Procesamiento del Lenguaje Natural, Revista nº 61, septiembre de 2018, pp. 171-174 recibido 04-04-2018 revisado 26-04-2018 aceptado 04-05-2018

enetCollect: A New European Network for combining Language Learning with Crowdsourcing Techniques

enetCollect: Una nueva red europea para el aprendizaje de idiomas y el crowdsourcing

Rodrigo Agerri[*], Montse Maritxalar[*], Verena Lyding[**], Lionel Nicolas[**]
[*]IXA NLP Group, University of the Basque Country UPV/EHU
[**]Eurac Research Bolzano, Italy
{rodrigo.agerri,montse.maritxalar}@ehu.eus,{verena.lyding,lionel.nicolas}@eurac.edu

Abstract: We present enetCollect, a large European COST action network set up with the aim of promoting a research trend combining the well-established domain of Language Learning with recent and successful crowdsourcing approaches. More specifically, the challenge of enetCollect is to foster the language skills of all citizens regardless of their backgrounds by enhancing the production of language learning material using Crowdsourcing techniques. In order to do so, the action will create a balanced interdisciplinary community of active stakeholders related to content-creation, content-usage, and Learning/Content Management Systems to create a theoretical framework for achieving a shared understanding of Language Learning and Crowdsourcing. This will allow to unlock the crowdsourcing potential available for language learning and to facilitate the development of prototypical experiments for the production of language learning material, such as lesson or exercise content. These activities would potentially benefit a wide range of users and languages.
Keywords: Crowdsourcing, Language Learning, Language Resources, COST Action

Resumen: En este artículo presentamos enetCollect, una extensa acción europea COST diseñada con el objetivo de promover una nueva línea de investigación que combine el dominio del aprendizaje de idiomas con recientes y exitosos enfoques basados en crowdsourcing. Más específicamente, el reto de enetCollect es fomentar el aprendizaje de idiomas para toda la ciudadanía europea mediante la mejora en la producción de materiales para el aprendizaje de idiomas usando técnicas de crowdsourcing. Para ello, la acción creará una comunidad interdisciplinar de agentes activos relacionados con la creación, uso y gestión de contenidos para el aprendizaje de idiomas que permita generar un marco teórico común en el cual investigar sobre el uso de crowdsourcing para la generación de contenidos y tecnología relacionada con el aprendizaje de idiomas. La idea es liberar el potencial de usar crowdsourcing para el aprendizaje de idiomas y facilitar el desarrollo de experimentos y prototipos para la generación de materiales de aprendizaje, tales como ejercicios, lecciones, etc. Estas actividades beneficiarán a la gran mayoría de las personas en proceso de aprender un nuevo idioma.
Palabras clave: Crowdsourcing, Aprendizaje de idiomas, Recursos Lingüísticos, acción COST

1 Introduction

enetCollect[1] is a COST Action supported by the EU Framework Programme Horizon 2020. COST Actions are European networking initiatives, which aim at creating new research communities around emerging research subjects with ground-breaking potential. The essence of a COST Action is the generation of a durable network of participants that meet regularly and exchange ideas, approaches and methods, and aim at building new research cooperations. COST funding covers expenses related to travel, meeting organization, and exchange of re-

[1] `https://enetcollect.net`

ISSN 1135-5948. DOI 10.26342/2018-61-25

searchers, while it purposefully does not fund research work in itself. Nonetheless, COST Actions have an excellent record of building consortia and follow up projects that do fund personnel. At this moment, enetCollect includes 36 European countries in the Management Committee. enetCollect started in March 2017 and it will run until March 2021. The authors of the present paper are the Management Committee members representing Spain.

The main challenge of enetCollect, the European Network for Combining Language Learning with Crowdsourcing Techniques is to foster the language skills of all citizens regardless of their backgrounds by enhancing the production of language learning material using Crowdsourcing techniques. Specifically, the Action aims at enhancing the production of learning material combining the well-established domain of language learning with recent and successful crowdsourcing approaches. While primarily focusing on Language Learning, EnetCollect also involves a variety of research players working on language-related topics and having tedious and/or complex tasks to perform that may be approached by crowdsourcing (e.g. Language Resources creation).

enetCollect will research both implicit and crowdsourcing approaches. Briefly, explicit crowdsourcing usually refers to the fact that the crowd intentionally participates in the crowdsourcing task whereas in an implicit crowdsourcing task the crowd is not necessarily aware of the fact that the results of its activity will be used for other, not explicitly explained, objective. The action is also interested in researching issues such as user-orientation and usability of technological applications for language learning driven by ethical, legal and commercial dimensions of developing such technology.

