EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language.
Its use in SMT-RBMT-EBMT hybridization

PhD. Candidate: Gorka Labaka Intxauspe
Supervisors: Arantza Díaz de Ilarraza Sánchez
             Kepa Sarasola Gabiola

Lengoaia eta Sistema Informatikoak/Lenguajes y Sistemas Informáticos
Euskar Herriko Unibertsitatea/Universidad del País Vasco

March 29, 2010
Basque Language

- Basque is a pre-Indo-European language [Trask, 1997] with no demonstrable genealogical relationship with other languages.
- There have been many unsuccessful attempts to relate Basque to other languages (Caucasian, Iberian, Berber).
- Most of the features present in Basque (agglutinative, ergative case system) are not unique, but their combination makes Basque a real challenge for Human Language Technologies (HLT).
Sociological Status

- There are few fluent speakers of Basque.
- Basque speakers are distributed between Spain and France and it is in a diglossic situation in all its territories.
- There are not many linguistic resources for Basque:
  - Few corpora, both parallel and monolingual.
  - Syntactic and semantic processors are still on development.
  - But high quality morphological processors (analyzer and generator).
- **This mentioned lack of resources makes the application of HLT and Machine Translation even harder.**
Due to the co-official language status of Basque in some Spanish regions, many administrative texts have to be translated.

Spanish-to-Basque translation is a real need.

The Ixa group has already developed a Rule-Based Machine Translation system [Mayor, 2007], and attempts on EBMT have been also done [Alegria et al., 2008b].

During the last years some SMT attempts have been developed by different authors [Sanchís and Casacuberta, 2007], [Pérez et al., 2008]. Most of them based on Stochastic finite-state transducers and synthetic corpora.

Other RBMT systems: Erderatu [Ginestí-Rosell et al., 2009] or the system available in the website of the Instituto Cervantes (http://oesi.cervantes.es/traduccionAutomatica.html).
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
   - [Agirre et al., 2006] - SEPLN 2006
   - [Labaka et al., 2007] - MT Summit 2007
   - [Labaka et al., 2008] - JTH 2008
   - [Díaz de Ilarraza et al., 2009] - EAMT 2008
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
2. To implement different techniques to deal with word order differences in SMT
   - [Díaz de Ilarraza et al., 2009] - MT Summit 2009
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
2. To implement different techniques to deal with word order differences in SMT
3. To combine by means of a Multi-Engine system SMT with previously developed RBMT and EBMT systems
   - [Alegria et al., 2008a] - MATMT 2008
   - [Alegria et al., 2008b] - AMTA 2008
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
2. To implement different techniques to deal with word order differences in SMT
3. To combine by means of a Multi-Engine system SMT with previously developed RBMT and EBMT systems
   - [Díaz de Ilarraza et al., 2008] - MATMT 2008
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
2. To implement different techniques to deal with word order differences in SMT
3. To combine by means of a Multi-Engine system SMT with previously developed RBMT and EBMT systems
5. To collect larger bilingual corpora and measure the impact of the size and nature of the corpora on the different techniques developed.
Objectives of this PhD thesis

Adaptation of SMT to Basque & First Hybridization Attempts

1. To deal with the agglutinative nature of Basque
2. To implement different techniques to deal with word order differences in SMT
3. To combine by means of a Multi-Engine system SMT with previously developed RBMT and EBMT systems
5. To collect larger bilingual corpora and measure the impact of the size and nature of the corpora on the different techniques developed.
6. To carry out a final evaluation of the work done in this thesis.
Outline

1. General experimental setup
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts
1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation
Outline

1. General experimental setup
2. Treatment of the morphological divergence between Spanish and Basque
3. Treatment of the syntactic divergence between Spanish and Basque
4. Hybridization attempts
5. Overall evaluation
6. Contributions and Further Work
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation

6. Contributions and Further Work
Statistical Machine Translation

- We develop our systems using freely available tools (Moses, GIZA and SRILM)
- We use the same feature combination in all our experiments:
  - phrase translation probabilities (in both directions)
  - word-based translation probabilities (in both directions)
  - a phrase length penalty
  - a 4-gram target language model
  - lexicalized reordering (except on those cases where we specifically deactivate it)
Parallel corpus for Basque: Consumer

<table>
<thead>
<tr>
<th></th>
<th>sentence</th>
<th>tokens</th>
<th>vocabulary</th>
<th>singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>training</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>58,202</td>
<td>1,284,089</td>
<td>46,636</td>
<td>19,256</td>
</tr>
<tr>
<td>Basque</td>
<td></td>
<td>1,010,545</td>
<td>87,763</td>
<td>46,929</td>
</tr>
<tr>
<td>development</td>
<td>1,456</td>
<td>32,740</td>
<td>7,074</td>
<td>4,351</td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td>25,778</td>
<td>9,030</td>
<td>6,339</td>
</tr>
<tr>
<td>Basque</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>test</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>1,446</td>
<td>31,002</td>
<td>6,838</td>
<td>4,281</td>
</tr>
<tr>
<td>Basque</td>
<td></td>
<td>24,372</td>
<td>8,695</td>
<td>6,077</td>
</tr>
</tbody>
</table>

Table: Some statistics of the corpus (Eroski Consumer).

- It is a collection of 1036 articles written in Spanish Consumer Eroski magazine, along with their Basque, Catalan and Galician translations.
- It contains more than 1,200,000 Spanish words and more than 1,000,000 Basque words.
- It was automatically aligned at sentence level [Alcázar, 2005].
- We have divided this corpus into three sets: training, development and test.
In order to assess the quality of the systems developed in this thesis, we used metrics that compare the translation with human references.

