and actions on certain contextual/structural properties

Therefore, additional and robust means for lexical selection need to be developed
Pre-editing: is tried to prepare the data; the most important steps are morphology, POS information/syntactic information, and word reordering.

System core modifications are tried by adding RBMT information to the phrase tables, and by using factored translation.
1. SMT – Pre-editing

a. **Morphology:** compounding (of english)/decompounding (of german words) to parallelise alignment. In languages with agglutinative behaviour like Turkish, Hungarian or Arabic, preprocessing is required to split complex word strings into meaningful parts to be able to align them.

b. **Syntax:** the idea is to parse source and target side of a corpus, and only let syntactically well-formed phrases enter the phrase table.
c. **Reorderings**: is a major challenge for SMT not just because languages have different word and constituent order (SVO vs. SOV) but also because the constituent order is meaning-bearing (e.g. case marking in English).

A proposal would be to extend the input word sequence into a lattice containing different reorderings of the input words.

Distortion rules can be set up manually or automatically, for contiguous and discontiguous POS sequences, by matching them on source and target side of the training corpus.

The input lattice contains the respective distorted strings, with weights on the probability of the distortion.
2. SMT core system modification

There are 3 approaches to incorporate RBMT resources into an SMT architecture:

a. Extension of the Phrase Table

b. Rule-based control of the Language-Model-based generation

c. Factored translation
a. **Importing RBMT resources into the phrase table**

Run RBMT systems in addition to SMT systems, and enrich the SMT phrase tables by terms and phrases produced by the RBMT.

**Result:** The coverage of the system can be increased.

However, as the SMT decoder runs last, the output can be less grammatical than the one of the original RBMT.
b. **Improving decoding using target grammars**

Several proposals exist how to learn grammar and transfer rules from bilingual corpora.

e.g. The hierarchical translation (Chiang 2007) uses synchronous context-free grammar in decoding.
c. Factored Translation

provides more information at word Level:

1. It treats words not just as simple text forms but as vectors of features
2. The approach decomposes phrase translations into a sequence of mapping steps, with translation steps operating on phrase level, and generation steps on word level.
Genuine hybrid architectures

They do not just use addons to their system architecture but combine whole system components of the respective approaches into novel systems.

They use three basic components:

1. Identification of source language „chunks“
2. Transformation of such chunks into the target language by means of a bilingual resource
3. Generation of a target language sentence
Rule-based analysis, bilingual dictionary, target language model

Analysis is done using available NL tools (lemmatisers, taggers, chunkers).

Transfer is based on existing dictionaries.

Generation uses a language model (based on a tokenized and tagged English corpus).
Data driven analysis and generation, bilingual dictionary

The required resources are:

a. a bilingual dictionary
b. an n-gram indexed target language corpus

In analysis an n-gram window is moved over the sentence, and all words in the window are translated using the bilingual dictionary. The result is a lattice of n-gram translations.
Domain Adaptation

All kinds of MT systems must cope with the fact that they will be used not only in the domain for which they had been developed but also for other domains.

While RBMT systems support adaptation by dictionary import and coding, the situation is less obvious for SMT-based systems, and a significant drop of quality had been observed.
Conclusion

The selection of the „best“ architecture for a practical MT system depends on three basic factors:

1. the intended use case e.g. the translation domain(s)
2. the translation quality which can be achieved (both for the domain in which the system was trained, and other domains in which the systems are supposed to be used)

According to Thurmair there are three types of HMT:

1. coupling
2. architecture extension
3. genuine hybrid architectures
References

Eisele A., 2008: *Hybrid Architectures for Machine Translation*

Lagarda A.-L., 2009: *Statistical Post-Editing of a Rule-Based Machine*

Thurmair Gr., 2006: *Comparing different architectures of hybrid Machine Translation systems*