

MOSES CORE

2014

Annual Public Report

 TAUS



Charles University,
Prague



THE UNIVERSITY
of EDINBURGH

CAPITA



FONDAZIONE
BRUNO KESSLER

1 Overview

The diversity of languages in Europe makes translation vitally important to the economic, cultural and social lives of Europeans. Machine translation (MT) provides a way of fully or partially automating the translation process, and hence reducing the costs and enabling more text and speech to be translated.

Machine translation, however, is a complex field and presents many substantial barriers for entry to potential researchers, and users of the technologies. The principal aim of MosesCore is to reduce these barriers, making it easier to join and participate in the MT research community, and to become an MT user.

MosesCore achieves these aims by organising a variety of events targeted at users, developers and researchers of MT, and by promoting and coordinating the development and use of open-source MT tools, in particular the Moses toolkit.

In this report we will describe the events organised by MosesCore during 2014, as well as the main developments in Moses during this timeframe.

1.1 Key Facts

Project type	FP7 Coordination Action
Duration	February 1st 2012 - January 31st 2015
Financing	€1.2M
Contact	Barry Haddow (info@mosescore.eu)

1.2 Partners

University of Edinburgh	United Kingdom
TAUS	Netherlands
Charles University, Prague	Czech Republic
Fondazione Bruno Kessler	Italy
Capita Translation and Interpreting (formerly Applied Language Solutions)	United Kingdom

1.3 Beneficiaries

Researchers have events in which they can showcase their research, compare their systems with others, and gather to implement new MT tech-

niques. They also have a state-of-the-art open-source platform to test out their ideas on.

Users and Developers have a stable and well supported open-source MT toolkit, and have forums to learn about new research developments in MT and share system building and deployment experience.

Everyone benefits from improved information exchange between developers, users and researchers.

2 Events

2.1 Machine Translation Marathon

After successfully running Machine Translation Marathons in Edinburgh (2012) and Prague (2013), the final MTM of MosesCore was held in Trento, organised by FBK. In this week-long event, MT researchers, developers and users gather for a mixture of talks, tutorials, hacking and informal discussions. The Trento MTM was the ninth edition of this event, and attracted around 90 participants.

The Marathon program (see <http://www.statmt.org/mtm14> for details) consisted of a variety of elements targeted at attendees with different levels of experience, and different types of interest in MT.

The “summer school” offered a comprehensive introduction to SMT, with lectures from world-leading researchers in the field. This year the summer school supplemented the lectures on standard SMT models with others on more advanced topics such as deep learning, discourse, morphology and post-editing. The lectures were also accompanied by practical labs on topics such as evaluation, morphology and decoding.

The daily keynote talks had a strong practical element, with three given by organisations that deploy MT systems for real-world translation problems (Joao Graça of Unbabel, Marco Trombetti of Translated.net and Bruno Pouliquen of WIPO). The academic keynotes came from Marine Carpuat of NRC Canada and Francisco Casacuberta of UPV, who spoke on domain adaptation and interactive MT, respectively.

In the open-source convention, MT researchers and developers had an opportunity to publish papers describing new tools in MT, and extensions

to existing tools. We accepted 9 papers covering topics such as MT models, post-editing, evaluation and MT education, which were presented at a poster/demo session.

Finally, but most importantly for the experienced MTM attendees, were the open-source hacking projects. The idea here was that any participant could propose an open-source SMT project at the start of the week, and try to attract a team of developers for the week. Each small group worked on the project through the week, culminating in the final project presentations at the end of the MTM. Projects covered a range of MT problems, including language modelling, new decoding algorithms, crowd-sourcing and computer-aided translation. Some aimed to kick off new research directions and collaborations, whilst others sought to add missing features to established MT systems like Moses and Joshua.



2.2 Industrial Outreach Events

2.2.1 MT Showcases

In 2014 MosesCore partner TAUS organised three events to showcase the use of MT in general and Moses specifically.

