KNOWLEDGE ACQUISITION THROUGH NATURAL LANGUAGE CONVERSATION AND CROWDSOURCING

Luka Bradeško

Doctoral Dissertation
Jožef Stefan International Postgraduate School
Ljubljana, Slovenia

Supervisor: Doc. Dunja Mladenić, Jožef Stefan Institute, Ljubljana, Slovenia

Evaluation Board:
Dr. Michael Witbrock, Chair, IBM, New York, New York
Prof. Erjavec, Member, Jožef Stefan Institute, Ljubljana, Slovenia
Prof. Iztok Savnik, Member, Univerza v Novi Gorici, Nova Gorica, Slovenia

MEDNARODNA PODIPLOMSKA ŠOLA JOŽEFA STEFANA
JOŽEF STEFAN INTERNATIONAL POSTGRADUATE SCHOOL

Luka Bradeško

KNOWLEDGE ACQUISITION THROUGH NATURAL LANGUAGE CONVERSATION AND CROWDSOURCING

Doctoral Dissertation

PRIDOBIVANJE STRUKTURIRANEGA ZNANJA SKOZI POGOVOR TER S POMOČJO MNOŽIČENJA

Doktorska disertacija

Supervisor: Doc. Dunja Mladenić

Ljubljana, Slovenia, April 2017

To the world...

Acknowledgments

Thank everyone who contributed to the thesis: - EU Projects - Cyc - Dave - Michael - Vanessa - Dunja - Coworkers

Abstract

The English abstract should not take up more than one page.

Povzetek

Povzetek v slovenščini naj ne bo daljši od ene strani.

Contents

List of Figures

List of Tables

List of Algorithms

Abbreviations

CC ... Curious Cat (a name of the knowledge acquisition application and platform that is a side result of this thesis)
CYC ... An AI system (Inference Engine and Ontology), developed by Cycorp Inc.
CycKB.. Cyc Knowledge Base (Ontology part of Cyc system)
JSI ... Jožef Stefan Institute
KA ... Knowledge Acquisition
NL ... Natural Language

Symbols

$j^\star$ ... black-body irradiance
σ ... Stefan's (or Stefan-Boltzmann) constant

Glossary

Glossary of terms, dada, bada

Chapter 1

Introduction

An intelligent being or machine solving any kind of a problem needs knowledge to which it can apply its intelligence while coming up with an appropriate solution. This is especially true for the knowledge-driven AI systems which constitute a significant fraction of general AI research. For these applications, getting and formalizing the right amount of knowledge is crucial. This knowledge is acquired by some sort of Knowledge Acquisition (KA) process, which can be manual, automatic or semi-automatic. Knowledge acquisition, using an appropriate representation and subsequent knowledge maintenance are two of the fundamental and as-yet unsolved challenges of AI. Knowledge is still expensive to retrieve and to maintain. This is becoming increasingly obvious, with the rise of chat-bots and other conversational agents and AI assistants. The most developed of these (Siri, Cortana, Google Now, Alexa), are backed by huge financial support from their producing companies, and the lesser-known ones still result from 7 or more person-years of effort by individuals Finish :

Knowledge acquisition and subsequent knowledge maintenance, are two of the fundamental and as-yet not-completely-solved challenges of Artificial Intelligence (AI).

1.1 Scientific Contributions

This section gives an overview of scientific and other contributions of this thesis to the knowledge acquisition approaches.

1.1.1 Novel Approach Towards Knowledge Acquisition

Traditionally KA (knowledge acquisition) approach focuses on one type of acquisition process, which can be either Labor, Interaction, Mining or Reasoning(**Zang2013**). In this thesis we propose a novel, previously untried approach that intervenes all aforementioned types with current user context and crowdsourcing into a coherent, collaborative and autonomous KA system. It uses existing knowledge and user context, to automatically deduce and detect missing or unconfirmed knowledge(reasoning) and uses this info to generate crowdsourcing tasks for the right audience at the right time(labor). These tasks are presented to users in natural language (NL) as part of the contextual conversation (interaction) and the answers parsed (mining) and placed into the KB after consistency checks(reasoning). The approach contribution can be summed up as a) definition of the framework for autonomous and collaborative knowledge acquisition with the help of contextual knowledge (chapter X), and b) demonstrate and evaluate the contributions of contextual knowledge and approach in general chapter X.

1.1.2 Knowledge Acquisition Platform Implementation as Technical Contribution

Implementation of the KA framework as a working real-world prototype which shows the feasibility of the approach and a way to connect many independent and complex subsystems. Sensor data, natural language, inference engine, huge pre-existing knowledge base (Cyc)CycRef, textual patterns and crowdsourcing mechanisms are connected and interlinked into a coherent interactive application (Chapter X).

