“Abu-MaTran project: tools for teaching machine translation”. A practical workshop on how to easily obtain parallel data from the web and train statistical machine translation systems.

Víctor M. Sánchez-Cartagena
Prompsit Language Engineering, S.L.
www.prompsit.com
Campus UMH. Edifici Quórum III.
Av. de la Universitat, s/n. 03203. Elx (Alacant). Spain

16th November 2016. Dublin City University (Ireland)

Contents

About this guide

Overview

This guide is intended to be your best friend during this workshop for the hands-on practical exercises you'll be working on to meet the following objectives:

1. Get translation memories from parallel websites
2. Build and test statistical machine translation systems from a web interface

For each objective we will work on several tasks. Each topic will be briefly introduced and then you will be putting your hands on it. Before starting, please download and unzip the workshop materials from `http://www.abumatran.eu/dcu-nov-2016-materials.zip`. You will need them later.

1 Getting translation memories from multilingual websites

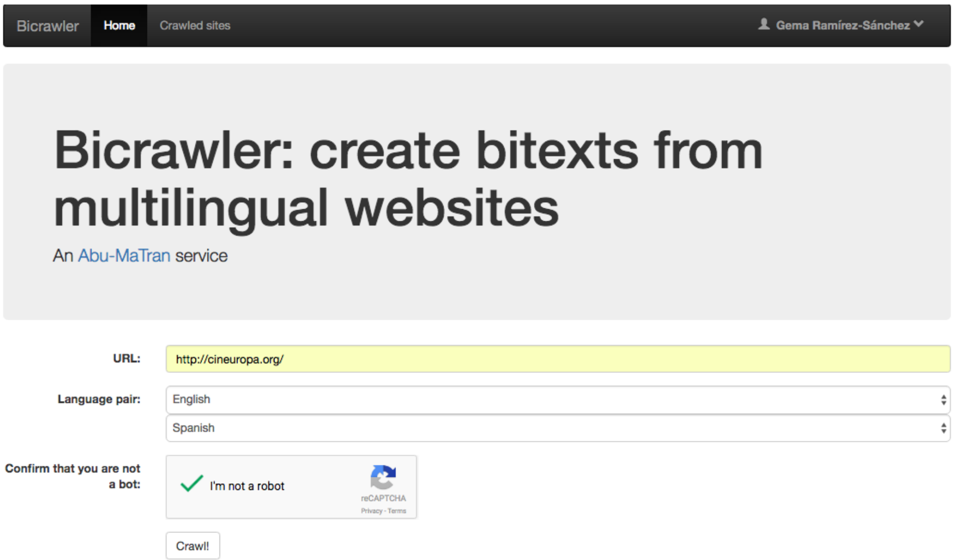
Translation memories are essential for the daily work of professional translators. How to get them when we do not have one? Bicrawler can help!

Task 1. Gathering translation memories with Bicrawler [20 min.]

Bicrawler is a web-based service to create translation memories (or bitexts) from multilingual websites. It allows for 1-hour crawling of a website and delivers ready-to-use bitexts. Let's take a look to how it works:

- Go to `http://bicrawler.com`
- If possible, login with a gmail account by clicking on the right upper menu called **Log in/Sign up**. Otherwise, Bicrawler will ask you for an e-mail address to communicate with you. Login allows you to see a dashboard with all your bitexts.
- Insert the following URL and languages in the **URL** field of the main page:
  - `http://sngular.team/` from English to Spanish

- Now click on **I'm not a robot**. If asked, please complete the tasks required to show that you are actually not a robot.
- Finally, click on **Crawl!** to launch Bicrawler.



Bicrawler: create bitexts from multilingual websites

An Abu-MaTran service

URL: http://cineuropa.org/

Language pair: English
Spanish

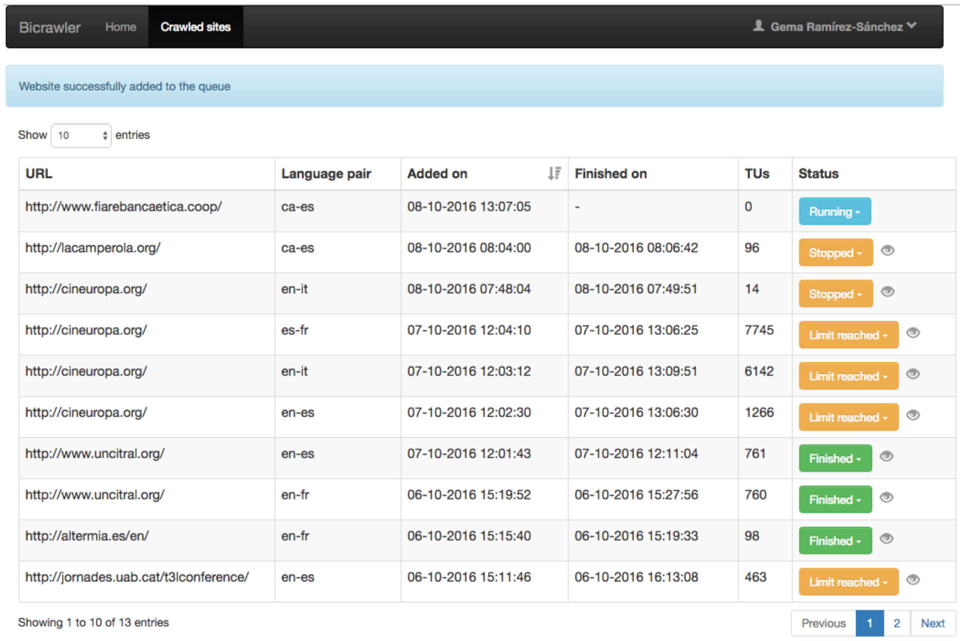
Confirm that you are not a bot: ☑ I'm not a robot

Figure 1: Bicrawler home page.

