Quantitative Evaluation of Machine Translation Systems: Sentence Level

Palmira Marrafa

António Ribeiro

Outline

- Motivation
- ISO Characteristics to Measure
- Draft Proposals
- Future Work

Motivation

- **Evaluate the Quality of Translated Sentences**
- **Assumption**
  - It is possible to quantify the quality of Translations

- **Design Evaluation Measures**

ISO Characteristics to Measure

2.2 System external characteristics

1 Functionality

2 Accuracy

…

ISO Characteristics to Measure

2 Individual sentence level

1 Morphology

2 Syntax (sentence and phrase structure)

3 Types of errors

3 Lexical errors

4 Syntax errors

5 Stylistic errors

Characteristics to Measure

- **Types of Errors**
  - Lexicon
  - Syntax
  - Morphology

MT Systems Used

- **Systems**
  - S1
  - S2

Lexicon

- « «Lexical» errors refer words or phrases that are inappropriate … »

- Example
  - Input: Foot-and-mouth disease
  - Output: Febre aftosa
  - Systran: *Doenca de pé-e-boca
  - Intertran: *Pé-e-boca doenca

Lexicon: Evaluation

- Collocations
- Fixed Expressions
- Semi-fixed Expressions

Lexicon: Evaluation

- 4-point scale
  - 0 wrong (Doenca de pé-e-boca)
  - 1 marginal (Febre dos pés e da boca)
  - 2 inadequate (Doenca aftosa)
  - 3 correct (Febre aftosa)

Lexicon: Evaluation

■ Scores

$$\sum_{i=1}^{n} \frac{score(word_i)}{n}, n = \#Words$$

– Average of individual word scores
– Order to be included
  ■ Foot-and-mouth disease vs
  ■ Disease of foot-and-mouth

Syntax

- ««Syntax» is concerned with grammaticality»
  - Previous work:
    - DARPA: 5-point scale

Syntax: Evaluation

- **NPs**
  - Specifiers
    - todos os estudantes
    - all *the students
  - Modifiers
  - Co-occurrence Restrictions
  - Order
  - …

Syntax: Evaluation

- **VPs**
  - Tense
  - Mode
  - Aspect
  - Modality (should vs might)
  - Negation
  - Complements order
  - Preposition Selection
  - ...

Future Work

- Generalise across the Categories
  - to build the metrics
- Assign scores to each error
- Test the metrics