CROWDSOURCING LANGUAGE RESOURCES FOR SPEECH RECOGNITION

Ing. Daniel Hládek PhD., Ing. Ján Staš PhD., prof. Jozef Juhár CSc.
Technical University of Košice, Letná 9, 040 01 Košice, Slovakia
E-mail: daniel.hladek@tuke.sk, jan.stas@tuke.sk, jozef.juhar@tuke.sk

Abstract:
An important part of any speech recognition system is a language model. Creation of a language model requires proper processing of large quantities of textual data. Part-of-speech tags, named entities or semantic roles in the text help with precise statistical language modeling. The natural language processing methods are usually trained on annotated text corpora. Annotation of text corpora or dictionaries is a difficult process that requires a lot of human work involved. Crowdsourcing is a specific sourcing model in which individuals or organizations use contributions of Internet users to create a specific knowledge base.

KEYWORDS: CROWDSOURCING, SPEECH RECOGNITION, LANGUAGE RESOURCES, SPEECH CORPORA

1. Introduction

1.1. Preparation of Data for Statistical Processing

Statistical models are a important part of contemporary systems of human-machine interaction. Construction of a statistical model requires large amount of manually annotated data. For the sake of building good language model it would be very helpful to have a sufficiently large database of text that would unify various sources in one place. The database can be easily used to easily construct domain-specific corpora from the already collected and prepared data.

Annotation of speech and language corpora is timely and costly process. The first step is to find interesting patterns that will be subject of data modeling. The research team must design annotation conventions and train a group of contributors. The prepared data are distributed to annotators that mark the interesting phenomena in the database. Partials results from the annotation team are gathered and a training corpus is constructed.

The corpus is the input of statistical model training. Adequate method is selected and the model is composed from the prepared training data. The model is then able to predict events, even if they were not seen in the training corpus. It is able to generalize implicit knowledge, written by the human annotators during database creation.

The annotation team is often not available. The first problem is financing the required people to annotate a database. Employing a person demands administrative work and sufficient financial resources that are often not sufficient. Even if financing is available, finding a skilled personnel is a hard task. The whole process is timely, because proper annotation requires lot of focus and effort.

The whole process of training database preparation can be described as a sequence of actions:

1. Task preparation: This step includes introduction of the task to the annotator and learning of the annotation conventions.

2. Annotation of data: Observing a part of data and marking interesting parts with proper tags.

3. Partial results gathering: Results from annotators are put in a common place. Missing and contradicting annotations have to be resolved.

4. Database finalization: giving the final form to the database. Data should be in a form that is feasible for statistical processing.

1.2.What is Crowdsourcing

There is a strong need to find an easier of annotated database creation methods. Annotation of data should be:

1. easier,

2. cost less,

3. be faster.

The solution is to automate the process of data annotation, and partial results gathering.

A task is split into several smaller sub-tasks that are easy to comprehend by a random user.

Jeff Howe defined “crowdsourcing” as “an idea of outsourcing a task that is traditionally performed by an employee to a large group of people in the form of an open call” [1].

1.3.State of the Art in Crowdsourcing

There are several works that give survey of the current literature in the field of corowsoucing, such as [2] or [3]

The most common method of crowdsourcing utilizes a market of micro-services. The Amazon Mechanical Turk platform offers a potential paradigm for engaging many users for low time and monetary costs [4].

The process of crowdsourcing is described as [3]:

“The crowdsourcing site exhibits a list of available tasks, associating with reward and time period, that are presented by requesters; and during the period, workers compete to provide the best submission. Meanwhile, a worker selects a task from the task list and completes the task because the worker wants to earn the associated reward. At the end of the period, a subset of submissions are selected, and the corresponding workers are granted the reward by the requesters. In addition to monetary reward, a worker gains credibility when his task accepted by the requester. Sometimes, the task requester is obligated to pay every worker who has fulfilled the task according to the requirements. In some cases, workers are not motivated by rewards, but they work for fun or altruis. “

These voting tasks require a crowdsourcing worker to select his answer from a number of choices. The answer that the majority selected is considered to be correct. Voting can be used as a tool to evaluate the correctness of an answer from the crowd [3].

Crowdsourcing is used as a source of annotated data for various tasks. Named entity annotation using Amazon Mechanical Turk is proposed in [5] Email annotation for NE is proposed in [6]□, work [7]□ designs a system for word sense disambiguation.

2. Prerequisites for solving the problem

Each crowdsourcing process consists of the following steps:

1. Short survey form – gathers basic information about participant.

2. Task form – displays partial task to the participant.

The task form is displayed after new question item is requested from the participant. The task form has several states:

1. Question requested. The application script requests question item from the database and displays it to the participant.

2. Question displayed: The participant fills the required answer.

3. Question submitted: The application script sends the answer back to the database. A new item can be requested by the participant.

An important factor in crowdsourcing is motivation. A participant can be motivated by a good will, or money. Participation in speech and language resources annotation can be made a part of educational process and student assessment.