You will see a message stating that your task has been sucessfully added to the Bicrawler queue. If you are logged in, you will also see dashboard in a page called Crawled websites (see screenshot below) with some information about the launched task: website URL, language pair, date and time in which the task was added, date and time in which the task finished, number of translation units and a button that indicates the status of the task and gives you several options at each stage:

- Stage 1: Bicrawler is running (or just about to do it)!
  - **Running**: it means that Bicrawler is doing its job. Option Stop: You can let it go until it is finished or reaches the 1-hour crawling limit or you can stop it by accessing the drop-down menu in the button and clicking **Stop**.
  - **In queue x/y**: it means that there are too many concurrent tasks and that yours is in position X in a queue of Y tasks.
  - **Stopping**: it means that you stopped Bicrawler. It is calculating a translation memory.

- Stage 2: Bicrawler has finished!
  - **Limit reached**: it means that Bicrawler has reached 1-hour crawling for the website and created a translation memory.
  - **Stopped**: it means that at some point you decided to stop the crawling. Bicrawler will generate a translation memory with the crawled text until the moment you stopped it.
  - **Finished**: it means that Bicrawler took all texts from the website before reaching the 1-hour limit and created a translation memory.
  - **Failed**: something went wrong. Check the URL and the language pairs. If they look good, probably Bicrawler experienced some technical problem. Try again from the beginning or contact us.



Website successfully added to the queue

Show 10 entries

URL	Language pair	Added on	Finished on	TUs	Status
http://www.fiarebancaetica.coop/	ca-es	08-10-2016 13:07:05	-	0	Running
http://lacamperola.org/	ca-es	08-10-2016 08:04:00	08-10-2016 08:06:42	96	Stopped
http://cineuropa.org/	en-it	08-10-2016 07:48:04	08-10-2016 07:49:51	14	Stopped
http://cineuropa.org/	es-fr	07-10-2016 12:04:10	07-10-2016 13:06:25	7745	Limit reached
http://cineuropa.org/	en-it	07-10-2016 12:03:12	07-10-2016 13:09:51	6142	Limit reached
http://cineuropa.org/	en-es	07-10-2016 12:02:30	07-10-2016 13:06:30	1266	Limit reached
http://www.uncitral.org/	en-es	07-10-2016 12:01:43	07-10-2016 12:11:04	761	Finished
http://www.uncitral.org/	en-fr	06-10-2016 15:19:52	06-10-2016 15:27:56	760	Finished
http://altermia.es/en/	en-fr	06-10-2016 15:15:40	06-10-2016 15:19:33	98	Finished
http://jornades.uab.cat/t3lconference/	en-es	06-10-2016 15:11:46	06-10-2016 16:13:08	463	Limit reached

Showing 1 to 10 of 13 entries

Figure 2: Bicrawler crawled websites dashboard.

Also, by clicking on the **Status button** a you will have 3 options:

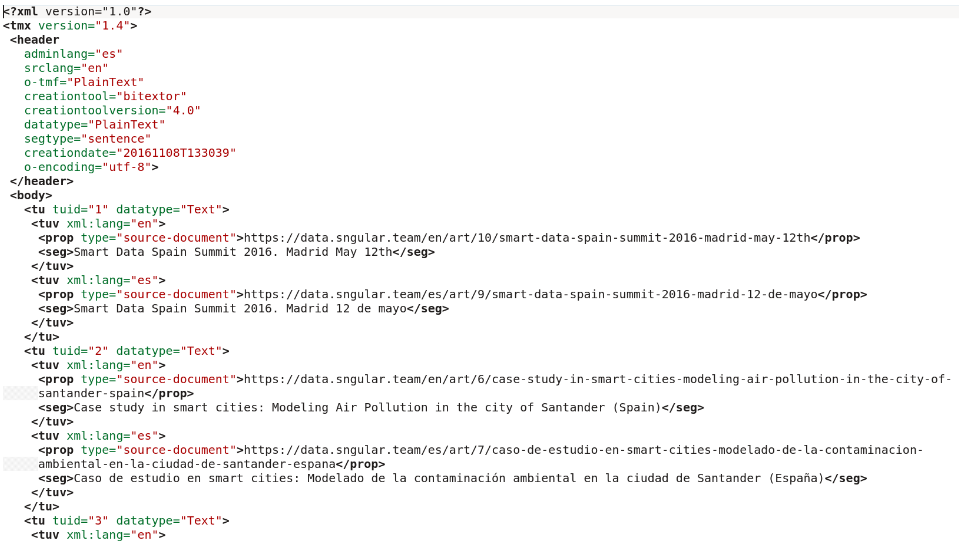
- **Download (TMX)**: you will get a translation memory in the standard Translation Memory Exchange (TMX) format.
- **Download (Moses format)**: you will get a compressed file with two texts with exactly the same amount of lines for each of the languages

of the crawling task. They are useful, among other uses, as texts to train Moses-based SMT systems. You will use them in the next section of this workshop.

- **Delete**: you will delete the current task, both the info and the files.

Please, download the bitexts in both TMX and Moses format and take a look to them (with a text editor or the browser) once you get to a finished status. Otherwise, download them from the `bicrawler` folder in the workshop materials (`sngular.team-en-es.tmx/en/es`) .