Accuracy metrics based on n-grams (higher values imply higher translation quality):
- BLEU [Papineni et al., 2002]
- NIST [Doddington, 2002]

Error metrics (lower values imply higher translation quality).
- Word Error Rate (WER) [Nießen et al., 2000]
- Position-independent word Error Rate (PER) [Tillmann et al., 1997]

Statistical Significance test by means of Paired Bootstrap Resampling [Koehn, 2004].
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque
   - Use of segmentation to adapt SMT to Basque
   - Different segmentation options
   - Experimental Results

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation

6. Contributions and Further Work
Morphological divergences between Spanish and Basque

- Basque is agglutinative: words are formed by joining several morphemes together:
  - Each postpositional case has four different variants.
  - For a lemma more than 360 forms are possible.
  - In the case of ellipsis more than one suffix can be added to the same lemma, increasing the word forms that can be generated from a lemma.
- Postpositions are added to the last word of each phrase.
### Basque morphological generation

<table>
<thead>
<tr>
<th>Basque</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>etxe</td>
<td>/house/</td>
</tr>
<tr>
<td>etxeoa</td>
<td>/the house/</td>
</tr>
<tr>
<td>etxeak</td>
<td>/the houses/</td>
</tr>
<tr>
<td>etxeok</td>
<td>/these houses/</td>
</tr>
<tr>
<td>etxeetara</td>
<td>/to [any] house/</td>
</tr>
<tr>
<td>etxera</td>
<td>/to the house/</td>
</tr>
<tr>
<td>etxeetara</td>
<td>/to the houses/</td>
</tr>
<tr>
<td>etxeotara</td>
<td>/to these houses/</td>
</tr>
<tr>
<td>etxeko</td>
<td>/of the house/</td>
</tr>
<tr>
<td>etxekoa</td>
<td>/the one of the house/</td>
</tr>
<tr>
<td>etxekoak</td>
<td>/the ones of the house/</td>
</tr>
<tr>
<td>etxeetako</td>
<td>/of the houses/</td>
</tr>
<tr>
<td>etxeetakoa</td>
<td>/the one of the houses/</td>
</tr>
<tr>
<td>etxeetakoak</td>
<td>/the ones of the houses/</td>
</tr>
<tr>
<td>etxeotako</td>
<td>/of these houses/</td>
</tr>
<tr>
<td>etxeotakoa</td>
<td>/the one of these houses/</td>
</tr>
<tr>
<td>etxeotakoak</td>
<td>/the ones of these houses/</td>
</tr>
</tbody>
</table>

**Figure:** Illustration of the Basque inflectional morphology.
Effect of morphology in the translation

- Sparseness (each Basque word appears few times in the corpus).
- Being Basque less-resourced, the sparseness problem is intensified.
- The agglutinative nature of Basque causes many 1:n alignments. Those alignments, even being allowed in the IBM models, harm the alignment quality.

<table>
<thead>
<tr>
<th></th>
<th>tokens</th>
<th>vocabulary</th>
<th>singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spanish</td>
<td>1,284,089</td>
<td>46,636</td>
<td>19,256</td>
</tr>
<tr>
<td>Basque</td>
<td>1,010,545</td>
<td>87,763</td>
<td>46,929</td>
</tr>
</tbody>
</table>

**Table:** Figures on the Consumer training corpus.
Different approaches for other highly inflected languages

- **Segmentation.** Words of the highly inflected languages are divided into several tokens [Goldwater and McClosky, 2005], [Oflazer and El-Kahlout, 2007], [Ramanathan et al., 2008].

- **Factored models.** Each word is tagged at different linguistic levels. Each level can be translated independently [Koehn and Hoang, 2007], [Bojar, 2007].

- **Morphology generation model.** The translation is carried out into target lemmas, and, then, their inflection is decided in a separated generation step [Minkov et al., 2007], [Toutanova et al., 2008], [Pérez et al., 2008].
Selected approach: Morphological segmentation

- Taking into account the work done for other highly inflected languages, we have chosen segmentation in order to adapt SMT to Basque.
  - High-precision morphological analyzer and generator are available for Basque.
  - The use of segmentation allows the generation of unseen words, unlike the factored model and the morphology generation model.
  - Complex translation steps make factored translation computationally unmanageable.
  - The biggest gains using factored models come from the incorporation of language models on different factors (lemmas, PoS or morphological information). This can also be combined with the segmentation.
Use of segmentation to adapt SMT to Basque

- Basque text is segmented before training, dividing each word into a set of tokens.
Use of segmentation to adapt SMT to Basque

- Basque text is segmented before training, dividing each word into a set of tokens.
- An SMT system is trained over the segmented text.
Use of segmentation to adapt SMT to Basque

- Basque text is segmented before training, dividing each word into a set of tokens.
- An SMT system is trained over the segmented text.
- After translation, the final Basque word has to be generated. At generation, Basque morpho-phonologic rules have to be taken into account.
Use of segmentation to adapt SMT to Basque

- Basque text is segmented before training, dividing each word into a set of tokens.
- An SMT system is trained over the segmented text.
- After translation, the final Basque word has to be generated. At generation, Basque morpho-phonologic rules have to be taken into account.
- No word-level language model is used at decoding. It is incorporated by means of n-best lists.
Eustagger segmentation

- We based our segmentation of the analysis obtained by the Eustagger Basque lemmatizer [Aduriz and Díaz de Ilarraza, 2003].
- Straightforward segmentation, creating a new token for each morpheme recognized by Eustagger.
- We compare the performance of this segmentation with a baseline (out-of-the-box Moses trained on the tokenized corpus).
Eustagger segmentation

- We based our segmentation of the analysis obtained by the Eustagger Basque lemmatizer [Aduriz and Díaz de Ilarraza, 2003].
- Straightforward segmentation, creating a new token for each morpheme recognized by Eustagger.
- We compare the performance of this segmentation with a baseline (out-of-the-box Moses trained on the tokenized corpus).
- Automatic evaluation metrics did not show significant improvement. Worst BLEU scores, slightly better for the rest of the metrics.

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>10.78</td>
<td>4.52</td>
<td>80.46</td>
<td>61.34</td>
</tr>
<tr>
<td>Eustagger segm.</td>
<td>10.52</td>
<td>4.55</td>
<td>79.18</td>
<td>61.03</td>
</tr>
</tbody>
</table>

Table: Evaluation of SMT systems.
Different segmentation options

- The lexicon of the Eustagger analyzer is too fine-grained.
- It defines morphemes according to the linguistic theories.
- This fine-grained morpheme definition does not agree with the functional usage.
- We conclude that, in case of using the segmentation, it is very important the way that the segmentation is carried out.
Different segmentation options

- We look for the best segmentation based on the analysis obtained by Eustagger.
- We define different ways to group the morphemes, giving rise to different segmentation options:
  1. **OneSuffix**: Groups all suffixes in a unique token.
  2. **AutoGrouping**: Groups those morpheme pairs scored over a threshold according to Pairwise Mutual Information.
  3. **ManualGrouping**: Morphemes are grouped according to hand-defined heuristics.
We look for the best segmentation based on the analysis obtained by Eustagger.

We define different ways to group the morphemes, giving rise to different segmentation options:

1. **OneSuffix**: Groups all suffixes in a unique token.
2. **AutoGrouping**: Groups those morpheme pairs scored over a threshold according to Pairwise Mutual Information.
3. **ManualGrouping**: Morphemes are grouped according to hand-defined heuristics.