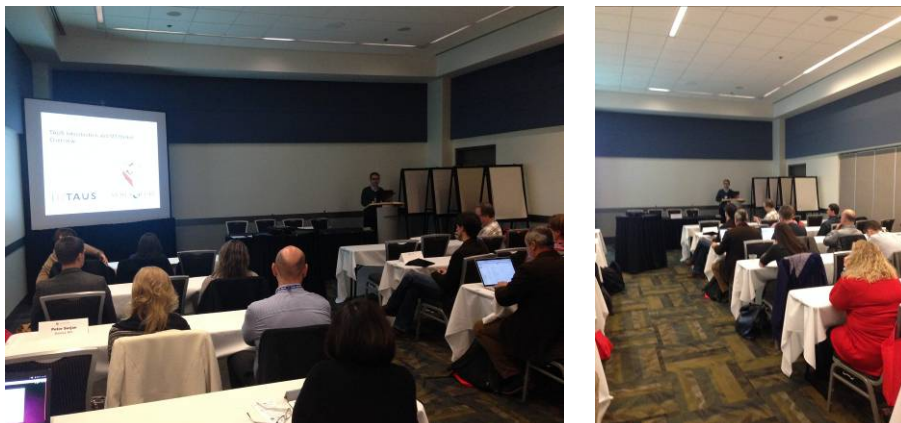
MT Showcase Dublin On 4 June, TAUS had a successful gathering of industry peers and academic representatives in Dublin (Ireland). 60 participants took part in this event and had an opportunity to discuss different MT customisation and optimisation techniques during the question and answer session of the showcase.

The following use cases were presented at the Dublin event: European Commission, Iconic Translation Machines, KantanMT, Sovee and Tilde. All presenters use Moses as the basis of their MT systems.



MT Showcase Vancouver The second MT Showcase took place in Vancouver (Canada) at the pre-conference day of Localization World on 29 October with 41 registered attendees.

TAUS invited a diverse set of presenters showing the use of MT for cross-border commerce (eBay), on-premise MT solutions (Precision Translation Tools), a post-editing tool with online learning of corrections (MateCat presented by Translated) and distributed crowd-sourced post-editing (Unbabel). As in Dublin all offerings are based on Moses, showing the vibrancy of the market of Moses-based solutions. TAUS presented a preview of research into the Moses MT market, research which will be published in an upcoming MosesCore report.

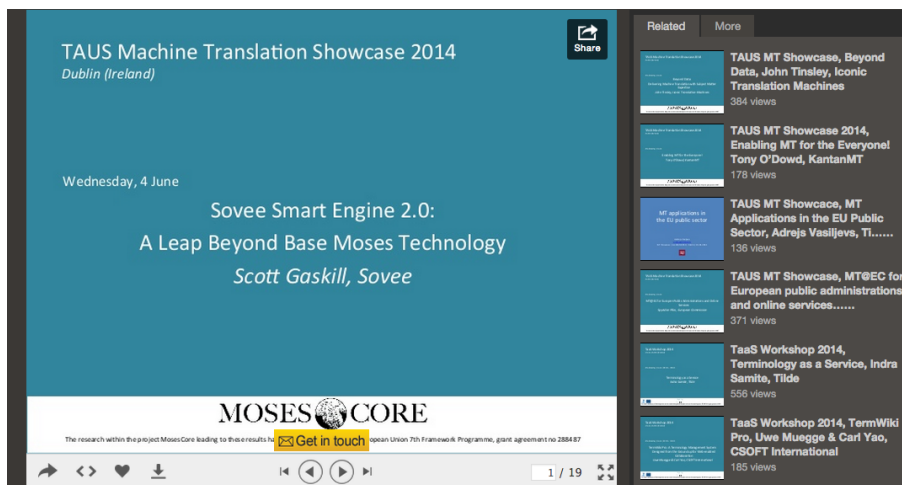


MT Discussion at the VViN Conference To meet the wishes of the participants of the previous MT Showcases about the shortage of similar events in Europe, on 19 September TAUS took part in the VViN Conference in The Hague (Netherlands). The MT discussion initiated by the TAUS team attracted 20 attendees, who were eager to chat about open-source MT, Moses and adaptation of MT in general.

Feedback and Results To capture feedback and questions and in turn to assess the impact of the events, we started to record audio at the MT showcase events. We also asked attendees to fill in a paper survey. This attendee input will be included in summary in the report on industry outreach events (D4.7).

The uses cases presented at the MT Showcase events are publicly available on the MosesCore project website, and from TAUS Labs and the TAUS Slideshare

Since the initial publication of MT use cases in June 2014, these presentations have been viewed a total of 1,669 times (7 November 2014), with an average of 166 views per case.



