X-plain: A Game that Collects Common Sense Propositions

Zuzana Nevěřilová

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

Abstract. Common sense knowledge is very important for some NLP tasks, but it is hard to extract from existing linguistic resources. Thus specialized collections of common sense propositions are created. This paper presents one of the ways of making such collection w.r.t. Czech language. We have created a cooperative game, where computer program plays together with human. The purpose of the game is to describe a word with short sentences to the co-player. While the human player is expected to use his/her common sense, the computer program uses word sketches. The paper describes in detail the game, its background and discusses the need for motivation and game policy. It also discusses the quality and coverage of the collection.

1 Introduction

Common sense knowledge is considered to be crucial for some NLP tasks. In principle there are two approaches on how to collect common sense data: collection made by experts, collection made by volunteers. Both approaches and many variants between them differ in several aspects such as cost, quality, coverage.

We present a project, where common sense propositions are collected by means of a game. In this article a Czech version of the game is presented. However the principle can be used for different languages as well.

The game presented in this paper is named X-plain. Players are at the same time contributors to the database of common sense propositions. Section 2 describes the common sense and explains the need for collecting it. In section 3 we describe the principle of the game. Section 4 describes closely how computer program can play together with human. In section 5 we discuss the quality of collected data and contribution policy. We have to expect that the database will be error prone and different contributors have different reliability.We propose some work in future in section 6.

2 Common Sense and How to Collect it

Common sense is often described as a huge set of processes of natural cognition and system of beliefs that people share. Common sense does not always correspond to scientific or even real world observation, rather it is a set of assumptions about the real world [6].

Inherently common sense propositions are not easy to collect. Therefore specialized collections of common sense exist. Well-known projects include CyC [4], Thought

NevÄŽÅŽilovÃą Z.
X-plain: A Game that Collects Common Sense Propositions.
DOI: 10.5220/0003015200470052
In *Proceedings of the 7th International Workshop on Natural Language Processing and Cognitive Science (ICEIS 2010)*, page
ISBN: 978-989-8425-13-3

Treasure [5] (both expert-made) or Open Mind Common Sense Initiative [8] (volunteer-made).

The game Verbosity [10] proposes another way of collecting common sense propositions. All mentioned projects contain mainly data in English language. This paper refers about a game similar to Verbosity, but with different engine. Its main purpose is to create a collection of common sense propositions in Czech language.

3 Game Principle

X-plain has analogy in board games such as “Taboo$^{\mathrm{TM}}$”. It is a cooperative game for two players. 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 say (or write down) the exact word.

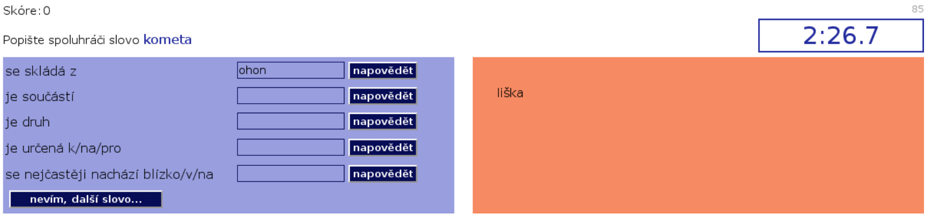
In X-plain the guesser tries to guess the word with apparently unlimited number of tries. When s/he is successful, the score is increased and next turn the roles swap. When the narrator is not able to describe the *secret word* or the guesser is not able to reveal it, they can pass on the word. Next turn the roles swap but the score stays unchanged. The game is time limited to 3 minutes.

In X-plain there are different *relation types* that together with the *secret word* and the *object* make sentence templates, e.g. X `is_kind_of` Y. Currently there are following relation types:

- `can_have_property`
- `has_part`
- `is_part_of`
- `is_a_type_of`
- `is_used_for`
- `can_be_used_for`
- `can_be_likely_found`
- `is_the_opposite_of`
- `is_similar_to`
- `is_related_to`
- `using`

At first relations were selected according to Verbosity. Afterwards the list was adapted to Sketch Engine outputs (see subsection 4.2). These relations are considered to be easy to understand, however it seems that players attach significant importance to *secret words* and *objects* and not to *relations* (see section 5).

