Verbs as Predicates: Towards Inference in a Discourse

Zuzana Nevěřilová, Marek Grác

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

Abstract. Generally, it is usual that communication partners mention only facts that are not believed to be known to both parties of the communication. This shared implicit knowledge is also known as common sense knowledge. Because of not mentioning everything important it is difficult (if not impossible) to analyze a discourse for computer programs without any background knowledge.
In this paper a relationship between logic and language is discussed. In NLP systems verbs are often seen as predicates and verb valencies are seen as arguments. According to this approach new sentences (propositions) can be inferred from the discourse.
Second, a system for inference from verb frames was created and a evaluation proposal is described. We have picked up 174 verbs occurring in Czech cooking recipes. For these verbs 232 inference rules were manually created. The inference process was tested on a corpus of 37 thousands tokens (2 400 sentences). As the result 253 new sentences were generated.

Key words: inference, common sense inference, corpus annotation, verb valency lexicon

1 Introduction

Much information in a discourse is not explicit. For example, the cookbook story "fry the onion till it looks glassy" actually means peel a fresh, uncooked onion, chop it, put grease into a cooking pot and heat it, put the onion into the pot and wait until the onion looks glassy. In natural language processing (NLP) systems we have to deal with this phenomenon to resolve stories such as: fry the onion till it looks glassy, reduce heat and cover. Where the heat comes from? What to cover?

Texts in natural languages usually contain "facts" (also known as common sense propositions or common sense facts) that are considered to be true in "normal" situations (also referred as stereotypical information [6]), e.g. fried onion looks glassy. From such facts some other propositions can be inferred, e.g. the glassy onion was fried in a cooking pot.

Henry Lieberman argues that common sense inference (CSI) differs from mathematical inference (MI). While MI operates with exact definitions, universally true statements and provides complete reasoning, CSI operates with

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2011, pp. 19–25, 2011.

vague definitions, therefore it infers contingent statements that are considered true until the opposite cannot be proved (also known as non-monotonic reasoning) [8]. In [13] a broader definition of logic is provided: "any precise notation for expressing statements that can be judged true or false". In the same context an inference rule is defined as "a truth-preserving transformation: when applied to a true statement, the result is guaranteed to be true".

In this paper we concentrate on inferring new propositions, obvious for humans, but unreachable for computer programs. The method is based on transformations on syntactic level and evaluation on semantic level. This means that the system works with syntactic units such as noun phrase (NP) and verb phrase (VP) and during the evaluation the meaning of the proposition is examined. As an example domain the Czech cooking recipes corpus was created and processed.

Within this domain we have constructed 232 inference rules for 174 verbs. The inference process was tested on a corpus of 37 thousands tokens (2 400 sentences).

2 Logics and inference

Mathematical inference takes place in several logics. Propositional (Boolean) logic is considered to be the basic logic. It uses well known reasoning patterns such as modus ponens or and-elimination [11, p. 239]. Propositional logic is useful since it is easy to implement, but for complex reasoning tasks it is not enough expressive. First-order logic introduces quantifiers, predicates and variables.

In knowledge representation mathematical logic (such as propositional or first-order logic) is not used directly. Usually knowledge is stored as objects in categories [11, p. 350] and the basic reasoning pattern is the inheritance. "Description logics provides a formal language for constructing and combining category definitions and efficient algorithms for deciding subset and superset relationships between categories." [11, p. 377]

Unlike MI principal inference tasks in DL are subsumption (checking if one category is a subset of another by comparing their definitions) and classification (checking if an object belongs to a category) [11, p. 381]. The DL is usually represented by a set of inference rules describing stereotypes ("normal" situations).

3 Verb Frames

Verb frames are closely related to the argument structure of sentences, but also to the lexical meaning of the VP itself and its dependents. "Argument structure is an interface between the semantics and syntax of predicators (which we may take to be verbs in the general case). Its function is to link lexical semantics to syntactic structures." [1]

Verbs mostly describe an action or state. Since the verb "is a hook upon which the rest of a sentence hangs" [12], it is often seen as a predicate (for example $to\_fry(x, y)$ means that x fries y). Verb valency refers to the number of arguments of a verbal predicate — the capacity of a verb to bind a certain number and type of syntactically dependent language units [9]. Syntactic valencies describe the syntactic properties (such as subject or object) of an argument. In Czech (as well as other Slavic languages) syntactic properties are expressed by the case and possibly a preposition (e.g. syntactic subject is in nominative).