The image shows a Slideshare interface. The main slide is titled "TAUS Machine Translation Showcase 2014" and "Dublin (Ireland)". It features the text "Wednesday, 4 June", "Solve Smart Engine 2.0: A Leap Beyond Base Moses Technology", and "Scott Gaskill, Solve". The MosesCore logo is at the bottom. To the right, a "Related" list shows several other presentations with their view counts:

- TAUS MT Showcase, Beyond Data, John Tinsley, Iconic Translation Machines (384 Views)
- TAUS MT Showcase 2014, Enabling MT for the Everyone! Tony O'Dowd, KantanMT (178 Views)
- MT applications in the EU public sector (136 Views)
- TAUS MT Showcase, MT@EC for European public administrations and online services..... (371 Views)
- TasS Workshop 2014, Terminology as a Service, Indra Samite, Tilde (556 Views)
- TasS Workshop 2014, TermWiki Pro, Uwe Muegge & Carl Yao, CSOFT International (185 Views)

The findings of the question and answer sessions of the MT Showcases and the MT discussion in The Hague will be reflected in the Moses MT Market report to be published by MosesCore in December 2014.

2.2.2 Moses Industry Round Table

At the first Moses Industry Round Table in 2013 at the Machine Translation Marathon in Prague, TAUS brought together the Moses developer community and Moses users from industry and governments to discuss common challenges and opportunities for cooperation to tackle common issues. With direct public funding for Moses ending next year we wanted to continue the conversation and explore stewardship options to ensure continued sustainable research, development and distribution for this popular open source MT toolkit.

To enable a broad mix of stakeholders from industry, government and academia to attend, we collocated the second Moses Industry Roundtable with the AMTA 2014 conference. This conference was followed by the large language industry conferences TAUS and Localization World. We had 37 registered attendees.

The Round Table started with short presentations by TAUS about the current state of the market for Moses-backed MT solutions and the use of Moses in the industry, Hieu Hoang from the University of Edinburgh detailed the improvements in Moses since the start of the MosesCore project and Ulrich Germann (also University of Edinburgh) outlined future plans for Moses use and improvements in upcoming projects.

The main focus of the Round Table however was discussions among the stakeholders. After discussing organisational and technical challenges in two separate breakout groups the Round Table came together again to discuss options for stewardship of this essential MT resource. TAUS as a discussion facilitator captured the breakout notes and audio recorded the discussions. We hope that the discussion results provide a stepping stone for continued maintenance, support and development of Moses, also with additional non-governmental funding given the increased use of Moses by industry.

The results of the Moses Industry Round Table discussions will be included in the upcoming Moses MT Market Report.

2.3 Workshop on Machine Translation

The Workshop in Machine Translation (WMT) and its associated shared tasks collectively provide an important synchronisation point for the MT research community. MosesCore plays a crucial role in WMT, providing overall coordination of the workshop and tasks, and funding the management

and test set creation for the main translation task.

The shared tasks allow researchers from around the world to compare their techniques, using standard benchmarks, and to report their results at the workshop. The data used for the shared tasks, as well as all the outputs from the task participants, is made freely available for ongoing research.

The shared task campaign ran from December 2013 to April 2014, and this year featured 4 different shared tasks:

News Translation This is the MosesCore-sponsored translation task, where participants had to use their systems to translate common test sets, consisting of published news text. The language pairs for this year were English to and from Czech, French, German, Hindi and Russian, with Russian test sets supplied by Yandex. The translations are submitted to extensive human evaluation.

Metrics In this MosesCore-sponsored task, participants build automated systems for assessing the quality of a translation, using a reference, and their judgements are compared to those of humans.

Quality Estimation This task consisted of several subtasks, all of which were concerned with measuring the quality of an MT system in the absence of a reference. The task was supported by [QTLaunchPad](#).

Medical Translation In this translation task, participants translated medical texts, including multilingual queries. The task was supported by [Khresmoi](#).

In all we had 23 teams participating in the news translation task, and 11 teams in the metrics task, drawn from North America, Europe, the Middle East and Asia. There were a total of 70 submissions for news translation, spread across all language pairs, plus a further 22 for the medical translation task.

