Building Evaluation Dataset for Textual Entailment in Czech

Zuzana Nevěřilová

NLP Centre, Faculty of Informatics,
Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic
xpopelk@fi.muni.cz

Abstract. Recognizing textual entailment (RTE) is a subfield of natural language processing (NLP). Currently several RTE systems exist in which some of the subtasks are language independent but some are not. Moreover, large datasets for evaluation are prepared almost exclusively for English language.
In this paper we describe methods for obtaining test dataset for RTE in Czech. We have used methods for extracting facts from texts based on corpus templates as well as syntactic parser. Moreover, we have used reading comprehension tests for children and students. The main contribution of this article is the classification of "difficulty levels" for particular RTE questions.

Key words: textual entailment

1 Introduction

Automatic reasoning systems are currently a promising application of Natural Language Processing (NLP). Since automatic natural language understanding (NLU) is a topic difficult to grasp and formalize scholars try to resolve sub-problems of it. Hence recognizing textual entailment (RTE) is a good application of NLU methods.

Basically RTE solves a yes/no question: whether a text T entails a hypothesis H. In most cases H is a sentence a T is a coherent text – a "story". *T entails H* if the meaning of H, as interpreted in the context of T, can be deduced from the meaning of T [1]. In this context *deduction* is not equal to logical deduction and has to be understood in a broader context. It is considered that systems with high precision on deciding the RTE question "understand" a text in natural language. Apart from being a good evaluation measure RTE can aim for several applications such as intelligent searching or automatic summarization.

However, large resources for testing RTE systems are needed. In this paper we describe the process of building a gold-standard for evaluation of a RTE system. Currently a RTE system is being developed and its preliminary results were promising. However, we cannot evaluate further improvements if have no testing data.

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 61–66, 2012. © Tribun EU 2012

In section 2 we describe the state-of-the-art both in developing RTE systems and creating test data sets. Section 3 presents several methods for creating test data sets from corpus and other resources as well. We present and discuss the "difficulty levels" of RTE and their evaluation.

2 State-of-the-art

Recognizing textual entailment represents an important domain of research. Since 2004 RTE Pascal challenges started with manually annotated datasets of texts and hypotheses (H-T pairs) that covered seven different tasks:

- information retrieval
- comparable documents
- reading comprehension
- question answering
- information extraction
- machine translation
- paraphrase acquisition

In each of the subsets the annotators either generated hypotheses or identified H-T pairs from texts. Afterwards, annotators decided whether given text entailed the hypothesis or not. Thus the datasets contain both positive and negative examples. Moreover annotators were asked to replace anaphora by their appropriate references so that the RTE task would not concern anaphora resolution [3].

Pascal Challenges took place from 2004 to present time (last challenge was RTE-7 in 2011) and the datasets are available. The data is stored in an XML format describing `pairs` and their sub-elements: `text` and `hypothesis`. We adopted this format for our new resource of Czech H-T pairs.

Recent RTE systems use different techniques how to decide whether *T* entails *H*. Apart from the ad-hoc and shallow approaches the sound approaches (e.g. [11]) use

- tree transformation operations that generate the hypothesis from the given text
- knowledge based operations

Tree transformation operations concern computing tree edit distance (insertion, deletion, substitution) as well as rules for entailment and contradiction. For example replacing a token (word or word expression in the parse tree of the sentence) by its antonym leads (in most cases) in contradiction.

Knowledge based operations concern generalization using a knowledge base (e.g. WordNet [5] or dbPedia [9]) or antonymy processing. Missing knowledge is considered to be a bottleneck of RTE.

Representants of working systems are: BIUTEE[1], EDITS[2], VENSES[3] or Nutcracker[4].

3 Collection H-T pairs

RTE applications use several methods for automated entailment judgment. We have to reflect this fact when preparing RTE datasets. We also wanted to keep the information about extraction of the H-T pairs as well as the "difficulty level" of the entailment. The latter is not easy to obtain. However, we propose a classification of the pairs in the following subsection.

