Acquiring Data for Textual Entailment Recognition

Zuzana Nevěřilová

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

Abstract. Language resources are hardly ever large enough. Building language resources that can be used as a gold standard for semantic analysis requires effort and investment. We present a prototype for acquiring language resources by means of a language game which is a cheap but long-term method.
Games employed to acquire language resources are not new. For example games with a purpose are used for collecting common sense knowledge. The game presented in this paper is a work in progress. It collects annotated pairs text–hypothesis suitable for recognizing textual entailment in Czech.
The game narrative is based on Sherlock Holmes and dr. Watson dialogues. For generating the dialogue line we use rule-based approaches such as syntactic analysis, anaphora resolution, synonym and hypernym replacement, word order rearrangement and verb frame based inference. To generate natural sounding sentences we added a language model score (based on n-gram frequencies in a corpus).

Key words: textual entailment, language game, games with a purpose, GWAP

1 Language Resources and Data Complexity

Although Czech is spoken only by about 10 million people it cannot be considered as a less resourced language. However, Czech language resources (LR) follow the typical distribution when sorted by their complexity: the more complex a resource is the smaller it is.

N.B. that the term *language resource complexity* is not defined but it is often mentioned when describing a LR. According to [16] LR complexity means the data size as well as its characteristics relevant to annotation.

Currently no LR for recognizing textual entailment is available for Czech.

2 Recognizing Textual Entailment

"A fundamental phenomenon of natural language is the variability of semantic expression, where the same meaning can be expressed by, or inferred from, different texts." [2, p. 2]

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2013, pp. 29–37, 2013.

Recognizing Textual Entailment (RTE) is defined as follows: "A text t entails a hypothesis h ($t \Rightarrow h$) if humans reading t will infer that h is most likely true." [3, p. 18]

Although RTE seems to be defined imprecisely ("humans will infer", "most likely"), it is one of the most well defined problems in semantic analysis. RTE systems are evaluated by a collection of pairs text–hypothesis (h–t pairs). Each pair can be (repeatedly) annotated either as true (if t entails h) or false (if t does not entail h).

Building a collection of h–t pairs of a considerable size and diversity is a challenging task. Possible resources include (manually prepared) reading comprehension tests for children and for adults (such as PISA[1] or PIAAC[2]) as well as automated techniques. [2] describes four scenarios that lead to creation of h–t pairs in RTE2 challenge[3]. These scenarios are less applicable to Czech since many tools are in development (e.g. information retrieval system for Czech described in [7]). We therefore propose an alternative method for obtaining annotated h–t pairs by means of a game.

3 Acquiring Annotated Data

Games with a purpose (GWAP) is a new concept [1] in the field of *collaboratively constructed language resources* (CCLR). The idea is based on collective "human computation" where peoples' brains are used for solving problems that are difficult for computer programs (such as natural language understanding or image content recognition). Because GWAPs are games, the main motivation for contributors is the fun.

X-plain [12] is a game for one player whose purpose is to collect common sense propositions. It can be played in Czech and Slovak [8]. Three years after its first release X-plain collected 14,898 unique assertions in Czech and 5,703 unique assertions in Slovak. Some of the assertions are entered repeatedly, for example the assertion "South is the opposite of North" was entered six times in Czech version. The average ratio of repeated assertion is 1.1025 for Czech version and 1.2372 for Slovak version. The number of assertions is increasing every month depending on players' interest in the game[4].

4 The Game

The results of this one-player game and its power to acquire LRs (in middle-term and long-term) encouraged us to develop a new game which purpose is to collect annotated h–t pairs. Both true and false entailments are needed,

[1] `http://www.oecd.org/pisa/`

[2] `http://www.piaac.cz/`

[3] `http://pascallin2.ecs.soton.ac.uk/Challenges/RTE2`

[4] The provided numbers date back to 2013-11-07.

however users are likely to annotate clear true entailments and feel many false entailments annoying.

The game is based on several existing modules for natural language analysis and generation such as morphological analyzer and syntactic parser.

4.1 The Game Narrative

The game narrative refers to one element of detective stories: a dialogue between the detective and his/her assistant. The purpose of the dialogue is to explain the detective's reasoning that lead to the criminal case solving to readers.

In the game, the dialogue starts with a (short) criminal case the detective (human player) presents to his assistant (the program). The assistant tries to reformulate the story and to infer new facts. The basic screen with sample dialog is shown in Figure 1. The player can judge assistant's effort true or false or mark a sentence (syntactically) wrong.

