

Foreword

Following two successful EBMT workshops in 2001 at the MT Summit VIII in Santiago de Compostela, Spain, and in 2005 at the MT Summit X in Phuket, Thailand, this is the third workshop of its kind. Many things have happened since 2005. The last few years have witnessed a decline in example-based machine translation (EBMT) research and statistical machine translation (SMT) has almost completely taken over the corpus-based machine translation arena, with many EBMT practitioners moving into hybrid approaches integrating EBMT with other approaches, mostly (but not only) SMT. Not having a clear definition of what EBMT is has also contributed to this lack of visibility. In fact, research that would have been considered EBMT has been published without the EBMT label.

Is the success of SMT due to the fact that it is the best way to do corpus-based machine translation or is it because many SMT software packages are readily available to researchers under free/open-source licences that allow use as well as collaborative improvement? Shouldn't EBMT practitioners start to think about putting together their tools, their engines and their data and releasing them under open licenses to extend their use both in academia and industry?

The pressure on machine translation researchers to prove their results through detailed empirical evaluation is growing. But the validity of empirical results hinges on reproducibility. Turning our experimental research into packages and tools that other researchers can use and improve is a challenge but it is not infeasible as SMT practitioners have shown.

The response of the community to the call for papers for this conference was very encouraging. We received fifteen papers addressing the main theme or other aspects related to a potential strengthening of EBMT and its real-world applications. Of these, 11 have been accepted for presentation at the workshop.

The papers have been grouped in four sessions.

In the first session, three papers present hybrid approaches to EBMT. James Smith and Stephen Clark present a new EBMT–SMT hybrid, Felipe Sánchez-Martínez and colleagues show an EBMT–RBMT hybrid, and Declan Groves and colleagues evaluate syntax-driven approaches to phrase extraction for machine translation.

The second session groups papers addressing the main theme of the Workshop, featuring free/open-source EBMT software: Aaron Phillips and Ralf Brown describe their Cunei MT platform, David Farwell and Lluís Padró describe the Freeling analyser suite and outline possible applications in EBMT, and Adrien Lardilleux and colleagues discuss the sampling-based alignment performed by their software.

The third session deals with what one could call “Pure EBMT”. Harold Somers and colleagues review the use of proportional analogies in EBMT, Maarten van Gompel and colleagues describe the extension of their memory-based translation formalism to phrases, and Vincent Vandeghinste and Scott Martens describe their top-down transfer strategy to EBMT.

Finally, the fourth session deals with applications of EBMT. Julien Gosme and colleagues discuss the translation of sublanguage using subgrammars and Marian Flanagan describes the use of EBMT to translate film subtitles.

The papers in these proceedings will therefore give a wide but focussed view of the current status of EBMT research and seem to hint at a revival of the whole field after years of languishing.

I will finish this foreword by thanking the authors who sent papers; the reviewers, who did an excellent work and gave many suggestions to improve the papers that were finally accepted; and the local organizers. I thank also the European Association for Machine Translation, the Office of the Vice-President for Research and the Centre for Next Generation Localisation of Dublin City University, and Science Foundation Ireland for their crucial support.

Dublin, November 2009

Mikel L. Forcada

Workshop co-chair

CNGL, School of Computing,

Dublin City University,

Dublin, Ireland.

Salute

The European Association for Machine Translation (EAMT) has been supporting workshops and conferences on at least an annual basis for more than 13 years now, bringing together users, developers, and vendors who have a shared interest in machine translation in its widest sense.

In 2005, the annual workshop was rebranded as a conference, given the quality and quantity of papers that EAMT events were able to attract, as well as a big increase in the number of attendees at our events.

However, the EAMT Committee is, as it has always been, receptive to any members who wish to run workshops on particular themes. We were delighted to be approached by Mikel Forcada with a view to endorsing, and supporting, his idea for a 3rd International Workshop on Example-Based Machine Translation (EBMT), following on from previous successful workshops at the MT Summits in 2001 and 2005.

By holding the 3rd EBMT Workshop in 2009, not only do we maintain the 4-year cycle of such events, but more importantly the time is ripe to address the sub-topic of the workshop, namely "Going open-source to revive EBMT". Since the last EBMT Workshop in Phuket, it's clear that relatively few published papers have appeared on EBMT, and still no open-source toolkits exist to entice interested researchers to become active in our area. When we distributed the call for papers, then, we wondered firstly whether we would be able to attract enough papers, and secondly whether any of the papers would address the theme of the workshop.

