Acquiring NLP Data by means of Games

Marek Grác and Zuzana Nevěřilová

NLP Centre, Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic

Abstract. For some of the NLP tasks, obtaining appropriate data is very difficult. In this paper we concentrate on acquiring NLP data by means of games. Two different projects are presented and various aspects of these games are discussed.
First, we discuss public-made collections of linguistic data generally: the quality and reliability of contributors and collections, and the high dependence of number of contributions on motivation and contribution policy. Second, we describe creation of games for acquiring NLP data in detail. As example, two existing games are presented. Finally, evaluation techniques for both projects are discussed.

1 Introduction

Language resources suitable for natural language processing are one of the keys to develop a successful project. Data of higher quality necessary for some tasks tends to be obtained only with difficulty and a lot of time and manpower is needed. Using large corpora can help, but still data has to be verified by a human. There are two basic approaches on how to collect and/or verify this data. Work can be done by experts or by non-experts (usually volunteers). Both approaches and many in-between variants differ in several aspects such as cost, quality and coverage.

This article focuses on basic aspects of obtaining data from non-expert volunteers. According to [1], we expect to acquire plausible data with lower costs than in the case of expert annotators. We test this expectation on two different projects that both use game as a tool for obtaining data.

In Section 2 we introduce data collections made by general public and discuss the advantages and disadvantages of this approach. We formulate the terms under which a public-made collection can be useful. In Section 3 we present two games that were designed to collect linguistic data. We discuss common aspects and differences of the games. In Section 4 we outline the evaluation of collected data.

2 Public-Made Collections

Internet proves to be very useful tool for grouping people willing to help. They do not need to be at same place, in same timezone or even be willing to

Petr Sojka, Aleš Horák (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2010, pp. 109–114, 2010. © Tribun EU 2010

gather at same time. Research of crowdsourcing (crowd + outsourcing) became very popular in psychology and sociology. Crowdsourcing is still something that cannot be well defined but we can present some of the advantages and disadvantages.

One of the main advantages is that we can work with people who do not belong to one specific social group (e.g. academic) and so we can receive various views on same problem. Also we can speak to people who are not interested in helping continuously, but they just wish to correct some of our data.

The main disadvantage of this approach is the fact that contributors' expertise may differ a lot. Also we are not able to focus the crowd to work on a schedule or on subsets that are in the worst shape. Crowd will work on what it believes is best for it or tries to offer something better. There are several projects based on content delivered by volunteer work: Wikipedia[1], OpenMind Initiative [2], Games with a Purpose (GWAP Portal) [3], several games including Amazon Mechanical Turk [4] or OnlineWord Games for Semantic Data Collection [1].

This concept can bring plausible results if it fulfils the conditions described in the subsections below.

2.1 Motivation for Players

Since we can only consider the data valuable if it is of sufficient volume, we emphasise the motivation for players to play. People will play a game because it is enjoyable, not because it helps computational linguistics.

The motivation for players to play is fun. First, the design of the game is significant. Second, players playing a game have to beat high scores or advance to new levels, which is often a good motivation. In future, we may consider other motivation such as monthly prizes for best players.

2.2 Formulation of the Problem

The game has to be understandable. Although it can be difficult to play, it should not be difficult to understand the rules. Both games presented in this paper are quite difficult to play but take advantage of the fact that they are only slightly modified but well-known games.

Playing a difficult game is also motivating. Among thousands or even millions of web users, it becomes a challenge to get to the high score list.

From the computational linguistics point of view, we need the game rules to correspond to a problem under consideration. We need the highest possible number of contributors to input the 'right' data. Sometimes it is useful to put restrictions on game rules. Since we always work with semantic data we can set up the rules so that the obtained data will be semantically disambiguated.

[1] `http://en.wikipedia.org`

2.3 Game Policy and Quality of Contributions

There are several measure for quality of contributions depending of the game type. The most used and most straightforward is the agreement of several players.