The underlying capacity of crowdsourcing to achieve ground-breaking results has been proven in several impressive ways. For example, Wikipedia[2] completely redefined the well-established domain of encyclopedias while reCAPTCHA[3] tackled the highly laborious task of manually transcribing vast amounts of text by obtaining an unparalleled and continuous workforce from the crowd. In enetCollect, similar approaches will be developed or adapted to facilitate the creation of language learning materials and language-related datasets.

2 Objectives

COST Actions distinguish objectives related to the creation of knowledge, termed Research Coordination (RC) Objectives, and objectives related to creating and empowering a community, termed Capacity-Building (CB) Objectives.

Regarding the RC objectives, enetCollect aims to review the state-of-the-art in order to gather and compile an overview of relevant approaches and techniques for crowdsourcing in order to obtain ashared understanding by creating a theoretical framework defining its terminology, key concepts, objectives and opportunities.

With respect to the CB objectives, the Action will create a community of active stakeholders and communication means allowing the easy exchange of information for, among other things, pursuing research experimentations and targeting new funded initiatives.

In pursuing these objectives, it is expected to create the following short and long term impacts: In the short term, the Action will build a balanced interdisciplinary community of experts that will initiate the R&I trend by creating a theoretical framework and evaluation data to complement it. Current members mostly come from the areas of Crowdsourcing, Computer Assisted Language Learning, Natural Language Processing, E-Lexicography, Learner Corpora, Corpus Linguistics and Learning Management Systems. In the short-to-long term, major impacts will consist of the enhanced creation of language learning material and of language-related data.

With regards to language learning material, the Action will foster the continuous creation and improvement of materials. This implies the participation of language learning participants, which will build an enormous crowdsourcing potential. Achievements even remotely comparable to the ones of Zooniverse[4] would significantly impact the domain. As a side effect, enetCollect will help balancing the coverage across languages by benefiting also less-resourced languages.

[2] `https://www.wikipedia.org/`

[3] `https://www.google.com/recaptcha`

[4] `https://www.zooniverse.org/`

3 Structure

The Grant Holder of enetCollect is EURAC, the European Academy of Bolzano[5]. The Management Committee includes 36 European countries[6]. enetCollect is structured in five Working Groups (WG) and three supportive coordination groups.

WG1: Explicit Crowdsourcing Its objective is developing or adapting explicit crowdsourcing approaches. For example, WG1 will research the most effective ways to collaboratively devise lesson content (e.g. grammar) and to assess its effectiveness by observing how different samples of users confronted to accordingly generated content perform subsequently.

WG2 Implicit Crowdsourcing WG2 aims at developing or adapting implicit crowdsourcing approaches. For instance, WG2 will research ways to generate exercises from language resources (e.g. lexica) and to crowdsource manual validation of automatically generated new entries (e.g neologisms) by cross-matching the learners' answers for exercises generated from such new entries.

WG3: User-oriented design strategies for a competitive solution. WG3 will create design strategies fostering the user-orientation of a language learning solution and ensuring its capacity to attract and retain a crowd. For example, WG3 will study the relevance and attractiveness of learner profiling for vocabulary training.

WG4 Technology-oriented specifications for a flexible and robust solution. WG4 targets the creation of technical specifications to support the functional demands of WG1, WG2, and WG3. For example, WG4 will study technical solutions for the scalability of crowdsourcing methods.

WG5: Application-oriented specifications for an ethical, legal and profitable solution. It will devise the application-oriented part of the theoretical framework related to (1) ethical questions regarding the user involvement and data collection, (2) legal regulations, and (3) opportunities and models for commercialization.

Dissemination, Exploitation and Outreach coordinations are overarching groups supporting the WGs with transversal tasks and standardizing the ways such transversal tasks are tackled.

