Lang Resources & Evaluation (2017) 51:891–895
DOI 10.1007/s10579-017-9405-8

EDITORIAL

Introduction to the special issue

Laurette Pretorius[1] · **Claudia Soria**[2]

Published online: 13 November 2017

The importance of the technological development of under-resourced languages is rooted in the imperative of cultural and language diversity and in the basic right of all communities, all languages and all cultures to be "first class citizens" in an age driven by information, knowledge and understanding. Under-resourced languages suffer from a chronic lack of available resources (human-, financial-, time-, data- and technology-wise), and from the fragmentation of efforts in resource development. This often leads to small resources, only usable for limited purposes or developed in isolation, without much connection with other resources and initiatives. The benefits of reusability, accessibility and data sustainability are, more often than not, out of the reach of such languages. Yet, these languages stand to benefit most from emergent collaborative approaches and technologies for language resource development.

This situation is exacerbated by the realization that as technology progresses and the demand for localised language services over digital devices increases, the divide between adequately- and under-resourced languages keeps widening. Moreover, given the high cost of language resource production, and given the fact that in many cases it is impossible to avoid the manual construction of resources, it is worth considering collaborative approaches and technologies for data collection, annotation and sharing. Indeed, these approaches seem particularly well-suited for collecting the data needed for the development of language technology applications

✉ Laurette Pretorius
pretol@unisa.ac.za

[1] School of Interdisciplinary Research and Graduate Studies, University of South Africa, Pretoria, South Africa

[2] Institute for Computational Linguistics A. Zampolli, National Research Council of Italy, Pisa, Italy

for under-resourced languages. It is also a good way to approach a small population of speakers who live in remote countries, or are scattered in diaspora all over the world.

In this spirit, the focus of this special issue of *Language Resources and Evaluation* is on two strategic approaches by which under-resourced languages can elevate themselves to levels of development that are potentially comparable to well-resourced, technologically advanced languages, viz. using the crowd and collaborative computational platforms, and using technologies of interoperability with well-developed languages. The following questions are addressed:

- How can collaborative approaches and technologies be fruitfully applied to the development and sharing of resources for under-resourced languages?
- How can (small) language resources for under-resourced languages be re-used, reach larger audiences, interoperate with well-resourced languages and be integrated into applications?

Each of the articles addresses at least one of these questions, thereby making a noteworthy contribution to the relevant scholarly literature and to the technological development of a wide variety of under-resourced languages.
The six papers are as follows:

1. *Ebaluatoia: crowd evaluation for English–Basque machine translation* by Nora Aranberri, Gorka Labaka, Arantza Díaz de Ilarraza and Kepa Sarasola.
2. *Assisting non-expert speakers of under-resourced languages in assigning stems and inflectional paradigms to new word entries of morphological dictionaries* by Miquel Esplà-Gomis, Rafael C. Carrasco, Víctor M. Sánchez-Cartagena, Mikel L. Forcada, Felipe Sánchez-Martínez and Juan Antonio Pérez-Ortiz.
3. *Nenek: a cloud-based collaboration platform for the management of Amerindian language resources* by J. L. Gonzalez, Anuschka van't Hooft, Jesus Carretero and Victor J. Sosa-Sosa.
4. *Crawl and crowd to bring machine translation to under-resourced languages* by Antonio Toral, Miquel Esplá-Gomis, Filip Klubička, Nikola Ljubešić, Vassilis Papavassiliou, Prokopis Prokopidis, Raphael Rubino and Andy Way.
5. *Reassessing the value of resources for cross-lingual transfer of POS tagging models* by Nicolas Pécheux, Guillaume Wisniewski and François Yvon.
6. *Modeling under-resourced languages for speech recognition* by Mikko Kurimo, Seppo Enarvi, Ottokar Tilk, Matti Varjokallio, Andre Mansikkaniemi and Tanel Alumäe.

We now take a closer look at each paper in terms of which question(s) they relate to, what the core contribution is and which languages they discuss.

The paper by Aranberri et al. primarily relates to the first question. As one of the key technologies to help preserve and promote linguistic diversity within the emerging information society, machine translation (MT) requires numerous natural language processing (NLP) tools and/or vast quantities of parallel texts of the

working languages. Developing MT systems is hard and becomes even more challenging for under-resourced languages. Indeed, with the scarce resources invested in development, there is little left for MT evaluation, let alone human evaluation, which still remains the most reliable source to check progress on translation quality.

The paper explores the feasibility of a crowd-based pair-wise comparison evaluation to get feedback on machine translation progress for under-resourced languages by proposing a task based on simple work units to compare the outputs of five English-to-Basque systems, all implemented in a web application. In the design and methodology specific attention is given to attracting and retaining participants. The authors note that "[c]ommunities from small, under-resourced or endangered languages tend to display a marked awareness and willingness to honestly contribute to initiatives that will allow their languages to survive the technological age." The approach may be readily generalised to other language communities and to other NLP-related tasks.

