Новости са трећег годишњег састанка enetCollect COST акције

Бранислава Шандрих
Филолошки факултет Универзитета у Београду
Branislava.Sandrih@fil.bg.ac.rs

Шта је enetCollect?

- European Network for Combining Language Learning with Crowdsourcing Techniques
- Веб презентације акције
  - http://enetcollect.eurac.edu/
- Велика интернационална мрежа финансирана као кост акција CA16105
  - Тренутно више од 120 учесника

Циљеви акције

- Подстицање истраживања и иновација у домену учења језика
  - у приступима у којима се примењује crowdsourcing
- Развијање материјала за учење језика
  - софтверска решења
  - игрице са наменом
  - скупови података
  - корпуси...

Акција је отворена за нове учеснике

- Посебно су пожељни људи који имају искуство у настави језика
- Пријављивање у оквиру веб странице и улазак у одговарајућу гугл групу

Пет радних група

- [WG1] R&I on Explicit Crowdsourcing for Language Learning material production,
- [WG2] R&I on Implicit Crowdsourcing for Language Learning material production,
- [WG3] User-oriented design strategies for a competitive solution,
- [WG4] Technology-oriented specifications for a flexible and robust solution,
- [WG5] Application-oriented specifications for an ethical, legal and profitable solution.

Укратко…

- *Whereas WG1 and WG2 focus on the combination of Language Learning with explicit and implicit Crowdsourcing techniques,* ***WG3 focus on the user-orientation of an online language learning solution to ensure its capacity to attract and retain a crowd,*** *WG4 focus on technical specifications to support the functional demands of Language Learning and Crowdsourcing approaches and WG5 focus on ethical questions, legal regulations and business opportunities that such a combination between Language Learning and Crowdsourcing can imply…*

Откуд ја ту?

- Дво и по дневни хакатон у Лисабону крајем јануара


SPA
SPA

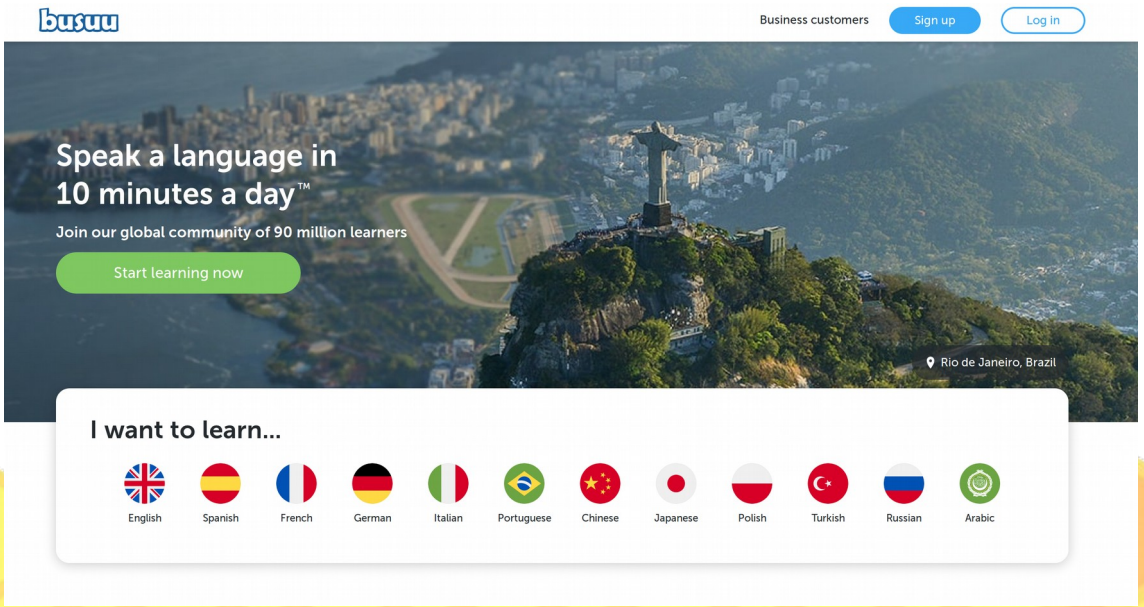
Трећи годишњи састанак

- У Лисабону, 14. и 15. март 2019.
- Институт Instituto Superior Técnico
- Додатни дан за обуку или интерне састанке радних подгрупа


@tecnico

Програм

- Реч добродошлице
- Уводно представљање компаније busuu
- Представљање постера
- До краја: представљање резултата радних група