Original word: `aukeratzerakoan` *when choosing*

Analysis: `aukeratu+<adize>+<ala>+<gel>+<ine>`
`aukeratu+tze` `+ra` `+ko` `+an`

Eustagger segm.: `aukeratu` `+<adize>` `+<ala>` `+<gel>` `+<ine>`
We look for the best segmentation based on the analysis obtained by Eustagger.

We define different ways to group the morphemes, giving rise to different segmentation options:

1. **OneSuffix**: Groups all suffixes in a unique token.
2. **AutoGrouping**: Groups those morpheme pairs scored over a threshold according to Pairwise Mutual Information.
3. **ManualGrouping**: Morphemes are grouped according to hand-defined heuristics.

Original word: aukeratzerakoan /when choosing/

Analysis: aukeratu+<adize>+<ala>+<gel>+<ine>
aukeratu+tze +ra +ko +an

OneSuffix: aukeratu +<adize>+<ala>+<gel>+<ine>
Different segmentation options

- We look for the best segmentation based on the analysis obtained by Eustagger.
- We define different ways to group the morphemes, giving rise to different segmentation options:
  1. **OneSuffix**: Groups all suffixes in a unique token.
  2. **AutoGrouping**: Groups those morpheme pairs scored over a threshold according to Pairwise Mutual Information.
  3. **ManualGrouping**: Morphemes are grouped according to hand-defined heuristics.

Original word: aukeratzerakoan /when choosing/

Analysis: aukeratu+<adize>+<ala>+<gel>+<ine>
aukeratu+tze +ra +ko +an

AutoGrouping: aukeratu +<adize>+<ala> +<gel> +<ine>
Different segmentation options

- We look for the best segmentation based on the analysis obtained by Eustagger.
- We define different ways to group the morphemes, giving rise to different segmentation options:
  1. **OneSuffix**: Groups all suffixes in a unique token.
  2. **AutoGrouping**: Groups those morpheme pairs scored over a threshold according to Pairwise Mutual Information.
  3. **ManualGrouping**: Morphemes are grouped according to hand-defined heuristics.

Original word: aukeratzerakoan /when choosing/

Analysis: aukeratu+<adize>+<ala>+<gel>+<ine>
          aukeratu+tze     +ra     +ko     +an

ManualGrouping: aukeratu+<adize>     +<ala>+<gel>+<ine>
Experimental results: Different segmentations

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>10.78</td>
<td>4.52</td>
<td>80.46</td>
<td>61.34</td>
</tr>
<tr>
<td>Eustagger segm.</td>
<td>10.52</td>
<td>4.55</td>
<td>79.18</td>
<td>61.03</td>
</tr>
<tr>
<td>OneSuffix segm.</td>
<td>11.24</td>
<td>4.74</td>
<td>78.07</td>
<td>59.35</td>
</tr>
<tr>
<td>AutoGrouping segm.</td>
<td>11.24</td>
<td>4.66</td>
<td>79.15</td>
<td>60.42</td>
</tr>
<tr>
<td>ManualGrouping segm.</td>
<td>11.36</td>
<td>4.67</td>
<td>78.92</td>
<td>60.23</td>
</tr>
</tbody>
</table>

Table: Evaluation of SMT systems with five different segmentation options.

- All the segmentations that group morphemes outperform both the baseline and the Eustagger segmentation.
- There are not big differences between grouping techniques, but according to BLEU the **improvement of the ManualGrouping segmentation is statistically significant over the others.**
**Experimental results: Vocabulary size vs. BLEU score**

<table>
<thead>
<tr>
<th>Segmentation option</th>
<th>Running tokens</th>
<th>Vocabulary size</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tokenized Spanish</td>
<td>1,284,089</td>
<td>46,636</td>
<td>-</td>
</tr>
<tr>
<td>Tokenized Basque</td>
<td>1,010,545</td>
<td>87,763</td>
<td>10.78</td>
</tr>
<tr>
<td>Eustagger segm.</td>
<td>1,699,988</td>
<td>35,316</td>
<td>10.52</td>
</tr>
<tr>
<td>AutoGrouping segm.</td>
<td>1,580,551</td>
<td>35,549</td>
<td>11.24</td>
</tr>
<tr>
<td>OneSuffix segm.</td>
<td>1,558,927</td>
<td>36,122</td>
<td>11.24</td>
</tr>
<tr>
<td>ManualGrouping segm.</td>
<td>1,546,304</td>
<td>40,288</td>
<td>11.36</td>
</tr>
</tbody>
</table>

Table: Correlation between token number in the training corpus and BLEU evaluation results

- There seems to be a correlation between the size of the vocabulary generated after segmentation and the BLEU score:
  - The closer the size of the vocabularies the bigger the obtained BLEU score.
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque
   - Moses’ Lexicalized Reordering
   - Syntax-Based Reordering
   - Statistical Reordering
   - Experimental Results

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation

6. Contributions and Further Work
Syntactic divergences between Spanish and Basque.

- The order of sentence constituents is very flexible, and mainly depends on the focus.
- Basque mainly follows the SOV sentence order.
- Spanish prepositions have to be translated into Basque postpositions (at the end of the phrase).
- Postpositional phrases attached to nouns are placed before nouns (instead of following them).
Effect of those divergences in the translation.

- SMT systems mainly follow a distance-based distortion method (both in word alignment and decoding).
- This method favour short-distance reordering, strongly penalize long-distance reordering.
- Spanish-to-Basque translation needs a high amount of long-distance reordering, and, as we will see, distance-based reordering produces worse translations.
Different approaches used in the literature

- **Lexicalized reordering**: reordering method integrated in Moses [Koehn et al., 2007].
- Methods based on pre-processing: they modify word order in source language to harmonize it with the target language’s word order.
  - **Syntax-based**: based on source syntactic analysis and hand-defined reordering rules [Collins et al., 2005], [Popović and Ney, 2006], [Ramanathan et al., 2008].
  - **Statistical reordering**: based on word alignments and pure statistical information [Chen et al., 2006, Zhang et al., 2007, Sanchís and Casacuberta, 2007, Costa-Jussà and Fonollosa, 2006].
Moses’ Lexicalized Reordering

- Reordering method implemented in Moses [Koehn et al., 2007].
- It adds new features to the log-linear framework.
- The orientation of each phrase occurrence is extracted at training, and their probability distribution is estimated.
- Those probability distributions are used to score each translation hypothesis at decoding.
Moses’ Lexicalized Reordering: Possible Orientations

Three different orientations are defined:

- **monotone**: continuous phrases occur in the same order in both languages. There is an alignment point to the top left.
- **swap**: continuous phrases are swapped in the target language. There is an alignment point to the top right.
- **discontinuous**: continuous phrases in the source language are not continuous in the target language. No alignment points to the top left or the top right.

