sloWCrowd: A crowdsourcing tool for lexicographic tasks

Darja Fišer[1], Aleš Tavčar[2], Tomaž Erjavec[2]
[1]University of Ljubljana
Aškerčeva 2, SI-1000 Ljubljana
[2]Jožef Stefan Institute
Jamova cesta 39, SI-1000 Ljubljana
darja.fiser@ff.uni-lj.si, ales.tavcar@ijs.si, tomaz.erjavec@ijs.si

Abstract

The paper presents sloWCrowd, a simple tool developed to facilitate crowdsourcing lexicographic tasks, such as error correction in automatically generated wordnets and semantic annotation of corpora. The tool is open-source, language-independent and can be adapted to a broad range of crowdsourcing tasks. Since volunteers who participate in our crowdsourcing tasks are not trained lexicographers, the tool has been designed to obtain multiple answers to the same question and compute the majority vote, making sure individual unreliable answers are discarded. We also make sure unreliable volunteers, who systematically provide unreliable answers, are not taken into account. This is achieved by measuring their accuracy against a gold standard, the questions from which are posed to the annotators on a regular basis in between the real question. We tested the tool in an extensive crowdsourcing task, i.e. error correction of the Slovene wordnet, the results of which are encouraging, motivating us to use the tool in other annotation tasks in the future as well.

Keywords: crowdsourcing, annotation, open-source software

1. Introduction

One of the best-known crowdsourcing platforms is the Amazon's Mechanical Turk[1] which is mostly suitable for non-linguistic or English tasks but, unfortunately, there are very few Turkers who would be able to solve tasks for Slovene. Other crowdsourcing tools, such as Word Detectives[2] that is used for anaphora annotation in texts (Chamberlain et al. 2008), or Wordrobe[3] which facilitates semantic corpus annotation (Venhuizen 2013), are either specialized for a particular task and cannot be easily adapted or not freely available. This is why we have developed a simple, adaptable, language-independent and open-source tool called sloWCrowd, suitable for a wide range of crowdsourcing lexicographic tasks which can be expressed as yes/no or multiple-choice questions. We tested the tool on the task of error correction in sloWNet (Fišer 2009), a WordNet based automatically developed semantic lexicon for Slovene and obtained very good results.

2. Description of the tool

The main purpose of the sloWCrowd[4] tool is to offer a micro-task to the annotator which in most cases is to answer whether a particular automatically assigned annotation is correct or wrong, but the tool also supports multiple choice questions, e.g. several annotations can be presented in a micro-task, and the annotator chooses the correct one. Each task also has the skip option.

The tool can handle various types of lexicographic tasks which require a large amount of human judgments. By assigning a lexicographic task to a crowd of annotators we decrease the time needed to solve the task considerably while keeping the quality of their answers at a high level as the same question is answered by multiple annotators.

2.1 Implementation of the tool

The tool uses PHP and MySQL, is simple to install and works on all popular web browsers, with the system designed so that it is easy to install and add new projects. The annotators are required to log in to the system, which allows tracking annotator answers, computing annotator accuracy and annotator administration. Log-in is via the PHP HybridAuth library.

Creating new projects consists of uploading three files:

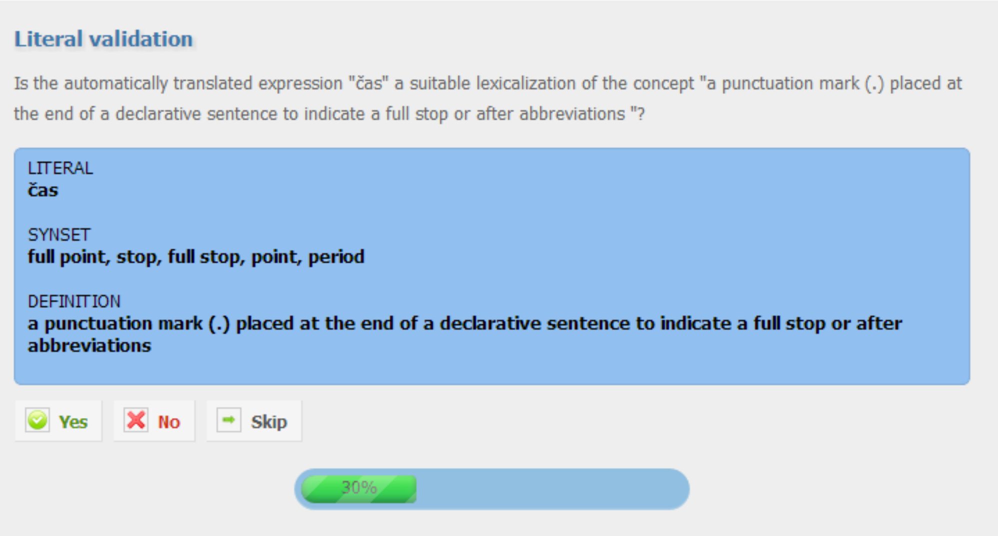
a) the dataset, a tab-delimited file in which each line contains one micro-task to be solved by the annotators;
b) the reference dataset containing solved tasks that are used for computing the accuracy and thus reliability of an annotator;
c) the interface file, containing the project-specific strings in the desired language to be d a ed in the user interface.