Secret words were selected randomly, one-meaning words are preferred. The list is continuously adapted, as the words are examined by human players and Sketch Engine (“difficult” words are rejected), see section 5.



Skóre: 0

85

Popište spoluhráči slovo kometa

2:26.7

se skládá z ohon

je součástí

je druh

je určená k/na/pro

se nejčastěji nachází blízko/v/na

liška

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 (has part); je součástí (is part of); je druh (is a type of); je určená pro/k/na (is used for); se nejčastěji nachází blízko/v/na (can be likely found). S/he types: "… se skládá z ohonu" (… has part tail). On the right the guesser (computer) tries to guess the *secret word*: "liška" (fox), "kůň" (horse).

4 Game Background

There is a significant difference between X-plain and Verbosity: in Verbosity two human players (that are chosen randomly from on-line players) play together, whether in X-plain human plays with computer program. The program has to take role of the second player. The program's "knowledge" is based upon two resources: previous contributions and word sketches.

X-plain is a web-based application where server side is programmed in PHP[1]. Client side uses Javascript and AJAX[2] for better comfort. Thus players do not have to install special software. Contributions from human narrators are stored in MySQL[3] database in form of triple (subject, relation, object) together with its number of occurrences. Explanations given by computer program are not stored because they result from the database itself or from the Sketch Engine (see subsection 4.1).

4.1 Word Sketches

Word sketch [2] is made from corpus using grammar patterns. It groups together words playing the same grammatical role in sentences. The Sketch Engine [3] is supplied with grammatical relations for the requested language.

Grammatical relations for Czech include three types: symmetric, dual and trinary (explained in detail in [1]). For Czech language the words are in grammatical relations such as:

- `coord` – words in coordination, typically nouns connected by conjunctions "and", "or". This relation is symmetric.
- `prec_<preposition>` – the word followed by `<preposition>` and `X`. This relation is trinary.
- `a_modifier` – adjective word modifier. This relation is dual to `modifies`.

[1] http://www.php.net
[2] Asynchronous Javascript And XML
[3] http://www.mysql.com

4.2 From Grammar to Semantics

In X-plain the relations in sentence templates are semantic, but in word sketches only grammatical relations exist. Therefore, we propose a set of rules that link grammatical and semantic relations. The idea is similar to grammatical relations in Sketch Engine: the rules are quite straightforward and the results do not tend to be perfect, but plausible. Currently there are grammar-to-semantics rules such as:

```
is_related_to  ⟹  ["coord"]
is_part_of  ⟹  ["gen_1"]
has_part_of  ⟹  ["gen_2"]
can_have_property  ⟹  ["a_modifier"]
```

The first rule is interpreted as "relation type `is_related_to` relates the *secret word* to all words from word sketch `coord` (coordination)". Similarly reverse grammar-to-semantics rules exist:

```
is_related_to  ⟹  ["coord"]
is_part_of  ⟹  ["gen_2"]
has_part_of  ⟹  ["gen_1"]
can_have_property  ⟹  ["modifies"]
```

4.3 Use of Word Sketches in the Game

In the role of narrator, X-plain looks for *object*s in the database of contributions (the result is a set), creates word sketch for the *secret word* and obtains a set of words depending on *grammar-to-semantics rules*. As explanation of the *secret word* it chooses randomly some word from the union of the two sets.

Conversely, in the role of guesser, the program looks for *subject*s in the database of contributions, creates word sketches for the *object* and obtains set of words depending on *reverse grammar-to-semantics rules*. X-plain tries to guess the *secret word* from words randomly chosen from the union of the two sets.

Example The *secret word* is "kometa" (comet)

- narrator (human) fills template: … souvisí s vesmírem (… is related to space)
- guesser (computer) gets "hvězda" (star) from the database and science, solar, astronomy … from word sketchesGuesser chooses following words: astronomie (astronomy), věda (science), hvězda (star)
- narrator (human) fills template: … může mít vlastnost Halleyova (… can have property Halley's)
- guesser (computer) gets no results from the database, but one result from word sketches: "kometa" (comet).
- success! (players score points)

So far human players score points 1,24× more often than computer. For making computer program more successful we can arrange the results from database and word sketches and do not choose randomly but consider the frequency. On the other hand the more successful the computer is less the propositions we collect.

