Detecting Machine-translated Subtitles in Large Parallel Corpora

Pierre Lison

Norwegian Computing Center (NR)
plison@nr.no

A. Seza Doğruöz

Independent Researcher
a.s.dogruoz@gmail.com

11th Workshop on Building and Using Comparable Corpora, Miyazaki

08/05/2018

Introduction

Movie and TV subtitles are a great resource for compiling parallel corpora:

1. Wide breadth of *linguistic genres*, from colloquial language to narrative and expository discourse.
2. Large databases with millions of subtitles available online, in a wide range of languages
3. Tight coupling between subtitles and their "source material" (a movie or TV episode)

Introduction

- However, the *quality* of the subtitles is uneven
  - Often created by movie and TV fans
  - Problems with linguistic fluency, faithfulness and adherence to formatting guidelines
- Some subtitles not created by humans, but produced by translating subtitles in other languages via **MT**
  - Often low quality, with frequent translation errors
  - Many generated through older MT engines (e.g. Babelfish)

Research question

Can we automatically detect whether a subtitle has been generated through machine translation?

Caveats:

- We do not know which *subtitle* might have been the source of the translation
- We do not even know which *language* could be the source
- And we do not know which *MT system* might have been used to produce the translation

Outline

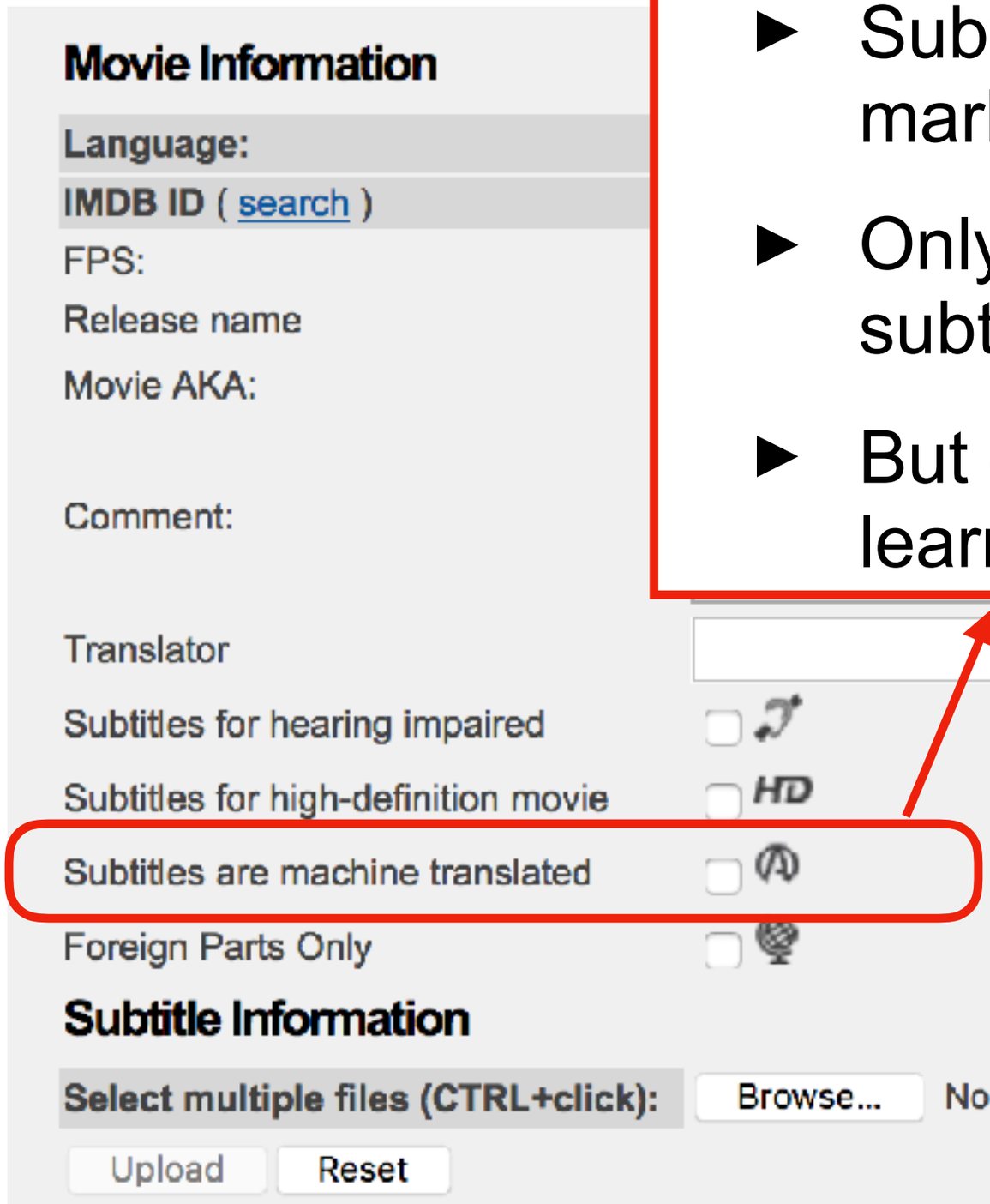
- Source corpus
- Approach
- Evaluation
- Discussion

