SMT error analysis and mapping to syntactic, semantic and structural fixes

Nora Aranberri

IXA Group
University of the Basque Country
Introduction

Error classification schemes

Our approach

The SMT systems

Error analysis results

Fixing possibilities with SSS

Conclusions
Error analysis for SMT

- Does error analysis make sense for SMT?
Error analysis for SMT

- Does error analysis make sense for SMT?
- Which aspects should it cover?
Introduction

Error classification schemes

Our approach

The SMT systems

Error analysis results

Fixing possibilities with SSS

Conclusions
Dynamic Quality Framework - TAUS

- Dimensions: attributes, grammatical and localization issues
Dynamic Quality Framework - TAUS

- Dimensions: attributes, grammatical and localization issues

**Disadvantages**

- High-level annotation only
- Mixed dimensions
MQM Core: a hierarchy of 22 issues
- Dimensions: quality attributes, grammar/linguistic and edit-types
  - Top level dichotomy: accuracy vs fluency
  - Lower-levels: grammatical/linguistic and edit-type errors
Multidimensional Quality Metric - QTLaunchPad

MQM Core: a hierarchy of 22 issues
- Dimensions: quality attributes, grammar/linguistic and edit-types
  - Top level dichotomy: accuracy vs fluency
  - Lower-levels: grammatical/linguistic and edit-type errors

Disadvantages
- Mixed dimensions
- Issues often too broad to identify SSS solutions
SMT-oriented classification - Vilar et al. 2006

- Dimensions: edit-types and linguistic issues
  - Top level: edit-types
  - Lower-levels: edit-types, spans and grammatical issues
SMT-oriented classification - Vilar et al. 2006

- Dimensions: edit-types and linguistic issues
  - Top level: edit-types
  - Lower-levels: edit-types, spans and grammatical issues

Disadvantages
- Not informative enough linguistically for SSS solutions
Proposed bidimensional error scheme

dynamic, extensible bidimensional scheme

<table>
<thead>
<tr>
<th>Top-level category</th>
<th>Subclasses</th>
<th>Incorrect</th>
<th>Missing</th>
<th>Additional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphosyntax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untranslated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Complementary dimensions: linguistic issues and edit-types
  - Six top linguistic categories
    - Dynamic extensible hierarchy
  - Three edit-type categories
## Proposed bidimensional error scheme

A dynamic, extensible bidimensional scheme is proposed, which categorizes errors based on linguistic issues and edit-types. The top-level categories include:

- **Lexis**
- **Morphosyntax**
- **Verbs**
- **Order**
- **Punctuation**
- **Untranslated**

### Table of Subclasses

<table>
<thead>
<tr>
<th>Top-level category</th>
<th>Subclasses</th>
<th>Incorrect</th>
<th>Missing</th>
<th>Additional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexis</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphosyntax</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbs</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Order</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Punctuation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Untranslated</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Complementary dimensions include linguistic issues and edit-types:

- **Six top linguistic categories**
  - Dynamic extensible hierarchy
- **Three edit-type categories**

### To be considered

Further dimensions (e.g., severity)
Introduction

Error classification schemes

Our approach

The SMT systems

Error analysis results

Fixing possibilities with SSS

Conclusions
### Description of the SMT systems

#### English-Spanish SMT
- **Standard phrase-based Moses system**
- **Training data:**
  - Bilingual: europarl, UN corpus, News Commentary and Common Crawl (~335M words)
  - Monolingual: Spanish part of europarl, News Commentary and Common Crawl (~60M words)
- **In-domain tuning data:** 1,000 QA interactions
- **BLEU score (in-domain):** 45.86

#### English-Basque SMT
- **Phrase-based Moses system**
- **Alignment at lemma-level**
- **Training data:**
  - Bilingual: academic books, software manuals and UI strings, web-crawled data (~13.5M words)
  - Monolingual: Basque part of the above + administrative text (~21M words)
- **In-domain tuning data:** 1,000 QA interactions
- **BLEU score:** 20.24
Error analysis for the English-Spanish pair

137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors

Example of a lexical error

Click *run* where it says vulnerabilities.
Pulse *correr* donde dice vulnerabilidades.
137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%

Example of a morphosyntactic error:

Yes, you can share files and folders with one or more users on MEO Cloud. Sí, puede compartir archivos y carpetas con uno o más usuarios sobre MEO Cloud.
Error analysis for the English-Spanish pair

137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%
- Verbs: 18%

example of a verb error

**Connect** your computer to the ZON HUB via Ethernet cable.
**Conectar** su ordenador a la HUB af a través de cable Ethernet. (infinitive to connect)
Error analysis for the English-Spanish pair

137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%
- Verbs: 18%
- Order: 11%

Example of an ordering error:
Tap Import to copy your Android Browser Favourites.
Toca Importar para copiar su navegador de Android favoritos.
137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%
- Verbs: 18%
- Order: 11%
- Punctuation: 6%

Example of a punctuation error:

If I buy a computer abroad, will it work in Portugal?
Si compro un ordenador en el extranjero, funcionará en Portugal? (missing ').
Error analysis for the English-Spanish pair

137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%
- Verbs: 18%
- Order: 11%
- Punctuation: 6%
- Untranslated: 5%

example of an untranslated unit

Then click on the yellow disc with a green tick.
Then haga clic en el disco de color amarillo con una marca verde.
Error analysis for the English-Spanish pair

137 sentences evaluated with a total of 169 errors.