**Figure**: Possible orientations of phrases defined on the lexicalized reordering
Moses’ Lexicalized Reordering: Training Example
Moses’ Lexicalized Reordering: Training Example

/el/precio
/no/influye
/en/lacalidad
de/el/agua
/que/se/consume

/prize/

/precio

/prezio

mon. swap disc.
0.01 0.79 0.20
Moses’ Lexicalized Reordering: Training Example

/el precio no influye en la calidad de el agua que se consume/

/prize/
/prize/

/does not influence/
/does not influence/

/prezio
/prezio

ez du eragin +nik

/world
/world

/no influye
/no influye

/0.01
/0.01

/swap
/swap

/0.79
/0.79

/disc.
disc.

/0.20
/0.20

/0.60
/0.60
Moses’ Lexicalized Reordering: Training Example

<table>
<thead>
<tr>
<th>/prize/</th>
<th>precio</th>
<th>prezio</th>
<th>mon.</th>
<th>swap</th>
<th>disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>/does not influence/</td>
<td>no influye</td>
<td>ez du eragin +nik</td>
<td>0.20</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>/influence/</td>
<td>influye</td>
<td>du eragin +nik</td>
<td><strong>0.60</strong></td>
<td>0.20</td>
<td>0.20</td>
</tr>
</tbody>
</table>
Moses’ Lexicalized Reordering: Training Example

```
/el/    precio
/no/    no
/influye/ influye
/en/     en
/la/     la
calidad/ de
/el/     el
/agua/   agua
/que/    que
/se/     se
/consume/ consume

/prize/    precio    prezzo
/does not influence/    no influye    ez du eragin +nik
/influence/    influye    du eragin +nik
/the price/    el precio    prezzo +ak
```

---

<table>
<thead>
<tr>
<th>mon.</th>
<th>swap</th>
<th>disc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>0.79</td>
<td>0.20</td>
</tr>
<tr>
<td>0.20</td>
<td>0.20</td>
<td>0.60</td>
</tr>
<tr>
<td>0.60</td>
<td>0.20</td>
<td>0.20</td>
</tr>
<tr>
<td>0.17</td>
<td>0.43</td>
<td>0.40</td>
</tr>
</tbody>
</table>
Moses’ Lexicalized Reordering: Training Example

| Lexicalized Reordering | --- | --- | --- |
| /prize/ | precio | prezio | 0.01 | 0.79 | 0.20 |
| /does not influence/ | no influye | ez du eragin +nik | 0.20 | 0.20 | 0.60 |
| /influence/ | influye | du eragin +nik | 0.60 | 0.20 | 0.20 |
| /the price/ | el precio | prezio +ak | 0.17 | 0.43 | 0.40 |
| /not/ | no | ez | 0.30 | 0.10 | 0.60 |
| /does not influence in the/ | no influye en la | +an ez du eraginik | 0.08 | 0.79 | 0.13 |
| /in the/ | en la | +an | 0.01 | 0.83 | 0.16 |
| /in the quality/ | en la calidad | kalitate +an | 0.04 | 0.56 | 0.40 |
| /in the quality of the/ | en la calidad de el | +ren kalitate +an | 0.14 | 0.71 | 0.15 |
| /quality of the water/ | calidad de el agua | ura +ren kalitate | 0.01 | 0.31 | 0.68 |
| /quality of the water that/ | calidad de el agua que | +n ura +ren kalitate | 0.03 | 0.86 | 0.11 |
This method tries to reorder the source sentence before SMT translation, harmonizing the source word order to the target one.

To reorder the source, we defined a set of rules that make use of syntactic analysis.

Those rules have been defined to deal with the most important word order differences between both languages.

They are divided into two sets: local reordering and long-range reordering.
Syntax-Based Reordering: Local Reordering

- Deals with word order differences in phrases (Spanish noun and prepositional phrases).
- Uses Freeling [Carreras et al., 2004] to mark each word’s PoS and phrase boundaries.
- Moves Spanish prepositions and articles to the end of the phrase, where Basque postpositions appear.

```plaintext
/price/ /no/ /has-influence/ /on/ /quality/ /of/ /water/ /that/ /is/ /consumed/
El precio no influye en la calidad de el agua que se consume
```
Syntax-Based Reordering: Local Reordering

- Deals with word order differences in phrases (Spanish noun and prepositional phrases).
- Uses Freeling [Carreras et al., 2004] to mark each word’s PoS and phrase boundaries.
- Moves Spanish prepositions and articles to the end of the phrase, where Basque postpositions appear.

```
/price/ /no/ /has-influence/ /on/ /quality/ /of/ /water/ /that/ /is/ /consumed/
El precio no influye en la calidad de el agua que se consume
precio El no influye calidad la en agua el de que se consume
```
Syntax-Based Reordering: Long-range Reordering

- Based on the dependency tree of the source.
- Manually-defined rules move entire subtrees along the sentence.
- Allows longer reorderings which are the ones that most severely affect the translation.
Syntax-Based Reordering: Long-range Reordering

- We have defined four reordering rules which deal with the most important word order differences.
Syntax-Based Reordering: Long-range Reordering

We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.
Syntax-Based Reordering: Long-range Reordering

We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.
(b) In negative sentences, the particle ‘no’ is moved together with the verb to the end of the clause.
We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.
(b) In negative sentences, the particle ’no’ is moved together with the verb to the end of the clause.
(c) Prepositional phrases and subordinate relative clauses which are attached to nouns are placed at the beginning of the whole noun phrase where they are included.
Syntax-Based Reordering: Long-range Reordering

We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.

(b) In negative sentences, the particle 'no' is moved together with the verb to the end of the clause.

(c) Prepositional phrases and subordinate relative clauses which are attached to nouns are placed at the beginning of the whole noun phrase where they are included.
We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.
(b) In negative sentences, the particle 'no' is moved together with the verb to the end of the clause.
(c) Prepositional phrases and subordinate relative clauses which are attached to nouns are placed at the beginning of the whole noun phrase where they are included.
(d) Conjunctions and relative pronouns placed at the beginning of Spanish subordinate (or relative) clauses are moved to the end of the clause, after the subordinate verb.
Syntax-Based Reordering: Long-range Reordering

We have defined four reordering rules which deal with the most important word order differences.