Human evaluation for the news translation task submissions used the [Appraise](#) tool to collect comparative judgements. This year we managed to collect sufficient judgements from researchers, to obviate the need for (sometimes lower quality) crowd-sourced judgements.

The workshop itself took place in June 2014, attached to the Association for Computational Linguistics (ACL) conference in Baltimore. As in previous years, WMT was one of the largest workshops at ACL with an attendance

of around 100. As well as the poster presentations from shared tasks, there were 12 research paper presentations, an invited talk from Alon Lavie (CMU and Safaba) and a panel discussion on the latest successful MT techniques. The workshop program is available at <http://www.statmt.org/wmt14>.

Next year's WMT will take place after MosesCore ends, but MosesCore is again providing the test sets for the shared (news) translation task, which will be kicked off in January 2015. For this task we will be making the following changes:

- Finnish will replace Hindi, to provide an example of a challenging European language which has poor MT support.
- The French-English test sets will be drawn from the reader comments on news articles, to provide a different, and perhaps more challenging, text genre.

3 The Moses Toolkit

3.1 Background

Moses¹ is an open-source toolkit for building statistical machine translation systems. It provides tools to train such systems from parallel data, and a decoder to translate sentences using models trained with the toolkit. The two main statistical MT paradigms (phrase-based and hierarchical/syntactic) are both implemented in Moses, and its extensible architecture means that it is able to absorb many of the advances in MT published in the literature. Being licensed under the liberal [LGPL](#) makes it easy to incorporate Moses into commercial applications, whilst preserving the ability to redistribute its source code, making it attractive for both academic and commercial users.

The MosesCore project aims to retain Moses' place as (arguably) the most popular open-source SMT toolkit by continuing to incorporate new research, whilst improving stability and support. It has funded the appointment of a "Moses Coordinator" (Hieu Hoang) to oversee Moses development.

¹<http://www.statmt.org/moses>

3.2 Releases

In January 2014, thanks to the support of MosesCore, we made the second major release of Moses (v2.1.1). This release incorporated many new features from the research community (fully described in the [Release Notes](#)) as well as extensive refactoring by the Moses Coordinator.

As well as the source-code version available in github, we make available several binary packages for popular Linux distributions, OSX and Windows (Cygwin). Furthermore, the MosesCore partner Capita has created a native [Windows version of Moses](#) with an installer and a GUI, mainly for demo purposes.

3.3 Current Development

Moses has continued to incorporate new research, useful features for SMT system builders, as well as bug fixes and tweaks. In 2014 we have seen 1400 commits by the time this report was prepared. Some highlights of the year are:

- **Performance Improvements** Machine translation is a resource-hungry application, and as new features are merged into Moses, there is a danger that the speed of the decoder could be adversely affected. This happened in v2.1, so to help reduce the risk of such regression in the future we have developed a suite of performance tests. We also tracked down and fixed a particularly troublesome multi-thread performance problem caused by certain versions of the C++ standard libraries, and fixed a separate threading problem in the Moses server. Decoding with syntax-based models has been speeded-up, as we explain below.
- **Dynamic Suffix Arrays** This is a new facility in Moses which enables you to incorporate extra training data into your translation models, without running the expensive batch retraining. It enables on-the-fly updates of Moses models, for example in post-editing scenarios.
- **Neural Network Language Models** There has been a resurgence of interest in neural networks in the natural language research community, and Moses now supports two different neural network LMs: OXLM (from Oxford University) and NPLM (from ISI in California).