We needn't have worried, as the programme is strong, diverse, and directly addresses the issue of open-sourcing EBMT software for reuse by the wider community. I am very grateful to Mikel for driving this forward, and I'm sure his initiative will be successful, not just for the EBMT community, but for all researchers in the area of MT.

Before closing, I'd like to thank the programme committee for their invaluable contributions, the invited speaker, panellists and authors of refereed papers, and our sponsors: Science Foundation Ireland, DCU, the CNGL, and the EAMT.

Finally, I wish you all a successful and enjoyable workshop, and also look forward to seeing you all at next year's EAMT Conference in St. Raphaël, May 27–28 2010!

Dublin, November 2009

Andy Way

Workshop co-chair and EAMT President
CNGL, School of Computing,
Dublin City University,
Dublin, Ireland.

Welcome

Localisation is the industrial process of adapting digital content to culture, locale and linguistic environment. Localisation is a value-adding component in the global content distribution, services, ICT and product industries, opening up markets that are otherwise inaccessible.

Currently localisation is facing three challenges: volume, access and personalisation. The amount of content that needs to be localised into increasing numbers of languages massively outstrips the supply of human translators. Traditional content access modalities cover print based media and/or assume electronic access using a full keyboard and screen. New devices such as smart phones or PDAs enable access to digital content on the fly. Novel access modalities need to be supported by localisation efforts. Traditional localisation is coarse-grained, e.g. we localise for the Middle-East, ignoring information that cuts across linguistic and geographical boundaries: nowadays a manager in Cairo has much more in common with their counterparts in Shanghai and New York than their parents and grandparents had 30 years ago. In terms of a slogan: the person is the ultimate locale.

The Centre for Next Generation Localisation (CNGL) addresses the three challenges of volume, access and personalisation. CNGL is a large collaborative research centre with four university and eleven industry partners with a strong research and commercialisation brief.

Translation is one of the core activities in localisation and the localisation industry is a strong and early large scale user of translation automation technology such as translation memories (TMs) and machine translation (MT).

Because of this it is very opportune that CNGL has the privilege of hosting the 3rd Workshop on Example-Based Machine Translation. Example-Based Machine Translation (EBMT) is one of the major paradigms in MT. In fact both TM and current statistical SMT technologies can be understood as memorising, retrieving and (in the case of SMT) recombining examples from previous translations.

I welcome you all to the 3-rd Workshop on EBMT in Dublin and I am looking forward to the presentations and discussions at the workshop.

Dublin, November 2009

Prof. Josef van Genabith

Director

Centre for Next Generation Localisation,

Dublin City University,

Dublin, Ireland.

Editors and Programme Committee Chairs:

Mikel L. Forcada, Universitat d'Alacant and Dublin City University
Andy Way, Dublin City University

Programme Committee:

Sivaji Bandyopadhyay, Jadavpur University, Kolkata, India
Ralf Brown, Carnegie Mellon University, Pittsburgh, USA.
Michael Carl, Copenhagen Business School, Denmark
David Farwell, ICREA-UPC, Barcelona, Spain
Declan Groves, Traslán, Ireland
John Hutchins, Norwich, U.K.
Sadao Kurohashi, Kyoto University, Japan.
Yves Lepage, Université de Caen, France.
Harold Somers, Dublin City University, Dublin, Ireland
Eiichiro Sumita, ATR, Kyoto, Japan.

Additional Reviewer:

Felipe Sánchez Martínez, Universitat d'Alacant

Invited Speaker:

Sadao Kurohashi, Kyoto University, Japan.

