# Thot Toolkit for Statistical Machine Translation

## User Manual

Daniel Ortiz Martínez
dortiz@dsic.upv.es
Universitat Politècnica de València

April 2014

# CONTENTS

CHAPTER 1

# INTRODUCTION

Thot is an open source toolkit for statistical machine translation. Originally, Thot incorporated tools to train phrase-based models. The new version of Thot now includes a state-of-the-art phrase-based translation decoder as well as tools to estimate all of the models involved in the translation process. In addition to this, Thot is also able to incrementally update its models in real time after presenting an individual sentence pair.

## 1.1 Features

The toolkit includes the following features:

- Phrase-based statistical machine translation decoder.
- Computer-aided translation (post-edition and interactive machine translation).
- Incremental estimation of all of the models involved in the translation process.
- Client-server implementation of the translation functionality.
- Single word alignment model estimation using the incremental EM algorithm.
- Scalable implementation of the different estimation algorithms using Map-Reduce.
- Compiles on Unix-like and Windows (using Cygwin) systems.
- Integration with the CasMaCat Workbench developed in the CasMaCat project[a]
- ...

[a] `http://www.casmacat.eu/`

## 1.2 Distribution Details

Thot has been coded using C, C++ and shell scripting. Thot is known to compile on Unix-like and Windows (using cygwin) systems. As future work we plan to port the code to other platforms. See Section 1.4 section of this file if you experience problems during compilation.

It is released under the GNU Lesser General Public License (LGPL)[b].

## 1.3 Current Status

The Thot toolkit is under development. Original public versions of Thot date back to 2005 [1] and did only include estimation of phrase-based models. By contrast, current version offers several new features that had not been previously incorporated.

A basic usage manual is currently being developed. In addition to this, a set specific tools to ease the process of making SMT experiments has been created.

In addition to the basic usage manual, there are some toolkit extensions that will be incorporated in the next few months:

- Virtualized language models (i.e. accessing language model parameters from disk).
- Interpolation of language and translation models.
- Improved concurrency in the Thot translation server (translation process is not still concurrent).

Finally, here is a list of known issues with the Thot toolkit that are currently being addressed:

- Phrase model training is based on HMM-based alignments models estimated by means of incremental EM. This estimation process is computationally demanding and currently constitutes a bottleneck when training phrase models. One already implemented solution is to carry out the estimation in multiple processors. However we are also investigating to improve the efficiency of the estimation algorithm.
- Log-linear model weight adjustment is carried out by means of the downhill simplex algorithm, which is very slow. Downhill simplex will be replaced by a more efficient technique.
- Non-monotonic translation is not yet sufficiently tested, specially with complex corpora such as Europarl.

## 1.4 Support

Project documentation is being developed. Such documentation include:

- README file included with the Thot package.

[b] `http://www.gnu.org/copyleft/lgpl.html`

- The Thot manual ("thot_manual.pdf" under the "doc" directory).

If you need additional help, you can:

- use the github issue tracker[c].
- send an e-mail to the author[d].
- join the CasMaCat support group[e].

Additional information about the theoretical foundations of Thot can be found in:

- Daniel Ortiz-Martínez. Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation. *PhD Thesis*. Universitat Politècnica de València. Advisors: Ismael Garca Varea and Francisco Casacuberta. 2011.

One interesting feature of Thot, incremental (or online) estimation of statistical models, is also described in the following paper:

- Daniel Ortiz-Martínez, Ismael García-Varea, Francisco Casacuberta. Online learning for interactive statistical machine translation. In *Proceedings of the North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pages 546–554, Los Angeles, USA, June 2010.

## 1.5 Citation

You are welcome to use the code under the terms of the license for research or commercial purposes, however please acknowledge its use with a citation:

- Daniel Ortiz-Martínez, Francisco Casacuberta. The New Thot Toolkit for Fully Automatic and Interactive Statistical Machine Translation. In *14th Annual Meeting of the European Association for Computational Linguistics: System Demonstrations*, Gothenburg, Sweden, April 2014.