Amazon Mechanical Turk is not suitable for our purposes. The first reason is that participation in Amazon Mechanical Turk market requires financing as a motivation. The second problem is that understanding the language is a prerequisite and there are not enough Slovak speaking people on Mechanical Turk. It was necessary to implement our own approach to crowdsourcing.

3. Solution of the examined problem

3.1 Our crowdsourcing system

The crowdsourcing system consists of the following components:

1. Database – Stores answers from the participants.

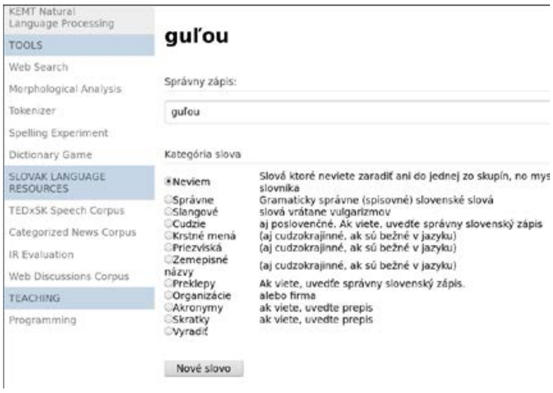
2. Web Application – displays survey and task forms and serves supporting files

3. Client-Side Application Script: Processes task forms, requests task data from the server and sends answers back.

We have implemented the following crowdsourcing modules to gather annotated data for training models for a speech recognition system or natural language processing system.

3.2 Crowdsourcing Modules

The crowdsourcing system is divided into three parts, each part is focused on gathering a certain type of language or speech data. The gathered data types are summarized in the Table 1.



guľou

Správny zápis:

guľou

Kategória slova

● Neviem Slová ktoré neviete zaradiť ani do jednej zo skupín, no mys slovníka
○ Správne Gramaticky správne (spisovné) slovenské slová
○ Slangové slová vrátane vulgarizmov
○ Cudzie aj poslovenčné. Ak viete, uveďte správny slovenský zápis
○ Krstné mená (aj cudzokrajinné, ak sú bežné v jazyku)
○ Priezviská (aj cudzokrajinné, ak sú bežné v jazyku)
○ Zemepisné názvy (aj cudzokrajinné, ak sú bežné v jazyku)
○ Preklepy Ak viete, uveďte správny slovenský zápis.
○ Organizácie alebo firma
○ Akronymy ak viete, uveďte prepis
○ Skratky ak viete, uveďte prepis
○ Vyradiť

Figure 1: Dictionary Game

Table 1: Modules Summary

Module Name	Type of the Gathered Data
Dictionary Game	Dictionary
Spelling Experiment	Text
Your Voice	Audio data

Dictionary Game – purpose of the crowdsourcing is to create manually checked dictionary. Dictionary of good words is an important part of both acoustic and language model of a speech recognition system. A good dictionary helps to restrict a search space and increase a precision of the automatic speech recognition.

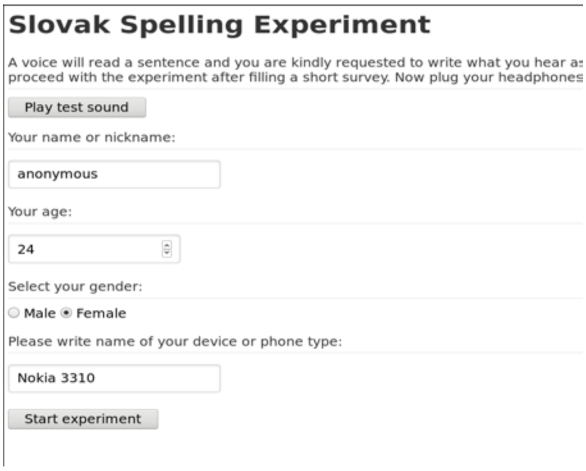
The task of the Dictionary game is to assign one of possible classes to a word. Some possible classes for a word are:

- Unable to Answer
- Unknown word
- Correct word
- Spelling error
- Foreign Word
- Proper Name

The participant selects one of the classes for the proposed word. The Class "Unable to Answer" is always pre-selected to filter out cases when participant marks the word without considering it. An example of a word category selection is in Fig. 1.

Result of the game is a set of decisions by the participants. The voting is used to select a class for a word that has most decisions. Output of the game is a dictionary, where each word has the most probable class.

Spelling Experiment – Purpose of this experiment is to observe common spelling errors. Language data from the internet that are



Slovak Spelling Experiment

A voice will read a sentence and you are kindly requested to write what you hear a: proceed with the experiment after filling a short survey. Now plug your headphones

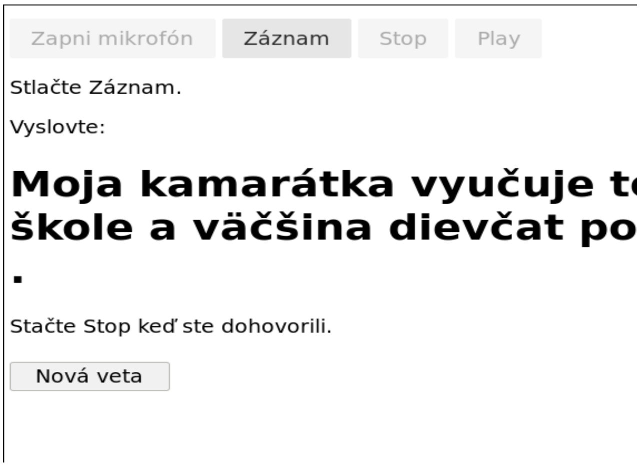
Your name or nickname: anonymous

Your age: 24

Select your gender: ○ Male ● Female

Please write name of your device or phone type: Nokia 3310

Figure 2: A Spelling Experiment Survey Form



Stlačte Záznam.