3rd Annual Meeting
of enetCollect

Програм

- Уједно и
  - прва ужина
  - ручак
  - друга ужина
  - вечера

Наш постер

- Crowdsourcing corpus cleaning for language learning - an approach proposal
  - Tanara Zingano Kuhn, CELGA-ILTEC, University of Coimbra, Portugal
  - Peter Dekker, Dutch Language Institute, The Netherlands
  - Branislava Šandrih, University of Belgrade, Serbia
  - Rina Zviel-Girshin, Ruppin Academic Center, Israel

Crowdsourcing corpus cleaning for language learning - an approach proposal

enetCollect

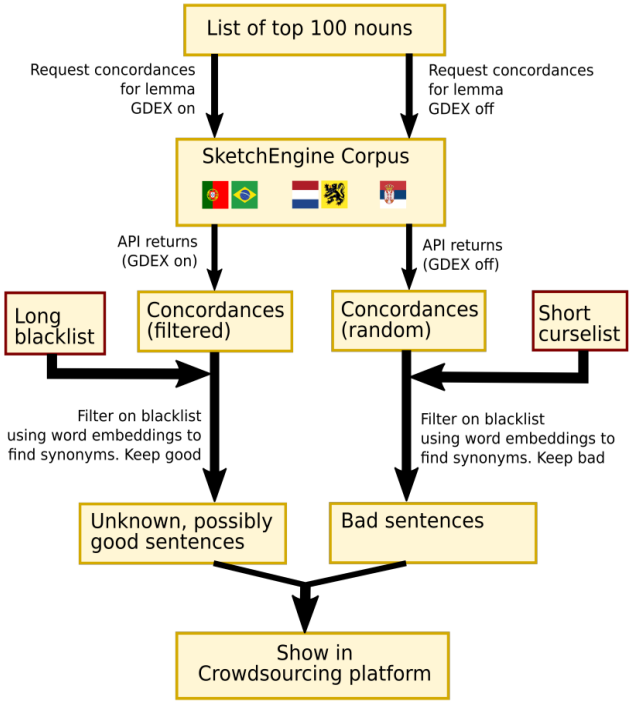
Tanara Zingano Kuhn, CELGA-ILTEC, University of Coimbra, Portugal
tanarazingano@outlook.com
Peter Dekker, Dutch Language Institute, The Netherlands
peter.dekker@ivdnt.org
Branislava Šandrih, University of Belgrade, Serbia
branislava.sandrih@fil.bg.ac.rs
Rina Zviel-Girshin, Ruppin Academic Center, Israel
rinazg@gmail.com

Introduction

Web corpora are valuable sources for the development of **language learning exercises**. However, the data may contain inappropriate or even **offensive language**, thus requiring data checking and filtering before pedagogical use. **We propose a language-independent crowdsourcing approach to clean up such corpora**, which we apply to Portuguese, Dutch and Serbian as case studies.

Approach proposal

- Sentences to be judged by native speakers are selected from a sample corpus and consist of **potentially good and "bad" (offensive) sentences**.
- Inappropriate sentences are included as ground truth for analysis.
- Potentially good sentences are extracted from the corpus, with **Sketch Engine GDEX filtering** on, and filtered using **long blacklist** of offensive and controversial words.
- Bad sentences are obtained from the corpus, without GDEX filtering, and filtered using **short blacklist** of offensive words, remainder is kept.
- Both blacklists are automatically extended with synonyms using semantic similarities of words from a **word embeddings** model.
- After performing the crowdsourcing experiment, contributor judgments can be fed to a **machine learning classification** model, for automatic cleanup of the remaining corpus.



Challenges

- Evaluation of the efficiency of crowdsourcing for **large-scale data processing**.
- Proper design of the project, so that not only valuable and reliable results can be collected, but the crowd also feels **motivated to participate**.

Future work

- Create learner's dictionary from cleaned web corpora
- Create a Machine Learning classifier that is able to classify sentence according to an appropriateness of the content, based on the sentence obtained after crowdsourcing step

enetCollect
COST Action CA16105
http://enetcollect.eurac.edu/
enetcollect@gmail.si

Исход

- Нове идеје
- Сугестије
- Предлози
- Коментари
- Два нова члана
- Један нови језик

Највише изазова за српски

- Ћирилични веб корпус на SketchEngine-у
  - налог?
- Конфигурација GDEX
- Обучени модел векторских угњеждења (енгл. WordEmbedding)
- Црна листа
- Орни *краудорсери*…

Али...

- Од бриге нема вајде, већ у се и у своје клъусе!
- А до тад...


Fastelaria Sornola
CASCAIS
KEBAB
RESTAURANTE - PASTELARIA
Coca-Cola
TAKE-AWAY


FÁCIA

obrigada e adeus!