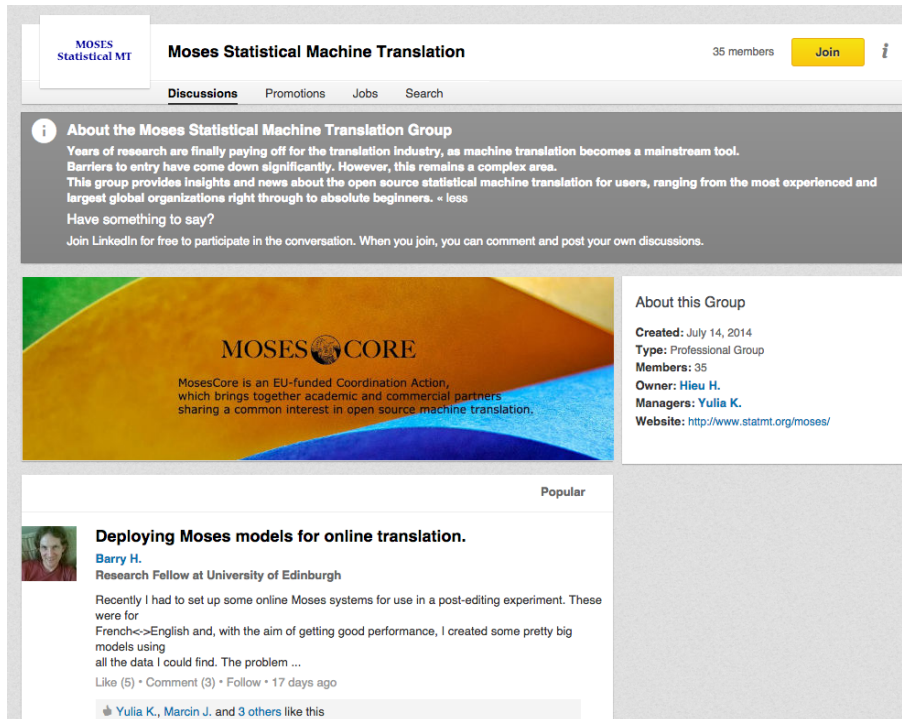
- **Improved Syntactic Models and Decoding** There have been a lot of improvements to syntax-based models in Moses, enabling new variants that, for example, can incorporate both source and target syntax. We have also have a new syntax-based decoder into the toolkit, which has been shown to be 40% faster on standard data sets.
- **Transliteration** This is necessary when translating between languages that do not share a common script. A transliteration feature has been integrating into Moses which is able to translate unknown words character-by-character.

4 Communications

4.1 Social Media

Moses has presences on [twitter](#) (@MosesSMT) , [LinkedIn](#), [Facebook](#) and [Google+](#).

In 2014 we have focused our social media strategy, concentrating on building up the Moses Linked group (now at 35 members) and posting relevant announcements on twitter (479 followers). There had been an effectively dormant Moses LinkedIn group for a few years, but in 2014 MosesCore created a new one, with an open membership policy.



MOSES Statistical MT **Moses Statistical Machine Translation** 35 members [Join](#)

[Discussions](#) [Promotions](#) [Jobs](#) [Search](#)

About the Moses Statistical Machine Translation Group
 Years of research are finally paying off for the translation industry, as machine translation becomes a mainstream tool. Barriers to entry have come down significantly. However, this remains a complex area. This group provides insights and news about the open source statistical machine translation for users, ranging from the most experienced and largest global organizations right through to absolute beginners. < less
 Have something to say?
 Join LinkedIn for free to participate in the conversation. When you join, you can comment and post your own discussions.

About this Group
Created: July 14, 2014
Type: Professional Group
Members: 35
Owner: Hieu H.
Managers: Yulia K.
Website: <http://www.statmt.org/moses/>

Popular

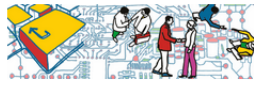
Deploying Moses models for online translation.
Barry H.
 Research Fellow at University of Edinburgh
 Recently I had to set up some online Moses systems for use in a post-editing experiment. These were for French->English and, with the aim of getting good performance, I created some pretty big models using all the data I could find. The problem ...
 Like (5) • Comment (3) • Follow • 17 days ago
 Yulia K., Marcin J. and 3 others like this

4.2 Digital Outreach

In the summer 2014 TAUS reassessed the communication strategy in relation to the Moses e-bulletins. From now on these newsletters are not only about the news, but they also include inspiring MT use cases and technical updates of the Moses toolkit, when possible. Combining various newsworthy aspects in one monthly digest, TAUS targets an audience who is interested in multiple aspects of MT adoption (development, operation, research).

MOSES CORE

Moses Travels Around the World



Moses in the Spotlight

In September MosesCore team took part in the **MT Marathon in Trento**. The Marathon was filled with both serious and fun moments. Have a look at some pictures

[Photo Gallery](#)

Moses User Survey

The deadline for the Moses User Survey is extended till **31 October**. Please help to make a difference in improving and growing the Moses SMT toolkit.