(a) The verb is moved to the end of the clause, after all its modifiers.
(b) In negative sentences, the particle 'no' is moved together with the verb to the end of the clause.
(c) Prepositional phrases and subordinate relative clauses which are attached to nouns are placed at the beginning of the whole noun phrase where they are included.
(d) Conjunctions and relative pronouns placed at the beginning of Spanish subordinate (or relative) clauses are moved to the end of the clause, after the subordinate verb.
Statistical Reordering

- As syntax-based reordering, this method tries to reorder the source sentence before the SMT translation, harmonizing the source word order to the target one.
- It does not use any kind of syntactic information, it relies on pure statistical information.
- Translation process is divided in two steps, each of those steps is carried out by an SMT system:
  1. The first system is trained to reorder source words, without any kind of lexical transference.
  2. The second one carries out the lexical transference, as well as minor order movements.
1. Align source and target training corpora in both directions and combine word alignments to obtain many-to-many word alignments.

2. Modify the many-to-many word alignments to many-to-one (keeping for each source word only the alignment with a higher IBM-1 probability)

3. Reorder source words in order to obtain a monotonous alignment.

4. Train a state-of-the-art SMT system to translate from original source sentences into the reordered source

5. A second SMT system is necessary to carry out the lexical transference.
Experimental Results: Reordering techniques

- All the systems use the best segmentation option (*ManualGrouping*).
- In order to measure the impact of each reordering technique, we train and evaluate six different systems.
  - **Baseline**: a simplification of the system called *ManualGrouping* in segmentation experiments (deactivating the Moses' lexicalized reordering).
  - Individual techniques: **lexicalized reordering** (*ManualGrouping* in previous experiment), **syntax-based reordering** and **statistical reordering**.
  - Combination of methods: **Statistical+Lexicalized** and **Syntax-based+Lexicalized**.
Experimental Results: Reordering techniques

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ManualGrouping w/o Lexicalized reord.)</td>
<td>10.37</td>
<td>4.54</td>
<td>79.47</td>
<td>60.59</td>
</tr>
<tr>
<td>Lexicalized reord. (ManualGrouping)</td>
<td>11.36</td>
<td>4.67</td>
<td>78.92</td>
<td>60.23</td>
</tr>
<tr>
<td>Syntax-based reord.</td>
<td>11.03</td>
<td>4.60</td>
<td>78.79</td>
<td>61.35</td>
</tr>
<tr>
<td>Statistical reord.</td>
<td>11.13</td>
<td>4.69</td>
<td>78.21</td>
<td>59.66</td>
</tr>
<tr>
<td>Statistical+Lexicalized reord.</td>
<td>11.12</td>
<td>4.66</td>
<td>78.69</td>
<td>60.19</td>
</tr>
<tr>
<td>Syntax-based+Lexicalized reord.</td>
<td>11.51</td>
<td>4.69</td>
<td>77.94</td>
<td>60.45</td>
</tr>
</tbody>
</table>

Table: BLEU, NIST, WER and PER evaluation metrics.

- All individual reordering techniques outperform the baseline.
- Best results are obtained by the lexicalized reordering.
- System combinations have different behaviours.
- Syntax-based+Lexicalized combination statistically significantly outperforms the all single systems.
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts
   Multi-Engine Combination
   Statistical Post-Editiion
   Experimental Results

5. Overall evaluation

6. Contributions and Further Work
Hybridization

- After the development of a SMT system to translate from Spanish to Basque.
- Improve the translation by system combination:
  - SMT (this PhD thesis)
  - RBMT and EBMT (previously developed in Ixa)
- We experimented with two combination approaches:
  - Multi-Engine combination.
Multi-Engine combination

- We translate each sentence using the three engines.
- We select one of the possible translations, dealing with the following facts:
  - Precision of the EBMT approach is very high, but its coverage is low.
  - The SMT engine provides us a confidence score.
  - N-gram based techniques penalize the RBMT systems, although its translations are more adequate for human post-edition [Labaka et al., 2007]
- We use a simple hierarchical selection criterion:
  - If the EBMT engine covers the sentence, we choose its translation.
  - We only choose the SMT translation if its confidence score was higher than a threshold, defined on the development text set.
  - Otherwise, we choose the output from the RBMT engine.
General architecture of the Statistical Post-Edition

It uses an SMT system to learn to post-edit the output of a RBMT system.

We do not have a real corpus of post-edited texts.

We create a synthetic post-edition corpus from a parallel corpus.

EUSMT: SMT for a Morphologically Rich Language

Gorka Labaka Intxauspe
It uses an SMT system to learn to post-edit the output of a RBMT system.
It uses an SMT system to learn to post-edit the output of a RBMT system.

We do not have a real corpus of post-edited texts.
It uses an SMT system to learn to post-edit the output of a RBMT system.

We do not have a real corpus of post-edited texts.

We create a synthetic post-edition corpus from a parallel corpus.
Experimental Results

**Experimental Results: General domain (Consumer corpus)**

<table>
<thead>
<tr>
<th>System Type</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-Based (Matxin)</td>
<td>6.87</td>
<td>3.78</td>
<td>81.68</td>
<td>66.06</td>
</tr>
<tr>
<td>SMT-Segmentation+Reorder</td>
<td>11.51</td>
<td>4.69</td>
<td>77.94</td>
<td>60.45</td>
</tr>
<tr>
<td>EBMT system (0%)</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Rule-Based + SPE</td>
<td>10.14</td>
<td>4.57</td>
<td>78.23</td>
<td>60.89</td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>11.16</td>
<td>4.56</td>
<td>79.83</td>
<td>62.31</td>
</tr>
</tbody>
</table>

**Table:** Scores for the automatic metrics for systems trained on the Consumer corpus.

- For a general domain corpus, both hybridization techniques outperform the RBMT system.
- But they do not improve the results obtained by the SMT system.
- The bias of the automatic metrics against RBMT system can penalize the hybrid systems.
- A human evaluation would be necessary.
Experimental Results

Labour Agreement corpus: Specific domain

<table>
<thead>
<tr>
<th>Subset</th>
<th>Lang.</th>
<th>Doc.</th>
<th>Senten.</th>
<th>Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train</td>
<td>Basque</td>
<td>81</td>
<td>51,740</td>
<td>839,393</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>81</td>
<td></td>
<td>585,361</td>
</tr>
<tr>
<td>Development</td>
<td>Basque</td>
<td>5</td>
<td>2,366</td>
<td>41,408</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>5</td>
<td></td>
<td>28,189</td>
</tr>
<tr>
<td>Test</td>
<td>Basque</td>
<td>5</td>
<td>1,945</td>
<td>39,350</td>
</tr>
<tr>
<td></td>
<td>Spanish</td>
<td>5</td>
<td></td>
<td>27,214</td>
</tr>
</tbody>
</table>