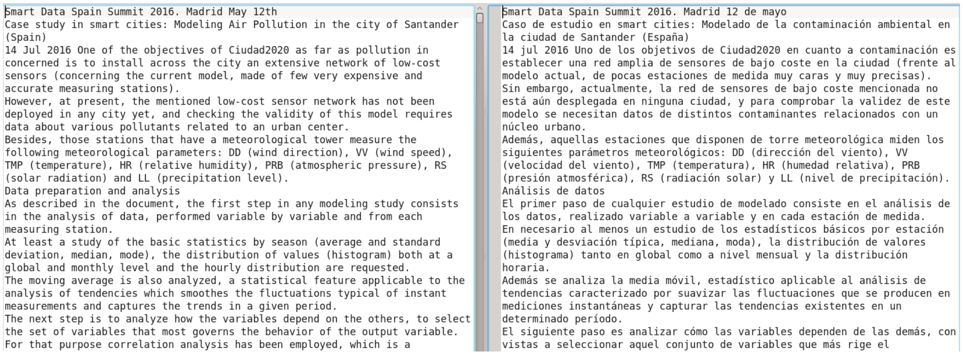
Once finished, the TMX should look like this:



```
<?xml version="1.0"?>
<tmx version="1.4">
 <header
   adminlang="es"
   srclang="en"
   o-tmf="PlainText"
   creationtool="bitextor"
   creationtoolversion="4.0"
   datatype="PlainText"
   segtype="sentence"
   creationdate="20161108T133039"
   o-encoding="utf-8">
 </header>
 <body>
   <tu tuid="1" datatype="Text">
    <tuv xml:lang="en">
     <prop type="source-document">https://data.sngular.team/en/art/10/smart-data-spain-summit-2016-madrid-may-12th</prop>
     <seg>Smart Data Spain Summit 2016. Madrid May 12th</seg>
    </tuv>
    <tuv xml:lang="es">
     <prop type="source-document">https://data.sngular.team/es/art/9/smart-data-spain-summit-2016-madrid-12-de-mayo</prop>
     <seg>Smart Data Spain Summit 2016. Madrid 12 de mayo</seg>
    </tuv>
   </tu>
   <tu tuid="2" datatype="Text">
    <tuv xml:lang="en">
     <prop type="source-document">https://data.sngular.team/en/art/6/case-study-in-smart-cities-modeling-air-pollution-in-the-city-of-
     santander-spain</prop>
     <seg>Case study in smart cities: Modeling Air Pollution in the city of Santander (Spain)</seg>
    </tuv>
    <tuv xml:lang="es">
     <prop type="source-document">https://data.sngular.team/es/art/7/caso-de-estudio-en-smart-cities-modelado-de-la-contaminacion-
     ambiental-en-la-ciudad-de-santander-espana</prop>
     <seg>Caso de estudio en smart cities: Modelado de la contaminación ambiental en la ciudad de Santander (España)</seg>
    </tuv>
   </tu>
   <tu tuid="3" datatype="Text">
    <tuv xml:lang="en">
```

Figure 3: Translation memory obtained from `http://sngular.team` in TMX format

And the text files should look like this:



```
Smart Data Spain Summit 2016. Madrid May 12th
Case study in smart cities: Modeling Air Pollution in the city of Santander
(Spain)
14 Jul 2016 One of the objectives of Ciudad2020 as far as pollution in
concerned is to install across the city an extensive network of low-cost
sensors (concerning the current model, made of few very expensive and
accurate measuring stations).
However, at present, the mentioned low-cost sensor network has not been
deployed in any city yet, and checking the validity of this model requires
data about various pollutants related to an urban center.
Besides, those stations that have a meteorological tower measure the
following meteorological parameters: DD (wind direction), VV (wind speed),
TMP (temperature), HR (relative humidity), PRB (atmospheric pressure), RS
(solar radiation) and LL (precipitation level).
Data preparation and analysis
As described in the document, the first step in any modeling study consists
in the analysis of data, performed variable by variable and from each
measuring station.
At least a study of the basic statistics by season (average and standard
deviation, median, mode), the distribution of values (histogram) both at a
global and monthly level and the hourly distribution are requested.
The moving average is also analyzed, a statistical feature applicable to the
analysis of tendencies which smoothes the fluctuations typical of instant
measurements and captures the trends in a given period.
The next step is to analyze how the variables depend on the others, to select
the set of variables that most governs the behavior of the output variable.
For that purpose correlation analysis has been employed, which is a
```

```
Smart Data Spain Summit 2016. Madrid 12 de mayo
Caso de estudio en smart cities: Modelado de la contaminación ambiental en
la ciudad de Santander (España)
14 jul 2016 Uno de los objetivos de Ciudad2020 en cuanto a contaminación es
establecer una red amplia de sensores de bajo coste en la ciudad (frente al
modelo actual, de pocas estaciones de medida muy caras y muy precisas).
Sin embargo, actualmente, la red de sensores de bajo coste mencionada no
está aún desplegada en ninguna ciudad, y para comprobar la validez de este
modelo se necesitan datos de distintos contaminantes relacionados con un
núcleo urbano.
Además, aquellas estaciones que disponen de torre meteorológica miden los
siguientes parámetros meteorológicos: DD (dirección del viento), VV
(velocidad del viento), TMP (temperatura), HR (humedad relativa), PRB
(presión atmosférica), RS (radiación solar) y LL (nivel de precipitación).
Análisis de datos
El primer paso de cualquier estudio de modelado consiste en el análisis de
los datos, realizado variable a variable y en cada estación de medida.
En necesario al menos un estudio de los estadísticos básicos por estación
(media y desviación típica, mediana, moda), la distribución de valores
(histograma) tanto en global como a nivel mensual y la distribución
horaria.
Además se analiza la media móvil, estadístico aplicable al análisis de
tendencias caracterizado por suavizar las fluctuaciones que se producen en
mediciones instantáneas y capturar las tendencias existentes en un
determinado período.
El siguiente paso es analizar cómo las variables dependen de las demás, con
vistas a seleccionar aquel conjunto de variables que más rige el
```

Figure 4: Translation memory obtained from `http://sngular.team` in two text files

Now that you know how Bicrawler works, feel free to experiment with other multilingual websites that come your mind.

Task 2. Finding new multilingual websites [15 min.]

Just type the URL, select the pair of languages and click on *Crawl!* as you did previously. A rule of thumb: open the website that you want to crawl in a browser, copy the url of the home page from the browser field (as it might be different from the one you entered) and paste it into Bicrawler URL field. For the moment, there are only a few language combinations available (the language selector will guide you). We will be adding more in the next future. As for this workhsop we are to many concurrent users, once you are in the Running status, please **Stop** Bicrawler if the task does not finish after 15 minutes. Once finished, you'll be able to see an excerpt of the crawled bitext by clicking on the **eye button** next to the **Status button**.