These configuration files enable the creation of a large variety of projects that are based on yes / no or multiple-choice answers. The first line in the dataset file defines the structure of the project and the first two columns in the table are mandatory. The first one contains the id of the micro-task and the second the string to be validated. Additional columns are optional and can be used to further explain the task and contain text with HTML tags.

[1] https://www.mturk.com/mturk/welcome
[2] http://anawiki.essex.ac.uk/phrasedetectives/
[3] http://wordrobe.housing.rug.nl/
[4] http://nl.ijs.si/slowcrowd/

ID	LITERAL	SYNSET	DEFINITION
7470671	vžigalica	match	a formal contest in which two or more persons or teams compete
511817	kapra	gambol, romp, play, caper, frolic	gay or light-hearted recreational activity for diversion or amusement
7560652	voznina	fare	the food and drink that are regularly served or consumed

Figure 1: Dataset file



Literal validation

Is the automatically translated expression "čas" a suitable lexicalization of the concept "a punctuation mark (.) placed at the end of a declarative sentence to indicate a full stop or after abbreviations "?

LITERAL
čas

SYNSET
full point, stop, full stop, point, period

DEFINITION
a punctuation mark (.) placed at the end of a declarative sentence to indicate a full stop or after abbreviations

Figure 2: Validation of literals

The dataset file for the task presented in the next section, i.e. validating sloWNet (wordnet) literals is given in Figure 1 and contains four columns: ID, LITERAL, SYNSET and DEFINITION. The literal and the optional columns are shown to the user during the annotation task. The goldstandard contains an additional column LABEL that states whether a definition for a literal is correct or wrong.

The interface file contains all the translations of the tool in a variable, which is a definition tuple. This makes the adaptation of the tool for different types of projects and different languages trivial.

The main MySQL database table holds information on particular projects, and each project is stored in two tables, one for annotator-related data, and the other for the collected answers to the tasks.

sloWCrowd implements a mechanism to validate the quality of the annotators. Every project allows uploading not only the dataset to be annotated, but also a gold standard, which contains micro-tasks with the correct answers. Annotators receive a mix of tasks from both datasets, allowing the system to evaluate their accuracy. Their score is on the fly to determine the ratio of tasks from both datasets and also allows post-hoc removal of answers of unreliable annotators.

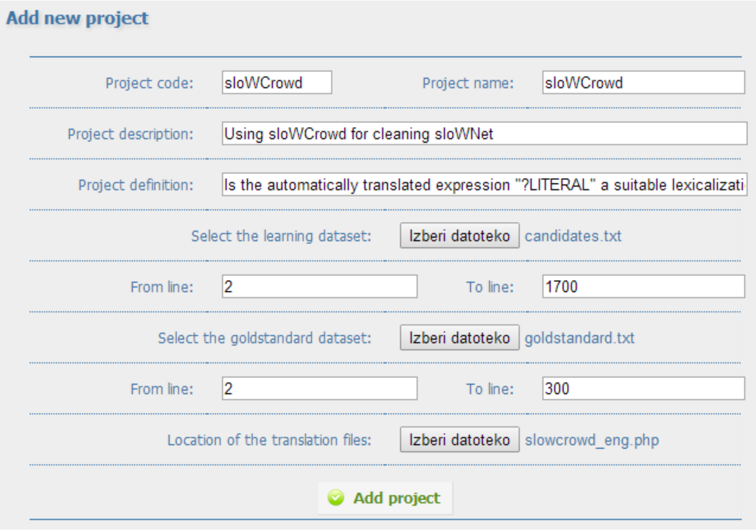
The tool incorporates strategies to make the tasks more interesting. For example, the annotator receives a batch of 10 randomly chosen micro-tasks from the project with the progress bar displayed, motivating them to finish the batch. The answers of the annotators are scored, and the top-scoring annotators are displayed in the Hall-of-Fame of the project. The scores are assigned not only regarding the reference dataset but also by computing the most common answer of all annotators.