Table: Some statistics of the Labour Agreements Corpus

- We rerun the hybridization experiments on a specific domain corpus (Labour Agreement corpus).
- Administrative texts that contain many formal patterns that allow the EBMT system to extract them.
Experimental Results: Specific domain

<table>
<thead>
<tr>
<th>Method</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule-Based (Matxin)</td>
<td>4.27</td>
<td>2.76</td>
<td>89.17</td>
<td>74.18</td>
</tr>
<tr>
<td>SMT-Segmentation+Reorder</td>
<td>12.27</td>
<td>4.63</td>
<td>77.44</td>
<td>58.17</td>
</tr>
<tr>
<td>EBMT system (64.92%)</td>
<td>32.42</td>
<td>5.76</td>
<td>60.02</td>
<td>54.75</td>
</tr>
<tr>
<td>Rule-Based + SPE</td>
<td>17.11</td>
<td>5.01</td>
<td>75.53</td>
<td>57.24</td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>37.24</td>
<td>7.17</td>
<td>56.84</td>
<td>45.27</td>
</tr>
</tbody>
</table>

Table: Evaluation on domain specific corpus.

- Both hybridization techniques entail important improvements.
- Statistical Post-Edit successfully corrects the RBMT output, outperforming the results of the SMT system.
- The higher contribution to the Multi-Engine system comes by the inclusion of EBMT systems.
- The inclusion of the RBMT engine causes a slightly negative effect (1% relative decrease for BLEU).
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation
   - Doubts about BLEU & evaluation alternatives
   - Systems selected to Human-targeted evaluation
   - Automatic Evaluation
   - Human-Targeted evaluation results

6. Contributions and Further Work
Overall Evaluation

- So far, we have evaluated each approach in isolation and by means of automatic metrics.
- But we only have one reference to calculate automatic metrics.
- The scores obtained in this situation could be biased.
- In order to corroborate the results obtained, we have carried out a final evaluation based on human-targeted metrics.
Doubts about BLEU measure

- In recent years many doubts have arisen about the validity of BLEU:
  - It is extremely difficult to interpret what is being expressed in \textit{BLEU} [Melamed et al., 2003]
  - Improving \textit{BLEU} does not guarantee an improvement in the translation quality [Callison-Burch et al., 2006]
  - It does not offer as much correlation with human judgement as was believed [Koehn and Monz, 2006]
- Those problems are intensified since we only have one reference per sentence.
Recent researches have present new metrics that computes the similarity according to linguistic features [Liu and Gildea, 2007], [Albrecht and Hwa, 2007], [Padó et al., 2007], [Giménez and Màrquez, 2008]

Two main reasons have led us to reject the use of metrics based on linguistic similarity:

- The applicability of these deep evaluation techniques are strongly conditioned by the accessibility to the linguistic processors required and their accuracy.
- Just like BLEU does, these metrics compare the automatic translations with human-defined references, and the evaluation is not so precise when we have only one reference.
Human-targeted metrics compare the automatic hypothesis with the closest human post-edited references.

We can use the post-edited references to calculate metrics, such as BLEU, NIST or TER, giving rise to human-targeted metrics such as HBLEU, HNIST or HTER.

HTER metric is particularly interesting, since TER (Translation Error Rate) measures the number of post-editions done by the human translator.
Overall Evaluation: Human-Targeted evaluation

- This method requires human post-edited references, and its high cost prevented us from evaluating many systems using this method.
- We have chosen the 5 systems we consider the most representative ones:
  - Rule-Based (Matxin)
  - SMT baseline
  - SMT systems that use segmentation and reordering
  - Multi-Engine combination
  - Statistical Post-Edition
- In order to evaluate all the systems properly we incorporate two variations:
  - A bigger corpus for training.
  - Matrex instead of Moses.
Training corpora used for the final evaluation

<table>
<thead>
<tr>
<th></th>
<th>tokens</th>
<th>vocabulary</th>
<th>singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial Bilingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>1,284,089</td>
<td>46,636</td>
<td>19,256</td>
</tr>
<tr>
<td>Basque</td>
<td>1,010,545</td>
<td>87,763</td>
<td>46,929</td>
</tr>
<tr>
<td><strong>Initial Monolingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>1,010,545</td>
<td>87,763</td>
<td>46,929</td>
</tr>
<tr>
<td><strong>Final Bilingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spanish</td>
<td>9,167,987</td>
<td>219,472</td>
<td>97,576</td>
</tr>
<tr>
<td>Basque</td>
<td>6,928,907</td>
<td>438,491</td>
<td>236,238</td>
</tr>
<tr>
<td><strong>Final Monolingual</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Basque</td>
<td>27,950,113</td>
<td>1,057,237</td>
<td>580,477</td>
</tr>
</tbody>
</table>

Table: Statistics on the final training corpora.

- 7 times larger bilingual corpus.
- 27 times larger monolingual corpus.
- Heterogeneous corpora that cover different topics and styles:
  - News
  - Administrative texts
  - Popular science texts
  - ...

EUSMT: SMT for a Morphologically Rich Language
Gorka Labaka Intxauspe
Matrex

MaTrEx is a data-driven MT system which combines both EBMT and SMT techniques.

- It aligns linguistic chunks using EBMT techniques and incorporates them into the SMT phrase table.
- The translation is carried out by a phrase-based decoder (Moses).

Figure: General design of the Matrex system [Stroppa and Way, 2006].
## Automatic Evaluation

### Table: Scores for the automatic metrics for systems trained on the Consumer corpus.

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin (RBMT)</td>
<td>6.87</td>
<td>3.78</td>
<td>81.68</td>
<td>66.06</td>
</tr>
<tr>
<td>SMT-baseline</td>
<td>10.78</td>
<td>4.52</td>
<td>80.46</td>
<td>61.34</td>
</tr>
<tr>
<td>SMT-Segmented</td>
<td>11.36</td>
<td>4.67</td>
<td>78.92</td>
<td>60.23</td>
</tr>
<tr>
<td>SMT-Segmented+Reorder</td>
<td>11.51</td>
<td>4.69</td>
<td>77.94</td>
<td>60.45</td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>11.16</td>
<td>4.56</td>
<td>79.83</td>
<td>62.31</td>
</tr>
<tr>
<td>Statistical Post-Edition</td>
<td>10.14</td>
<td>4.57</td>
<td>78.23</td>
<td>60.89</td>
</tr>
</tbody>
</table>

Automatic Evaluation: Reminder of previous evaluation.
Automatic Evaluation

Automatic Evaluation: larger training corpus

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin (RBMT)</td>
<td>6.87 (=)</td>
<td>3.78 (=)</td>
<td>81.68 (=)</td>
<td>66.06 (=)</td>
</tr>
<tr>
<td>SMT-baseline</td>
<td>11.12 (+0.34)</td>
<td>4.71 (+0.19)</td>
<td>78.13 (-2.33)</td>
<td>59.48 (-1.86)</td>
</tr>
<tr>
<td>SMT-Segmented</td>
<td><strong>11.56</strong> (+0.20)</td>
<td><strong>4.83</strong> (+0.16)</td>
<td>77.83 (-1.09)</td>
<td><strong>58.94</strong> (-1.29)</td>
</tr>
<tr>
<td>SMT-Segmented+Reorder</td>
<td>11.19 (-0.32)</td>
<td>4.69 (=)</td>
<td>77.44 (-0.50)</td>
<td>60.09 (-0.36)</td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>11.29 (+0.13)</td>
<td>4.73 (+0.17)</td>
<td><strong>76.99</strong> (-2.84)</td>
<td>59.63 (-2.68)</td>
</tr>
<tr>
<td>Statistical Post-Edition</td>
<td>10.85 (+0.71)</td>
<td>4.67 (+0.10)</td>
<td>77.45 (-0.78)</td>
<td>60.42 (-0.47)</td>
</tr>
</tbody>
</table>

**Table:** Scores for the automatic metrics for all systems trained on the larger training corpus.

- **Increasing the training corpus.**
  - RBMT does not change, since it does not use the corpora for training.
  - All systems improve their scores, except the one we consider the best one (SMT-Segmented+Reorder).
  - The contribution of Syntax-based reordering is questioned.
### Automatic Evaluation: MaTrEx vs. SMT

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin (RBMT)</td>
<td>6.87 (=)</td>
<td>3.78 (=)</td>
<td>81.68 (=)</td>
<td>66.06 (=)</td>
</tr>
<tr>
<td>MaTrEx-baseline</td>
<td>11.23 (+0.11)</td>
<td>4.75 (+0.04)</td>
<td>78.21 (+0.08)</td>
<td>59.66 (+0.18)</td>
</tr>
<tr>
<td>MaTrEx-Segmented</td>
<td>11.71 (+0.15)</td>
<td>4.82 (-0.01)</td>
<td>77.69 (-0.14)</td>
<td>58.99 (+0.04)</td>
</tr>
<tr>
<td>MaTrEx-Segmented+Reorder</td>
<td>11.52 (+0.33)</td>
<td>4.82 (+0.13)</td>
<td>76.35 (-1.09)</td>
<td>58.94 (-1.15)</td>
</tr>
<tr>
<td>Multi-Engine Hybridization</td>
<td>11.29 (=)</td>
<td>4.73 (=)</td>
<td>76.99 (=)</td>
<td>59.63 (=)</td>
</tr>
<tr>
<td>Statistical Post-Edition</td>
<td>10.85 (=)</td>
<td>4.67 (=)</td>
<td>77.45 (=)</td>
<td>60.42 (=)</td>
</tr>
</tbody>
</table>

**Table:** Scores for the automatic metrics for Matrex systems trained on the larger training corpus.

- The incorporation of EBMT phrases to SMT phrase-table consistently improves the results of the three SMT systems.
### Automatic Evaluation: MaTrEx vs. SMT

<table>
<thead>
<tr>
<th>System</th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin (RBMT)*</td>
<td>6.87 (=)</td>
<td>3.78 (=)</td>
<td>81.68 (=)</td>
<td>66.06 (=)</td>
</tr>
<tr>
<td>MaTrEx-baseline*</td>
<td>11.23 (+0.11)</td>
<td>4.75 (+0.04)</td>
<td>78.21 (+0.08)</td>
<td>59.66 (+0.18)</td>
</tr>
<tr>
<td>MaTrEx-Segmented</td>
<td>11.71 (+0.15)</td>
<td>4.82 (-0.01)</td>
<td>77.69 (-0.14)</td>
<td>58.99 (+0.04)</td>
</tr>
<tr>
<td>MaTrEx-Segmented+Reorder*</td>
<td>11.52 (+0.33)</td>
<td>4.82 (+0.13)</td>
<td>76.35 (-1.09)</td>
<td>58.94 (-1.15)</td>
</tr>
<tr>
<td>Multi-Engine Hybridization*</td>
<td>11.29 (=)</td>
<td>4.73 (=)</td>
<td>76.99 (=)</td>
<td>59.63 (=)</td>
</tr>
<tr>
<td>Statistical Post-Edition*</td>
<td>10.85 (=)</td>
<td>4.67 (=)</td>
<td>77.45 (=)</td>
<td>60.42 (=)</td>
</tr>
</tbody>
</table>

**Table:** Scores for the automatic metrics for Matrex systems trained on the larger training corpus.

- The incorporation of EBMT phrases to SMT phrase-table consistently improves the results of the three SMT systems.
- The systems evaluated by means of human-targeted metrics are those marked with a *.

---

**EUSMT: SMT for a Morphologically Rich Language**

Gorka Labaka Intxauspe
Automatic Evaluation: MaTrEx vs. SMT

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin (RBMT)*</td>
<td>6.87 (=)</td>
<td>3.78 (=)</td>
<td>81.68 (=)</td>
<td>66.06 (=)</td>
</tr>
<tr>
<td>MaTrEx-baseline*</td>
<td>11.23 (+0.11)</td>
<td>4.75 (+0.04)</td>
<td>78.21 (+0.08)</td>
<td>59.66 (+0.18)</td>
</tr>
<tr>
<td>MaTrEx-Segmented</td>
<td><strong>11.71 (+0.15)</strong></td>
<td><strong>4.82 (-0.01)</strong></td>
<td><strong>77.69 (-0.14)</strong></td>
<td><strong>58.99 (+0.04)</strong></td>
</tr>
<tr>
<td>MaTrEx-Segmented+Reorder*</td>
<td>11.52 (+0.33)</td>
<td><strong>4.82 (+0.13)</strong></td>
<td><strong>76.35 (-1.09)</strong></td>
<td>58.94 (-1.15)</td>
</tr>
<tr>
<td>Multi-Engine Hybridization*</td>
<td>11.29 (=)</td>
<td>4.73 (=)</td>
<td>76.99 (=)</td>
<td>59.63 (=)</td>
</tr>
<tr>
<td>Statistical Post-Edition*</td>
<td>10.85 (=)</td>
<td>4.67 (=)</td>
<td>77.45 (=)</td>
<td>60.42 (=)</td>
</tr>
</tbody>
</table>