Local Organising Committee:

Sudip Naskar
Ankit Srivastava
Sandipan Dandapat
Cara Greene
Eithne McCann

Sponsoring Institutions:

European Association for Machine Translation
Dublin City University
Centre for Next Generation Localisation
Science Foundation Ireland

Table of Contents

Invited talk

<i>Fully Syntactic Example-based Machine Translation</i> Sadao Kurohashi	1
-----------------------------------------------------------------------------------	---

Hybrid Approaches to EBMT

<i>EBMT for SMT: A New EBMT-SMT Hybrid</i> James Smith and Stephen Clark	3
-----------------------------------------------------------------------------------	---

<i>Hybrid Rule-Based – Example-Based MT: Feeding Apertium with Sub-sentential Translation Units</i> Felipe Sánchez-Martínez, Mikel L. Forcada and Andy Way	11
---------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

<i>Evaluating Syntax-Driven Approaches to Phrase Extraction for MT</i> Ankit Srivastava, Sergio Penkale, Declan Groves and John Tinsley	19
--------------------------------------------------------------------------------------------------------------------------------------------------	----

Open-Source EBMT packages and tools

<i>Cunei Machine Translation Platform: System Description</i> Aaron B. Phillips and Ralf D. Brown	29
------------------------------------------------------------------------------------------------------------	----

<i>FreeLing: From a multilingual open-source analyzer suite to an EBMT platform</i> David Farwell and Lluís Padró	37
----------------------------------------------------------------------------------------------------------------------------	----

<i>Lexicons or phrase tables? An investigation in sampling-based multilingual alignment</i> Adrien Lardilleux, Jonathan Chevelu, Yves Lepage, Julien Gosme, and Ghislain Putois	45
------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	----

“Pure” EBMT

<i>A review of EBMT using proportional analogies</i>	
Harold Somers, Sandipan Dandapat and Sudip Kumar Naskar	53

<i>Extending Memory-Based Machine Translation to Phrases</i>	
Maarten van Gompel, Antal van den Bosch, Peter Berck.	61

<i>Top-Down Transfer in Example-based MT</i>	
Vincent Vandeghinste and Scott Martens.	69

Applications

<i>Translation of sublanguages by subgrammars</i>	
Julien Gosme, Yves Lepage, and Adrien Lardilleux.	77

<i>Using Example-Based Machine Translation to translate DVD Subtitles</i>	
Marian Flanagan	85

Workshop Programme

Thursday, 12th November 2009

08.30 **Registration**

09.15 **Opening address**

09.30 **Invited talk:** *Fully Syntactic Example-Based Machine Translation*
Sadao Kurohashi

10.30 **Coffee break**

Session 1: Hybrid approaches to EBMT

11.00 *EBMT for SMT: A New EBMT-SMT Hybrid*
James Smith and Stephen Clark

11.30 *Hybrid Rule-Based – Example-Based MT: Feeding Apertium with Sub-sentential Translation Units*
Felipe Sánchez-Martínez, Mikel L. Forcada and Andy Way

12:00 *Evaluating Syntax-Driven Approaches to Phrase Extraction for MT*
Ankit Srivastava, Sergio Penkale, Declan Groves, and John Tinsley

12.30 **Lunch break**

Session 2: Open-source EBMT packages and tools

14.00 *Cunei Machine Translation Platform: System Description*
Aaron B. Phillips and Ralf D. Brown.

14.30 *FreeLing: From a multilingual open-source analyzer suite to an EBMT platform*
David Farwell, Lluís Padró

15.00 *Lexicons or phrase tables? An investigation in sampling-based multilingual alignment*
Adrien Lardilleux, Jonathan Chevelu, Yves Lepage, Julien Gosme, and Ghislain Putois

15.30 **Coffee break**

Roundtable: Going open-source to revive EBMT?

16.30

Seeding talks:

Andy Way: *Open research questions for EBMT*

Mikel L. Forcada: *Why free/open-source EBMT?*

16.20

Panel: Ralf D. Brown, Yves LePage, Sadao Kurohashi

16.40

Open discussion

17.40 **End of day 1**

19.30 **Conference dinner**

Friday, 13th November 2009

Session 3: “Pure” EBMT

- 09.30 *A review of EBMT using proportional analogies*
Harold Somers, Sandipan Dandapat, and Sudip Naskar
- 10.00 *Extending Memory-Based Machine Translation to Phrases*
Maarten van Gompel, Antal van den Bosch, and Peter Berck
- 10.30 *Top-down Transfer in Example-Based MT*
Vincent Vandeghinste and Scott Martens

11.00 **Coffee break**

Session 4: Applications

- 11.30 *Translation of sublanguage by subgrammars*
Julien Gosme, Yves LePage, and Adrien Lardilleaux
- 12.00 *Using Example-Based Machine Translation to translate DVD subtitles*
Marian Flanagan
- 12.30 **Closing address**
- 13.00 **Workshop close**