Czech Knowledge-Based System with Temporal Reasoning

Andrej Gardoň

Faculty of Informatics, Masaryk University
Brno, Czech Repulic
xgardon@fi.muni.cz

Abstract. In this paper we discuss recent advancements in the field of knowledge representation and question-answering using natural language interfaces. The focus is to reveal projects' weak capabilities in temporal processing, namely time modeling and reasoning over temporal facts. With these deficiencies in mind, we propose new knowledge-based system called Dolphin suitable to parse sentences in Czech language, identify and model temporal information and infer answers to questions regarding time information. Fundamental theory behind Dolphin is Transparent Intensional Logic, which is high-order intensional logic calculus. Finally, we discuss Dolphin evaluation and future work.

Key words: question-answering, time representation, Dolphin

1 Introduction

Advancements in computer vision [1], artificial muscle fiber [2] and state-of-the-art systems including IBM Watson [3] have fundamentally influenced interaction between computers and humans. Projects like Watson, TrueKnowledge [4] or AURA [5]) aim at providing natural language interface to a user. They answer questions about factual or even scientific knowledge. Expansion of smart-phones, equipped with powerful processors, allows research in personal assistants that interact with the user and provide necessary information in real-time. Siri is an assistant for Apple phones and it focuses on question-answering in given domains. It waits for activation words and then uses so called active ontologies to answer the question. Siri even tells jokes [6]! Google's huge database of information, stored within its search engine, provides factual answers on natural language inputs to Google Now assistant[1]. It can also find navigation guidelines. Evi[2] is based on TrueKnowledge technology and is especially built for everyday information retrieval including answering the request *"give me a list of restaurants in my area"*. Aforementioned projects prove the possibility to move from standard key-driven approach in information retrieval

[1] `http://www.google.com/landing/now`

[2] `http://www.evi.com`

Aleš Horák, Pavel Rychlý (Eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2012, pp. 43–50, 2012.

towards more human-like methods. While processing of natural language input is well established, processing of temporal aspect is almost new feature in the projects.

2 Time representation and temporal reasoning

Time is an inseparable part of the space. Both components form so called time-space, the World in which our lives take place. Almost any utterance in natural language regards time information either in explicit or implicit form. For search engines like Google, representing temporal information is not mandatory as seeking for factual information is nearly time independent. However, assistants like Siri or Evi must realize time dimension to become human-like secretary. It is surprising that the most advanced question-answering system Watson is only able to detect basic temporal expressions. Its processing engine detects TLink (defined in TimeML [7]) relation in the input sentence and candidate answers. That relation specifies the end points of a general temporal relationship. To eliminate candidates that are temporally incompatible with the question, Watson computes *before* feature of the clue to determine whether it is chronologically before the entity came into the existence [8]. TrueKnowledge defines Temporal partner as a fact that references other fact and makes assertion about its validity. Key element of temporal partner is a Timepoint, single moment on the time line. Timeperiod is a class of two time points and is equivalent to an interval with start and end points. To cover infinite and indefinite time periods there are special objects including *iafter=unknown point in the future*, *timezore=point in time infinitely ago*, *forever=infinite future* [9]. Aura is a project that aims to provide scientific knowledge to everyone by natural language interface to books. HaloBook is a next generation electronic book based on Aura technology. The reader can interactively read the text (self-explanatory tooltips), ask questions and get explanations or learn the content through individualized tutoring [10]. Besides these advanced features, AURA supports just basic event ordering and partitioning of knowledge into situations. A situation represents changes of facts during time. Ordering of situations is also possible [11]. Advanced model of time is available in Cyc ontology. It utilizes two abstractions of time. The first one is interval-based and defines relations between intervals thanks to the primitives: *before* and *simultaneousWith*. The second one is set-based and it can state axioms like *"people do not eat and sleep at the same time"*. Temporal projection axioms associate default periods of persistence with propositions. With their use, it is possible to infer the name of a person in the future [12]. ThoughtTreasure is a comprehensive platform for natural language processing and common sense reasoning. It can keep a calendar up-to-date based on natural language input. Timestamps are basic objects for time representation and their combinations form timestamps ranges. Durations define the length of a given interval and Extended timestamps ranges represents