Did you manage to obtain a translation memory from the website(s) you chose? What would you like to see as a feature in Bicrawler in the future (or to remove)? Let's share our findings together!

2 Building and testing statistical machine translation systems from a web interface

Although the emerging neural machine translation paradigm is currently the most popular option among machine translation researchers, previous state-of-the-art statistical machine translation is still the most widespread paradigm in translation industry. Hence, statistical machine translation is an important topic in translation studies. Allowing students to train and query statistical machine translation systems is an effective way to make them understand how statistical machine translation works. However, installing and running *Moses*, the most popular statistical machine translation toolkit can be a tedious task.

In this part of the workshop, you will learn to install and use *MTradumàtica*, an open-source toolkit that allows anyone to easily train and use Moses-based statistical machine translation systems from a web interface. Throughout the remainder of the document, you will find some questions addressed to you. Think about them and you will find the answer to each question a couple of paragraphs below the place it is raised.

2.1 Installing and running MTradumàtica

MTradumàtica is shipped as a Docker[1] image. It is installation is straightforward since you do not need to worry about installing dozens of software dependencies. You just need to install Docker, download the image and run it. Please follow the instructions specific to your operating system below.

Task 3. Installing and running MTradumàtica [30 min.]

2.1.1 Linux

- Go to `https://docs.docker.com/engine/installation/linux/`.
- Select your Linux distribution and follow the instructions.
- If you are using Ubuntu, follow the instructions under the sections *Prerequisites*, *Install* and *Create a Docker group*. You can ignore the remainder of the section *Optional configurations*
- Download and unzip the Docker image from `http://abumatran.eu/mtradumatica.2.1.1.docker.tar.zip`.
- Run it by issuing the following commands from a terminal:

  ```
  docker load < /path/to/mtradumatica.2.1.1.docker.tar
  docker run mtradumatica:v2.1.1
  ```

  Please note that you should replace `/path/to` with the actual path of the file. If you successfully executed the commands, you will see an output similar to this one:

[1] https://www.docker.com/

```
Starting redis + celery + gunicorn...  [done]
127.0.0.1 localhost
::1 localhost ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
172.17.0.2 3d6a9c8716bf
```

- Copy the the 4 numbers from the last line into the address bar of your web browser, add `:8080` (e.g. `http://172.17.0.2:8080/`) and you will see *MTradumàtica* welcome page.

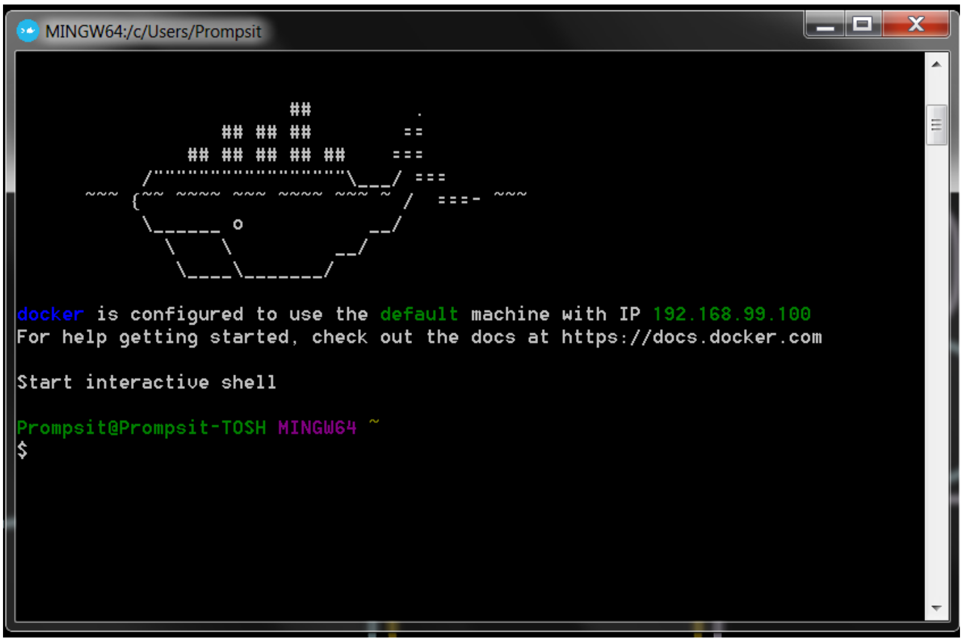
2.1.2 Windows

- Go to `https://www.docker.com/products/docker-toolbox` and click on the Download button.
- Click on the file *DockerToolbox-1.12.2.exe* and save it to your computer.
- Double click on it and follow the installation instructions.
- Download and unzip the Docker image from `http://abumatran.eu/mtradumatica.2.1.1.docker.tar.zip`.
- Open the *Docker Quickstart Console* shortcut that has just been created in your desktop. The first time you open it, it will carry out some initializations that may take some time. Be patient.
- A Docker terminal will be shown to you. Write down (in a piece of paper/text file, not in the terminal!!) the IP address (the four numbers separated by a dot) after the text *docker is configured to use the default machine with IP*. In the screenshot in Figure 5, the IP address is `192.168.99.100`.
- Issue the following commands in the Docker terminal:

  ```
  docker load < /path/to/mtradumatica.2.1.1.docker.tar
  ```

  ```
  docker run -it -p 8080:8080 mtradumatica:v2.1.1
  ```

  Please note that you should replace `/path/to` with the actual path of the file. If you unzipped the file to your desktop, `./Desktop/mtradumatica.2.1.1.docker.tar` will probably work. If you successfully executed the commands, you will see an output similar to this one:



```
                        ##         .
                  ## ## ##        ==
               ## ## ## ## ##    ===
           /"""""""""""""""""\___/ ===
      ~~~ {~~ ~~~~ ~~~ ~~~~ ~~~ ~ /  ===- ~~~
           \______ o           __/
             \    \         __/
              \____\_______/

docker is configured to use the default machine with IP 192.168.99.100
For help getting started, check out the docs at https://docs.docker.com

Start interactive shell

Prompsit@Prompsit-TOSH MINGW64 ~
$
```

