Parallel Corpus Creation

Pavan Mandava, Isaac Riley, Jasvinder Singh

30. Januar 2020

Overview

What is a parallel corpus?

A corpus containing the same text in two or more languages - especially valuable for machine translation, more useful when aligned.

Uses of parallel corpora

- machine translation
- translation analysis & studies
- comparative linguistics
- digital humanities
- technology-assisted language learning

Examples of parallel corpora

- Rosetta Stone
- OPUS (collection - incudes OpenSubtitles, TED Talks, etc.)
- EuroParl Corpus
- EUR-Lex Corpus
- Tatoeba

Our contribution

We create a robust toolkit for creating and querying parallel corpora, with support for XML, HTML, and SQL.

Project overview

Collect

Not too much automatic collection. A more advanced implementation could make use of crawling, but we gathered data by hand.

Prepare

1. find texts
2. segment texts into chapters
3. align chapters at sentence level
4. write to XML files
5. create SQL database from XML files

Bitext alignment in general

Problem: Given two parallel texts, identify corresponding units (usually chapters, paragraphs, sentences, potentially even words)

Alignment algorithm

```
def master_align(text0, text1, lang0, lang1):
    """ Takes two equivalent texts (original, translation);
         returns dictionaries containing {sent#: sent}. """
    df0 = frame_from_text(text0, lang0, lang1)
    df1 = frame_from_text(text1, lang1, lang0, is1=True)
    # columns [‘sent’, ‘trans’, ‘rellen’, ‘relpos’]
    anchors = anchors_from_frames(df0, df1, window=2)
    aligns = inter_align(df0, df1, anchors, lookahead=4)
    dict0, dict1 = dicts_from_aligns(df0, df1, aligns)
    return dict0, dict1
```

Example

Alice fing an sich zu langweilen; sie saß schon lange bei ihrer Schwester am Ufer und hatte nichts zu thun. Das Buch, das ihre Schwester las, gefiel ihr nicht; denn es waren weder Bilder noch Gespräche darin. „Und was nützen Bücher," dachte Alice, „ohne Bilder und Gespräche?"

Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

Example: Sentence Tokenization

1. Alice fing an sich zu langweilen; sie saß schon lange bei ihrer Schwester am Ufer und hatte nichts zu thun.
2. Das Buch, das ihre Schwester las, gefiel ihr nicht; denn es waren weder Bilder noch Gespräche darin.
3. „Und was nützen Bücher,“ dachte Alice, „ohne Bilder und Gespräche?“

1. Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, ‘and what is the use of a book,’ thought Alice ‘without pictures or conversation?’

Example: Translate German

1. Alice started to get bored; she had been sitting with her sister on the bank for a long time and had nothing to do.
2. She didn't like the book her sister read; because there were no pictures or conversations in it.
3. „And what use are books," thought Alice, „without pictures and conversations?"

1. Alice was beginning to get very tired of sitting by her sister on the bank, and of having nothing to do: once or twice she had peeped into the book her sister was reading, but it had no pictures or conversations in it, 'and what is the use of a book,' thought Alice 'without pictures or conversation?'

Example: Alignment Step

$$trsim(a, b) = 1 - levenshtein(a, b)/max(length(a), length(b))$$
$$trsim(a1, b1) = 0.272$$
$$trsim(a1 + a2, b1) = 0.395$$
$$trsim(a1 + a2 + a3, b1) = 0.591 > trsim(a1 + a2 + a3 + a4, b1)$$

Data format and query types

Example

Potential Enhancements / Applications

Enhancements:

- Expand corpus
- Alignment at word level
- Robust chapter segmentation and pre-processing to take noisier and less-structured text formats

Applications:

- Compare different translation styles
- Investigate translation variance of terms
- Paremiology: how are proverbs and idiomatic expressions rendered in other languages?
- Language learning: parallel reading, translation tests

Summary