VerbNet (English language) [12] and VerbaLex (Czech language) [4] are examples of current collections of verb frames and their arguments (in the frame lexicons often called slots). These collections capture the syntactic information (e.g. information about prepositions and cases of the arguments in VerbaLex) as well as semantics (reference to semantic roles and Princeton WordNet [2] (PWN) hypernym in VerbaLex, flat representation using predicates in VerbNet). In our work we proceeded from VerbaLex and its relation to PWN.

4 Inferring New Propositions

In this section we describe the whole process from preparing the data and the inference rules to applying the rules to corpus data and generating new sentences.

4.1 The Language of Cooking Recipes

The language of cooking recipes differs from the general language in the following attributes:

- use of imperative. In Czech cooking recipes most cooking recipes authors use first person plural instead of imperative (literally "we fry the onion…"). Sometimes, infinitive (literally "to fry the onion…") or imperative forms are used.
- frequent use of coordinations of NPs and of VPs
- use of adverbs describing duration and manner

4.2 Corpus Annotation

The annotation method was that of the BushBank project [3]. The corpus was annotated on several language levels: tokens (words and sentence boundary marks), morphology (lemma and morphologic tag for tokens), syntactic structures (NPs, VPs, coordinations and clauses), syntactic structures relations (dependencies). Annotation itself was done purely by automatic tools (desamb [10], SET [5]) and manual annotation was used for confirmation of syntactic structures and relations between them. It means that structures that were not identified by automatic tools could not be added by annotators.

This was done contrary to traditional requirements in which we tried to obtain completness of annotation. BushBank ideas put greater impact on *simplicity* of annotation (without definition of all border-line cases), *usability* (proved by this project of inference) and *rapid-development* (annotation itself was done in 40 (wo)man-hours). As we are working on concept, data were manually checked by just one annotator.

4.3 Creating Inference Rules

Verb valencies refer to the arguments of a verbal predicate (see 3. Therefore valencies play a critical role in inference. In this section we will outline the inference:

Let C_i be the input clause, I be the inference rule and C_o be the output clause. Then the inference is a function $I(C_i) = C_o$.

Both C_i and C_o are sets of NPs and a VP: $C_i = \{N_{i1}, N_{i2}, \ldots, N_{im}, V_i\}$ and $C_i = \{N_{o1}, N_{o2}, \ldots, N_{on}, V_o\}$. The inference rule describes:

- the type of the inference (see below)
- what NPs will participate the transformation
- what syntactic changes are needed for each transformed NP

Therefore the inference rule I is a tuple $(\mathbb{S}, t)$, where $\mathbb{S}$ is a set of syntactic transformation rules and t is the inference type (see below). Each transformation rule $S \subset \mathbb{S}$ defines a transformation for an input pair $(preposition_i, case_i)$ to an output pair $(preposition_o, case_o)$. Prepositions can be either none (direct case) or prepositions agreeing to a case. Case is marked be a number[1].

The system covers the following inference types t: effect, precondition, decomposition, equivalence. These relations are often used in discourse planning and are therefore present in common sense knowledge bases such as CyC [7]. The use of inference types is more general than just saying *a implies b*. Moreover, it is relatively straightforward for humans to describe inference rules of these types.

4.4 Generating New Propositions

In this section the whole process of inferring new propositions is described. First, the input text is parsed by the SET syntactic parser [5]. The output of this parsing consists of decomposition on sentences and clauses, identification of constituents of the clause and grammatical properties of NPs.

Second step consists of annotation of valencies. This step is necessary to eliminate NPs independent of the verb (e.g. parts of adverbial phrases). This process is so far done manually but with the use of verb valency lexicon VerbaLex [4]. Annotators had to detect if a NP is a valency of the verb according to VerbaLex. They could benefit from VerbaLex binding to PWN (e.g. they can easily detect what meaning of the verb to choose).

[1] 1 – nominative, 2 – genitive, 3 – dative, 4 – accusative, 6 – locative or 7 – instrumental

```
<title>'opéct' has precondition 'rozpálit pánev'</title>
<verbalex:inference type="precondition" verb="opéct">
   <verbalex:ruleset id="heat_pan" inferred_verb="rozpálit pánev"
      negation="False">
     <verbalex:rule case="c1" prep="" inferred_case="c1"
        inferred_prep="" />
     <verbalex:rule case="c4" prep="" inferred_case="c4"
        inferred_prep="na" />
   </verbalex:ruleset>
</verbalex:inference>
```

Fig. 1. Inference rule that says that *roast* (opéct) has precondition of *heating the pan* (rozpálit pánev)

Third step is the application of inference rules. Each NP identified as a verb valency and the VP are the input of inference rules (such as in Figure 1. If the grammatical agreement in case and preposition is fulfilled the NP is transformed according the particular syntactic transformation. The morphological parser majka [14] is used for generating the new NP (the grammatical number is the one property not affected by the transformation). Note that not all NPs contained in the inference rules has to be present. In such cases only those NPs that are present are transformed. This can sometimes lead to semantically incomplete sentences such as "we have the meat".