The paper by Esplá-Gomis et al. also relates to the first question. Researchers dealing with under-resourced languages cannot usually afford the skilled labour required to create linguistic resources. Under this assumption, methods that ease the involvement of a broader group of non-expert people can significantly reduce the development costs and speed up the creation of high-quality linguistic data for these under-resourced languages. This paper concerns the enlargement of monolingual morphological dictionaries (such as those used by the morphological analysers of many natural language processing applications) by non-expert mother-tongue speakers under the assumption that such a speaker can correctly answer the polar question "is x a valid form of the word w to be inserted?", where x represents tentative alternative (inflected) forms of the new word w. Indeed, when there is a source word form that the (rule-based) MT system is not able to analyse because it is not present in its source language morphological dictionary, this approach will help the user to insert the corresponding entry, its stem and a suitable inflection paradigm.

The obvious questions that arise are "Which polar questions should be presented to the user?", "How many of these questions are necessary to arrive at a correct dictionary entry?" and "How well does this approach work?". The authors propose a computational approach to the first two questions and then test it with both an oracle and non-expert users. The languages considered were Spanish, Catalan, Basque and Maltese, but the approach generalises to other languages as well.

The paper by Gonzalez et al. is the third paper that relates to the first question by focusing on cloud-based collaboration between language communities of under-resourced languages to perform a range of language resource development tasks. The proposed *Nenek* cloud-based platform is said to serve both as an archive and a platform for collaborative documentation activities. In Nenek it is possible to store resources, to generate resources with the help of the speakers, as well as return the resources to the speakers through an easily accessible monolingual website. The paper describes the different stages (acquisition, manufacture, diffusion) of the web-based resource development life cycle. It then continues to report on a case study for

the Huastec language and concludes that the Nenek-approach also "works well for larger under-resourced languages".

The paper by Toral et al. addresses the first, and, to some extent, also the second question. Since the vast majority of language pairs and domains are under-resourced with respect to MT, this paper is relevant as it considers approaches to tackle this resource bottleneck so that these languages and domains can be equipped with MT support in a way that is both cost-effective and rapid. The authors propose the use of two widely used techniques to this end: web crawling and crowdsourcing. More specifically, they propose to use web crawling to automatically acquire parallel data to train statistical MT (SMT) systems, if any such data can be found for the language pair and domain of interest. If that is not the case, they resort to (1) crowdsourcing to translate small amounts of text (hundreds of sentences), which can then be used to tune SMT models, and to (2) monolingual crawling of vast amounts of monolingual data (millions of sentences), which are then used to build language models for SMT. The focus is on two use-cases involving Croatian, viz. the tourism domain and FIFA world cup tweets. In both cases improvements on baseline results are demonstrated. Although the focus is on English and Croatian, the approach has also been successfully applied to a number of other European languages.

The paper by Pécheux et al. addresses the second question, more specifically showing how to resort to less complete forms of annotation obtained from crawled dictionaries and/or through cross-lingual transfer when linguistically annotated data is scarce, as is the case for many under-resourced languages. The paper reviews two existing proposals for learning with ambiguous labels, both of which extend conventional supervised approaches to a weakly supervised setting: a history-based model using a variant of the perceptron, on the one hand; an extension of the Conditional Random Fields model on the other hand. Focusing on the part-of-speech tagging task, but considering a large set of ten languages for different language families, they show that (a) good performance can be achieved even in the presence of ambiguity, provided however that both monolingual and bilingual resources are available; (b) their two learners exploit different characteristics of the training set, and are successful in different situations; and (c) in addition to the choice of an adequate learning algorithm, many other factors are critical for achieving good performance in a cross-lingual transfer setting. All in all, they confirm earlier work that there are probably more gains to be achieved by improving the available resources than by designing more sophisticated learners.

The paper by Kurimo et al. also concerns the second question and is contextualised in the field of large-vocabulary continuous speech recognition for agglutinative languages, a field which heavily relies on a variety of methods related to language modeling. The focus is on four language modeling topics in which the authors claim to have been able to show significant progress: selecting conversational language modeling data from the Internet, adapting pronunciation and language models for foreign words, multi-domain and adapted neural network language modeling for improving performance in target topic and style, and decoding with subword lexical units. The article is methodologically interesting in that it uses Finnish and Estonian, both under-resourced languages with moderate amounts of resources available, to test and evaluate the mentioned language

modeling methods. These results are then used to inform the choice of methods and the better collection of data for related severely under-resourced languages, such as Northern Sami.

It is well-known that the destiny of a language is primarily determined by its native speakers and their broader cultural context.[1] It, therefore, comes as no surprise that the same is true for under-resourced languages in the digital realm. The first three articles focus in some way or another on human participants and their collaborative, technology-supported efforts for language resource and language technology development. Even the fourth article, which deals with machine learning approaches, also concludes that the importance of high quality language resources, rather than more sophisticated computational algorithms and techniques, are necessary for better results. By implication, it underscores the role of the human as the primary producer of authentic, high quality natural language data. However, the continued technological development of under-resourced languages also requires advanced computational approaches to the use and application of scarce language resources, as is evident from the last three articles.

In closing, it is our hope that this special issue will contribute to the study of the almost 7000 under-resourced languages of the world, drawing the attention to the role that both digital language resources and language technology can play in protecting and enhancing language diversity.

[1] https://www.sil.org/language-assessment/language-vitality.