Figure 5: Docker Quickstart Console

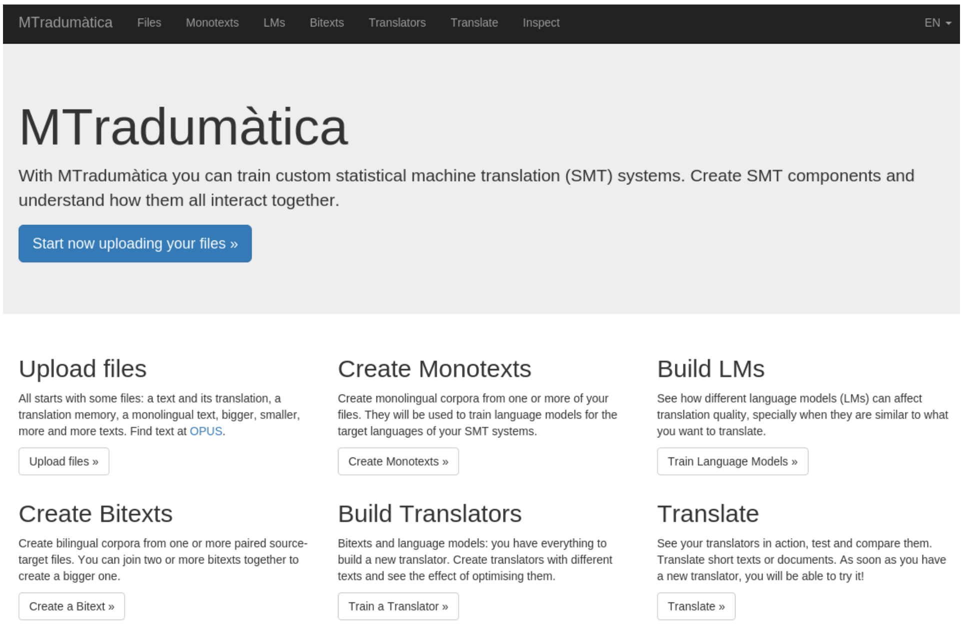
```
Starting redis + celery + gunicorn...  [done]
127.0.0.1 localhost
::1 localhost ip6-localhost ip6-loopback
fe00::0 ip6-localnet
ff00::0 ip6-mcastprefix
ff02::1 ip6-allnodes
ff02::2 ip6-allrouters
172.17.0.2 3d6a9c8716bf
```

- Type the 4 numbers you wrote down in step 6 into the address bar of your web browser, add :8080 (e.g. http://192.168.99.100:8080/) and you will see *MTradumàtica* welcome page.

2.2 Training, querying and inspecting statistical machine translation systems with MTradumàtica

MTradumàtica website contains different sections. You can access them from menu at the top and the purpose of each one is listed below.

- **Files**: upload text files. This is the starting point for training systems.



MTradumàtica

With MTradumàtica you can train custom statistical machine translation (SMT) systems. Create SMT components and understand how them all interact together.

Upload files

All starts with some files: a text and its translation, a translation memory, a monolingual text, bigger, smaller, more and more texts. Find text at OPUS.

Create Monotexts

Create monolingual corpora from one or more of your files. They will be used to train language models for the target languages of your SMT systems.

Build LMs

See how different language models (LMs) can affect translation quality, specially when they are similar to what you want to translate.

Create Bitexts

Create bilingual corpora from one or more paired source-target files. You can join two or more bitexts together to create a bigger one.

Build Translators

Bitexts and language models: you have everything to build a new translator. Create translators with different texts and see the effect of optimising them.

Translate

See your translators in action, test and compare them. Translate short texts or documents. As soon as you have a new translator, you will be able to try it!

Figure 6: *MTradumàtica* welcome page.

- **Monotexts**: create monolingual corpora from the text files you uploaded.
- **LMs**: train language models from the monolingual corpora you created.
- **Bitexts**: create parallel corpora from the text files you uploaded.
- **Translators**: train SMT systems from the parallel corpora and language models you created.
- **Translate**: translate sentences with the SMT systems you trained.
- **Inspect**: study the behaviour of the language models and the content of the phrase table of the SMT systems.

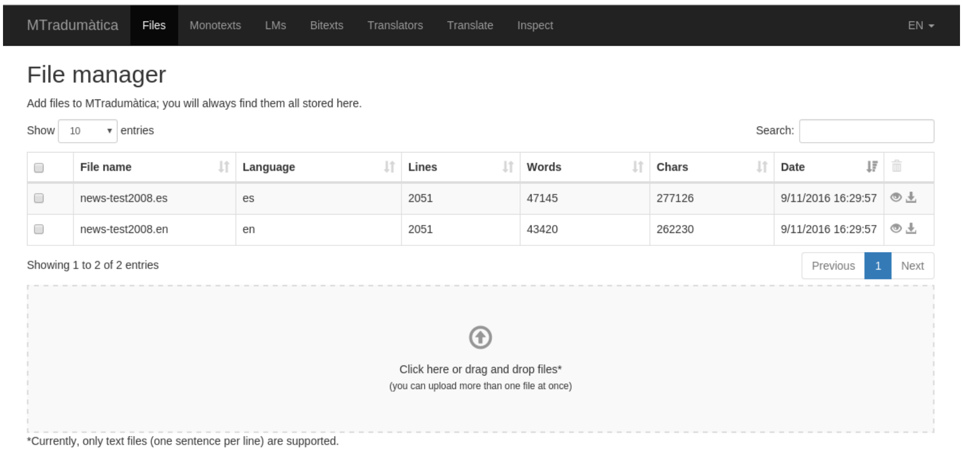
You will learn how to use them all if you complete the tasks described in this section. Let's start by simply training an SMT system from a parallel corpus.