3.1 Reading comprehension tests for children/adults

We have analyzed reading comprehension tests for children and secondary school students. The classification reflects common reading comprehension problems w.r.t. reader's age.

- subsequence – the easiest entailment, most often it is a true entailment
- synonyms – replace a word in H by its synonym, obtain H' and then $H' \in T$, true entailment
- siblings – a word w_h in H is a sibling of a word w_t in T (w_h and w_t have common (direct) hypernym), but w_h and w_t are not synonyms, false entailment
- specification – a word w_h in H is a hyponym of a word w_t in T, false entailment
- syntactic rearrangement – H is a reformulation of a sentence in T, e.g. active–passive transformation or subordinate clause–object transformation
- interchanged arguments – all words from H are present in a sentence from T but their order or syntactic arrangement is different, false entailment
- qualifiers – the meaning of H is modified by a qualifier, judgment or by hedging
- anaphora – H is a paraphrase of a sentence s in T, but s contains anaphora and H contains reference, entailment value depends on anaphora resolution
- other – the meaning of H is not present in the context of T, other knowledge (encyclopedic, mathematical etc.) is needed or H is off-topic (and then the entailment is negative)

We have started with tests for 7-years-old children [10]. So far we have collected 12 documents with tests. We did a classification on 34 H-T pairs. Afterwards we have classified 24 H-T pairs extracted from secondary school leaving exam. Table 3.1 shows classification results. In tests for 7-years-old children

[1] `http://u.cs.biu.ac.il/~nlp/downloads/biutee/protected-biutee.html`

[2] `http://edits.fbk.eu/`

[3] `http://project.cgm.unive.it/venses.html`

[4] `http://svn.ask.it.usyd.edu.au/trac/candc/wiki/nutcracker`

Table 1. Classification of reading comprehension tests. Question types that are frequent in tests for 7-years-old children (left column) are expected to be easier to solve than questions frequent in tests for 18-years-old students.

question type	7-years	18-years
subsequence	20 %	0 %
synonyms	17 %	12 %
siblings	35 %	8 %
specification	2 %	4 %
syntactic rearrangement	5 %	50 %
interchanged arguments	5 %	3 %
qualifiers	0 %	17 %
anaphora	0 %	4 %
off-topic	16 %	21 %

each question is in one class, in final exam test several techniques are used at the same time (e.g. syntactic rearrangement together with hedging). Therefore the sum of the rightmost column is greater than 100 %. The classification was done by one annotator since it is a preliminary phase of the dataset development.

3.2 Corpus patterns

While observing questions in reading comprehension tests we have proposed several templates for extracting facts from corpora. Some parts of the templates are language independent while other are language dependent. We have used Corpus Query Language (CQL) in The Sketch Engine corpus tool [6]. This task is inspired by information extraction applications.

We were working with the Czech morphologically annotated and disambiguated corpus czes that contains 465,102,710 tokens[5].

Enumeration We have extracted nouns and adjectives following a noun in accusative with the column sign and delimited by commas and conjunctions a, nebo, ani (and, or, neither–nor). The hypothesis is then built rearranging the enumeration items, e.g. for a text “Každý objekt obsahuje tři základní datové komponenty: data, metody a atributy.” (Each object contains three basic data components: data, methods and attributes.) we obtain three hypotheses such as “Metody jsou komponenty objektu.” (Methods are components of the object.). This method extracted 738 hypotheses.

Passive We have extracted sentences with the verb *to be*, a passive verb and noun in instrumental. This is a typical passive construction in Czech and it is relatively easy to transform such sentences to active. We have

[5] 2012-06-21 size

obtained hypotheses such as “Lesy obklopují obec” (Forests surrond the village) from passive constructions such as “Obec je obklopena lesy.” (The village is surrounded by forests). This method extracted 12.573 hypotheses.

Aliases We have extracted sentences containing “also known as”. The hypothesis is created by stating that the alias is other name for an object, e.g. “Václava Zapletalová, jinak zvaná Wendy” (Vaclava Zapletalova, also known as Wendy) resulted to the hypothesis “Václavě Zapletalové se říká Wendy” (Wendy is a different name for Vaclava Zapletalova). This method extracted 26 hypotheses.

