Hybrid Machine Translation Guided by a Rule–Based System

Cristina España-Bonet, Gorka Labaka, Lluís Màrquez, Arantza Díaz de Ilarraza, Kepa Sarasola

Universitat Politècnica de Catalunya, University of the Basque Country
Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT
- Statistical MT
- We would like to get the best of each world.
All the current MT approaches have their pros and cons.

- **Rule-Based MT**
  - Syntactically better translations: long distance reordering, agreement
  - Worse lexical selection
  - Performance degradation for unexpected syntactic structures

- **Statistical MT**

- We would like to get the best of each world.
Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT: +syntax, −lexical selection, −unexpected structures
- Statistical MT
  - Better lexical selection and fluency
  - Structurally worse translations
  - Performance degradation for *out-domain* texts.
- We would like to get the best of each world.
Motivation

- All the current MT approaches have their pros and cons.
- Rule-Based MT: +syntax, −lexical selection, −unexpected structures
- Statistical MT: +lexical selection, −long distance reordering, −out-domain performance
- We would like to get the best of each world.
  - RBMT’s grammatical correctness.
  - SMT’s lexical selection
  - SMT tolerance to unexpected structures.
Outline

1. SMatxinT: RBMT leaded hybrid system
2. Experimental Results
3. Conclusions
Outline

1. SMatxinT: RBMT leaded hybrid system
   Individual Systems
   SMatxinT architecture

2. Experimental Results

3. Conclusions
Statistical machine translator(s)

- **SMTb (Moses with default configuration):**
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...

- **SMTm:**
  - Uses segmentation of Basque
  - Default Moses configuration: phrase and lexical prob., 3-gram LM, ...
  - Word generation phase to go from segmented text to final words
  - Word level LM added using n-best list
  - ...
Statistical machine translator(s)

- **SMTb (Moses with default configuration):**
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...

- **SMTm:**
  - Uses segmentation of Basque
  - Default Moses configuration: phrase and lexical prob., 3-gram LM, ...
  - Word generation phase to go from segmented text to final words
  - Word level LM added using n-best list
Statistical machine translator(s)

- **SMTb (Moses with default configuration):**
  - phrase and lexical translation probabilities
  - 3-gram LM
  - lexicalized reordering
  - ...

- **SMTm:**
  - Uses segmentation of Basque
  - Default Moses configuration: phrase and lexical prob., 3-gram LM, ...
  - Word generation phase to go from segmented text to final words
  - Word level LM added using n-best list

- ...

Hybrid Machine Translation Guided by a Rule–Based System

Gorka Labaka
Matxin, a rule-based translator

- In-house developed Open-Source Rule-Based MT system.
- Classical transfer-based approach: analysis, transfer and generation.
- Chunk-based dependency tree:
  - Dependency trees + chunk boundaries.
SMatxinT: RBMT guided Hybrid MT

- Translation process is guide by the Rule-Based system.
  - Ensures syntactic correctness.
  - Takes care of long distance reordering.

- Allow substitution of RBMT partial translations with their SMT counterparts.
  - Substitutions of short strings improve lexical selection.
  - Longer substitutions allow to overcome wrong syntactic analysis.
Two new modules are added to the RBMT architecture.
- Tree enrichment (between analysis and transfer).
- Linear decoding (after generation).
Tree enrichment

- After analysis, and before transfer.
- Each phrase in the tree is enriched with one (or several) SMT translation counterpart.
- Two types of SMT correspondences:
Tree enrichment

- After analysis, and before transfer.
- Each phrase in the tree is enriched with one (or several) SMT translation counterpart.
- Two types of SMT correspondences:
  - local: Allows SMatxinT to use SMT lexical selection
Tree enrichment

- After analysis, and before transfer.
- Each phrase in the tree is enriched with one (or several) SMT translation counterparts.
- Two types of SMT correspondences:
  - local: Allows SMatxinT to use SMT lexical selection
  - full subtree: Allows SMatxinT to overcome analysis errors.
Tree enrichment: Context Discriminate translations

- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.
Tree enrichment: Context Discriminate translations

- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.
Tree enrichment: Context Discriminate translations

- Many chunks are too small to SMT
  - Few context to get proper translation.
- We also extract corresponding local translation form longer SMT translations.
  - Based on alignment provided by SMT.
  - Allows the SMT to use the context.
Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.

emanaldiak ez dituzte aurreikusten arauz kontrako armekin , barne sailburua baiez tetu zuen
Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
- A Linear decoding module is used to choose the best candidate for each phrase.
Linear decoding