OpenSubtitles

- Latest version of OpenSubtitles (2018 release) contains **3.73 million subtitles** in 60 languages
  - Total of **3.35 billion** sentences (22 billion tokens)
  - Alignment at both document- and sentence-level for all language pairs (1782 bitexts), based on timestamps
- The subtitles may have various origins:
  - Creation from scratch by fans, rips from DVD releases or TV streams, translations from existing subtitles, etc.
  - But this origin is typically unknown

Corpus available on OPUS:
http://opus.nlpl.eu/OpenSubtitles2018.php

OpenSubtitles



- Subtitles to upload can be marked as machine-translated
- Only a small number (4 999 subtitles) marked as such so far
- But enough to train a machine learning model!

Translation issues

- Wrong lexical choices, grammatical errors:

* *Come, you will see well.*
(**French**): Venez, vous verrez bien.
'Come, you'll see.'

* *How are you take you?*
(**French**): Comment vas-tu t'y prendre?
'How will you go about it?'

- Literal translations, unknown tokens:

* *Hij is gonna verkopen ons allen langs de rivier.*
(**English**): 'He's gonna sell us all down the river'

Translation issues

- Subtitles are conversational in nature, with many short segments and a tight dependence to context
- This is lost when applying MT engines at sentence level:

* *And Michael? It must come back, you hear?*
(**French**): Et Michael? Il doit revenir, vous entendez?
'And Michael? He must come back, you understand?'

- Translations into pro-drop languages also problematic

Approach

- Machine learning approach using the 4,999 subtitles marked as MT-generated as training set
- Two types of features:
  - **Monolingual features**, extracted from the subtitle itself.
  - **Similarity features**, extracted by determining the most likely source subtitle and extracting similarity features between the source and target sentences.
- Features must be as language-independent as possible

Monolingual features

1. Occurrence of rare or unknown tokens
   - ► According to statistical language models (bigrams)
   - ► Thresholds adjusted for every language
2. Meta-data: movie genre, release type, original language of the movie or TV episode, etc.
3. Surface cues at start or end of the subtitle:
   - ► For instance, the occurrence of the word "Google"

Similarity features

- First step: identify a plausible *source* for the translation
- The subtitle that served as source can sometimes be inferred from the **display times**
  - Intuition: if a subtitle is MT-generated, these display times (timestamps in milliseconds) will be left unchanged
  - For each subtitle, we look for subtitles for the same movie but in another language (preferably a pivot language)
  - The subtitle with the most similar timestamps is then considered as the most likely source subtitle

Similarity features

- **Surface-level features:**
  - Ratios of tokens in the "source" and target sentences (literal translations more likely when MT-generated)
  - (Also adjusted language by language)
- **Syntactic features:**
  - Intuition: MT-generated subtitles are more likely to follow the syntactic structure of its "source" subtitle
  - Captured by *k*-gram precision scores on POS sequences and dependent relations

Evaluation

- **Experimental design**:
  - Dataset: 4 999 MT-generated subtitles + 50 000 subtitles with high user ratings (assumed to be human-created)
  - 10-fold cross validation, with class reweighting
- **Baseline 1**: Occurrence of the word "Google" (and similar tokens) at the start and end of the subtitle
- **Baseline 2**: Timestamps that are identical or near-identical (Jaccard coefficient > 0.99) to another subtitle