**Table:** Scores for the automatic metrics for Matrex systems trained on the larger training corpus.

- The incorporation of EBMT phrases to SMT phrase-table consistently improves the results of the three SMT systems.
- The systems evaluated by means of human-targeted metrics are those marked with a *.
- As a consequence of the unexpected behaviour at increasing the training corpus, we have not evaluated the system that gets the highest BLEU score.
Human-Targeted evaluation results

<table>
<thead>
<tr>
<th></th>
<th>HTER</th>
<th>HBLEU</th>
<th>HNIST</th>
<th>HWER</th>
<th>HPER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin</td>
<td>54.74</td>
<td>26.88</td>
<td>6.84</td>
<td>58.51</td>
<td>42.98</td>
</tr>
<tr>
<td>MaTrEx-baseline</td>
<td>53.59</td>
<td>27.86</td>
<td>7.23</td>
<td>58.48</td>
<td>40.23</td>
</tr>
<tr>
<td>MaTrEx-Segmented+Reorder</td>
<td>48.10</td>
<td>33.29</td>
<td>7.60</td>
<td>54.52</td>
<td>35.45</td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>47.62</td>
<td>34.71</td>
<td>7.64</td>
<td>53.74</td>
<td>35.27</td>
</tr>
<tr>
<td>Statistical Post-Edition</td>
<td>47.41</td>
<td>34.80</td>
<td>7.74</td>
<td>52.04</td>
<td>36.05</td>
</tr>
</tbody>
</table>

Table: Scores for the human-targeted metrics for selected systems.

- The Matrex system that uses the improvements proposed in this PhD thesis outperform the Matrex baseline consistently.
- The two hybridization attempts obtain even better results, showing up as an interesting field in which to continue our investigation.
- All the differences between systems are statistically significant except those between Multi-Engine and Statistical Post-edition systems.
Human-Targeted evaluation results vs. BLEU

<table>
<thead>
<tr>
<th></th>
<th>HTER</th>
<th>HBLEU</th>
<th>HNIST</th>
<th>HWER</th>
<th>HPER</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matxin</td>
<td>54.74</td>
<td>26.88</td>
<td>6.84</td>
<td>58.51</td>
<td>42.98</td>
<td>6.87</td>
</tr>
<tr>
<td>MaTrEx-baseline</td>
<td>53.59</td>
<td>27.86</td>
<td>7.23</td>
<td>58.48</td>
<td>40.23</td>
<td>11.23</td>
</tr>
<tr>
<td>MaTrEx-Segmented+Reorder</td>
<td>48.10</td>
<td>33.29</td>
<td>7.60</td>
<td>54.52</td>
<td>35.45</td>
<td><strong>11.52</strong></td>
</tr>
<tr>
<td>Multi-Engine</td>
<td>47.62</td>
<td>34.71</td>
<td>7.64</td>
<td>53.74</td>
<td>35.27</td>
<td>11.29</td>
</tr>
<tr>
<td>Statistical Post-Edition</td>
<td><strong>47.41</strong></td>
<td><strong>34.80</strong></td>
<td><strong>7.74</strong></td>
<td><strong>52.04</strong></td>
<td><strong>36.05</strong></td>
<td>10.85</td>
</tr>
</tbody>
</table>

**Table**: Scores for human-targeted metrics and BLEU.

- The automatic evaluation penalizes the RBMT system and the hybrid systems that use it.
Comparison with other systems

<table>
<thead>
<tr>
<th></th>
<th>BLEU</th>
<th>NIST</th>
<th>WER</th>
<th>PER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UPV-PRHLT</td>
<td>7.11</td>
<td>3.65</td>
<td>82.64</td>
<td>65.56</td>
</tr>
<tr>
<td>Avivavoz</td>
<td>8.12</td>
<td>3.90</td>
<td>81.60</td>
<td>64.22</td>
</tr>
<tr>
<td>EHU-IXA (MaTrEx-Segmented)</td>
<td>8.10</td>
<td>3.98</td>
<td>78.70</td>
<td>62.25</td>
</tr>
</tbody>
</table>

Table: Official results provided by the Albayzin evaluation organizers.

- We obtained the best results in Albayzin evaluation campaign:
  - Our system gets the best results by means of NIST, WER and PER.
  - The difference between our system and the Avivavoz system were not significant regarding BLEU.
- It was the only occasion that we could directly compare our work with other translation systems for Basque.
- The system we presented to the evaluation was the one called *MaTrEx-Segmented* in this thesis.
Outline

1. General experimental setup

2. Treatment of the morphological divergence between Spanish and Basque

3. Treatment of the syntactic divergence between Spanish and Basque

4. Hybridization attempts

5. Overall evaluation

6. Contributions and Further Work
Contributions: SMT to Basque

- Development of a state-of-the-art SMT system for Basque.
- Improvement of that baseline by means of segmentation.
  - Better scores in automatic evaluation for small and large corpora.
  - Definition of a hand-defined heuristic for morpheme-grouping that outperforms automatic segmentations.
- Combination of syntax-based reordering and lexicalized reordering.
  - Statistically significant improvement in 1M words corpus.
  - Those results are not corroborated at enlarging the training corpus.
- The combination of segmentation and syntax-based reordering clearly outperforms the baseline.
  - Statistically significant improvements in human-targeted evaluation.
  - 10% relative improvement in HTER and 16% in HBLEU.
Contributions: System combination

- Development of Multi-Engine and Statistical Post-Edition systems.
  - Both systems considerably outperform single systems in a specialized text like Labour Agreement corpus.
  - For a general domain corpus those gains are not perceived by automatic metrics.
  - But human-targeted evaluation shows statistically significant improvement.
Further work

- Investigate segmentation based on Bootstrapping and Word-Packing [Ma et al., 2007].
- Clarify, by means of human evaluation, the contribution of the syntax-based reordering method.
- Go deeper into Multi-Engine hybridization, creating new translation hypothesis combining phrases from the translation proposed by the different engines.
- Make use of factored machine translation implemented in Moses to integrate bilingual information at Statistical Post-Editing.
- Collect a real post-edition corpus to rerun post-edition experiments.
Thanks for your Attention

Thank you!
Eskerrik asko!
EUSMT: Incorporating Linguistic Information into SMT for a Morphologically Rich Language.
Its use in SMT-RBMT-EBMT hybridization

PhD. Candidate: Gorka Labaka Intxauspe
Supervisors: Arantza Díaz de Ilarraza Sánchez
             Kepa Sarasola Gabiola

Lengoaia eta Sistema Informatikoak/Lenguajes y Sistemas Informáticos
Euskal Herrikoe Unibertsitatea/Universidad del País Vasco

March 29, 2010
7. Bibliography


Bibliography


Trask (1997).  
*The History of Basque.*  