Task 4. Train an SMT system from a couple of files [10 min.]

1. Go to the *Files* section and click on the rectangle labeled with *Click here or drag and drop files*. Select the files `news-test2008.es` and

`news-test2008.en` from the `mtradumatica` folder in the workshop materials.

2. Information about the files you uploaded will appear in the grid. You will see their name, language (automatically detected; you can change it by clicking on it), number of lines, words and characters and date of upload. You can see an snapshot of the file by clicking on the eye icon and download it by clicking on the arrow. Check that the detected language matches the language of the file and change if necessary.



File manager

Add files to MTradumàtica; you will always find them all stored here.

Show 10 entries

Search:

□	File name	Language	Lines	Words	Chars	Date	
□	news-test2008.es	es	2051	47145	277126	9/11/2016 16:29:57	
□	news-test2008.en	en	2051	43420	262230	9/11/2016 16:29:57	

Showing 1 to 2 of 2 entries

Previous 1 Next

Click here or drag and drop files*
(you can upload more than one file at once)

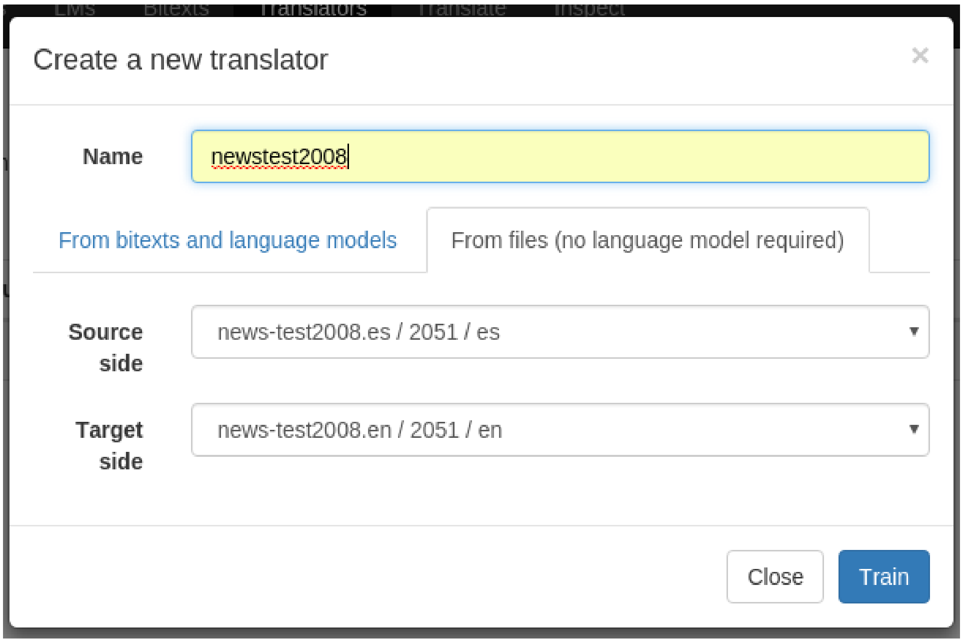
*Currently, only text files (one sentence per line) are supported.

Figure 7: *Files* section.

3. Go to the *Translators* section and click on the plus sign at the top right corner of the grid. Click on the tab labellled with the text *From files (no language model required)* and select the two files you have just uploaded. We are going to build an SMT system for translating from Spanish to English. Give it a name (for instance, *news-test2008*) and click on *Train*.

4. The information about the new translator will be displayed in the grid together with a time counter. When it turns green, the traslator will be ready to use. We have just trained our first SMT system with *MTradumàtica*! Since we have only defined a parallel corpus, the language model has been trained from its target language side.

5. Let's test our translator. Go to *Translate* section, select *newstest2008* in the drop-down list, type the text *La contaminación es un problema en una economía global.* in the text box and click on *Translate*.



Create a new translator

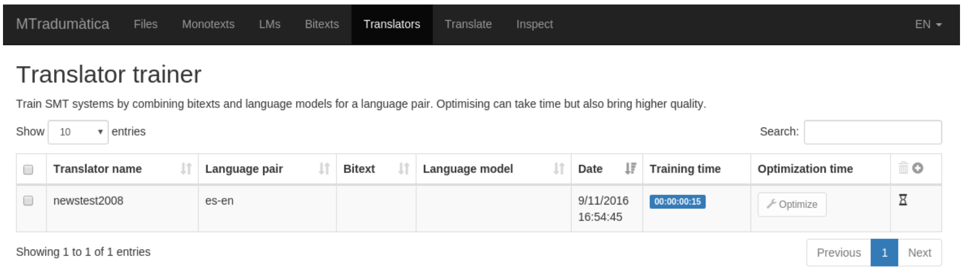
Name: newstest2008

From bitexts and language models | From files (no language model required)

Source side: news-test2008.es / 2051 / es

Target side: news-test2008.en / 2051 / en

Figure 8: Form for training an SMT system from two files.



Translator trainer

Train SMT systems by combining bitexts and language models for a language pair. Optimising can take time but also bring higher quality.

Show 10 entries

Search:

□	Translator name	Language pair	Bitext	Language model	Date	Training time	Optimization time	
□	newstest2008	es-en			9/11/2016 16:54:45	00:00:00:15	Optimize	⌛

Showing 1 to 1 of 1 entries

Figure 9: Status of a system being trained.