- After tree enrichment, transfer and generation are applied as usual.
- RBMT translation enriched with several candidates for each phrase.
- A Linear decoding module is used to choose the best candidate for each phrase.
  - We used Moses (in monotonous way) as linear decoder.
  - A wide range or features are defined.
Linear decoding: Features

**Standard SMT features**

- Language model
- Word penalty
- Phrase penalty
Linear decoding: Features

**Standard SMT features**
- Language model
- Word penalty
- Phrase penalty

**Source/consensus features**
- Counter (1...n)
- SMT (1/e)
- RBMT (1/e)
- Both (e#)
Linear decoding: Features

Standard SMT features
- Language model
- Word penalty
- Phrase penalty

Source/consensus features
- Counter (1...n)
- SMT (1/e)
- RBMT (1/e)
- Both (e#)

Lexical features
- Corpus lexical probabilities (eu2es & es2eu)
- Dictionary lexical probabilities (eu2es & es2eu)
Outline

1. SMatxinT: RBMT leaded hybrid system

2. Experimental Results
   - Corpora
   - Systems
   - Results

3. Conclusions
Corpora

Language pair
- Spanish–Basque

Training corpus
- Administrative documents and descriptions of TV programs
- 491,853 parallel sentences

Development and test corpora
- *Elhuyar dev & test*: Administrative documents (1500 sentences)
- *EITB*: News (1500 sentences, 1 reference)
- *NEWS*: News (1500 sentences, 2 references)
Individual systems

- SMTb
- SMTm
- Matxin
Systems

**Individual systems**
- SMTb
- SMTm
- Matxin

**Hybrid systems**
- SMatxinT₀: Hybrid system where only SMT translations are used
- SMatxinT
Systems

**Individual systems**
- SMTb
- SMTm
- Matxin

**Hybrid systems**
- SMatxinT₀
- SMatxinT

**Control system**
- Google
# Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th>EITB</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td><strong>Matxin</strong></td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
</tr>
<tr>
<td><strong>SMTb</strong></td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
</tr>
<tr>
<td><strong>SMTm</strong></td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
</tr>
<tr>
<td><strong>Google</strong></td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
</tr>
<tr>
<td><strong>SMatxinT</strong></td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
</tr>
<tr>
<td><strong>SMatxinT</strong></td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
</tr>
</tbody>
</table>
Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th>EITB</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Matxin</td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
</tr>
<tr>
<td>SMTb</td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
</tr>
<tr>
<td>SMTm</td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
</tr>
<tr>
<td>Google</td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
</tr>
<tr>
<td>SMatxinT</td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
</tr>
<tr>
<td>SMatxinT</td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
</tr>
</tbody>
</table>

BLEU and TER scores are generally low
### Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th></th>
<th>EITB</th>
<th></th>
<th>NEWS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
<td>TER</td>
</tr>
<tr>
<td>Matxin</td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
<td>87.73</td>
<td>11.72</td>
<td>82.04</td>
</tr>
<tr>
<td>SMTb</td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
<td>83.27</td>
<td>14.74</td>
<td>78.63</td>
</tr>
<tr>
<td>SMTm</td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
<td>85.59</td>
<td>14.58</td>
<td>78.90</td>
</tr>
<tr>
<td>Google</td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
<td>86.32</td>
<td>12.01</td>
<td>81.84</td>
</tr>
<tr>
<td><strong>SMatxinT</strong>&lt;sub&gt;0&lt;/sub&gt;</td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
<td>82.17</td>
<td>14.90</td>
<td>77.29</td>
</tr>
<tr>
<td><strong>SMatxinT</strong></td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
<td>81.33</td>
<td>15.31</td>
<td>76.54</td>
</tr>
</tbody>
</table>

All systems outperform the RBMT.
### Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th>EITB</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Matxin</td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
</tr>
<tr>
<td>SMTb</td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
</tr>
<tr>
<td>SMTm</td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
</tr>
<tr>
<td>Google</td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
</tr>
<tr>
<td>SMatxinT&lt;sub&gt;0&lt;/sub&gt;</td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
</tr>
<tr>
<td>SMatxinT</td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
</tr>
</tbody>
</table>

Results of hybrid system are comparable but not better than SMT subsystems.
## Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th>EITB</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Matxin</td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
</tr>
<tr>
<td>SMTb</td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
</tr>
<tr>
<td>SMTm</td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
</tr>
<tr>
<td>Google</td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
</tr>
<tr>
<td>SMatxinT₀</td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
</tr>
<tr>
<td>SMatxinT</td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
</tr>
</tbody>
</table>

SMatxinT and SMatxinT₀ are consistently better than single systems.
## Automatic Metrics

