OLANTO

myPREP v2.0

Installation Manual for the Alignment tool Server

CONTENT

1. Context

myPREP is a text aligner software, a tool which makes possible to automatically align two by two the documents in a multilingual corpus. The outcome of the alignment is a translation memory in TMX format. The alignment is done at the sentence level.

Thanks to myPREP, the production of training corpus for statistical translation is also possible (in Moses format). Such corpora are divided into several parts for the training, the tuning and the evaluation.

myPREP also makes possible the alignment of comparable corpora. The outcome of the alignment is a set of pair of sentences associated with a score, the number of aligned terms, and the length of sentences. These functions can control the alignments.

myPREP requires segmented documents corpora in UTF-8 format. The converter and the segmentation tool of the myCAT software are included in the installation of myPREP.

1.1. Windows/Unix

This installation manual is aimed for Windows and for GNU/Linux (tested on Ubuntu 12.04).

The default installation is done with the "olanto" user. So, the root of the installation will be **/home/olanto/**. After unpacking the archive, you can find in the **config** folder a **forUnix** folder that contains the configuration files suitable for Unix and a **shell** folder that contains the commands suitable for Unix.
The rest of the manual is identical for both systems.
Notations:
For Linux [root] = **/home/olanto**; **command.sh**
For windows [root] = **C:** ; **command.bat**

2. Installation Procedure

- Prerequisites: both Java7 and OpenOffice must be installed (see the procedure on Olanto)
- Uncompress the **Front_End.zip** file that you have downloaded from Olanto's website. This file contains a folder: MYPREP. The folder must be extracted to the [root].
- Move the **MYCAT_for_MYPREP** folder to [root] and rename it as MYCAT (in capital letters). This is not an installation of the myCAT application (which is another software distributed by Olanto), but only the storing of the folders that are shared by the conversion and the segmentation modules.

3. Alignment Procedure for BiText

The following languages are configured in this distributed version: **English, French, Spanish, Arabic, Russian, Chinese**. Alignment maps (built with the statistical machine translation tool called Moses) are only available for the following language pairs (in both directions) and all their combinations:
English – French English – Arabic
English – Spanish English – Russian
English – Portuguese

Some other language pairs are already available. You can download them from the MyCAT webpage.

- **Preparing the corpus**
  - Copy your training corpus to **[root]/MYCAT/corpus/docs** (with the language extensions _EN, _FR, …)
  - Run the **[root]/ MYPREP/run/conversion.sh** or **[root]/ MYPREP/run/conversion.bat** file to convert the corpus into TXT (UTF-8)
  - Run the **[root]/ MYPREP/run/segmentation.sh** or **[root]/ MYPREP/run/segmentation.bat** file to segment the TXT version of the corpus (**[root]/MYCAT/corpus/txt**); the language folders must be correct, and in **[root]/MYCAT/config/SEG_fix.xml** the following parameter must be in line with those language folders:
    `<entry key="LIST_OF_SEG_LANG">EN FR ES AR ZH RU</entry>`

- **Building the TMX**
  - The configuration file is **[root]/MYPREP/config/ALN_fix.xml**. The following parameter describes the extracted TMX:
    `<entry key="LIST_OF_BITEXT_LANG">ENFR ENES ENAR</entry>`
  - The XXYY pairs must correspond to an existing dictionary (those in **[root]/MYCAT/map**)
  - To build the TMX, use the **makeTMX.sh** or **makeTMX.bat** command in **[root]/MYPREP/run**
  - The TMX are generated in **[root]/MYPREP/TMX/XXYY** (language pair)
  - A TMX is generated for each document pair
  - In **[root]/MYPREP/logs**, you can find a log about the TMX generation "**makeTMX_logs.txt**"

- **Building training corpora for SMT**
  - The corpora are extracted from the TMX
  - For each so (source) ta (target) language, extract the corpus, tuning and eval files
  - The files are generated in **MYPREP/SMT/XXYY** (the folder must be created before starting the command)
  - To build the SMT, use the **makeSMT.sh** or **makeSMT.bat** command in **[root]/MYPREP/run**
  - The command must be appropriate depending on the desired languages and sizes for the different uses

    ```
    java -Xmx1000m -classpath "./Extraction.jar"
    org.olanto.smt.extraction.RUNExtraction 20 20 5
    "/home/olanto/MYPREP" ENAR  >
    /home/olanto/MYPREP/logs/extraction_ENAR_logs.txt
    ```

    or

    ```
    java -Xmx512m -jar C:\MYPREP\dist\Extraction.jar 20 20 5
    "C:/MYPREP" ENAR  > C:\MYPREP\logs\extraction_ENAR_logs.txt
    ```

  - The `20 20 5` parameters determine:
    - `20`: the maximum number of lines of the **tuning** file

- 20: the maximum number of lines of the **eval** file
- 5: a line is selected every 5 lines (of the corpus file). To spread the sample, this number requires an adjustment depending on the size of the corpus.
- Of course these figures have to be adapted regarding the size of the corpus

- To build the "reverse" training, use the **reverseTraining ENFR FREN** command (that renames the files)
- The produced corpora can be directly used for Moses.

4. Alignment Procedure for Comparable Corpora

- The comparable files are documents whose contents deal with the same topic, but are not mutual translations. Wikipedia's articles are a source of comparable documents.
- **Preparing the corpus**
  - The same way as for the Bitext corpora.
- **Comparable corpora alignment**
  - The configuration file is **[root]/MYPREP/config/COMP_fix.xml**. The following parameter describes the comparable files that will be extracted:
    `<entry key="LIST_OF_COMPARABLE_LANG">ENFR</entry>`
  - The XXYY pairs must correspond to an existing dictionary (those in **[root]/MYCAT/map**)
  - To build the COMP, use the **makeCOMP.sh** OR **makeCOMP.bat** command in **[root]/MYPREP/run**
  - The TMX are generated in **[root]/MYPREP/COMP/XXYY** (language pair)
  - The outcome is a set of files (each file contains 100,000 alignments) for the whole corpus. Each alignment is preceded by the score, the number of identical terms, the number of aligned terms, the length of the source sentence, the length of the target sentence, the source, and the target.

```
0.26432112     1     4     13    14    In Denmark, squatters
occupied a disused military base and declared the Freetown
Christiania, an autonomous haven in central Copenhagen.  La
commune libre Christiania à Copenhague au Danemark, expérience au
bénéfice d'aides gouvernementales d'un squat autonome/autogéré au
niveau d'un quartier
```

  - In **[root]/MYPREP/logs**, you can find a log about the TMX generation "makeTMX_logs.txt"

- **Building training corpora for SMT**
  - The corpora are extracted from the COMP
  - For each so (source) ta (target) language, extract the corpus file
  - The files are generated in **[root]/MYPREP/COMP/SMT/XXYY** (the folder must be created before starting the command)
  - To build the SMT, use the **makeSMTfromCOMP.sh** **or** **makeSMTfromCOMP.bat** command in **[root]/MYPREP/run**
  - The command must be appropriate depending on the desired languages and sizes for the different uses

```
java -Xmx1000m -classpath "./Extraction.jar"
org.olanto.comp.extraction.RUNExtraction 2 4 1.7f 0.3f
/home/olanto/MYPREP EN FR    >
/home/olanto/MYPREP/logs/extraction_from_COMP_ENFR_logs.txt
```

or

```
java -Xmx1000m -classpath "C:\MYPREP\dist\Extraction.jar"
org.olanto.comp.extraction.RUNExtraction 2 4 1.7f 0.3f
C:/MYPREP EN FR  >
C:\MYPREP\logs\extraction_from_COMP_ENFR_logs.txt
```

The `2 4 1.7f 0.3f` parameters determines the filtering:

- `2`: the minimum number of aligned terms
- `4`: the minimum length of the sentences
- `1.7f`: the maximum ratio between the source and the target
- `0.3f`: the minimum score
- Of course these figures have to be adapted regarding the size of the corpus

- These corpora can be added to existing corpora to improve the terminology, for example.

5. Source

The sources are published on Github in myPREP project: https://github.com/myPREP