Note that in many cases this structure responds to operational reasons. Thus, the five WGs each have their individual challenging tasks to address but they are interdependent in some aspects. For example, the boundary between explicit and implicit crowdsourcing (WG1 and WG2) can be fairly fuzzy for several approaches. Also, any crowdsourcing is meaningless if there is no crowd to rely on (WG3), no scalable solution to implement it (WG4), and no appropriate ethical or legal contexts (WG5).

4 enetCollect and NLP

The enetCollect action can be interesting for the Natural Language Processing (NLP) community to explore new avenues for crowdsourcing language resources (LR) which may be appropriate for language learning but also for domain specific learning (Science, History, etc). For example, implicit crowdsourcing could be done by means of simple language games where learners have to complete some challenges in order to improve their levels on the game. Similarly, explicit crowdsourcing could be implemented in collaborative applications where learners of different language levels or language communities help to each other by means of a peer-learning approach. Then, learners could tag or interpret the answers of lower level colleague mates as a peer collaboration, acting somehow as expert taggers of the data during the learning process.

enetCollect member statistics show that out of the five WGs, WG2 is the WG most followed by NLP-oriented Action members up to now. However, as there is a strong collaboration between WG1 and WG2, it is viable to organize NLP oriented Crowdsourcing challenges from both perspectives. This way language resources could be created to be used on learning approaches based on NLP.

5 Related Work

The online language learning platform Duolingo[7] follows a logic that is similar in many points to the one of enetCollect. It offers free language learning services for numer-

[5] http://www.eurac.edu

[6] http://www.cost.eu/COST_Actions/ca/CA16105

[7] www.duolingo.com

ous languages while explicitly crowdsourcing lessons from pro-active users and implicitly crowdsourcing translations through well-blended exercises.

The state-of-the-art regarding implicit crowdsourcing and NLP is mainly defined by "Games With A Purpose" (GWAPs) (Lafourcade, Brun, and Joubert, 2015). Some of the most well-known gamified interface for language resource production are JeuxDeMots (Lafourcade, Brun, and Joubert, 2015), Phrase Detectives (Poesio et al., 2012), ZombiLingo (Guillaume, Fort, and Lefebvre, 2016) and Wordrobe (Bos et al., 2017). Finally, two tools were designed for teaching and allow to crowdsource POS corpora (Sangati, Merlo, and Moretti, 2015) and syntactic dependencies (Hladká, Hana, and Lukšová, 2014).

6 Concluding Remarks

We have presented enetCollect, the European Network for Combining Language Learning with Crowdsourcing Techniques. The main objective of the action is to create a theoretical framework for achieving a shared understanding of Language Learning and Crowdsourcing. The network is well-balanced in terms of gender and includes both experienced as well as early-stage researchers (PhD students and post-docs), which enables the action to facilitate knowledge transfer and training of new generations of Crowdsourcing-focused researchers. The involvement of new and current members will be pursued through promotion of the Action through relevant communication channels of the research domains concerned. In addition, opportunities for short-term research stays will be advertised through open calls and workshops, and training schools related to relevant topics of the individual working groups will be organized in regular intervals.

Acknowledgements

The authors have been funded by the Horizon 2020 Framework Programme of the European Union under the enetCollect CA16105 COST action.

References

Bos, J., V. Basile, K. Evang, N. Venhuizen, and J. Bjerva. 2017. The groningen meaning bank. In N. Ide and J. Pustejovsky, editors, *Handbook of Linguistic Annotation*, volume 2. Springer, pages 463–496.

Guillaume, B., K. Fort, and N. Lefebvre. 2016. Crowdsourcing complex language resources: Playing to annotate dependency syntax. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, Osaka, Japan.

Hladká, B., J. Hana, and I. Lukšová. 2014. Crowdsourcing in language classes can help natural language processing. In *Second AAAI Conference on Human Computation and Crowdsourcing*.

Lafourcade, M., N. L. Brun, and A. Joubert. 2015. *Games with a Purpose (GWAPS)*. Wiley-ISTWiley-ISTE, July.

Poesio, M., J. Chamberlain, U. Kruschwitz, L. Robaldo, and L. Ducceschi. 2012. The phrase detective multilingual corpus, release 0.1. In *Collaborative Resource Development and Delivery Workshop Programme*, page 34.

Sangati, F., S. Merlo, and G. Moretti. 2015. School-tagging: interactive language exercises in classrooms. In *LTLT@ SLaTE*, pages 16–19.