EUROPEAN COOPERATION IN SCIENCE & TECHNOLOGY

Funded by the Horizon 2020 Framework Programme of the European Union

eman ta zabal zazu

UPV EHU

WG2 Implicit Crowdsourcing for Language materials production

enetCollect CA 16105

Bolzano, 2017-09-07

Rodrigo Agerri

Agenda

- Invited talk: Federico Sangati
- WG2 leaders intro
- Planning months 1-12
  - Task 2.1: Overview of existing resources, discussion groups, surveys, stakeholders (D2.1)
  - Task 2.2: Review of implicit crowdsourcing approaches (D2.2)
  - Task 2.3: Theoretical framework on producing learning material through implicit crowdsourcing (M6-24)
- Discussion

WG2 Objectives

- Developing or adapting implicit crowdsourcing approaches for producing language learning materials and language-related data.
  - Overview of existing materials, workflows and existing implicit crowdsourcing approaches
  - Generate exercise content from language resources (lexica)
  - Crowdsource manual validation of automatically generated content via cross-referencing.

Motivation for Implicit Crowdsourcing

“Implicit crowdsourcing involves users doing another task entirely where a third party gains information for another topic based on the user’s actions.” (Wikipedia)

- Duolingo: learn a language and help contributing in Machine Translation applications.
- News business: reader’s (paying members) feedback is taken into account
- ReCaptcha and Google robot detection: annotate image datasets to access another (free) service
- GWAP: ESP game, Zombilingo, Phrase Detectives, Wordrobe, etc.

Research

- Games4NLP@EACL2017:
- Player motivation
- Game design: reports of success and/or failure of designs
  - Generated data: minimizing noise and cheating
  - Usage of game generated data: How game-generated data has been used for NLP applications, objective evaluation on end-tasks (extrinsic)
  - Evaluation of games for NLP: Metrics for evaluating game performance, evaluating player performance, motivation and bias, and evaluating task difficulty
- Wordrobe for deep semantic representation (you can bet!)
- Phrase Detectives for anaphora resolution
- Ebaluatoia at IXA: competing with fellow users to help Basque-English Machine Translation
- GWAP used for teaching

EFCAMDAT

- NLP annotation of large learner corpus
  - Competitive performance for POS and dependency parsing
  - 33% of 1000 sample sentences contain error
  - Local morphosyntactic, word order, and semantic errors do not affect the main dependency relations
- The parser successfully captures syntactic patterns used by learners and provide valuable annotations for the investigation of a wide range of phenomena and SLA hypotheses
- NLP tools a starting point for development of tools that can accurately model the erroneous and untypical patterns of learner language

Beyond SOA

- Theoretical framework combining crowdsourcing and language learning
- In language learning there is a continuous renewal of users
- Generating exercise content from the results of the implicitly crowdsourced data

Tasks for year 1

- Information is scattered
- Across disciplines, across languages
- WG1 strategy: STSMs for extensive work
- WG3 strategy: responsible per country in WG2

Planning ahead

- AP1: responsible per country
- AP2: questionnaire (pending agreement WG1)
  - Distribute by 30/10/2017?
- AP3: STSM(s)
- AP4: Next meeting topics/actions
  - Specific to Implicit Crowdsourcing for LL