<table>
<thead>
<tr>
<th></th>
<th>Elhuyar</th>
<th>EITB</th>
<th>NEWS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU</td>
<td>TER</td>
<td>BLEU</td>
</tr>
<tr>
<td>Matxin</td>
<td>5.10</td>
<td>83.32</td>
<td>5.77</td>
</tr>
<tr>
<td>SMTb</td>
<td>14.96</td>
<td>70.20</td>
<td>8.03</td>
</tr>
<tr>
<td>SMTm</td>
<td>13.71</td>
<td>71.64</td>
<td>7.64</td>
</tr>
<tr>
<td>Google</td>
<td>7.32</td>
<td>78.43</td>
<td>6.73</td>
</tr>
<tr>
<td>SMatxinT\textsubscript{0}</td>
<td>14.50</td>
<td>69.73</td>
<td>8.45</td>
</tr>
<tr>
<td>SMatxinT</td>
<td>14.73</td>
<td>69.18</td>
<td>8.81</td>
</tr>
</tbody>
</table>

SMatxin\textsuperscript{T} results are slightly better than SMatxin\textsubscript{T\textsubscript{0}}
## Source systems

<table>
<thead>
<tr>
<th>System</th>
<th>SMatxinT</th>
<th>sBLEU</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chunks</td>
<td>tokens</td>
<td>chunks</td>
</tr>
<tr>
<td>SMT</td>
<td>2,682 (44.2%)</td>
<td>11,391 (65.4%)</td>
<td>3,202 (38.4%)</td>
</tr>
<tr>
<td>SMT-CD</td>
<td>523 (8.6%)</td>
<td>1,737 (10.0%)</td>
<td>779 (9.3%)</td>
</tr>
<tr>
<td>RBMT</td>
<td>401 (6.6%)</td>
<td>1,279 (7.3%)</td>
<td>969 (11.6%)</td>
</tr>
<tr>
<td>BOTH</td>
<td>2,454 (40.5%)</td>
<td>3,013 (17.3%)</td>
<td>3,389 (40.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>6,060 (100%)</td>
<td>17,420 (100%)</td>
<td>8,339 (100%)</td>
</tr>
</tbody>
</table>
Source systems

<table>
<thead>
<tr>
<th>System</th>
<th>SMatxinT</th>
<th>sBLEU</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chunks</td>
<td>tokens</td>
<td>chunks</td>
</tr>
<tr>
<td>SMT</td>
<td>2,682 (44.2%)</td>
<td>11,391 (65.4%)</td>
<td>3,202 (38.4%)</td>
</tr>
<tr>
<td>SMT-CD</td>
<td>523 (8.6%)</td>
<td>1,737 (10.0%)</td>
<td>779 (9.3%)</td>
</tr>
<tr>
<td>RBMT</td>
<td>401 (6.6%)</td>
<td>1,279 (7.3%)</td>
<td>969 (11.6%)</td>
</tr>
<tr>
<td>BOTH</td>
<td>2,454 (40.5%)</td>
<td>3,013 (17.3%)</td>
<td>3,389 (40.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>6,060 (100%)</td>
<td>17,420 (100%)</td>
<td>8,339 (100%)</td>
</tr>
</tbody>
</table>

Hybrid system use very few RBMT translations
### Source systems

<table>
<thead>
<tr>
<th>System</th>
<th>SMatxinT</th>
<th>sBLEU</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chunks</td>
<td>tokens</td>
<td>chunks</td>
</tr>
<tr>
<td>SMT</td>
<td>2,682 (44.2%)</td>
<td>11,391 (65.4%)</td>
<td>3,202 (38.4%)</td>
</tr>
<tr>
<td>SMT-CD</td>
<td>523 (8.6%)</td>
<td>1,737 (10.0%)</td>
<td>779 (9.3%)</td>
</tr>
<tr>
<td>RBMT</td>
<td>401 (6.6%)</td>
<td>1,279 (7.3%)</td>
<td><strong>969 (11.6%)</strong></td>
</tr>
<tr>
<td>BOTH</td>
<td>2,454 (40.5%)</td>
<td>3,013 (17.3%)</td>
<td>3,389 (40.6%)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6,060 (100%)</strong></td>
<td><strong>17,420 (100%)</strong></td>
<td><strong>8,339 (100%)</strong></td>
</tr>
</tbody>
</table>

