EU-BRIDGE Partners

University of Edinburgh

Contacts

The University of Edinburgh
South Bridge
UK-Edinburgh EH8 9YL
www.ilcc.inf.ed.ac.uk
Philipp Koehn
pkoehn@inf.ed.ac.uk

Open Source Statistical Machine Translation

Description and Exploitable Knowledge

Moses is an open source statistical machine translation (SMT) project, started in 2005 at the University of Edinburgh. Since then, hundreds of researchers have contributed towards the project, and it has become the most widely adopted translation engine both as a baseline research system and for commercial use in industry.

There is extensive online documentation (http://www.statmt.org/moses) and there is an active mailing list (http://mailman.mit.edu/mailman/listinfo/moses-support) for support.

Technology / Application Sectors:

SMT models are the dominant technology for automatic machine translation and this is because they can leverage large amounts of human translated text. SMT systems can be trained within a few days, for any language pair, delivering state-of-the-art performance when enough training data is available. SMT models win annual machine translation competitions, and are deployed by the likes of Google and Microsoft. SMT models are trained on parallel corpora, where human translated sentences in the source and target language are aligned. By using parallel corpora, Moses is able to discover which words or phrases are good translations of each other by looking at how often they co-occur. SMT models also use large amounts of monolingual text in the target language to learn what sentences in the target language should look like.

Components

The two main components in Moses are the training pipeline and the decoder. There are also a variety of contributed tools and utilities. The training pipeline is really a collection of tools (mainly written in perl, with some in C++) which take the raw data (parallel and monolingual) and turn it into a machine translation model and a model of the target language. The decoder is a single C++ application which, given a trained machine translation model and a source sentence, will translate the source sentence into the target language.

Development

Moses is an open-source project, licensed under the LGPL, which incorporates contributions from many sources. There is no formal management structure in Moses, so anyone is welcome to contribute. For those interested in getting involved, there is a list of possible project on the Moses website. The annual MT Marathon is also a good way to learn about the Moses project.

In general, the Moses administrators are fairly open about giving out push access to the git repository. This means that trunk occasionally breaks, but given the active Moses user community, it does not stay broken for long. The nightly builds and tests of trunk are reported on the cruise control web page, but for more stable versions, official releases are also available.

Moses in Use

The liberal licensing policy in Moses, together with its wide coverage of current SMT technology and complete tool chain, make it probably the most widely used open-source SMT system. It is used in teaching, research, and, increasingly, in commercial settings.
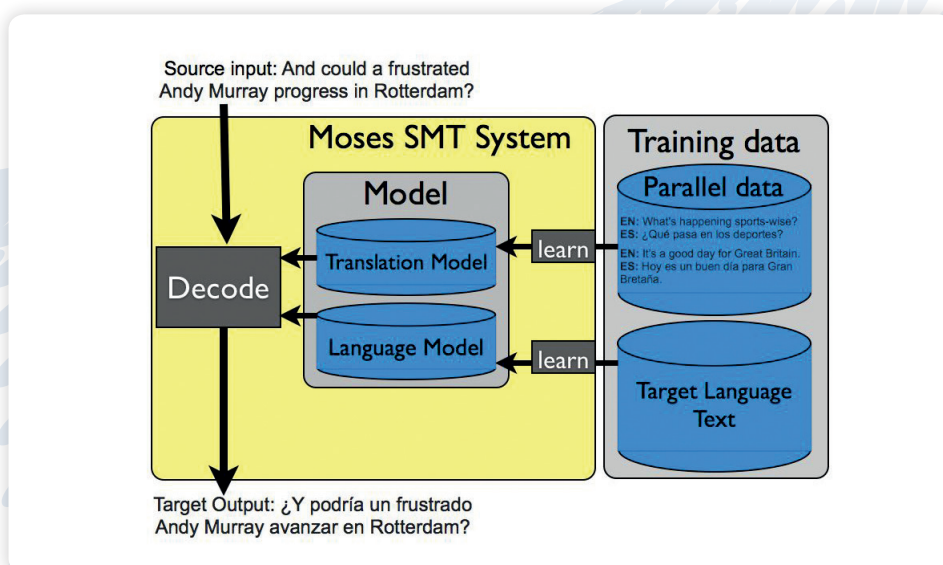
Commercial use of Moses is promoted and tracked by TAUS. The most common current use for SMT in commercial settings is post-editing where machine translation is used as a first-pass, with the results then being edited by human translators. This can often reduce the time (and hence total cost) of translation. There is also work on using SMT in computer-aided translation, which is the research topic of two current EU projects, Casmacat and MateCat. In the EU-Bridge project, Moses is used for spoken language translation and for punctuating and segmenting ASR output.

Terms of Availability

Can be inquired at the University of Edinburgh (Philipp Koehn)

IPR Protection

University of Edinburgh (Philipp Koehn)

Moses diagram