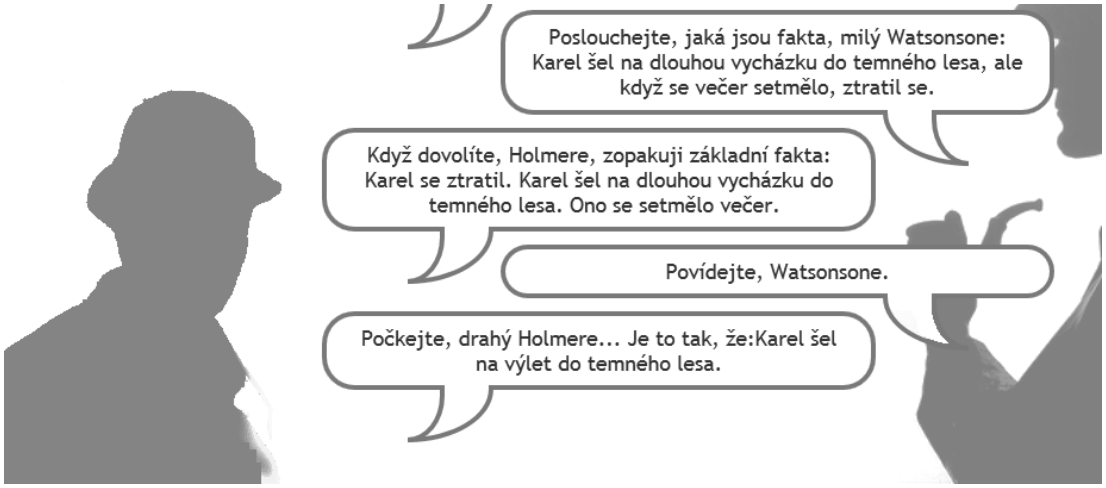

Fig. 1: The game environment is a dialogue between the detective Sherlock Holmer and his assistant dr. Watsonson.

From the RTE's point of view the human player enters a text t, the computer player proposes several hypotheses h and the human player annotates the entailment $t \Rightarrow h$. Depending on particular modules (see 4.3) h vary from simple paraphrases (i.e. syntactic rearrangements) to entailments.

Repeated annotations can be obtained as well. When a player is not about to tell a new story s/he can choose "get back to an old story". In this case a random story is selected from the story database.

4.2 User Experience and Data Complexity

Although the task of entailment is not even easy for humans (otherwise reading comprehension tests would not be used for testing people's understanding capabilities), the game is intended for non-expert users without any training. A progression loop is a typical game design element [17] and the game provides levels as usual. At each level more modules for paraphrasing and entailment are employed. We suppose the increase of complexity in each level reflects the intricacy of the entailments. Thus experienced players will be "trained" by the game itself.

The data complexity in relation with CCLRs is widely discussed in [16, p.10]. The players' task is somewhat similar to that in GWAPs but also to that of Wisdom of the Crowds (WotC). Unlike GWAPs no instant human feedback is present, but long-term feedback similar to Open Mind Common Sense [14] exists.

4.3 Modules

The application is based on several modules. The stories and entailments are not represented by a formalism (such as first order logic). Instead step-by-step the input sentences are transformed syntactically. For further processing, each transformation records its originator.

Parsing and partial anaphora resolution Players (in the detective's role) are asked to input a short story, thus a few sentences. The SET parser [9] divides each sentence on clauses if necessary and represents each clause as a set of sentence constituents (verb phrases, noun phrases, prepositional phrases, adverbial phrases, coordinations). Not individual words but phrases are subject of further processing.

At this phase the Czech anaphora resolution system Saara [13] supplements unexpressed subjects and replaces demonstrative pronouns with their antecedents. Table 1 outlines the processing of an example story. All other modules do not process sentences but these phrasal representations of individual clauses.

Individual phrases are marked according their syntactic roles, e.g. if the phrase's case is nominative it is marked as subject (SUBJ), if the phrase is an adverb it becomes the adverbial (ADV). At this level we cannot distinguish between an adverbial and an object therefore *do temného lesa* (in the dark forest) is marked as object (OBJ) although it is an adverbial.

Word reordering Czech is a (so called) free word order language i.e. *nearly* all orders of sentence constituents are allowed. Several aspects of word ordering in Czech (e.g. clitics, modal verbs in verb phrases, reflexive particles) are discussed in [5]. Different word orders signal different discourse structures but do not change the truth value. Thanks to this fact further processing leads to mostly correct results.

Table 1: Story representation: each sentence is divided in clauses, each clause is parsed on phrases. Phrases are marked according their syntactic roles: SUBJ(ect), VERB phrase, OBJ(ect), REFL(exive particle), ADV(erbial).

