Projects: Analysing Machine Translation Output and Compound Splitting

Liane Guillou and Alexander Fraser
{liane,fraser}@cis.uni-muenchen.de

CIS, Ludwig-Maximilians-Universität München

Morphology
2016
06.06.2016

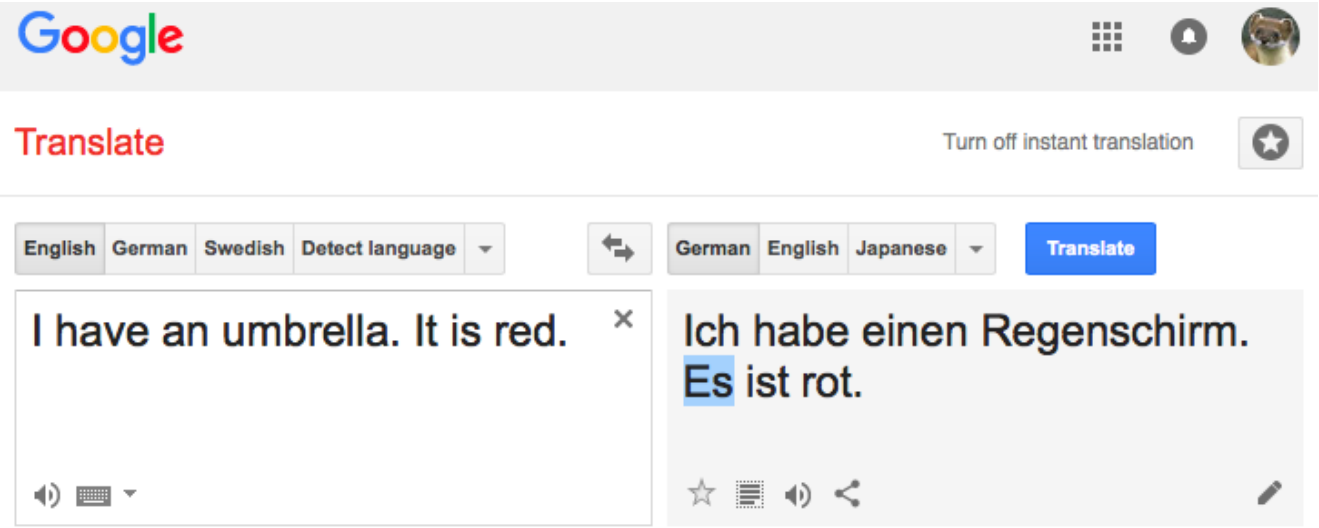
Outline

Outline

Analysing Machine Translation Output

The Problem

- Machine translation (e.g. Google Translate) is far from perfect
- For example in English → German translation
  - Incorrect verb inflections
  - Incorrect choice of pronoun
  - etc.



Analysing Machine Translation Output

The Task: Find and categorise morphology errors in MT

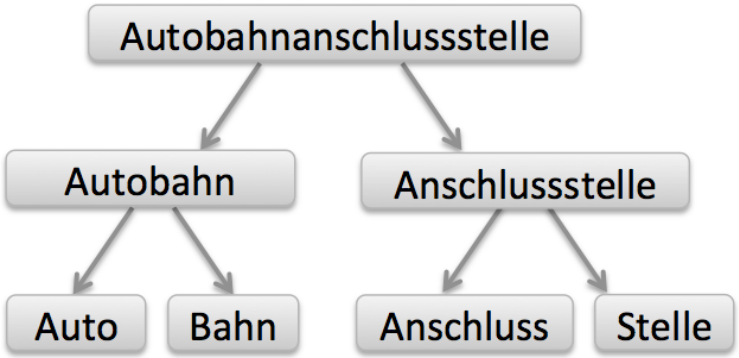
- **Preparation**: select a set of English texts
- **Translation**: translate the texts into German using a translation tool of your choice
- **Analysis**: identify errors in the German translations
- **Categorise**: construct a hierarchy / hierarchies of error categories
- **Write**: prepare guidelines for annotators to follow to label errors according to the categories
- **Assess**: follow the guidelines and annotate the translation of a test file
- **Assess**: assign a severity score to each error category
- **Code**: calculate document stats based on number of errors for each category: counts, average score over words in document, etc.

Outline

Compound Splitting

The Problem

- German has many compound words, such as:
  - Bananenbrot (Banana bread)
  - Autobahnanschlussstelle (Motorway junction)
  - Donaudampfschiffahrtsgesellschaftskapitän (Danube steamship company captain)
- Long compound words may occur infrequently in text
- In NLP we often want to split them into shorter words to make them easier to handle (e.g. Machine Translation)



Compound Splitting

The Task: Design and build a compound splitter

- **Analysis**: examine a corpus of text and identify some compound words (test set)
- **Research**: read grammar books / look up existing compound splitters
- **Planning**: devise a set of compound splitting rules (or a method of your choice)
- **Development**: code up the method
- **Testing**: apply the method to a corpus of text and analyse the output

- Possible corpus resources:
  - TED Talk corpus: https://wit3.fbk.eu (XML format)
  - Europarl corpus: http://www.statmt.org/europarl/ (text format)