Using TM Cleaner with Bing translation engine.

In this modality the software uses an extra-feature equal to the cosine similarity between the translation of the source segment and the target segment. The translation is done with Bing translation engine.

TM Cleaner Configuration File

The configuration file used by the scripts can be found under Parameters directory "p-Bing.txt".
Add your key for interrogating Bing API after the "appId=".

Resources

The resources that can be used with this tutorial are following one:

1. An *English-Italian* scikit-learn model trained on a sample of English-Italian positive and negative bi-segments extracted from MyMemory:
   Training/Bing/ full-English-Italian-Features.csv
2. An *English-Spanish* scikit-learn model trained on a sample of English-Spanish positive and negative bi-segments extracted from MyMemory:
   Training/Bing/ full-English-Spanish-Features.csv
3. An *English-Italian* test file obtained from automatically aligning a web site containing English and Italian parallel documents. The file contains positive and negative segments.
   a. The file to be classified:
      Resources/Examples/Bing/about-small-en-it.txt
   b. The file annotated with correct labels for evaluation:
      Resources/Examples/Bing/about-small-en-it-annotated.txt

4. An *English-Spanish* test file containing English and Spanish positive and negative bi-segments
   a. The file to be classified:
      Resources/Examples/Bing/testSet-en-es.txt

b. The file annotated with correct labels for evaluation

Resources/Examples/Bing/testSet-en-es-annotated.txt

Training:

The configuration parameters for training are in "*Parameters/Bing/p-Training-XXX.txt*" files. You should copy and edit the corresponding file to fit your purposes.

```
python generateFeaturesAndClassify.py --features --config Parameters/Bing/p-Training-Italian.txt
```

For this tutorial you do not need to train: we did the training for you and obtained the models presented in the previous section.

Classification:

English-Italian example.

1. Copy the English Italian test file inside the "TestFiles" directory taking care that the TestFiles directory is empty.

   ```
   cp Resources/Examples/Bing/about-small-en-it.txt TestFiles/
   ```

2. Classification using the default algorithm "SVM with linear kernel"

   ```
   python generateFeaturesAndClassify.py --classify --config Parameters/Bing/p-Batch-Italian.txt
   ```

3. Classification using the algorithm "Logistic regression" with the default class 0 and the threshold 0.7 (To see what this means read the configuration file)

   ```
   python generateFeaturesAndClassify.py --classify --config Parameters/Bing/p-Batch-Italian.txt --mlalgorithm LogisticRegression
   ```

English-Spanish example:

1. Copy the English Spanish test file inside the "TestFiles" directory taking care that the TestFiles directory is empty.

   ```
   cp Resources/Examples/Bing/testSet-en-es.txt TestFiles
   ```

2. Classification using the default algorithm "SVM with linear kernel"

   ```
   python generateFeaturesAndClassify.py --classify --config Parameters/Bing/p-Batch-Spanish.txt
   ```

3. Classification using the algorithm “Logistic regression” with the default class 0 and the threshold 0.7 (To see what this means read the configuration file)

```
python generateFeaturesAndClassify.py --classify --config Parameters/Bing/p-Batch-Spanish.txt --mlalgorithm LogisticRegression
```

A directory called “Translated” is created inside the directory where the files to be classified are and the translations obtained with Bing are stored in case you need them in the future.

Evaluation:

To see how well the algorithms performed look inside the directory: “Resources/Examples/Bing/Evaluation”.

To perform the evaluation for yourself and know about each file returned by the evaluation script read the tutorial about the Evaluation.