Oracle uses more RBMT fragments
### Source systems

<table>
<thead>
<tr>
<th>System</th>
<th>SMatxinT</th>
<th>sBLEU</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chunks</td>
<td>tokens</td>
<td>chunks</td>
</tr>
<tr>
<td>SMT</td>
<td>2,682 (44.2%)</td>
<td>11,391 (65.4%)</td>
<td>3,202 (38.4%)</td>
</tr>
<tr>
<td>SMT-CD</td>
<td>523 (8.6%)</td>
<td>1,737 (10.0%)</td>
<td>779 (9.3%)</td>
</tr>
<tr>
<td>RBMT</td>
<td>401 (6.6%)</td>
<td>1,279 (7.3%)</td>
<td>969 (11.6%)</td>
</tr>
<tr>
<td>BOTH</td>
<td>2,454 (40.5%)</td>
<td>3,013 (17.3%)</td>
<td>3,389 (40.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>6,060 (100%)</td>
<td>17,420 (100%)</td>
<td>8,339 (100%)</td>
</tr>
</tbody>
</table>

Oracle uses more RBMT fragments  
The fragments are in average shorter
Source systems

<table>
<thead>
<tr>
<th>System</th>
<th>SMatxinT</th>
<th>sBLEU</th>
<th>Oracle</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>chunks</td>
<td>tokens</td>
<td>chunks</td>
</tr>
<tr>
<td>SMT</td>
<td>2,682 (44.2%)</td>
<td>11,391 (65.4%)</td>
<td>3,202 (38.4%)</td>
</tr>
<tr>
<td>SMT-CD</td>
<td>523 (8.6%)</td>
<td>1,737 (10.0%)</td>
<td>779 (9.3%)</td>
</tr>
<tr>
<td>RBMT</td>
<td>401 (6.6%)</td>
<td>1,279 (7.3%)</td>
<td>969 (11.6%)</td>
</tr>
<tr>
<td>BOTH</td>
<td>2,454 (40.5%)</td>
<td>3,013 (17.3%)</td>
<td>3,389 (40.6%)</td>
</tr>
<tr>
<td>Total</td>
<td>6,060 (100%)</td>
<td>17,420 (100%)</td>
<td>8,339 (100%)</td>
</tr>
</tbody>
</table>

Linear decoding fails in use RBMT correct information.
Human Evaluation

- 100 random sentences from NEWStest
- Preference between SMT and SMatxinT (2 assessment per sentence)

<table>
<thead>
<tr>
<th>Assessments</th>
<th>SMT</th>
<th>Tied</th>
<th>SMatxinT</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>45 (22.5%)</td>
<td>64 (32.0%)</td>
<td>91 (45.5%)</td>
</tr>
<tr>
<td>Agreement</td>
<td>13 (21.3%)</td>
<td>17 (27.9%)</td>
<td>31 (50.8%)</td>
</tr>
</tbody>
</table>

Results confirm that SMatxinT overcomes the best SMT system.
Human Evaluation

- 100 random sentences from NEWStest
- Preference between SMT and SMatxinT (2 assessment per sentence)

<table>
<thead>
<tr>
<th>Assessments</th>
<th>SMT</th>
<th>Tied</th>
<th>SMatxinT</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>45 (22.5%)</td>
<td>64 (32.0%)</td>
<td>91 (45.5%)</td>
</tr>
<tr>
<td>Agreement</td>
<td>13 (21.3%)</td>
<td>17 (27.9%)</td>
<td>31 (50.8%)</td>
</tr>
</tbody>
</table>

Results confirms that SMatxinT overcomes best SMT system.
Outline

1. SMatxinT: RBMT leaded hybrid system

2. Experimental Results

3. Conclusions
We present a hybrid machine translation system that combines RBMT with phrase-based SMT.

The RBMT system leads the translation process and generates the syntactic structure in the target language.

The SMT system generates multiple candidate translations of any fragment in this tree. A posterior linear decoder selects the best combination to create the final output.

- Short translation alternatives correct RBMT lexical selection errors.
- Longer ones allow to overcome syntactic analysis error.
Conclusions

- SMatxinT achieves statistically significant improvements on out-of-domain test sets for the Spanish-to-Basque translation, according automatic metrics.
- This advantage has been corroborated by a manual evaluation conducted on a set of 100 samples.
- The analysis of the oracles shows that there is still a large room for improvement.
- Oracle translations tend to be composed by more and shorter chunks, and a larger proportion of chunks coming from RBMT.
Future work

- Define new linguistically based features for the linear decoder to identify correct RBMT translations.
- Alleviate the strong dependence of SMatxinT on the initial syntactic parsing incorporating multiple syntactic trees from the side of the rule based system.
- A more detailed manual comparison of the outputs of the different systems
- Broadening the study to other language pairs.
Thanks

Thanks for your attention.
Hybrid Machine Translation Guided by a Rule–Based System

Cristina España-Bonet, Gorka Labaka, Lluís Màrquez, Arantza Díaz de Ilarraza, Kepa Sarasola

Universitat Politècnica de Catalunya, University of the Basque Country