2.2 User interface

The annotator window for the task of annotating the Slovene Wordnet is given in Figure 2. At the top of the screen instructions for the task are given. The question is displayed in the central part and the possible answers at the bottom.

The tool uses a mechanism for the internal evaluation of the annotators. New users get a higher ratio of literals from the reference dataset in order to be able to quickly determine their annotation ability. As users progress through the micro-tasks, their accuracy stabilizes and a higher ration of dataset micro-task is given to them. Users with low annotation accuracy get around 50% of literals from the reference dataset, while the most reliable users get as low as 10%. The tool aims to provide as much dataset micro-tasks as possible to the reliable annotators and at the same time supervises less reliable annotators by introducing more micro-tasks from the reference dataset.

During the annotation, the tool keeps track of the used time by the user to solve each task. Micro-tasks where users require more time to solve are deemed hard. A different use of this feature is to identify neglectful users. Those users who consistently require only a few seconds to answer may not pay enough attention and usually do not provide useful answers. We also noticed that skipped literals in general require more decision time.



Add new project

Project code:	sloWCrowd	Project name:	sloWCrowd
Project description:	Using sloWCrowd for cleaning sloWNet		
Project definition:	Is the automatically translated expression "?LITERAL" a suitable lexicalizati		
Select the learning dataset:	Izberi datoteko candidates.txt		
From line:	2	To line:	1700
Select the goldstandard dataset:	Izberi datoteko goldstandard.txt		
From line:	2	To line:	300
Location of the translation files:	Izberi datoteko slowcrowd_eng.php		

Figure 3: Project definition

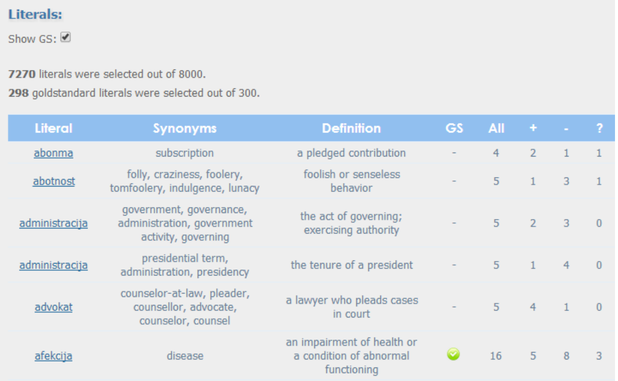
2.3 Administrator's interface

Monitoring and exporting active projects and adding new ones is possible in the administrator's interface. The administrator adds a new project by entering its name and description, uploads tabular files with the reference and task datasets, and chooses the file with the user interface of the project. On initialising the project the necessary database files are created by the tool. The project configuration window is given in Figure 3.

The project code field is an internal variable used for the creation of database tables used in the project. The project name field is an identifier for the project and is used during the annotation tasks. In the Project definition field it is possible to further explain the task in the project. It supports the use of variables (i.e., columns defined in the dataset file). A variable starts with question mark, followed by the column name (c.f. Figure 1). At runtime, the variable is replaced by the column value for the micro-task and shown to the annotator.

For active projects, the administrator can inspect the quality of the annotators and the number of tasks they have the option of disabling the answers of particular annotators.

A different screen gives an overview of all micro-tasks: how many annotators solved them, what is the ratio of different answers, and, if we choose to include them, whether a micro-task belongs to the reference dataset, as shown in Figure 4. It is also possible to focus on a particular micro-task and inspect all the answers of the individual annotators; we show this screen in Figure 5. In this screen all the relevant information is provided. Date and time of the annotation task, the annotators' choice, the actual value (for literals from the reference dataset), the used time by the user and his overall accuracy. Finally, the administrator interface allows downloading the answers, subject to several filters: whether to include non-active annotators, the reference dataset, etc.