- Lexis: 31% of the total errors
- Morphosyntax: 29%
- Verbs: 18%
- Order: 11%
- Punctuation: 6%
- Untranslated: 5%
Error analysis for the English-Basque pair

140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors

example of a lexical error

Go to WhatsApp > 'Menu Button' > 'Status'.
Joan menu botoia WhatsApp > " " > " egoera ".

140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors
- Morphosyntax: 39%

Example of a morphosyntactic error:

simply by dragging the profile of the person concerned to the various circles.

besterik gabe, arrastatu pertsonaren profila hainbat nahia zirkulu.
Error analysis for the English-Basque pair

140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors
- Morphosyntax: 39%
- Verbs: 18%

Example of a verb error

Choose a standard status or personalize one.
Egoera estandar bat edo pertsonalizatu bat.
140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors
- Morphosyntax: 39%
- Verbs: 18%
- Order: 14%

example of an ordering error

You can use the app iPP Podcast Player you find on Google Play.
Aplikazioa erabil dezakezu IPP podcast Player aurkitu duzu Google erreproduzitu.
Error analysis for the English-Basque pair

140 sentences evaluated with a total of 393 errors.
- Lexis: 23% of the total errors
- Morphosyntax: 39%
- Verbs: 18%
- Order: 14%
- Punctuation: 4%

Example of a punctuation error

How can I change the language to of Mega to Portuguese?
Nola aldatu hizkuntza of Mega, portugesa?
Error analysis for the English-Basque pair

140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors
- Morphosyntax: 39%
- Verbs: 18%
- Order: 14%
- Punctuation: 4%
- Untranslated: 1%

example of an untranslated unit

How much space do I have for free on Mega?
Zenbat leku ditut doan on Mega?
Error analysis for the English-Basque pair

140 sentences evaluated with a total of 393 errors.

- Lexis: 23% of the total errors
- Morphosyntax: 39%
- Verbs: 18%
- Order: 14%
- Punctuation: 4%
- Untranslated: 1%
Index

1 Introduction
2 Error classification schemes
3 Our approach
4 The SMT systems
5 Error analysis results
6 Fixing possibilities with SSS
7 Conclusions
Main sources for errors

- use-scenario-specific features
  - Q&A chat scenario in the information technology domain
- language pair-specific features
Fixing possibilities with syntax, semantics and structure

IT names not identified (CR)NER
Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>IT names not identified</th>
<th>(CR)NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
</tbody>
</table>
### Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>IT names not identified</th>
<th>(CR)NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>domain-specific terminology</td>
<td>WSD</td>
</tr>
</tbody>
</table>
### Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>IT names not identified</th>
<th>(CR)NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>domain-specific terminology</td>
<td>WSD</td>
</tr>
<tr>
<td>Basque postpositions</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
</tbody>
</table>
Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>Issue</th>
<th>Fixing Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT names not identified</td>
<td>(CR)NER</td>
</tr>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>Domain-specific terminology</td>
<td>WSD</td>
</tr>
<tr>
<td>Basque postpositions</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>Incorrectly ordered phrases</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
</tbody>
</table>
### Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>Issue</th>
<th>Fixation Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT names not identified</td>
<td>(CR)NER</td>
</tr>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>Domain-specific terminology</td>
<td>WSD</td>
</tr>
<tr>
<td>Basque postpositions</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>Incorrectly ordered phrases</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>Split phrases</td>
<td>syntax</td>
</tr>
</tbody>
</table>
Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>IT names not identified</th>
<th>(CR)NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>domain-specific terminology</td>
<td>WSD</td>
</tr>
<tr>
<td>Basque postpositions</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>incorrectly ordered phrases</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>split phrases</td>
<td>syntax</td>
</tr>
<tr>
<td>incorrect phrase-internal order</td>
<td>syntax</td>
</tr>
</tbody>
</table>
Fixing possibilities with syntax, semantics and structure

<table>
<thead>
<tr>
<th>IT names not identified</th>
<th>(CR)NER</th>
</tr>
</thead>
<tbody>
<tr>
<td>UIs not identified</td>
<td>NER</td>
</tr>
<tr>
<td>domain-specific terminology</td>
<td>WSD</td>
</tr>
<tr>
<td>Basque postpositions</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>incorrectly ordered phrases</td>
<td>predicate-argument structures and semantic roles</td>
</tr>
<tr>
<td>split phrases</td>
<td>syntax</td>
</tr>
<tr>
<td>incorrect phrase-internal order</td>
<td>syntax</td>
</tr>
<tr>
<td>incorrectly generated verb phrases</td>
<td></td>
</tr>
</tbody>
</table>
Index

1 Introduction
2 Error classification schemes
3 Our approach
4 The SMT systems
5 Error analysis results
6 Fixing possibilities with SSS
7 Conclusions
Conclusions

- proposed a bidimensional error classification
  - dynamic, extensible linguistically-informed dimension
  - edit-types
- evaluated English-Spanish and English-Basque SMT systems
  - domain-related issues:
    - terminology and UI strings
  - language-pair-specific issues:
    - English prepositions and subordinate markers for Basque
    - verb phrase generation
    - ordering at different levels
- revealed potential of SSS to solve errors
Thanks!
nora.aranberri@ehu.eus