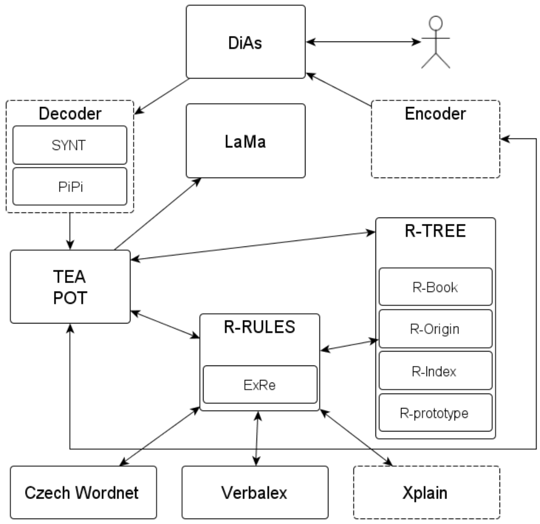

Fig. 1. Architecture of the Dolphin system

repeating or regularly scheduled events. ThouhtTreause looks to have the most advanced time model. On the other side it is ontology oriented. Set of specialized agents analyzes input texts and extract information that is encoded in the self-contained ontology. To utilize ThoughtTreasure on a new domain, the ontology has to be extended and agents must be added [13].

While temporal processing by knowledge systems is in early ages, support by formal languages is well established. TimeML is an example. This annotation language addresses time stamping of events, ordering of events, reasoning with underspecified temporal expressions (e.g. last week) and reasoning about the persistence of events. Events are defined as situations that happen or occur, and their detection is separated from identification of relations between them [7]. Interval temporal logic represents both propositional and first-order reasoning over period of time[3] . Transparent Intensional Logic (TIL) is a comprehensive high-order calculus that supports grammatical tenses, representation of events and episodes. It can process sentences like *"I go swimming every Friday"* and undefined time points and intervals like in *"I was there"*. Besides excellent temporal support, TIL can handle personal attitudes and incorporates the model of possible worlds. All these benefits make it good candidate for a knowledge-representation [14].

3 Dolphin system

Dolphin is a knowledge-based system that aims to process sentences in natural language, translate them into TIL formulas and provide reasoning capabilities to allow question-answering. Motivation for the project is to develop an

[3] `http://en.wikipedia.org/wiki/Interval_temporal_logic`

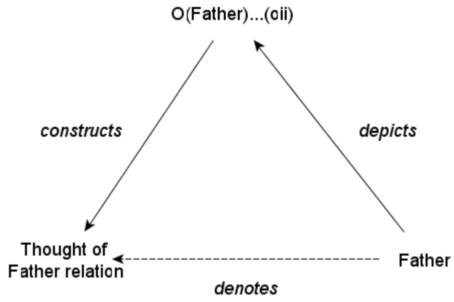

Fig. 2. Meaning of the Father relation

advanced knowledge-based system primarily understanding other language than English. In case of Dolphin, the mother language is Czech. As far as the authors known, there is no such system yet. Orientation on Czech is not limiting as architecture of the system is language independent. To support new language, one has to provide language specific decoder and encoder modules. Main goal of the project is advanced temporal processing. The system identifies most of temporal expressions in sentences and disposes by inference rules for their processing. Events and episodes are useful concepts from TIL that enable Dolphin to represent coherent view of subsequent situations and model relations between them. TIL supports deep analysis of grammatical tenses, undefined time intervals, repetitive activities and thus handles the most of time information in any sentence. Gardoň and Horák provide comprehensive introduction to temporal processing in Dolphin [15]. Following sections briefly describe essential modules from Dolphin's architecture depicted on Figure 1

- ***DiAs*** (Dialogue Assistant) handles the dialogue between a user and the system. It proposes questions to be asked and manages inputs/outputs from decoder and encoder modules.
- ***Decoder*** is a plug-in module that generally incorporates parser and PiPi (Post Processing). Input is the sentence in given language. After processing, there is a TIL construction on the output. Parser analyses the sentence to obtain syntactical, lexical, morphological and relational data from the sentence's parts. Based on the analysis, the parser translates the sentence into TIL construction. PiPi has to evaluate additional information from the sentence and prepare final output. For example, it utilizes anaphora resolution tool. In case of a question, it transforms simple TIL construction into TIL match that orders subsequent modules to treat the input as a query. In case of ambiguous parsing, decoder informs subsequent module about alternative inputs. For the purposes of Dolphin, we have utilized Synt parser that processes Czech sentences [16] and Saar a system for automatic anaphora resolution [17].
- ***LaMa*** (Language Manager) is a key component responsible for language independency of the system. TIL constructions encodes crucial information