We have to consider involuntary errors: For example, a time limit can lead to spelling errors. Players compelled by time limit often write the first idea that comes to their mind. For example: the task of *describing a frog* leads to descriptions such as 'frog is a princess' instead of 'frog is an animal that transforms to princess after you kiss it'. This is not necessarily a disadvantage.

We have to accept that not all contributors understood the game well. For this reason a reliability measure is considered useful for every game (and every set of contributors).

Besides involuntary errors, we have to cope with players contributing deliberately faulty inputs. Primarily we encourage players to register. Contributions by registered players are considered more reliable. Registration has to bring benefits such as higher levels or access to game statistics. In case of a large number of players, we have to find automatic or semi-automatic ways to discover 'hostile' contributors and filter out their contributions.

The games have to take into account language specific features. In case of Czech we have to deal with nominal inflection, by integrating a lemmatizer [5] for finding the appropriate basic form or (in case of X-plain) for generating appropriate word form.

Some web users are used to write *without* diacritics, even if they are normally used in Czech. We have observed the collected data for a period of time and decided that such users form a minority and words without diacritics can be disposed.

3 Games

The following subsections describe two existing games that were designed for collecting data for NLP.

They have several aspects in common such as:

- they refer to existing desktop games
- they are difficult to play
- they are extremely difficult to play for non-native speakers

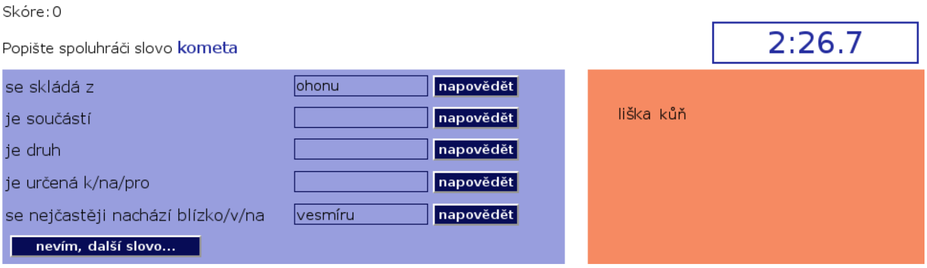
They differ in aspects such as:

- cooperative/competitive approach
- game based on human-computer or human-human interaction
- suitable for occasional/regular players.

3.1 X-Plain

X-plain [6] has analogy in board games or TV Shows. It is significantly inspired by Verbosity [7], but the engine is based upon word sketches provided by Sketch Engine [8]. It is a cooperative game for two players – a human and a computer. The principle is that a random word (called *secret word*) is displayed to one player (narrator) and s/he has to explain it to the second player (guesser). The guesser has to write down the exact word.

In X-plain the game is time-limited to 3 minutes, therefore it is suitable for occasional players. There are different relation types that together with the *secret word* and the *object* make sentence templates, e.g. `X is_kind_of Y`.



Skóre: 0

Popište spoluhráči slovo kometa

2:26.7

se skládá z	ohonu
je součástí	
je druh	
je určená k/na/pro	
se nejčastěji nachází blízko/v/na	vesmíru

liška kůň

Fig. 1. Screenshot (part) from X-plain: narrator (human) has to describe the word "kometa" (comet). On the left s/he has to fill the following sentence templates: se skláda z (consists of); je součástí (is part of); je druh (is a type of); je určena pro/k/na (is used for); se nejčastěji nachází blízko/v/na (can be likely found). They type: "...se skládá z ohonu (...has part tail). On the right the guesser (computer) tries to guess the secret word: "liška" (fox), "kůň" (horse). There is a countdown timer in the top right corner of the screen.

X-plain is a web-based application and its server side is programmed in PHP, while the client side uses Javascript and AJAX[2] for better user comfort. Contributions are stored in MySQL database.

Figure 1 shows the game interface. When human plays the role of the narrator, his descriptions are stored in form of triples (subject, relation, object) together with their number of occurrences. Triples contain words or word expression in their base forms (lemma), as provided by a lemmatizer [5]. The database is already quite large (nearly 5,000 unique triples in October 2010) and continuously grows. So far the only criterion of contribution quality is frequency: the more often a triple appears, the more probably it is a 'good' one.