Vyslovte:

**Moja kamarátka vyučuje t
škole a väčšina dievčat po
.**

Stačte Stop keď ste dohovorili.

Figure 3: Your Voice Task Form

part of the language model training are full of spelling errors. The annotation task is defined as follows:

1. A short survey form is filled
2. The spelling experiment plays a sentence read by artificial voice.
3. The participant writes the sentence in his smartphone and the application records key-presses.
4. Results are sent to the server and a new sentence is proposed.

A screenshot of the survey form for spelling experiment is in Fig. 2.

Result of the spelling experiment is a set of manually transcribed sentences that can be used to create a model of spelling errors int the Slovak language. The modeling method was proposed in [8]□ and our research [9]□.

Your Voice – this experiment gathers audio data from recording of a mobile phone. The process is reverse as it is in the Spelling experiment. A sentence is displayed and the participant is requested to read it into his or hers mobile phone. The gathered data then can be used to train an acoustic model for recognition of voice in mobile phone.

1. The user fills a survey form. This step is important to get information about acoustic environment where a sound is recorded.
2. Then a sentence is shown on screen and a participant is requested to record it into the system. The sound is recorded in 16bit resolution using Recorder.js script and sent in an uncompressed form to the server.
3. Result of the crowdsourcing process is a set of recordings with transcriptions that can be used to train or adapt an acoustic model of the speech recognition system.

An example of a task form for Your Voice game is depicted in the Fig. 3.

4. Results and discussion

Results of crowdsourcing language resources should be applicable in the field of statistical language modeling in speech recognition. Our design of a system for quick annotation is applicable in the academic environment. Students are involved in the research by participating in creation and processing of language resources. Participating student becomes an integral part of research team and project. What we gathered from the application is summarized in the Table.2.

Table 2: Results Summary

Number of Gathered Items	Number of participants
8,079	123

5. Conclusion

The main task is split into small portions that are easily comprehensible even for a person without preliminary training, often in a form of a game. The web interface is accessible and an annotator does not need to install additional software. The process of annotation of language resources is easier, faster and has a low learning curve. This paper describes a design of a system for quick annotation of language resources for under-resourced languages by a group of volunteer annotators using crowdsourcing.

In the case of the Slovak language, the natural language processing technologies are just in the beginning. This initiative aims to support cooperation and sharing research results by allowing to easily try the tools by other researchers, without the need to compile sources, train classifiers or gather textual data. On the website http:\\nlp.web.tuke.sk a set of natural language tools can be tried out, enables members of the scientific community to easily see capabilities of the proposed tools, compare it with their own results and discuss it.

Acknowledgement

The research in this paper was supported by the Ministry of Education, Science, Research and Sport of the Slovak Republic under the research project VEGA 1/0511/17 and the Slovak Research and Development Agency under the research project APVV-15-0517.

References

[1] J. Howe, “The Rise of Crowdsourcing,” *Wired Mag.*, vol. 14, no. 6, pp. 1–5, 2006.

[2] A. J. Quinn and B. B. Bederson, “Human Computation: A Survey and Taxonomy of a Growing Field,” *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, pp. 1403–1412, 2011.

[3] M. C. Yuen, I. King, and K. S. Leung, “A survey of crowdsourcing systems,” in *Proceedings - 2011 IEEE International Conference on Privacy, Security, Risk and Trust and IEEE International Conference on Social Computing, PASSAT/SocialCom 2011*, 2011, pp. 766–773.

[4] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with Mechanical Turk,” in *Acm*, 2008, pp. 453–456.

[5] H. Fromreide, D. Hovy, and A. Søgaard, “Crowdsourcing and annotating NER for Twitter #drift,” *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14)*. pp. 2544–2547, 2014.

[6] N. Lawson and K. Eustice, “Annotating Large Email Datasets for Named Entity Recognition with Mechanical Turk,” *Comput. Linguist.*, no. June, pp. 71–79, 2010.

[7] C. Akkaya, A. Conrad, J. Wiebe, and R. Mihalcea, “Amazon Mechanical Turk for Subjectivity Word Sense Disambiguation,” *Comput. Linguist.*, no. June, pp. 195–203, 2010.

[8] E. S. Ristad and P. N. Yianilos, “Learning string-edit distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 5, pp. 522–32, May 1998.

[9] D. Hládek, J. Staš, S. Ondáš, J. Juhár, and L. Kovács, “Learning string distance with smoothing for OCR spelling correction,” *Multimed. Tools Appl.*, pp. 1–19, 2016.