from the sentence in procedure-like fashion. For example *"John is a father of Andrew"* is represented as a procedure Father over arguments John and Andrew. LaMa maps any procedure or argument on a referent's ID. TIL is based on the three-fold expression-meaning relation as depicted on the Figure 2. The picture represents meaning of the relation *Father* that is mapped on corresponding TIL construction (sense). The construction is an abstract procedure that construct a referent. Dolphin system approximates every referent in object-oriented manner, and it assigns universal ID to such approximation that is called Referent.

– ***TEA POT*** is heart of the system. TEA (Til Execution Agent) first invokes LaMa to translate TIL construction into R-format (words are substituted for referent IDs). The result of TEA is named R-TIL (referential TIL). TEA passes the processing of R-TIL to POT (Processing and Organizing Tool) that parses the construction and identifies its components (sub constructions). POT utilizes R-TREE and R-RULES to propose a response on the input that is send to Encoder module.
– ***R-TREE*** is the knowledge-base. It stores R-TIL constructions and provides methods for their manipulation. R-TIL Construction is kind of a function with exactly one result. Thus every R-TIL construction stores referenced R-TIL construction. Knowledge is organized into possible worlds, fundamental concept of TIL. It allows one fact to have different truthfulness depending on the considered world. It also makes it possible to process dialogue with a user in a separate world and reflect new knowledge only after final checking proved the consistence with the general world. Knowledge-base incorporates four components:
  - *R-Book* – basic evidence of registered Referents. It allows answering questions like *"Is every object that you know red?"*
  - *R-Origin* – a list of R-TIL constructions that referent particular R-TIL construction. This component tracks every natural language representation that denotes given Referent.
  - *R-Index* – indexing component that provides rapid search.
  - *R-Statistics* – Automatic prototype generation. It can be turned on/off for every R-TIL construction. Constructions representing a set have prototype generation initially activated. Prototype is a statistical member of a set. It represents the distribution of properties over legitimate set members. This allows to infer answer for question *"Can every bird fly?"*. With the knowledge that pigeon cannot fly, corresponding distribution of fly property for every bird is under 100%. Thus, Dolphin can generate answer *"Almost all birds can fly"*.
– ***ROOT*** is brain of the system. It consists of the repository of inference rules and ExRe (External Resources) component. These rules can be invoked to prove a statement, find an answer for a question, keep the consistency of the R-TREE or generate new knowledge. Besides rules in R-TIL format (e.g. R-TIL representation of if-then statements), ROOT contains direct C++ functions (C-rules) mapped on particular Referents. C-rules allow Dolphin to be interactive with outside components or boost the performance of

heavy-load inference rules treated in R-TIL-way. ExRe component is an API for external resources. Dolphin system employs Czech WordNet and Verbalex[4] as essential resources of common-sense. When the system needs additional information about Referent (not found in R-TREE), it queries ExRe component to possible obtain it. In future, interactive games like X-plain[5] can be used to ask a user for required knowledge.
- ***Encoder*** plug-in module that transforms R-TIL constructions back to natural language sentences.

4 Evalution and future work

Dolphin system is the highlight of author's thesis. In three years horizon, the goal is to implement essential Dolphin modules namely LaMa, TEA POT, R-TREE and ROOT. Previous research and prototype implementations of Dolphin system [18], [19], and [20] already contain parts of aformentioned modules and are valuable sources of information for the planned work. We expect the ROOT component to contain inference rules for temporal processing and ExRe implementations for Czech WordNet and Verbalex. To evaluate the system capabilities we plan to build a corpus of condensed news articles. Lidovky.cz or Tyden.cz provides summarizations at the top of each article. These summarizations are ideal start texts for the corpus. Basic Czech decoder (based on Synt and Saara) and manual refinement will transform corpus knowledge into internal R-TIL representation. This combination will also handle mainly time oriented questions contained in the corpus. These questions will demonstrate the system abilities to handle undefined time intervals, repetitive actions and contiguous events grouped in episodes. Output will be presented in clear and readable form (Encoder is not a part of current effort). In future, we hope to complete the Decoder for Czech to automatically parse any input. With the finalization of Encoder and Dias, Dolphin will become full-value member of the question-answering community. Our long-time goal is to develop Siri like assistant oriented on time management and planning.

