Play the Language: An Alternative Manner Of Collecting Annotated Data

Abstract

We present the idea of the LGame language game portal. This language game portal is to be an online portal providing entertainment with natural language in the form of games with words, sentences and even documents. The entertaining games are designed to provide annotated textual data for the natural language tasks that either have not been implemented yet or have already been implemented with a performance lower than human performance.

1 Introduction

Machine learning methods find their place in the field of computational linguistics. However, building the data for the supervised methods is a very demanding, time-consuming and expensive activity. Let us exploit the capacity of Internet users who love to play games. Let us offer them language games with words, sentences or documents – language games that will be fun for the players and at the same time provide the valuable data needed for machine learning methods.

This approach to collect data quickly for free comes from the fields of image processing and computer vision. For instance, through the ESP Game (the ESP Game, 2005), (von Ahn, 2006) people help determine the content of images on the Web by providing meaningful labels for them; through the Peekaboom game (the Peekaboom Game, 2005), (von Ahn, Liu, Blum, 2006) people locate objects in images. Both games became very popular and the ESP Game has collected more than 10 million image labels.

The situation with texts seems to be different in some sense. While observing images, one can easily list possible labels for them. In the case of a text (a document), one has to read it to detail its topics. Certainly reading texts takes more time than observing images – the longer text, the worse. Since the game must be of a dynamic character it is unimaginable that the players will spend minutes reading an input text: No one will do it. From this point of view, a text must be open to the players 'role,' with a 'role' that will differ from game to game and mostly will be set up empirically.

We want to concentrate language games on one site. Thus we will have opened an on-line language game portal.

2 The Language Game Portal

The LGame language game portal (www.lgame.cz)[1] will concentrate on the language games designed to provide annotated textual data for those natural language tasks that either have not been implemented yet because of the lack of training data or have already been implemented, but their performance is lower than human performance. At the very beginning of designing the language games we formulated 10 basic features the games on the portal should satisfy:

1. Playing the game generates data that is valuable for natural language processing tasks.

[1] If not available, visit http://lgame.fuzzy.cz

2. During the game, the players are doing "something" until they achieve agreement. Like in the ESP game (the ESP Game, 2005), the players type in the possible image labels until they type the same label.

3. Playing the game requires either no knowledge of a given language grammar or basic knowledge of grammar acquired at school.

4. The game rules are language independent. No specific knowledge on the language under consideration is applied.

5. The game is provided for Czech and English by default.

6. The game offers different levels of difficulty.

7. The game is of a dynamic and quick character.

8. The game is of an interactive character. During the game, the players must have a general idea what the opponent(s) do.

9. The game is at least a two-player game.

10. The game has a separate top score list.

During the game's registration process, the information regarding year of birth, gender and mother tongue are collected. Once the player logs in, the games page is uploaded.

We have launched the language portal with the so-called Shannon Game, which can provide data for language modelling.

2.1 The Shannon Game

Mathematical background Shannon entropy (or information entropy) is a key term in information theory. It is a measure of the uncertainty associated with a random variable. The concept was introduced by Claude E. Shannon in his 1948 paper (Shannon, 1948). Mathematically speaking, the information entropy of a discrete random variable X, which can take on possible values $x_1, ..., x_n$ is

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log_2 p(x_i) \qquad (1)$$

Shannon (Shannon, 1950) estimated the entropy of written English based on the ability of human subjects to guess successive letters in a text.

Guessing the words, the size of the dictionary of the given language and the context information (looking back to n words) employed during the predicting of the words are the parameters that influence it the most.

Basic Game Play The beta version of the Shannon Game is designed as a two-player game. The players do not know each other and they are paired randomly by the system. Once the players are paired, the session can start. The goal is to guess words from the input sentence in the shortest time possible and using the smallest number of guesses.

Figure 1 illustrates the main interface of the Shannon Game. The input sentence is being selected from the sentence database. There is only one criterion restricting sentence selection – both players have not seen the given sentence during the sessions they have played so far, i.e. the input sentence is selected randomly from sentences unknown to both players. Based on the chosen difficulty level, some words of the input sentence do not appear and the pencils are displayed instead of them. Above the pencils, a circular target line of the length corresponding to the number of hidden words is displayed. Target is visualised in the same way as targets in the biathlon competition. While the biathlete can only see his/her targets, the Shannon Game players can see the target of his/her opponent as well – see line with the heading "Opponent". The targets are in grey by default. As the session goes on, the colour of the targets changes.

While guessing the word, a player has a limited number of guesses and each guess is limited by time. The course of the session is pictured by the colour of the given words, by the number of guesses already completed and by the elapsed time:

- Green - A player has already guessed the word.
- Gray - A player is still guessing. The number of guesses left is displayed.
- Red - A player has not guessed the word, i.e. all guesses have failed.

Since the players can see the targets of his/her opponent, they have an actual idea of how the opponent is doing at each time. The total game score for a given player is calculated with penalisations for incorrect guesses and for exceeding the time limit. Given that, being the first to reach the end of the sentence does not necessary mean that you are the winner. That is why the players get points when both of them reach the end of the sentence. The exact number of points depends on the total playing time and on the total number of guesses. Then, for both players, the overall score in the Shannon Game list is recalculated and their current position in the list is displayed.