Results

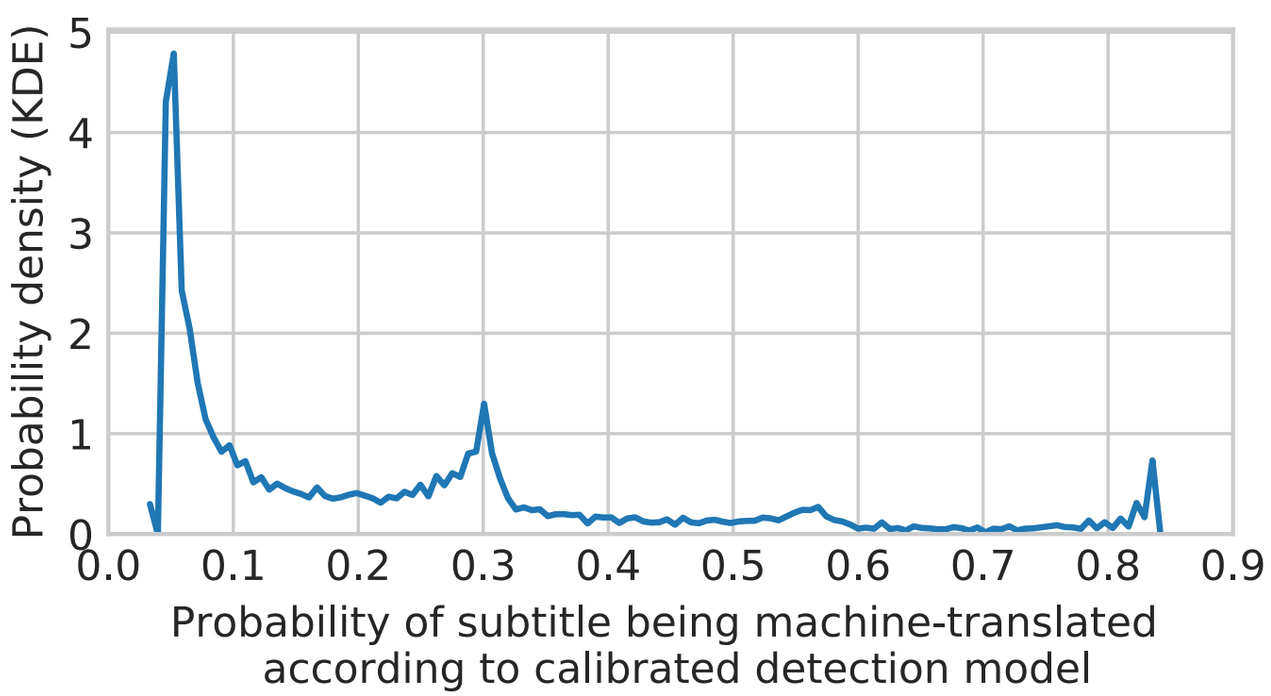
Model	**P**	**R**	F_1	**Acc**
Keyword baseline ("Google" at start/end of subtitle)	1.000	0.017	0.030	0.910
Jaccard baseline (Jaccard coefficient ≥ 0.99)	0.360	0.248	0.294	0.841
Logistic regression (l_2 reg., C =1)	0.266	0.757	0.394	0.787
SVMs (RBF kernel, C =1)	0.372	0.803	0.508	0.858
K-nearest neighbours (k=1)	0.610	0.514	0.558	0.925
Decision tree (1 sample per leaf)	0.436	0.431	0.434	0.897
Random Forest (n=100)	0.772	0.448	0.567	**0.937**
Gradient Boosting (n=100)	0.762	0.444	0.561	0.936
Neural net (1 hidden layer, d=10)	0.377	0.808	0.513	0.860
(1 hidden layer, d=50)	0.506	0.697	0.585	0.909
(1 hidden layer, d=200)	0.622	0.657	**0.638**	0.932
(2 hidden layers, d_1=50, d_2=10)	0.504	0.685	0.580	0.909

Discussion

- **Feature contributions**:
  - Most discriminative features: Jaccard coefficient between the timings, occurrence of "Google", nb. of unknown tokens
  - All feature families are useful for the detection
- **Error analysis**:
  - Dataset is not error-free (misclassifications)
  - Influence of other types of errors (e.g. OCR errors)
  - Some MT-generated subtitles are post edited

Estimates on full corpus

- We can use the detection model to extrapolate the total number of MT-generated (or at least "low quality") subtitles
  - Probability calibration with Platt's sigmoid model
  - Poisson Binomial distribution estimated from the results of the calibrated detection model
- *Results*: about **9%** of the corpus is classified by the ML model as being MT-generated



Conclusion

- Subtitles are a great resource for corpus building, but they need to be **quality checked**
  - In particular for *low-quality, MT-generated subtitles*
- Machine learning approach to detect these subtitles
  - Features extracted from the subtitles itself and from comparisons with its closest subtitle(s)
  - Detection model is language independent
  - Can be used to filter out (or assign a lower weight to) subtitles below a certain quality threshold