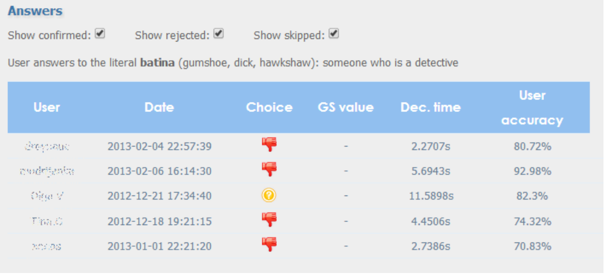
Literals:

Show GS: ☑

7270 literals were selected out of 8000.
298 goldstandard literals were selected out of 300.

Literal	Synonyms	Definition	GS	All	+	-	?
abonma	subscription	a pledged contribution	-	4	2	1	1
abotnost	folly, craziness, foolery, tomfoolery, indulgence, lunacy	foolish or senseless behavior	-	5	1	3	1
administracija	government, governance, administration, government activity, governing	the act of governing; exercising authority	-	5	2	3	0
administracija	presidential term, administration, presidency	the tenure of a president	-	5	1	4	0
advokat	counselor-at-law, pleader, counsellor, advocate, counselor, counsel	a lawyer who pleads cases in court	-	5	4	1	0
afekcija	disease	an impairment of health or a condition of abnormal functioning	✓	16	5	8	3

Figure 4: Tasks overview



Answers

Show confirmed: ☑ Show rejected: ☑ Show skipped: ☑

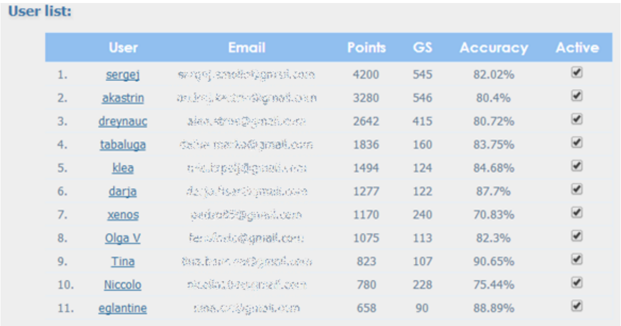
User answers to the literal **batina** (gumshoe, dick, hawkshaw): someone who is a detective

User	Date	Choice	GS value	Dec. time	User accuracy
[redacted]	2013-02-04 22:57:39	👎	-	2.2707s	80.72%
[redacted]	2013-02-06 16:14:30	👎	-	5.6943s	92.98%
[redacted]	2012-12-21 17:34:40	?	-	11.5898s	82.3%
[redacted]	2012-12-18 19:21:15	👎	-	4.4506s	74.32%
[redacted]	2013-01-01 22:21:20	👎	-	2.7386s	70.83%

Figure 5: Voting results

3. Correcting sloWNet

In our crowdsourcing project we used the developed sloWCrowd tool to correct errors in the automatically developed and therefore noisy WordNet for Slovene. The list of potentially erroneous literals was produced in our previous research (Sagot and Fišer, 2012) by using methods of distributional semantics. sloWCrowd annotators were asked to read the problematic literal, its definition and their English equivalents from the aligned Princeton WordNet (Fellbaum, 1998) and decide whether or not the problematic literal fits into the synset or not. If they could not decide, they could skip the question and move on to the next one.



User list:

	User	Email	Points	GS	Accuracy	Active
1.	sergej	[redacted]	4200	545	82.02%	☑
2.	akastrin	[redacted]	3280	546	80.4%	☑
3.	dreynauc	[redacted]	2642	415	80.72%	☑
4.	tabaluga	[redacted]	1836	160	83.75%	☑
5.	klea	[redacted]	1494	124	84.68%	☑
6.	darja	[redacted]	1277	122	87.7%	☑
7.	xenos	[redacted]	1170	240	70.83%	☑
8.	Olga V	[redacted]	1075	113	82.3%	☑
9.	Tina	[redacted]	823	107	90.65%	☑
10.	Niccolo	[redacted]	780	228	75.44%	☑
11.	eglantine	[redacted]	658	90	88.89%	☑

Figure 6: User management

In the period between 25-10-2012 and 2-10-2013 310 different annotators logged into sloWCrowd and solved 41,587 micro-tasks for 7,544 different literals. Without taking into account the literals from the reference dataset that were used to calculate the annotators' accuracy, we have collected 31,637 new answers for 7,246 literals, which means that we have 4.36 answers per literal on average. The annotators needed an average of 10 seconds per micro-task, which means that the project lasted for 115 full hours or an equivalent of a 2-week full-time position of a single person. Some annotators (12%) immediately realized they were not interested in the task as they did not provide even a single answer. Among the annotators who answered at least one question, we collected 152 questions per annotator on average but the distribution of the annotators' answers is very uneven as 100 annotators answered only 10 questions or fewer while the annotator who provided the most answers checked as many as 4197 literals. As can be seen in Figure 6, the number of annotators that provided more than 500 answers is 11. These annotators contributed almost 58% of all the answers collected in the project.