New sentences are generated as a sequence of new NPs and a VP. Since Czech language uses nearly free word order this method is correct. If the syntactic transformation is correct the method generates a syntactically correct phrase.

5 Evaluation Proposal

Since the project is a work-in-progress this section will only outline the evaluation steps. Evaluation will proceed on two levels: syntactic and semantic. At each level annotators will have to decide whether or not the new sentence is correct. On syntactic level, there is one possible source of ambiguity: mistaking nominative/accusative forms. This ambiguity is eliminated during the (manual) valency detection. Therefore a failure of the new sentence on this level points to either an error in the inference rule, or corpus annotation error, or morphological generator error.

On semantic level only syntactically correct (those with a good annotator agreement on syntactic level) sentences will appear. Here the annotators have to decide whether or not the sentence is "true". The truthfulness will be judged in the context between two clauses of the original text, first of them is the inference input clause. Here we expect low coverage but high precision.

We have chosen a well delimited domain of cooking recipes. Therefore we should avoid most problems with lexical ambiguity. However, we expect that

problems with lexical ambiguity arise in case of highly polysemous verbs such as put, give or bring. In this case further evaluation of the results will be needed.

6 Conclusion and Future Work

In this paper we have discussed the relation between natural language and logic. We suppose that logic and inference are widely used within a discourse but we have to broaden the definition of inference. With the notion of inference as a transformation we did experiments on a cooking recipes corpus. The result of applying inference rules on a 37 000 size corpus is that 253 new sentences were generated. Since the paper is a preliminary research in this area the evaluation is not finished, but a evaluation method is proposed.

In future, we plan to evaluate in detail the "quality" of the new sentences and examine the failures on each level. Moreover we plan to add adverbial constituents to inference rules because in the case of cooking recipes we have observed that the adverbs strongly influence the semantic of the inferred proposition. Next step will consist of picking up other objects that are not mentioned in the discourse. In case of cooking recipes these objects will be pans, pots, spoons, grease etc.

Acknowledgements This work has been partly supported by the Ministry of Education of CR within the Center of basic research LC536 by the Czech Science Foundation under the project 407/07/0679.

References

1. Bresnan, J.: Lexicality and argument structure. In: Minimal Ideas. pp. 283–304. John Benjamins Publishing Company, Paris (Apr 1996)
2. Fellbaum, C.: WordNet: An Electronic Lexical Database (Language, Speech, and Communication). The MIT Press (May 1998), published: Hardcover
3. Grác, M.: Case study of BushBank concept. In: PACLIC 25 25th Pacific Asia Conference on Language, Information and Computation. p. 9 (2011), [release Dec 2011]
4. Hlaváčková, D., Horák, A.: Verbalex – new comprehensive lexicon of verb valencies for czech. In: Proceedings of the Slovko Conference (2005)
5. 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)
6. Lehmann, D.J.: Stereotypical reasoning: Logical properties. CoRR cs.AI/0203004 (2002)
7. Lenat, D.B.: CYC: a Large-Scale investment in knowledge infrastructure. Communications of the ACM 38(11), 33–38 (1995)
8. Lieberman, H.: MAS.964 Common Sense Reasoning for Interactive Applications. Massachusetts Institute of Technology: MIT OpenCourseWare (2002), [Retrieved on-line from `http://ocw.mit.edu`; accessed 11-January-2006]. License: Creative Commons BY-NC-SA

9. Lopatková, M.: Valency Lexicon of Czech Verbs: Towards Formal Description of Valency and Its Modeling in an Electronic Language Resource. Habilitation thesis, Charles University in Prague, Prague (2010)
10. Pavel Šmerk: Towards Morphological Disambiguation of Czech. Ph.d. thesis proposal, Faculty of Informatics, Masaryk University (2007)
11. Russell, S.J., Norvig, P., Candy, J.F., Malik, J.M., Edwards, D.D.: Artificial intelligence: a modern approach. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1996)
12. Schuler, K.K.: VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon. Ph.D. thesis, Faculties of the University of Pennsylvania (2005)
13. Sowa, J.: Fads and fallacies about logic. IEEE Intelligent Systems p. 84–87 (2007)
14. Šmerk, P.: Fast morphological analysis of czech. In: Proceedings of the Raslan Workshop 2009. Masarykova univerzita (2009)