For sentence generation we used a system for Czech noun phrases declension [8]. This system is built upon the morphological analyser/generator `majka`.

In the last stage we are planning to use the tool `efa` for fact extraction [2]. It is based on syntactic parser SET [7] but moreover modules for recognizing information about time, location and manner are implemented.

3.3 Annotation

All these methods are used to extract H-T pairs from Czech texts. We plan to annotate each pair at least by two annotators independently. Moreover, in secondary school final exams correct answers are available. We expect high inter-annotator agreement in case of 7-years-old children and low inter-annotator agreement in case of secondary school final exam. In other methods we expect high coverage and high inter-annotator agreement since the methods are quite straightforward. According to [4] we plan to compute inter-annotator agreement. However, we plan to exclude H-T pairs where annotators will not agree on. The aim is to build a dataset with clear distinction what is a valid entailment and what is not.

4 Conclusion and Future Work

We have presented building an evaluation dataset for a system for recognizing textual entailment. We propose several resources of H-T pairs: reading comprehension tests, corpus querying using templates and fact extraction software. We have also presented an approach for judging the difficulty level of particular H-T pairs.

Future work concerns retrieval of more data from corpus. We will observe reading comprehension tests a create more patterns for paraphrasing sentences extracted from corpus.

In future we have to annotate the data by multiple annotators and to evaluate the inter-annotator agreement.

Acknowledgments

This work has been partly supported by the Czech Science Foundation under the project P401/10/0792 and by the Ministry of Education of CR within the LINDAT-Clarin project LM2010013.

References

1. Akhmatova, E.: Textual entailment resolution via atomic propositions. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment (April 2005)
2. Baisa, V., Kovář, V.: Information extraction for czech based on syntactic analysis. In: Vetulani, Z. (ed.) Human Language Technologies as a Challenge for Computer Science and Linguistics, Proceedings of 5th Language and Technology Conference. pp. 466–470 (2011)
3. Challenges, P.: Recognising textual entailment challenge. online at `http://pascallin.ecs.soton.ac.uk/`
4. Dagan, I., Dolan, B., Magnini, B., Roth, D.: Recognizing textual entailment: Rational, evaluation and approaches. Natural Language Engineering 15(Special Issue 04), i–xvii (2009), `http://dx.doi.org/10.1017/S1351324909990209`
5. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998), published: Hardcover
6. Kilgarriff, A., Rychlý, P., Smrž, P., Tugwell, D.: The sketch engine. In: Proceedings of the Eleventh EURALEX International Congress. p. 105–116 (2004), `http://www.fit.vutbr.cz/research/view_pub.php?id=7703`
7. 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: 4th Language and Technology Conference, LTC 2009, Poznan, Poland, November 6-8, 2009. p. 161 (2011)
8. Neverilová, Z.: Declension of czech noun phrases. In: Actes du 31e Colloque International sur le Lexique et la Grammaire. pp. 134–138. České Budějovice (2012)
9. Orlandi, F., Passant, A.: Modelling provenance of DBpedia resources using wikipedia contributions. Web Semantics: Science, Services and Agents on the World Wide Web 9(2), 149 – 164 (2011), `http://www.sciencedirect.com/science/article/pii/S1570826811000175`, <ce:title>Provenance in the Semantic Web</ce:title>
10. Střední odborná škola Otrokovice, s.s.v.a.z.p.d.v.p.p.: Pracovní listy k nácviku porozumění čtenému textu a nápravě čtení. `http://www.zkola.cz/zkedu/pedagogictipracovnici/kabinetpro1stupenzsamaterskeskoly/metodickematerialyvyukoveprogramy/pracovnilistykporozumenitextuanapravecteni/default.aspx` (2003),
11. Stern, A., Dagan, I.: A confidence model for syntactically-motivated entailment proofs. In: Proceedings of the International Conference Recent Advances in Natural Language Processing 2011. pp. 455–462. RANLP 2011 Organising Committee, Hissar, Bulgaria (September 2011), `http://www.aclweb.org/anthology/R11-1063`