crowdMT.

Proceedings of the

1st Workshop on Open Community-Driven Machine Translation

June 15 2023
Tampere, Finland

Edited by

Miquel Esplà-Gomis (Universitat d'Alacant, Spain), Mikel L. Forcada (Universitat d'Alacant, Spain), Taja Kuzman (Jožef Stefan Institute, Slovenia), Nikola Ljubešić (University of Ljubljana, Slovenia), Rik van Noord (University of Groningen, The Netherlands), Gema Ramírez-Sánchez (Prompsit Language Engineering, Spain), Jörg Tiedemann (University of Helsinki, Finland), Antonio Toral (University of Groningen, The Netherlands)

Organised by

macocu

ISBN: 978-84-1302-228-4

Contents

Invited Speeches **1**
Mikel L. Forcada. *Apertium: empowering vulnerable language communities through free/open source rule-based machine translation.* 1
Santhosh Thottingal and Niklas Laxström. *Machine Translation at Wikipedia* 2

Full papers **3**
Antoni Oliver and Sergi Álvarez. *Training and integration of neural machine translation with MTUOC* 5
Séamus Lankford, Haithem Afli, and Andy Way. *Design of an Open-Source Architecture for Neural Machine Translation* 15
Anna Dmitrieva and Aleksandra Konovalova. *Creating a parallel Finnish–Easy Finnish dataset from news articles* 21

Extended abstracts **27**
Javad Pourmostafa Roshan Sharami, Dimitar Shterionov, and Pieter Spronck. *A Python Tool for Selecting Domain-Specific Data in Machine Translation* 29
Gema Ramírez-Sánchez. *MutNMT, an open-source NMT tool for educational purposes* 31

Organising committee

Miquel Esplà-Gomis (Universitat d'Alacant, Spain)
Mikel L. Forcada (Universitat d'Alacant, Spain)
Taja Kuzman (Jožef Stefan Institute, Slovenia)
Nikola Ljubešić (University of Ljubljana, Slovenia)
Rik van Noord (University of Groningen, The Netherlands)
Gema Ramírez-Sánchez (Prompsit Language Engineering, Spain)
Jörg Tiedemann (University of Helsinki, Finland)
Antonio Toral (University of Groningen, The Netherlands)

Programme Committee

Miquel Esplà Gomis (Universitat d'Alacant, Spain) Mikel L. Forcada (Universitat d'Alacant, Spain) Taja Kuzman (Jožef Stefan Institute, Slovenia) Nikola Ljubešić (Jožef Stefan Institute, Slovenia) Rik van Noord (University of Groningen, The Netherlands) Juan Antonio Pérez-Ortiz (Universitat d'Alacant, Spain) Gema Ramírez Sánchez (Prompsit Language Engineering, Spain) Peter Rupnik (Jožef Stefan Institute, Slovenia) Felipe Sánchez-Martínez (Universitat d'Alacant, Spain Víctor Manuel Sánchez-Cartagena (Universitat d'Alacant, Spain) Jörg Tiedemann (University of Helsinki, Finland) Antonio Toral (University of Groningen, The Netherlands) Jaume Zaragoza-Bernabeu (Prompsit Language Engineering, Spain)

Invited Speeches

Apertium: empowering vulnerable language communities through free/open source rule-based machine translation

Mikel L. Forcada, Universitat d'Alacant (Alacant, Spain) and Prompsit Language Engineering (Elx, Spain)

Language technologies, including machine translation, are crucial in our multilingual world, particularly since a growing fraction of communication takes place online. For many languages, reasonably useful machine translation does not yet exist, or if there is, it is often in the hands of one or a few companies —with honourable exceptions. In the age of neural machine translation and deep learning, a concentration of translation power occurs: only a few privileged companies (a) possess the necessary resources to collect and curate bilingual corpora which they do not publish, (b) are able to train and execute neural machine translation models which they usually do not publish, and (c) do so on massive computers that only they can afford, generating large amounts of greenhouse gases and heat and leaving behind the waste generated when building their computing facilities. This generates a kind of technological language injustice through dynamics of technological disempowerment in the communities of the affected languages. As a result, speakers in technology-deprived or dependent communities experience an incomplete citizenship, as citizenship is built and articulated through communication. Apertium, a free/open-source rule-based machine translation platform, was born in 2005 when statistical machine translation, the precursor of neural machine translation, was blooming. Apertium was originally designed to deal with closely related languages such as Spanish and Catalan, but the free/open-source licensing attracted a community that started to create systems for other language pairs, encoding in open dictionaries and rules explicit knowledge about their languages, knowledge that can be used to create new machine translation systems or other language technologies. In this talk, I will argue in favour of a free/open-source incarnation of such a vintage but frugal and sustainable technology as rule-based machine translation as a way for vulnerable language communities to technologically empower themselves.

Machine Translation at Wikipedia

Santhosh Thottingal and Niklas Laxström, Language team, Wikimedia Foundation

Wikipedia, the multilingual encyclopedia available in over 320 languages, uses machine translation technology primarily for article translation. The translation process involves an integrated tool that utilizes various machine translation services to provide initial translations, which are then refined by editors before publication. To date, approximately 1.5 million articles have been translated. This presentation aims to introduce a human-in-the-loop product design, highlighting the provision of high-quality rich text translations through text-only machine translation, coupled with manual curation facilitated by human edits. Additionally, we will share insights and analytics pertaining to translation quality and translators. The discussion will encompass the machine translation engines employed, ranging from free and open-source systems to self-hosted services and external paid APIs. Lastly, we will present the optimization techniques employed to scale machine translation models in order to meet the performance requirements of Wikipedia.