6. You should have obtained a translation similar to *The contaminación is a problem in a economy global*.

The translation *The contaminación is a problem in a economy global* is far from being perfect. It contains two important errors:

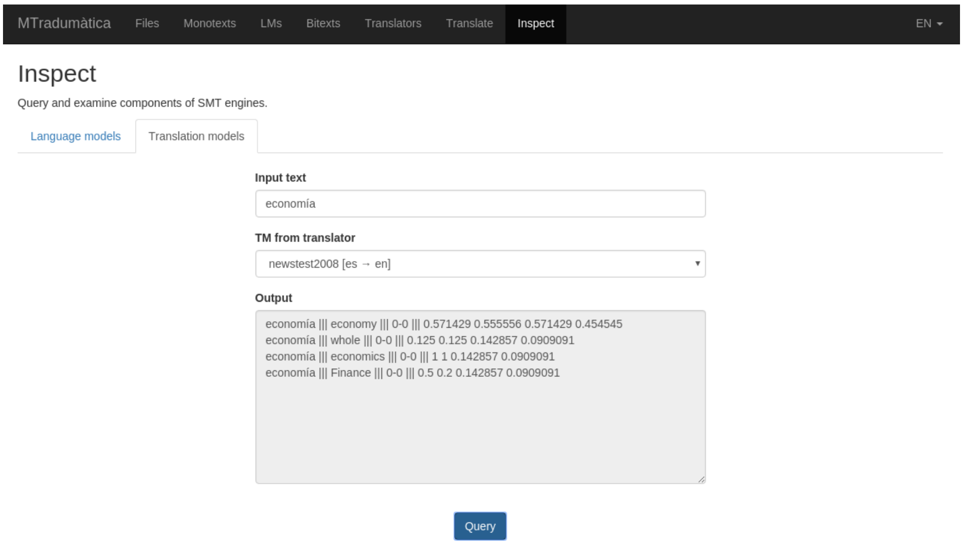
- *contaminación* has not been translated from Spanish [it means *pollution*].
- The phrase *economy global* is not written with right word order. It should be *global economy* instead.

We are going to use the tools in the *Inspect* section in order to figure out how to improve it.

Task 5. Inspect a system in order to identify the cause of translation errors [10 min.]

1. Go to the *Inspect* section and click on *Translation models*. In this page, we can check whether some words/phrases are in the phrase table and which are their translations and their probabilities. Type *contaminación* under *Input text*, select our *newstest2008* translator and click on *Query*. Why *contaminación* was not translated?

2. Now check the phrase table entries for the source phrases *economía global*, *economía* and *global*. We can clearly see that the translation model has not enough information to produce *global economy*, since it only knows how to translate *economía*. It has copied the word *global* verbatim from the input because it is not in the phrase table. Of course, the whole phrase *economía global* is not in the phrase table either.



Input text: economía

TM from translator: newstest2008 [es → en]

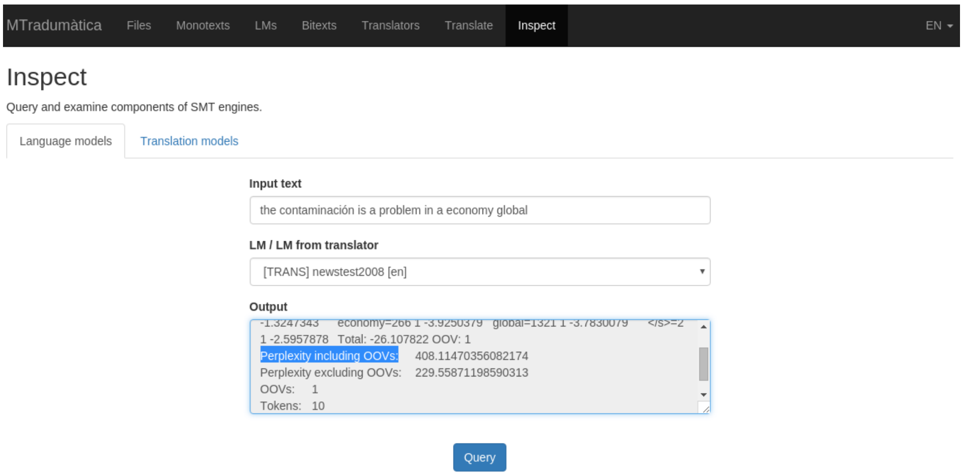
Output

```
economía ||| economy ||| 0-0 ||| 0.571429 0.555556 0.571429 0.454545
economía ||| whole ||| 0-0 ||| 0.125 0.125 0.142857 0.0909091
economía ||| economics ||| 0-0 ||| 1 1 0.142857 0.0909091
economía ||| Finance ||| 0-0 ||| 0.5 0.2 0.142857 0.0909091
```

Figure 10: Phrase table inspection.

3. Consequently, the only way to produce the right translation *global economy* is making the target language model assign a lower perplexity to the translation hypotheses that contain it. Recall that the

lower the value of the perplexity, the more likely is a target language sentence for the language model. Click on *Language models*, type the translation *the contaminación is a problem in a economy global*, click on *Query* and check the value of the perplexity (*Perplexity including OOVs*). Now type the corrected translation *the contaminación is a problem in a global economy* and check the perplexity again. Why couln't the language model help the SMT system to produce *global economy*?



Inspect

Query and examine components of SMT engines.

Input text

the contaminación is a problem in a economy global

LM / LM from translator

[TRANS] newstest2008 [en]

Output

```
-1.3247343      economy=266 1 -3.9250379   global=1321 1 -3.7830079     </s>=2
1 -2.5957878   Total: -26.107822 OOV: 1
Perplexity including OOVs:     408.11470356082174
Perplexity excluding OOVs:     229.55871198590313
OOVs:     1
Tokens:   10
```

Figure 11: Language model inspection.

Once we know the origin of the errors produced by the SMT system, we are going to fix them. Since the language model was not able to assign a lower perplexity to *global economy*, we are going to use an additional monolingual corpus to build a more powerful LM for our SMT system.

Task 6. Train an SMT system from a bitext and a language model [10 min.]

1. Go to the *Files* section and upload the file `news-commentary-subset.en` from the `mtradumatica` folder in the workshop materials. Check that the language (English) has been correctly detected and change it if necessary.

