LEXACC: Lucene-based parallel sentence EXtractor from Comparable Corpora

Table of Contents

1 BASIC INFORMATION

1.1 Tool name

LEXACC: Lucene-based parallel sentence EXtractor from Comparable Corpora

1.2 Overview and purpose of the tool

LEXACC identifies and extracts parallel sentences from bilingual comparable corpora, no matter the comparability level of that corpora. Each extracted sentence pair is assigned a translation similarity score which is a measure of the degree of paralelism between the sentences forming the pair. The tool is a stand-alone application developed in C#.

1.3 A short description of the algorithm

The algorithm has been thoroughly described in Ştefănescu et al. (2012). It uses the Lucene search engine to drastically reduce the search space of possible sentence pairs given by the Cartesian product between the sentences in the source and target texts. Then, an optional filtering step reduces this search space even more, making the algorithm fast, even on large data. The final selection of the sentence pairs that are corresponding translations is made with the help of a translation similarity measure computed using 5 features: (i) content words translation strength, (ii) functional words translation strength, (iii) alignment obliqueness, (iv) strong translation sentinels and (v) same punctuation ending.

LEXACC has 2 versions for both 32-bit and 64-bit machines. It already contains resources for running it for the following language pairs (in both directions): English-Croatian, English-German, English-Greek, English-Estonian, English-Latvian, English-Lithuanian, English-Romanian, English-Slovene and English-Spanish.

2 TECHNICAL INFORMATION

2.1 Software dependencies and system requirements

This tool requires Linux machines with Mono, or Windows machines with Microsoft .Net Framework 4 installed and 3Gb of RAM. When using big dictionaries (e.g. English-Romanian (~300Mb)) the users will need more memory. In this case, it is recommended to use x64 machines.

2.2 Installation

This tool requires Microsoft .Net Framework 4. No other installation is required.

2.3 Execution instructions

LEXACC can be run in two modes: with available document alignments (the recommended usage) and without document alignments (only if document alignments are hard to obtain/do not exist for whatever reason).

The switches controlling the I/O data for LEXACC are:

- "--source" and "--target" specify the source language and the target language respectively. These languages are to be specified by a 2 letter code (e.g.: en, et, el, de, lt, lv, ro, sl, es);
- "--docalign" gives the document alignments list in a format similar to that produced by EMACC or DictMetric (see their entries in this document). This is the run mode with available document alignments;
- "--input" and "--input" (always 2 input switches) give the source and the target document lists in the case that the document alignment list is not available. The formats of these lists are the same as in case of DictMetric or EMACC. This is the run mode without the document alignments. If these switches are present then "--docalign" must NOT be given and vice versa;
- "--output" specifies the file to write the found parallel sentence pairs to;
- "--param seg=true" specifies that the text in the source and target documents is already sentence split and tokenized (default "false");
- "--param maxrep=" specifies the maximum number of target sentences to align to one source sentence (default "1");
- "--param kif=true" instructs LEXACC to not delete the intermediary files it produces (i.e. to keep intermediary files). Useful for debugging purposes; default "false". When processing very large corpora it is recommended to set this parameter to "true" because LEXACC may crash when trying to sort (in memory) the extracted pairs by the translation similarity measure;
- "--param t=" causes LEXACC to output only those sentence pairs that have a translation similarity measure above the specified real value (default "0.2");
- "--param filter=false" causes LEXACC to NOT perform a pre-filtering step of the candidate sentence pairs before computing the PEXACC translation similarity measure (default "true"). Filtering greatly reduces the running time but it also reduces the recall of LEXACC.

For instance, running LEXACC on an English-Romanian comparable corpus with available document alignments, requesting at most 2:2 sentence alignments with at least 0.3 translation similarity score, with filtering and LEXACC-supplied sentence splitting and tokenization, the command line would be:

lexacc32.exe --source en --target ro --docalign en-ro-docalign-list.txt --param seg=false --param filter=true --param maxrep=2 --param t=0.3 --output results.txt

or, using the defaults:

lexacc32.exe --source en --target ro --docalign en-ro-docalign-list.txt --param maxrep=2 --param t=0.3 --output results.txt

2.4 Input / Output data formats

Input:

If the user has aligned documents, each line in the input file should contain tab separated paths corresponding to a document pair. Otherwise, the user must use two input files, each containing the paths to the documents in the source, and respectively the target language. All the documents the paths in the input file refer to, should be UTF-8 encoded.

Output:

The output file is UTF-8 encoded and contains pairs of sentences considered reciprocal translations, along with their assigned translation similarity score value (see Fig. 1).

```
Florence Wysinger Allen nació en Oakland, California, en 1913.
Florence Wysinger Allen was born in Oakland, California in 1913.
0.975578665733337

Julio Vallejo Ruiloba nació en Barcelona, España, en 1945.
Julio Vallejo Ruiloba was born in Barcelona, Spain, in 1945.
0.975578665733337

Álbum mix de colaboración entre Coldcut, DJ Krush y DJ Food
A mix album collaboration between Coldcut, DJ Krush and DJ Food
0.974823534488678

Fue el fundador del Larry Aldrich Museum en 1964.
He founded the Larry Aldrich Museum in 1964.
0.97479522228241

A pesar - (1987) Canta: Morist Jimenez
A pesar - (1987) Singer: Morist Jimenez
0.974738240242004
```

Figure 1: LEXACC output

2.5 Integration with external tools

LEXACC is fully self-contained.

3 CONTENT INFORMATION

3.1 Test input files

This application can be tested by using the example provided in the *runme.bat* file on data in the *test* folder containing already sentence-split Wikipedia comparable corpora for Spanish-English pair of languages. The comparable test corpus contains 1000 document pairs (3.5 Mb of data for English and 3.1 Mb for Spanish).

3.2 Output files

The main output file is (in our test case) the file *results_es_en.txt*. Optionally, the user may keep all the intermediate files the app generates (using the argument --param kif=true). From the input data LEXACC extracts 7,764 sentence pairs (2.4 Mb) having translation similarity scores from 0.99 (parallel) to 0.1 (very weakly comparable).

3.3 Running times

In order to report the running times for this tool, we run it on a 64bit 12-core Intel(R) Core(TM) i7 CPU 980 @ 3.33GHz and 16 GB of RAM. On the test input data LEXACC runs for 15.29 minutes and needs 638.38 Mb of RAM.

4 ADMINISTRATIVE INFORMATION

4.1 Contact

For further information, please contact: Dan ŞTEFĂNESCU (http://www.racai.ro/~danstef/; danstef@racai.ro; dstefanescu@gmail.com) or Radu ION (http://www.racai.ro/~radu/; radu@racai.ro).

5 REFERENCES

Dan Ştefănescu, Radu Ion, and Sabine Hunsicker. Hybrid Parallel Sentence Mining from Comparable Corpora. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 137—144, Trento, Italy, May 28-30, 2012.