5 Conclusions

This paper provides brief overview of current advancements in question-answering systems. It identifies a gap in their temporal abilities and proposes the architecture of Dolphin system that should enhance time processing by handling undefined intervals, repetitive actions and contiguous events. Being the long time effort, author reveals his intentions in project's implementation and uncovers further plans.

[4] `http://nlp.fi.muni.cz/cs/VerbaLex`

[5] `http://nlp.fi.muni.cz/projekty/x-plain/rules.php`

Acknowledgments

This work has been partly supported by the Czech Science Foundation under the project P401/10/0792.

References

1. Connolly, C.: A new integrated robot vision system from fanuc robotics. The Industrial Robot **34**(2) (2007) 103–106
2. Shahinpoor, M., Kim, K., Mojarrad, M.: Artificial Muscles: Applications of Advanced Polymeric Nanocomposites. Taylor & Francis (2007)
3. Ferrucci, D.A.: Introduction to "This is Watson". IBM Journal of Research and Development **56**(3/4) (2012)
4. Tunstall-Pedoe, W.: True Knowledge: Open-domain question answering using structured knowledge and inference. AI Magazine **31**(3) (2010) 80–92
5. Gunning, D., et al.: Project Halo update—progress toward digital Aristotle. AI Magazine **31**(3) (2010) 33–58
6. Aron, J.: How innovative is apple's new voice assistant, Siri? New Scientist **212**(2836) (2011) 24 –
7. Pustejovsky, J., et al.: The Specification Language TimeML. In Mani, I., Pustejovsky, J., Gaizauskas, R., eds.: The Language of Time: A Reader. Oxford University Press (2004)
8. Kalyanpur, A., et al.: Structured data and inference in deepqa. IBM Journal of Research and Development **56**(3/4) (2012)
9. Tunstall-Pedoe, W.: Knowledge storage and retrieval system and method (2006)
10. Gunning, D.: Halobook and progress towards digital aristotle. In: Innovative Applications of Artificial Intelligence, San Francisco, CA (2011) Invited talk.
11. Clark, P., Porter, B.: Km - situations, simulations, and possible worlds. Technical report, AI Lab, Univ Texas at Austin (1999)
12. Lenat, D.B., et al.: Cyc: toward programs with common sense. Commun. ACM **33**(8) (1990) 30–49
13. Mueller, E.T.: Natural Language Processing with Thought Treasure. Signiform (1998)
14. Horák, A.: The Normal Translation Algorithm in Transparent Intensional Logic for Czech. PhD thesis, Masaryk University (2002)
15. Gardoň, A., Horák, A.: Time dimension in the dolphin nick knowledge base using transparent intensional logic. In Habernal, I., Matoušek, V., eds.: Proceedings of the 14th international conference on Text, speech and dialogue. Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence, Springer (2011)
16. Horák, A.: Computer Processing of Czech Syntax and Semantics. Librix.eu, Brno, Czech Republic (1998)
17. Němčík, V.: The Saara framework: An anaphora resolution system for czech. In: RASLAN 2009: Recent Advances in Slavonic Natural Language Processing, Karlova Studánka, Czech Republic, Masaryk University, Brno, Czech Republic (2009) 49–54
18. Gardoň, A., Horák, A.: The learning and question answering modes in the dolphin system for the transparent intensional logic. In: RASLAN 2007 : Recent Advances in Slavonic Natural Language Processing, Karlova Studánka, Czech Republic, Masaryk University, Brno (2007) 29–36

19. Gardoň, A.: Dotazování s časovými informacemi nad znalostmi v transparentní intenzionální logice (in slovak) (2010)
20. Gardoň, A., Horák, A.: Knowledge base for transparent intensional logic and its use in automated daily news retrieval and answering machine. In: 3rd International Conference on Machine Learning and Computing (ICMLC 2011), Singapore, IEEE (2011) 59–63