According to the reference dataset, the annotators' average accuracy is 80.12 %, which is high for lexical semantics tasks. However, the fluctuation of accuracy among the annotators is very high. 72.16 % of all annotators pass the 75% threshold, which are the annotators who provided 85% of all the collected answers. Average accuracy of the ten annotators who have contributed the most answers is as high as 83.71 %. This means that the answers of the annotators who have provided the most answers are at the same time also the most reliable. All the annotators that do not pass the 75% threshold can be deactivated in sloWCrowd. By discarding their answers, we lose only 15% of the collected answers.

The annotators answered to most questions negatively (15,861 or just over 50%), which means that a good half of the potentially problematic literals were in fact incorrect and should be deleted from sloWNet. 14,984 or 47.36% of the questions were answered positively and only 792 or 2.5% of the questions were skipped. This means that the task was straightforward and the errors quite obvious as we had anticipated. In order to obtain the most reliable answers possible, the same question was repeated up to five times to different annotators and the final decision whether to delete the literal from sloWNet or not was reached by taking into account the majority answer. The annotators validated 1,476 nouns that had been automatically assigned to 2,901 different sloWNet synsets. By taking into account the majority answers obtained with sloWCrowd we deleted literals from 1,264 (44%) different synsets while 1,446 (50%) were validated as correct. 190 literals (6%) received the same number of positive and negative answers, which means that we need to collect more votes for those literals before reaching the final decision about deleting them from sloWNet.

4. Conclusion

We have presented the sloWCrowd tool that is an adaptable crowdsourcing tool for various lexicographic tasks. While we presented the test of the tool on the project for error correction of sloWNet, the tool is also being employed for various other projects, in particular for correcting the French wordnet WOLF[5], for choosing good corpus examples and collocations for Lexical Database of Slovene (Gantar and Krek, 2011), and for manually validating PoS tagging in a corpus of Croatian. In the full paper we will also introduce these projects in more detail and further explain the working of the tool, as well as introduce some new features, e.g. using time needed to solve a micro-task as an additional indicator of the reliability of the answer and annotator and computing the optimal number of answers for a given micro-task.

Some on-going sloWCrowd projects can be accessed at http://nl.ijs.si/slowcrowd/. The tool can be downloaded under CC-BY (Creative Commons Attribution) license from http://nl.ijs.si/slowcrowd/sloWCrowd.rar.

5. References

Adda, G., Sagot, B., Fort, K., Mariani, J. (2011). Crowdsourcing for Language Resource Development: Critical Analysis of Amazon Mechanical Turk Overpowering Uses. *Proceedings of the LTC 2011*, Poznań.

Chamberlain, J., Poesio, M., Kruschwitz, U. (2008) Phrase Detectives: A Web-based collaborative annotation game. *Proceeding of the conference iSemantics*, Graz.

Fišer, D. (2009). Pristopi za avtomatizirano gradnjo semantičnih zbirk. *Terminologija in sodobna terminografija*. Ljubljana: Založba ZRC, ZRC SAZU, 357--370.

Fišer, D., Novak, J. (2011). Visualizing sloWNet. *Proceedings of the conference eLEX2011*. Bled, Slovenia.

Gantar, P., Krek, S.. (2011). Slovene Lexical Database. *Proceedings of the sixth Slovko international conference on natural language processing, Modra, Slovakia, 20-21 October 2011* [Brno]: Tribun.

Sagot, B., Fišer, D. (2012). Cleaning noisy wordnets, 2012. *Proceeding of the conference LREC 2012*, Istanbul.

Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y. (2008). Cheap and Fast - But is it Good? Evaluating

[5] http://alpage.inria.fr/~sagot/wolf-en.html

Non-Expert Annotations for Natural Language Tasks. *Proceeding of the conference EMNLP 2008*, 254--263.

Venhuizen, N. J., Valerio, B., Evang, K., Johan, B. (2013). Gamification for word sense labeling. *Proceeding of the conference IWCS*.

von Ahn, L. (2006). Games with a Purpose. *Computer*, 39(6), 92--94.