Table 1. Relations that are used with same subjects and objects: X relation 1 Y and X relation 2 Y.

relation	relation	% of occurrences
is_similar_to	is_related_to	1.08
is_similar_to	is_a_type_of	0.51
is_related_to	is_part_of	0.47
is_related_to	is_a_type_of	0.47
can_be_likely_found	is_part_of	0.47

5 Game Policy and Quality of Contributions

A simple measure for quality of contributions is the agreement. Since common sense propositions are not a scientific approach we do not need to collect the "truth". All we need is the usage. Where a proposition repeats from different contributors, it means that several players think the same way about the *secret word*.

Players are playing with time limit, so they often write the first idea that comes to their mind. When collecting common sense propositions, this is rather an advantage. On the other hand the time limit can lead to many spelling errors.

In the data we have already collected (about 2200 propositions), the *relation type* is often misused. For example in the database we can find records such as: X is_similar_to Y and X is_opposite_of Y. This has not to be error in all cases, however we cannot weight the *relation type* same as the *secret word* or the *object*. Table 1 shows what types of relations (the most occuring cases) are used with the same subject and object and their occurrence ratio in the whole collection.

An important aspect of the collection is the coverage. We can observe that some words are passed very often with no propositions: either they are not understood by players or they are "hard" to explain. Table 2 shows words that are poorly covered and their categorization. The majority of them are abstract words and we can assume that these words are difficult to explain.

Table 2. Words difficult to explain for humans and their categorization. Number of unsuccessful guesses take in account only games where human player gives at least some clue.

word	translation	number of unsuccessful guesses	category
zpronevěra	fraud	5	abstract words
zkouška	exam/testing	4	abstract words, polysemes
myslivost	woodcraft	3	domain specific terms
nemocný	sick/invalid	3	polysemes
vztah	relation	3	abstract words
copyright	copyright	2	abstract words
demokracie	democracy	2	abstract words
guvernér	governor/proconsul	2	polysemes
hrana	edge/angle/knell	2	polysemes
lesák	woodlander	2	domain specific terms

6 Conclusions and Future Work

This paper describes another approach to linguistic data collecting. It is designed mainly for collecting common sense propositions within Czech language. Czech is a minor language thus we cannot expect millions of propositions within a few months like GWAP [9]. We are strongly interested to players' motivation.

Game history is available for each game, so we can identify the words that are hard to explain (many passes, few propositions) or conversely the words that are easy to explain (best scored guesses). Further analysis should answer the question *why* some words are "easy" and others are not. We have to carefully choose the words for each level so that players stay motivated.

The major contribution of this work is the method how to collect common sense propositions in Czech. We have to evaluate the reliability of the collection over time. We expect that a plausible number of common sense propositions will be collected over time.

Acknowledgements

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

References

1. *Corpus querying and grammar writing for the sketch engine*. Retrieved March 2, 2010 from http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying.
2. Kilgarriff, A. and Rundell, M. (2002). Lexical profiling software and its lexicographic applications - a case study. In *Proceedings of the Tenth EURALEX International Congress*, pages 807–818.
3. Kilgarriff, A., Rychlý, P., Smrž, P., and Tugwell, D. (2004). The sketch engine. In *Proceedings of the Eleventh EURALEX International Congress*, pages 105–116.
4. Lenat, D. B. (1995). CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38.
5. Mueller, E. T. (2003). *ThoughtTreasure: A natural language/commonsense platform*. Retrieved November 9, 2009 from http://alumni.media.mit.edu/ mueller/papers/tt.html.
6. Smith, B. (1995). Formal ontology, common sense and cognitive science. *International Journal of Human-Computer Studies*, pages 641–667.
7. Stork, D. G. (2001). Toward a computational theory of data acquisition and truthing. In *COLT '01/EuroCOLT '01: Proceedings of the 14th Annual Conference on Computational Learning Theory and and 5th European Conference on Computational Learning Theory*, pages 194–207, London, UK. Springer-Verlag.
8. Stork, D. G. (2007). *Open mind initiative — about*. Retrieved October 28, 2007 from http://openmind.org.
9. von Ahn, L. (2006). Games with a purpose. *Computer*, 39(6):92–94.
10. von Ahn, L., Kedia, M., and Blum, M. (2006). Verbosity: a game for collecting commonsense facts. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 75–78, New York, NY, USA. ACM.