Karel šel na dlouhou vycházku do temného lesa, ale když se večer setmělo, ztratil se. Karel went for a long walk in a dark forest but when it got dark in the evening he got lost.										
Karel šel na dlouhou vycházku do temného lesa Karel went for a long walk in a dark forest				ono se večer setmělo it got dark in the evening				Karel se ztratil Karel got lost		
Karel	jít	(na) dlouhá vycház-ka	(do) temný les	on	se	večer	setmět	Karel	se	ztratit
Karel	go	(for) long walk	(in) dark forest	it		in the evening	get dark	Karel		get lost
SUBJ	VERB	OBJ1	OBJ2	SUBJ	REFL	ADV	VERB	SUBJ	REFL	VERB

Synonym and hypernym replacement We use Czech WordNet [15] for synonym replacement. No word sense disambiguation method is used and therefore false paraphrases are generated as long as the module replaces all possible word expressions of the story with their synonyms in all senses registered in Czech WordNet.

Since all transformations originators are recorded we can later discover WordNet synonyms unlikely in stories. For example *pes* has two senses: one corresponds to the synset `dog:1, domestic dog:1, Canis familiaris:1` in Princeton WordNet [4], the other corresponds to `martinet:1, disciplinarian:1, moralist:2`. A preliminary search in existing h–t pairs indicates the unlikely occurrence of the second sense in stories. In fact, none of the hypotheses generated with the replacement *pes–moralista* (moralist) were judged true. An example synonym replacement is shown in Table 2.

Similarly to synonym replacement phrases or their parts are replaced by their hypernyms. In this case two restrictions apply. First, we do not replace word expression by all hypernyms but omit those from the WordNet Top Ontology. Such replacement (e.g. replace "student" by "living entity") will never generate a natural sounding expression. Second, we do not replace by hypernyms in sentences with negative polarity. While in positive sentences (such as "He came in his new coupe.") the hypernym replacement (replace "coupe" by "car") is valid, in negative sentences the replacement results always in false entailments ("He did not came in his new coupe." does not entail "He did not came in his new car.").

Table 2: Synonym replacement using Czech WordNet: “vycházka” (walk) was replaced by “výlet” (trip). N.b. that the modifier “dlouhý” (long) had to be modified to fulfill the grammatical agreement with “výlet” (trip).

Karel Karel	jít go	(na) dlouhá vycházka (for) long walk	(do) temný les (in) dark forest
SUBJ	VERB	OBJ1	OBJ2
Karel Karel	jít go	(na) dlouhý výlet (for) long trip	(do) temný les (in) dark forest

Verb frame inference Word reordering and synonym replacement result in paraphrases while verb frame inference can result into new facts. In this module we take advantage of the Czech verb valency lexicon VerbaLex [6] and use verb valency frames for inferences of three types: equality, effect, precondition. The inference process is described in detail in [11]. It consists of several transformations of all phrases matched by the verb frame as shown in Table 3.

Table 3: This verb frame inference corresponds to the common sense inference “If someone gets lost s/he becomes unhappy.”

Karel Karel	se ztratil got lost
SUBJ $\rightarrow$ SUBJ SUBJ $\rightarrow$ SUBJ	ztratit se $\rightarrow$ být nešťastný get lost $\rightarrow$ to be unhappy
Karel Karel	byl nešťastný was unhappy

Natural sounding sentences The system generates sentences using one, two or three modules. Even with only these three modules (word order, synonymy and hypernymy, verb frame inference) the application can produce tens to hundreds of sentences at one time. Since players find annotation of many sentences (that may be very similar) annoying we use a language model to select the most natural sounding sentences.

The appropriate n-gram frequencies were calculated using the Czes corpus, normalized (divided) by $corpsize$[5]. The resulting score is calculated according to Equation 1 where $ngrams$ means the n-gram raw frequency. n-grams of higher

[5] 465,102,710 tokens in 2013-11-07

n are more important for natural sounding sentences therefore they obtain a higher weight (by multiplication by a higher number).

$$score = 10^2 \sum \frac{2grams}{corpsize/2} + 10^3 \sum \frac{3grams}{corpsize/3} + \cdots + 10^5 \sum \frac{5grams}{corpsize/5} \quad (1)$$

Sentence Generation Each module uses an independent module for generating syntactically well-formed sentences. Sentence generation in Czech is a complex task because of grammar agreements between the verb and the subject and between noun phrase and its modifiers. The module for sentence generation declines all noun phrases and prepositional phrases using the application described in [10] and conjugates verb phrases as well.

4.4 The Game Loop

First, the player is asked to input a story or to choose a random existing story. Second, the story is scored (according to the number of its clauses, the number of different verbs and the number of named entities recognized). The player gains points for a new story or less points for choosing an existing story. Third, the parsed story is reproduced using the generation module. The player is asked to evaluate the reproduced story. Fourth, paraphrases and entailments ordered by its natural sounding score are proposed to the player for evaluation. For each annotation the player gets a point. In case the entailment (with the same annotation) is already in the database the player gains two points. Afterwards, the player can either add sentences to the story or begin a new game loop.