3.2 Game of Scrabble

Second game we wish to present is based on the well-known game Scrabble. In this game, players compete against each other to obtain the highest score. They

[2] Asynchronous Javascript and XML

are using letters with different point values and they use the letters to create new word(s) on a gameboard. This game is one of the most popular word games in the world.

There is no problem to play Scrabble on the Internet even for minor languages like Czech. Sites are available where you can meet other players and start a new two-player game. Unlike in X-plain game there is no direct way how to focus player efforts on a specified subset of our problem. Even though the data is more diffused, its quality is much higher because each player's turn has to be confirmed by another player. Some of the word forms are used quite rarely outside of Scrabble world, so developer of the game is extending dictionary of correct words and such words are confirmed automatically. Thus a good dictionary is in players' interest.

Another difference between Scrabble and X-plain is that Scrabble is time-constrained but one game usually takes at least half an hour. There are players who take this game quite seriously and they play more than 150 games per month. If a player plays as many games and they have good winning ratio, we can consider their data to be 'better' then average. It is in our best interest to keep these players interested in our games. To achieve this, we need to offer additional services, even though they do not give us useful data directly—however, they really help to make players more loyal.

We can use scrabble as a tool not only for obtaining new words but also for verifying our existing morphological database. This is not very useful for common words but there are a lot of word forms that are not widely used and some of them are not covered by existing corpora. Good scrabble players tend to know and discuss these forms and their verification can improve our morphological database too.

4 Conclusion and Future Work

This paper describes a different approach to collecting linguistic data. It is designed mainly for collecting linguistic data for Czech language. Czech is a minor language, therefore we cannot expect millions of contributions within a few months like GWAP [3] and have to attend strongly to players' motivation. On the other hand, we can assume only native-speakers will play and no foreigners' language errors will appear.

For each game we record a game history (in case of Scrabble with response latency). Therefore we can identify the pitfalls that players have to face. Further analysis should answer the question why some cases are 'easy' and others are not. We have to carefully choose the data for each level so that players stay motivated.

Beyond these practical questions concerning the games themselves we have to test the resulting collections. We expect that a reasonable number of contributions will be collected over time. We also expect that evaluation techniques to reduce noise in the data will need to be designed in the future.

So far, the data is still not so numerous for a serious evaluation. We plan to evaluate each collection by different means. In case of X-plain, the rising associative network can be compared to other associative networks such as Czech WordNet [9]. Some types of relation are expected to appear in both resources. In case of Scrabble, the evaluation is planned to be manual or semi-automatic and the method of multiple annotation will probably be used.

Acknowledgements

This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536.

References

1. Vickrey, D., Bronzan, A., Choi, W., Kumar, A., Turner-Maier, J., Wang, A., Koller, D.: Online word games for semantic data collection. In: EMNLP '08: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2008) 533–542.
2. Stork, D.G.: Open mind initiative — about (2007) Retrieved October 28, 2007 from `http://openmind.org`.
3. von Ahn, L.: Games with a purpose. Computer **39**(6) (2006) 92–94.
4. Snow, R., O'Connor, Jurafsky, D., Ng, A.: Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks. Proceedings of EMNLP-08 (2008).
5. Šmerk, P.: Fast morphological analysis of Czech. In: Proceedings of the Raslan Workshop 2009, Masarykova univerzita (2009).
6. Nevěřilová, Z.: X-plain – a game that collects common sense propositions. In: Proceedings of NLPCS, Funchal, Portugal, SciTePress (2010) 47–52.
7. von Ahn, L., Kedia, M., Blum, M.: Verbosity: a game for collecting common-sense facts. In: CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, New York, NY, USA, ACM (2006) 75–78.
8. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The Sketch engine. In: Proceedings of the Eleventh EURALEX International Congress. (2004) 105–116.
9. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (1998).