Here is a BiBTeX entry:

```
@InProceedings{Ortiz2014,
  author    = {D.~Ortiz-Mart\'{\i}nez and F.~Casacuberta},
  title     = {The New Thot Toolkit for Fully Automatic and
               Interactive Statistical Machine Translation},
  booktitle = {14th Annual Meeting of the European Association for Computational
               Linguistics: System Demonstrations},
  year      = {2014},
  month     = {April},
  address   = {Gothenburg, Sweden},
  pages     = "To appear",
}
```

[c] https://github.com/daormar/thot/issues
[d] daormar2@gmail.com
[e] http://groups.google.com/group/casmacat-support/boxsubscribe

## 1.6 Acknowledgements

Thot is currently supported by the European Union under the CasMaCat research project. Thot has also received support from the Spanish Government in a number of research projects, such as the MIPRCV project[f] that belongs to the CONSOLIDER programme [g].

[f] http://miprcv.iti.upv.es/
[g] http://www.ingenio2010.es/

CHAPTER 2

# INSTALLATION

## 2.1 Basic Installation

The code of the Thot toolkit is hosted on github [a].

To install Thot, first you need to install the autotools (autoconf, autoconf-archive, and automake packages in Ubuntu). If you are planning to use Thot on a Windows platform, you also need to install the Cygwin environment.

Once the autotools are available (as well as Cygwin if required), you can proceed with the installation of Thot by following the next sequence of steps:

1. Obtain the package using git:

```
$ git clone https://github.com/daormar/thot.git
```

   Additionally, Thot can be downloaded in a zip file[b].

2. `cd` to the directory containing the package's source code and type `./reconf`.
3. Type `./configure` to configure the package
4. Type `make` to compile the package
5. Type `make install` to install the programs and any data files and documentation
6. You can remove the program binaries and object files from the source code directory by typing `make clean`

[a] https://github.com/

[b] https://github.com/daormar/thot/archive/master.zip

By default the files are installed under the /usr/local/ directory (or similar, depending of the OS you use); however, since Step 5 requires root privileges, another directory can be specified during Step 3 by typing:

```
$ configure --prefix=<absolute-installation-path>
```

For example, if "user1" wants to install the Thot package in the directory "/home/user1/thot", the sequence of commands to execute should be the following:

```
$ ./reconf
$ configure --prefix=/home/user1/thot
$ make
$ make install
```

In order to build and use Thot on Windows platforms, the linux-like environment called Cygwin must be downloaded and installed previously.

See "INSTALL" file in the directory where Thot has been downloaded for more information.

## 2.2 Installation Including the CasMacat Workbench

Thot can be combined with the CasMaCatWorkbench that is being developed in the project of the same name. The specific installation instructions can be obtained at the project website[c].

[c] http://www.casmacat.eu/index.php?n=Workbench.Workbench

CHAPTER 3

# USER GUIDE

Prior to be able to perform fully-automatic or interactive machine translation, it is necessary to carry out a series of steps related to model estimation and parameter tuning. In the following sections we describe the different tools offered by the Thot toolkit to fully exploit its functionality.

## 3.1 Language Model Training

TBD (sorry for the inconvenience)

## 3.2 Translation Model Training

TBD (sorry for the inconvenience)

## 3.3 Parameter Tuning

TBD (sorry for the inconvenience)

## 3.4 Fully Automatic Translation

TBD (sorry for the inconvenience)

### 3.4.1 Command Line Tools

TBD (sorry for the inconvenience)

### 3.4.2 Client-Server tools

TBD (sorry for the inconvenience)

## 3.5 Interactive Machine Translation

TBD (sorry for the inconvenience)

# BIBLIOGRAPHY

[1] Ortiz, D., García-Varea, I., and Casacuberta, F. (2005). Thot: a toolkit to train phrase-based statistical translation models. In *Proceedings of the Machine Translation Summit X*, pages 141–148, Phuket, Thailand.