2. Go to the *Monotexts* section and create a new monolingual corpus by clicking on the plus sign at the top right corner of the grid. Give it the name *newstest2008-newscommentary* and select the English language.

3. The information about the new monolingual corpus will be displayed in the grid. Note that it is empty (the number of lines is 0). In *MTradumàtica*

you create monolingual and bilingual corpora by appending the contents of the files you upload.

4. Append to the monolingual corpus the contents of the two English text files you uploaded. You can append a file to corpus by clicking on the plus sign at the right edge of the row. At the end of the process, the monolingual corpus *newstest2008-newscommentary* should contain 5 051 lines. If you make a mistake, you can delete the monolingual corpus by cliking on the checkbox and on the trash icon afterwards.

5. Train an English language model from the monolingual corpus you have just created. Go to the *LMs* section, click on the plus sign at the top right corner of the grid, give a name (*newstest2008-newscommentary*), select the language (English) and the monotext, and click on *Train*. Training information is displayed in the same way as in the *Translators* section.

6. Now go to the *Inspect* section check again the perplexity of the two translation hypotheses that you tested previously. Do you think that an SMT system with this new language model will produce *global economy*?

7. Go to *Bitext* section and create a new Spanish–English parallel corpus. Give it the name *newstest2008* and append the contents of the files *news-test2008.en* and *news-test2008.es* to it. Be careful: *MTradumàtica* allows you to append a couple of files that do not contain the same number of lines to a parallel corpus (*Bitext*). It will just truncate the longer file before appending it so as to ensure that the parallel corpus contains the same number of lines in both sides.

8. Now we are ready to train our SMT system with an improved language model: go to the *Translators* section and create a new Spanish-to-English translator from the parallel corpus (*Bitext*) and language model that you have just created. Give it the name *newstest2008-newscommentary*. You will see the number of lines of the parallel corpus: make sure that it contains 2 051 lines.

9. Go to the *Translate* section and translate again the sentence *La contaminación es un problema en una economía global.* with the new translator. You should have obtained a translation similar to *The contaminación is a problem in a global economy*.

We identified the origin of the translation error *economy global*: there is not an entry for *economía global* in the phrase table and the language model assigns the same perplexity to translation hypotheses with *economy global* and with *global economy*. We fixed the problem by training a system with additional monolingual data and made the language model assign lower perplexities to the translation hypotheses that contain *global economy*. Now, we are going to use the parallel corpus we obtained with Bicrawler in order to allow the system to translate *contaminación*. We are going to create a system that contains *contaminación* in its phrase table.

Task 7. Train an SMT system from the concatenation of multiple files [10 min.]

1. Download from Bicrawler the parallel corpus obtained from `http://sngular.team` in Moses format.
2. Go to the *Files* section in *MTradumàtica* and upload the files `sngular.team-en-es.en` and `sngular.team-en-es.es`.
3. Create a English monolingual corpus called *newstest2008-newscommentary-sngular* from the concatenation of the files `news-test2008.en`, `newscommentary-subset.en` and `sngular.team-en-es.en`.[2]
4. Train a language model from the monolingual corpus and give it the same name. Make sure it contains 6 440 lines before training.
5. Go to the *Inspect* section and make sure that the new language model is still able assign lower perplexities to the hypotheses with *global economy*.
6. Create a parallel corpus (*Bitext*) called *newstest2008-sngular* from the concatenation of the files *news-test2008* and *sngular.team-en-es*. Make sure it contains 3 440 lines.
7. Train an SMT system called *newstest2008-newscommentary-sngular* from the parallel corpus and language model you have created.
8. Go to the *Inspect* section and check whether there is an entry for the source word *contaminación* in the phrase table of the new translator. Will the translation of our test sentence improve?
9. Translate again the sentence *La contaminación es un problema en una economía global.* with the new translator. You should have obtained a translation similar to *The pollution is a problem in a global economy*. The word *contaminación* has been translated as *pollution* thanks to a new entry in the phrase table.

[2] Concatenate the files in this particular order. Otherwise, obtaining the same translations as in this guide is not guaranteed.

We finally fixed the two main errors in our example source language sentence. However, there is an important feature of *MTradumàtica* we have not used in this process: tuning. When we tune an SMT system we set the weight of each model so as to maximise translation quality on a development set. In our last step in this guide about *MTradumàtica*, we are going to test the effect of tuning.

Task 8. Tune an SMT system [10 min.]

1. Train another SMT system called *newstest2008-newscommentary-sngular-tuned* from the language model *newstest2008-newscommentary-sngular* and the parallel corpus *newstest2008-sngular*.
2. Go to the *Files* section and upload the files `newstest2011.head.en` and `newstest2011.head.es` from the `mtradumatica` folder in the workshop materials. They are going to be our development set.
3. Create a parallel corpus called *newstest2011-development* from the two files.
4. Go to the *Translators* section and click on the *Optimize* button of the SMT system *newstest2008-newscommentary-sngular-tuned*. Select the *newstest2011-development* bitext. You will see a time counter with a blue background, as during the traing phase.
5. Once the system is optimized (tuned), you can check that its output differs from that of the untuned system. You will notice some changes using the source language sentence *La contaminación es un importante problema en una economía global*. You may extract more source language sentences from the document `Dossier-OLE-castellano.docx` that you can find in the `mtradumatica` folder in the workshop materials.[3]

Was training, tunning and inspecting SMT system easy to you with *MTradumàtica*? What would you like to see as a feature in *MTradumàtica* in the future (or to remove)? Let's share our findings together!

Recap and useful info

In this workshop we have introduced you to some useful tools to create translation memories from multilingual websites and easily train statistical machine translation systems.

[3] The translation of full documents will be enabled soon.

We thank you for your participation. We encourage you to use our tools and help us improving.

Your feedback will be very welcome. To send it, please contact us at info dot prompsit dot com.

License

Please contact Víctor M. Sánchez-Cartagena (vmsanchez at prompsit dot com) for a copy of the source files.

[4] © Prompsit Language Engineering.