[Go to Survey](#)

Vancouver is Calling

AMTA 2014 Tutorials / Pre-Conference day (Wednesday, 22 October)

If you are a researcher or a commercial user interested in open-source machine translation but lack hands-on experience, check the presentation about the use of statistical machine translation with the Moses toolkit by Hieu Hoang, Matthias Huck, and Philipp Koehn.

[Register for AMTA 2014](#)

It's less than a week till the TAUS Moses Industry Roundtable in Vancouver. There are over 20 registrations and counting. Representatives from Google, John Deer, Intervium, Unbabel, Safaba Translation Solutions and many more will get together on 26 October for highly interactive discussions to help agree on priorities for potential industry cooperation around Moses. Make sure to secure your spot.

TAUS Moses Industry Roundtable
Vancouver (Canada), 26 October

[Register for Moses Roundtable](#)

TAUS Machine Translation Showcase
Vancouver (Canada), 29 October

Free session at the pre-Conference day of Localization World.

[Register for MT Showcase](#)

Moses Use Case

We continue to explore the Moses possibilities and its implementation by some of the market leaders. This month we take a look behind the scenes of MT implementation at Moravia.

COMPANY NAME

USE SCENARIO

MT is used for projects with a minimum volume of 10,000 words, with immediate training of an engine. The more mature engines are updated at least twice a year, in case of SMT, or at the end of each project, in case of RBMT or hybrid engines. Linguists and translators cooperate to improve MT quality.

MARKET POSITIONING

LexWorks boasts experience and independence as two of its main strengths. In terms of MT deployment, the Company calls itself technology agnostic: "We deploy the right engine for the right content and the right need." Depending on a number of factors (such as language combination, quality required, file format, content type) LexWorks offers four different MT approaches: RBMT pure, SMT pure, Hybrid and online SMT (e.g. Customized Microsoft Translator Hub). All solutions are offered in combination with PEMT. LexWorks is engaged in finding ways to build engines more quickly and scalably. According to Lori Thicke, the main threat to the MT market for LexWorks and other LSPs with MT in their portfolios comes from human translation services available for extremely low rates. The future of MT lies in the general acceptance of MT and, at the same time, the opening up of a great number of opportunities for high volume content that is not today being translated. In terms of economically less viable languages, the Company advocates solutions for scalable MT engines because local languages are gaining in economic importance. LexWorks is the founder of the non-profit organization Translators without Borders, dedicated to helping NGOs extend their humanitarian work by providing free, professional translations. Translators without Borders also builds MT engines for local languages ignored by mainstream players.

VIEWS ON CURRENT STATE OF MT

MT is gaining momentum, partly thanks to the increasing number of solutions providers. In this scenario, independence, experience and confidentiality of data remain strong points of the Company's offering.

QUOTE

"We look for the best of breed."

For Real Techies - News on the Moses**LiveCD for Moses**

Based on the collected feedback, the following work has been done:

1. A new Windows installer has been produced.
2. We have created a Windows GUI which is used to run the Moses decoder, and download and manage model packages.
3. Specification of a model package format, that is supported by the GUI, has been designed. An example model, provided in year one, has been converted to a model package and is available for use with the GUI from the Moses web-site.
4. 32-bit Linux installation packages have been made available.
5. Testing of the Linux installation packages has been started, and is on going, with the following distributions:

- Ubuntu 12.04 LTS,
- Redhat EL 6.4, and
- Fedora 19.

Balancing the need to maintain contacts lists for future outreach we made more content available for download from the MosesCore web site without registration.

4.3 Tutorials

Hieu Hoang and Matthias Huck delivered their "Open Source Statistical Machine Translation" at AMTA in Vancouver, in October 2014. They also have been accepted to present a similar tutorial at ICON 2014, in Goa, India in December 2014.

We also published the videos of the Machine Translation and Moses Tutorial (developed by TAUS) on [YouTube](#). Whilst this removes some of the context of the online learning environment, it makes the tutorial content more discoverable and accessible. The videos average about 100 views since

the publication through this channel in July 2014 (this is in addition to views through the official tutorial website).

Bibliography

- [1] Ondrej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Herve Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics.