The New THOT Toolkit for Fully-Automatic and Interactive Statistical Machine Translation

Daniel Ortiz-Martínez
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46071 Valencia, Spain
dortiz@dsic.upv.es

Francisco Casacuberta
Dpto. de Sist. Inf. y Comp.
Univ. Politéc. de Valencia
46071 Valencia, Spain
fcn@dsic.upv.es

Abstract

We present the new THOT toolkit for fully-automatic and interactive statistical machine translation (SMT). Initial public versions of THOT date back to 2005 and did only include estimation of phrase-based models. By contrast, the new version offers several new features that had not been previously incorporated. The key innovations provided by the toolkit are computer-aided translation, including post-editing and interactive SMT, incremental learning and robust generation of alignments at phrase level. In addition to this, the toolkit also provides standard SMT features such as fully-automatic translation, scalable and parallel algorithms for model training, client-server implementation of the translation functionality, etc. The toolkit can be compiled in Unix-like and Windows platforms and it is released under the GNU Lesser General Public License (LGPL).

1 Introduction

Open-source software constitutes a valuable resource for researchers or companies. Due to the inherent difficulties of developing good quality software (correct, efficient, modular, extensible, well-documented, etc.), there are interesting research ideas that not always receive enough attention from the open-source software community.

We present the THOT toolkit for statistical machine translation (SMT). The first public version of THOT was initially created in 2005 (Ortiz et al., 2005) and its functionality was restricted to train phrase-based models (Koehn et al., 2003). Here we present a new version of THOT which includes several new features related to phrase-based translation. More specifically, the set of features provided by THOT can be classified into advanced features and standard features. Advanced features correspond to sophisticated functionality that has received poor or no attention in existing SMT toolkits. By contrast, standard features correspond to functionality already provided by popular tools such as Moses (Koehn et al., 2007). In this regard, THOT neither is based on Moses nor shares any source code with it.

THOT includes the following advanced features:

- Computer-aided translation, including post-editing and interactive machine translation (IMT). This functionality has been integrated in a translation tool developed in the CasMaCat project[1] (the so-called CasMaCat Workbench).
- Incremental estimation of all of the models involved in the translation process.
- Robust generation of phrase-based alignments.

Computer-aided translation and more specifically two of its applications, post-editing and IMT, constitute a field of increasing interest in SMT. In particular, IMT has been studied in numerous research papers during the last years. In spite of this, this application has not previously been implemented in open-source software tools.

Incremental (or online) learning is a hot research topic in SMT due to the great interest of quickly incorporating incoming data into existing translation systems. In spite of the fact that the Moses toolkit already implements incremental learning techniques, such techniques are designed to work by incrementally processing large blocks of data and not in a sentence-wise manner, as it is pointed out in (Mirking and Cancedda, 2013). By

[1] http://www.casmacat.eu/

Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 45–48,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

contrast, the incremental learning techniques implemented by THOT allows to process new training samples individually in real time.

Finally, the necessity of generating phrase-level alignments is present in a wide range of tasks, from multisource SMT to discriminative training. However, as far as we know this functionality also is not included in existing SMT tools.

In addition to the above mentioned advanced features, THOT offers a set of standard features:

- Phrase-based SMT decoder.
- Scalable training and search algorithms.
- Client-server implementation.
- Miscellaneous SMT tools

2 The THOT toolkit

THOT can be downloaded from GitHub[2] and is distributed under the GNU Lesser General Public License (LGPL). It has been developed using C++ and shell scripting. The design principles that have led the development process were:

- **Modularity**: The THOT code is organised into separate packages for each main functional component (training of phrase-based and language models, decoding, etc.). Each component can be treated as a stand-alone tool and does not rely on the rest of the code.
- **Extensibility**: The functionality provided by each package is structured into classes. Abstract classes are used when appropriate to define the basic behaviour of the functional components of the toolkit, allowing us to easily extend the toolkit functionality.
- **Scalability**: THOT is able to train statistical models from corpora of an arbitrary size. Moreover, the toolkit takes advantage of parallel and distributed computing to reduce the time cost of the implemented algorithms. Additionally, the parameters of the resulting models can be pruned or accessed from disk during the decoding process.
- **Portability**: It is known to compile on Unix-like and Windows (using Cygwin) systems.

In the rest of the paper we give additional details about the different toolkit features that have been mentioned above.

3 Computer-Aided Translation

Current MT systems are not able to produce ready-to-use texts. Indeed, they usually require human post-editing in order to achieve high-quality translations. This motivates an alternative application of MT in which the MT system collaborates with the user to generate the output translations. This alternative application receives the name of computer-assisted translation (CAT).

CAT can be instantiated in different ways. The THOT toolkit incorporates tools that are useful in two different CAT instantations, namely, post-editing and interactive machine translation.

3.1 Post-Editing

Post-editing (PE) involves making corrections and amendments to machine generated translations (see (TAUS, 2010) for a detailed study). In the PE scenario, the user only edits the output of the MT system without further system intervention.

3.2 Interactive Machine Translation

In the IMT framework (Foster et al., 1997; Langlais et al., 2002), the user obtains her desired translations in a series of interactions with an MT system. Specifically, the system initially generates a translation without human intervention and after that, the user validates a prefix of the translation and introduce the next correct character of it. With this information, the IMT system returns the suffix which best completes the user prefix. This process is repeated until the user gets the sentence she has in mind. In (Barrachina et al., 2009), SMT techniques were embedded within the interactive translation environment.

A common problem in IMT arises when the user sets a prefix which cannot be explained by the statistical models. This problem requires the introduction of specific techniques to guarantee that the suffixes can be generated. The majority of the IMT systems described in the literature use error-correcting techniques based on the concept of edit distance to solve the coverage problems. Such error-correction techniques, although they are not included in the statistical formulation of the IMT process, are crucial to ensure that the suffixes completing the user prefixes can be generated.

THOT implements an alternative formalisation that introduces stochastic error-correction models in the IMT statistical formulation. Such a formalisation was introduced in (Ortiz-Martínez, 2011)

[2] `https://github.com/daormar/thot`

and it generates the suffixes required in IMT by partially aligning a prefix of the target hypotheses with the user prefix. Once the partial alignment is determined, the suffix is given by the unaligned portion of the target sentence.

Experiments to test the above mentioned IMT proposal were carried out using THOT. The results showed that the proposed IMT system outperforms the results of other state-of-the-start IMT systems that are based on word graphs (see (Ortiz-Martínez, 2011) for more details).

3.3 Integration with the CasMaCat Workbench

THOT can be combined with the CasMaCat Workbench[3] that is being developed within the project of the same name. The CasMaCat Workbench offers novel types of assistance for human translators, using advanced computer aided translation technology that includes PE and IMT.

4 Incremental Learning for SMT

Thot incorporates techniques to incrementally update the parameters of the statistical models involved in the translation process. Model updates can be quickly executed in a sentence-wise manner allowing the system to be used in a real time scenario. For this purpose, a log-linear SMT model where all its score components are incrementally updateable is defined. The implemented proposal uses the incremental version of the EM algorithm (Neal and Hinton, 1998) and the specific details can be found in (Ortiz-Martínez et al., 2010; Ortiz-Martínez, 2011).

Empirical results obtained with THOT and reported in (Ortiz-Martínez et al., 2010; Ortiz-Martínez, 2011) show that incremental learning allows to significantly reduce the user effort in IMT tasks with respect to that required by a conventional IMT system.

Additionally, the incremental learning techniques provided by THOT are currently being used in other sophisticated applications such as active learning for SMT (González-Rubio et al., 2012).

5 Generation of Phrase-Based Alignments

The generation of phrase-level alignments is interesting due to its utility in a wide range of applications, including multi-source SMT, Viterbi-like estimation of phrase-based models or discriminative training, just to name a few.

A very straightforward technique can be proposed for finding the best phrase-alignment. Specifically, the search process only requires a regular SMT system which filters its phrase table in order to obtain those target translations for the source sentence that are compatible with the given target sentence. Unfortunately, this technique has no practical interest when applied on regular tasks due to problems with unseen events.

To overcome the above-mentioned difficulty, an alternative technique that is able to consider every source phrase of the source sentence as a possible translation of every target phrase of the target sentence can be defined. The THOT toolkit implements the proposal described in (Ortiz-Martínez et al., 2008), which combines a specific search algorithm with smoothing techniques to enable efficient exploration of the set of possible phrase-alignments for a sentence pair.

Phrase-based alignment quality was difficult to evaluate since there is not a gold standard for this task. One way to solve this problem consists in refining the phrase alignments to word alignments and compare them with those obtained in existing shared tasks on word alignment evaluation. Results obtained with THOT reported in (Ortiz-Martínez et al., 2008) clearly show the efficacy of the implemented method.

6 Standard Features

THOT incorporates a number of standard features that are present in existing translation tools. Such standard features are briefly enumerated and described in the following paragraphs.

Phrase-Based SMT Decoder The toolkit implements a state-of-the-art phrase-based SMT decoder. The decoder uses a log-linear model with a complete set of components similar to those implemented in other tools such as Moses. Results reported in (Ortiz-Martínez, 2011) show that the translation quality obtained by THOT is comparable to that obtained by means of Moses.

Scalable Training and Search Algorithms Due to the increasing availability of large training corpora, it is necessary to implement scalable training and search algorithms. THOT incorporates tools to train statistical models from corpora

[3] See installation instructions at `http://www.casmacat.eu/index.php?n=Workbench.Workbench`

of an arbitrary size. Such tools can take advantage of the availability of multiple processors or computer clusters. The parameters of the resulting models can be pruned or accessed from disk during the decoding stage.

Client-Server Implementation An important part of the functionality provided by the toolkit can be accessed using a client-server model. This is a useful feature to build web applications offering SMT services.

Miscellaneous SMT tools THOT reduces dependencies with third-party software by integrating most critical components of a typical machine translation pipeline, from the estimation of phrase-based and language models to the generation of translations and their automatic evaluation. The estimation of word-alignment models using the incremental EM algorithm is also implemented by the toolkit.

7 Conclusions

THOT is an open-source toolkit for SMT designed for its use in Unix-like and Windows systems. It has been developed using C++ and shell scripting, and it is released under LGPL license. THOT incorporates three advanced features that have received little attention in previous publicly-available SMT tools, namely, interactive machine translation, incremental learning and generation of phrase-based alignments. Additionally, THOT also implements standard features such as training of statistical models or decoding. The functionality of the toolkit has been empirically tested, showing its efficacy in different SMT-related tasks.

Acknowledgments

Work supported by the European Union 7th Framework Program (FP7/2007-2013) under the CasMaCat project (grant agreement nº 287576), by Spanish MICINN under grant TIN2012-31723, and by the Generalitat Valenciana under grant ALMPR (Prometeo/2009/014).

References

S. Barrachina, O. Bender, F. Casacuberta, J. Civera, E. Cubel, S. Khadivi, A. L. Lagarda, H. Ney, J. Tomás, E. Vidal, and J. M. Vilar. 2009. Statistical approaches to computer-assisted translation. *Computational Linguistics*, 35(1):3–28.

G. Foster, P. Isabelle, and P. Plamondon. 1997. Target-text mediated interactive machine translation. *Machine Translation*, 12(1):175–194.

Jesús González-Rubio, Daniel Ortiz-Martínez, and Francisco Casacuberta. 2012. Active learning for interactive machine translation. In *Procs. of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 245–254.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Procs. of the Human Language Technology and North American Association for Computational Linguistics Conference*, pages 48–54, Edmonton, Canada, May.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Procs. of the 45th Annual Meeting of the Association for Computational Linguistics*, pages 177–180, Prague, Czech Republic, June.

P. Langlais, G. Lapalme, and M. Loranger. 2002. Transtype: Development-evaluation cycles to boost translator's productivity. *Machine Translation*, 15(4):77–98.

S. Mirking and N. Cancedda. 2013. Assessing quick update methods of statistical translation models. In *Procs. of International Workshop of Spoken Language Translation*, pages 264–271, Heidelberg, Germany.

R.M. Neal and G.E. Hinton. 1998. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Procs. of the NATO-ASI on Learning in graphical models*, pages 355–368, Norwell, MA, USA.

D. Ortiz, I. García-Varea, and F. Casacuberta. 2005. Thot: a toolkit to train phrase-based statistical translation models. In *Machine Translation Summit*, pages 141–148, Phuket, Thailand, September.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2008. Phrase-level alignment generation using a smoothed loglinear phrase-based statistical alignment model. In *Procs. of the European Association for Machine Translation*.

D. Ortiz-Martínez, I. García-Varea, and F. Casacuberta. 2010. Online learning for interactive statistical machine translation. In *Procs. of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 546–554.

D. Ortiz-Martínez. 2011. *Advances in Fully-Automatic and Interactive Phrase-Based Statistical Machine Translation*. Ph.D. thesis, Universidad Politécnica de Valencia.

TAUS. 2010. Postediting in practice. a TAUS report. Technical report, March.