5 Conclusion and Future Work

We presented a new annotation game which aims to create a collection of h–t pairs for future Czech RTE system. The prototype is working although it misses modules for specific types of inference (see below). It can be found at `http://nlp.fi.muni.cz/projekty/watsonson/`. Our outlook is that in a few years we can obtain a large collection of stories, hypotheses and their annotations as well as information about the way the hypotheses were generated. The contribution of this work is therefore twofold: create a new CCLR and provide feedback to diverse software tools contributing to the generation process.

Feedback for existing software tools New sentences annotations provide information about:

- the distribution of correct and natural sounding word orders
- the distribution of Czech WordNet senses in stories
- the quality of syntactic parsing using SET
- the quality of anaphora resolution using Saara

Future Work We plan to add modules for entailments about time, locality, modality as well as involve the encyclopedic knowledge to the entailments.

With new modules transforming relative time to absolute time and vice versa can be entailed (e.g. transform "last Christmas" to "2012-12-25"). With encyclopedic knowledge module transformations like "Edvard Munch's The Scream" to "Edvard Munch painted The Scream" will be possible.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Ahn, L.v.: Games with a purpose. Computer 39(6), 92–94 (2006)
2. Dagan, I., Dolan, B., Magnini, B., Roth, D.: Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering 15, i–xvii (10 2009), http://journals.cambridge.org/article\char‘_S1351324909990209
3. Dagan, I., Roth, D., Zanzotto, F.M.: Tutorial notes. In: 45th Annual Meeting of the Association of Computational Linguistics. The Association of Computational Linguistics (2007)
4. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998)
5. Grepl, M., Karlík, P.: Skladba spisovné češtiny. Edice Učebnice pro vysoké školy, Státní naklad. (1986), http://books.google.cz/books?id=yV1iAAAAMAAJ
6. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for Czech. In: Proceedings of the Slovko Conference (2005)
7. Ircing, P., Pecina, P., Oard, D., Wang, J., White, R., Hoidekr, J.: Information retrieval test collection for searching spontaneous Czech speech. In: Matoušek, V., Mautner, P. (eds.) Text, Speech and Dialogue, Lecture Notes in Computer Science, vol. 4629, pp. 439–446. Springer Berlin Heidelberg (2007), http://dx.doi.org/10.1007/978-3-540-74628-7\char‘_57
8. Kostolná, M.: Vylepšení hry X-Plain (X-plain Enhancement). Bachelors thesis, Masarykova univerzita, Fakulta informatiky (2013)
9. Kovář, V., Horák, A., Jakubíček, M.: Syntactic analysis using finite patterns: A new parsing system for Czech. In: Human Language Technology. Challenges for Computer Science and Linguistics. pp. 161–171. Springer, Berlin/Heidelberg (2011), http://dx.doi.org/10.1007/978-3-642-20095-3\char‘_15
10. Nevěřilová, Z.: Declension of Czech noun phrases. In: Radimský, J. (ed.) Actes du 31e Colloque International sur le Lexique et la Grammaire. pp. 134–138. Université de Bohême du Sud à České Budějovice (République tchèque), České Budějovice (2012)
11. Nevěřilová, Z., Grác, M.: Common sense inference using verb valency frames. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) Proceedings of 15th International Conference on Text, Speech and Dialogue. pp. 328–335. Springer, Berlin / Heidelberg (2012)
12. Nevěřilová, Z.: X-plain – a game that collects common sense propositions. In: Sharp B., Zock M. (eds.) Proceedings of NLPCS. p. 47–52. SciTePress, Funchal, Portugal (2010)

13. Němčík, V.: Saara: Anaphora resolution on free text in Czech. In: Horák, A., Rychlý, P. (eds.) Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012. pp. 3–8. Tribun EU, Brno (2012)
14. Speer, R.: Open mind commons: An inquisitive approach to learning common sense. In: Workshop on Common Sense and Intelligent User Interfaces (2007)
15. Vossen, P.: EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Computers and the humanities, Springer (1998), http://books.google.cz/books?id=-qEep-1ib8UC
16. Wang, A., Hoang, C., Kan, M.Y.: Perspectives on crowdsourcing annotations for natural language processing. Language Resources and Evaluation 47(1), 9–31 (2013), http://dx.doi.org/10.1007/s10579-012-9176-1
17. Werbach, K.: Gamification: Course wiki. online (2013), [accessed 2013-11-07 from https://share.coursera.org/wiki/index.php/Gamification:Main]