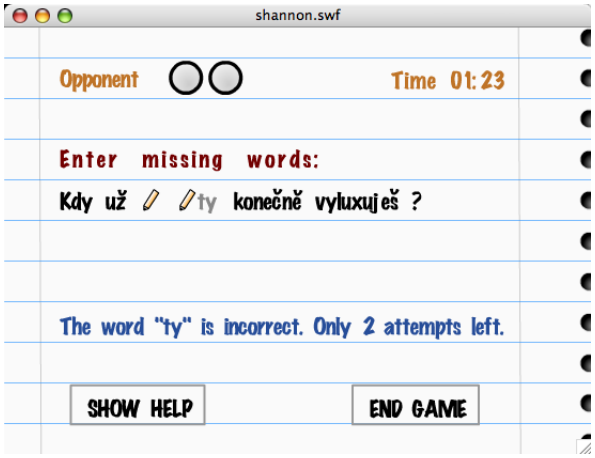

Figure 1: Main interface of the Shannon Game

Difficulty level In total, 3 difficulty levels (1-3) have been set up: simple, normal, hard. Each difficulty level corresponds to the proportional (to the chosen difficulty level) number of the input sentence words that are displayed to the players at the beginning of the session. The lower the difficulty level, the more words that are displayed. By default (no matter what difficulty level has been chosen), the first word of the sentence is displayed.

The sentence database Originally, we planned to have the Shannon Game session consisting of two rounds. During the first round, Player A writes a sentence and Player B guesses its words. In the second round, their roles change – Player B comes up with a sentence and Player A guesses. Since we were afraid of senseless and obscene input sentences, we decided to build a database of the 'safe' sentences from which the sentences for playing the Shannon Game will be picked up. A 'safe' sentence is recognised as a sentence that is syntactically correct. That is why the sentence databases were built from the syntactically annotated corpora – the Prague Dependency Treebank v. 2.0 (PDT, 2006) and the Prague Czech-English Dependency Treebank v. 1.0 (PCEDT, 2004). Both corpora are tokenised so the identification of a word was already solved earlier.

Mainly because of uncertainty during guessing (Shannon entropy – see above) sentences from the corpora had to pass through a filtering procedure that excludes sentences:

- of a length up to 2 words and more than 11 words, excluding final punctuation (i.e. the sentence databases consist of sentences of a length at least 3 words and up to 10 words),
- consisting of any punctuation marks, excluding the final punctuation,
- consisting of digit tokens,
- having either a preposition or conjunction or article at the beginning.

After the sentence filtering, the Czech sentences database consists of 10,000 sentences and the English database 10,000 sentences (not all sentences are included into the databases in the beta version).

Implementation The game server of the Shannon Game is implemented using PHP5 and the data is stored in the PostgreSQL database. The Flash communicates directly with the PHP scripts during the game as well.

What data is collected? For each session, the basic settings are stored: the language, the input sentence, the player's id, the difficulty level and the complexity parameters (number of guesses, time limit). The information about the course of the session is stored as well – see Table 1. For each word w_i of the input sentence and for each player, the guess(es)[2], time from when a guess was typed

[2]If the first guess is the correct one, then the other two are not needed.

	w_i		
$PlayerA$	$guess_{A1}$	$[guess_{A2}]$	$[guess_{A3}]$
	$time_{A1}$	$[time_{A2}]$	$[time_{A3}]$
	$0\|1$	$0\|1$	$0\|1$
$PlayerB$	$guess_{B1}$	$[guess_{B2}]$	$[guess_{B3}]$
	$time_{B1}$	$[time_{B2}]$	$[time_{B3}]$
	$0\|1$	$0\|1$	$0\|1$

Table 1: What data is stored?

and if it was correct (1) or incorrect (0) are stored. The number of such records for each sentence corresponds to the number of players who encounter the particular sentence while playing the Shannon Game, i.e. to the multiple of 2 (0, 2, 4, ...).

3 Conclusion

We have presented an idea of an online language game portal. At least to our knowledge, building the annotated textual data through the on-line games has not been experienced yet.

We are aware that the Shannon Game does not fulfill Feature 2. To start with the Shannon Game was motivated by the fact that we wanted to begin this project with the kind of game where the rules are already known, and we wanted to get an initial of how these sort of games would be accepted by the Internet game audience.

However, we cannot provide any feedback from users as of yet since the portal was launched at the same time as the submission of this paper.

Currently, we are working on games that will provide data for the tasks of coreference resolution and keywords detection.

References

Luis von Ahn. 2006. Games with A Purpose. *IEEE Computer Magazine June 2006*, 92-94.

Luis von Ahnand Ruoran Liu and Manuel Blum 2006. Peekaboom: A Game for Locating Objects in Images In *ACM CHI 2006*.

Claude E. Shannon. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal 27*, 379-423, 623-656.

Claude E. Shannon. 1950. Prediction and Entropy of Printed English. *Bell System Technical Journal 3*, 50-64.

The Peekaboom Game. 2005. http://www.peekaboom.org/

The ESP Game. 2005. http://www.espgame.org/

The Prague Depedency Treebank v. 2.0. 2006. http://ufal.mff.cuni.cz/pdt20

The Prague Czech-English Dependency Treebank v. 1.0. 2004. http://ufal.mff.cun.cz/pcedt