1.1.3 A Shift From NL Patterns to Logical Knowledge Representation in Conversational Agents

Besides the main contributions presented above, one aspect of the approach introduces a shift in the way how conversational agents are being developed. Normally the approach is to use textual patterns and corresponding textual responses, sometimes based on some variables, and thus encode the rules fro conversation. As a consequence of natural language interaction, the proposed KA framework is in some sense a conversational agent which is driven by the knowledge and inference rules and uses patterns only for conversion from NL to logic. This shows promise as an alternative approach to building non scripted conversational engines (Chapter X).

1.2 Thesis structure

The rest of the thesis is structured in to chapters covering specific topics. Chapter X introduces

Chapter 2

Background and Related Work

In this chapter we will give an overview of approaches and related works on broader knowledge acquisition research field, information extraction, crowdsourcing and geo-spatial context mining.

Knowledge Acquisition has been addressed from different perspectives by many researchers in Artificial Intelligence over decades, starting already in 1970 as a sub-discipline of AI research (Feigenbaum), and since then resulting in a big number of types and implementations of approaches and technologies/algorithms. In more recent survey of KA approaches (**Zang2013**), authors categorize all of the KA approaches into four main groups, regarding the source and way the knowledge is acquired:

- *Labour Acquisition.* This approach uses human minds as the knowledge source. This usually involves expert ontologists manually coding the knowledge.
- *Interaction Acquisition.* As in Labour Acquisition, the source of the knowledge is human minds, but in this case the KA is wrapped in a facilitated interaction with the system, and is sometimes implicit rather than explicit.
- *Reasoning Acquisition.* In this approach, new knowledge is automatically inferred from the existing knowledge using logical rules and machine inference.
- *Mining Acquisition.* In this approach, the knowledge is extracted from some large textual corpus or corpora.

We believe this categorization most accurately reflects the current state of machine (computer) based knowledge acquisition, and we decided to use the same classification when structuring our related work descriptions, while focusing more on closely related approaches and extending where necessary. According to this classification, our work which is presented in this thesis, fits into a hybrid approach combining all four groups, with main focus on interaction and reasoning. We address the problem by combining the labour and interaction acquisition (users answering questions as part of NL interaction aimed at some higher level goal, such as helping the user), adding unique features of using user context and existing knowledge in combination with reasoning to produce a practically unlimited number of potential interaction acquisition tasks, going into the field of crowd-sourcing because we can address many users simultaneously.

The existing work that can compare in some way to our solution can be divided into the systems that exploit existing knowledge (generated anew during acquisition or pre-existing in other sources) (**Singh2002a**; **Witbrock2003**; **Forbus2007**; **Kvo2010**; **Sharma2010**; **Mitchel2015**), reasoning (**Witbrock2003**; **Speer2007**; **Speer2008**; **Kuo2010**), crowdsourcing (**Singh2002**; **Speer2009**; **Kuo2010**; **Pedro2012a**; **Pedro2013**), acquisition

through interaction (**Speer2009**; **Pedro2012**; **Pedro2013**), acquisition through labour(add, probably rather refer to subsections) () and natural language conversation(**Pedro2012**; **Speer2007**; **Speer2009**; **Witbrock2003**; **Kuo2010**).

2.1 Labour Acquisition

This category consists of KA approaches which rely on explicit human work, to collect the knowledge. A number of expert (or also untrained) ontologists or knowledge engineers is employed to codify the knowledge by hand into the given knowledge representation (formal language).

2.2 Interaction Acquisition

adada

2.2.1 Games

adada

2.2.2 Natural Language Conversation

dada

2.3 Reasoning Acquisition

adad dada

2.4 Mining Acquisition

adad

2.5 Acquisition with the help of existing knowledge

adad

2.6 Crowdourcing Acquisition

adad

2.7 Acquistion of Geospatial Context

adad

Chapter 3

Knowledge Acquisition Approach

Chapter 4

Real World Knowledge Acquisition Implementation

4.1 Cyc

TBW

Chapter 5

Evaluation

TBW

Chapter 6

Conclusions

We came to the following conclusions ...

Chapter 7

Introduction

7.1 Thesis Structure

The thesis should be structured as follows (note that while some parts are optional, their order is not):

- *Front Matter*
  - Title pages
  - Dedication (optional)
  - Acknowledgments
  - Abstract
  - Povzetek
  - Contents
  - List of Figures (required if the thesis contains figures)
  - List of Tables (required if the thesis contains tables)
  - List of Algorithms (required if the thesis contains algorithms)
  - Abbreviations (optional)
  - Symbols (optional)
  - Glossary (optional)
- *Main Matter*[1]
  - 1 Introduction
  - $2, \ldots, n-1$ Chapters
  - n Conclusions

  The Introduction must clearly summarize the thesis from the topic application of the doctoral dissertation. The core text of the doctoral dissertation can be substituted by publications (or papers accepted for publication) in internationally recognized journals. In this case, the Introduction should clearly describe the scientific method and the candidate's contribution to any publication which has been produced by several authors. In the Discussion or Conclusions, the candidate should summarize coherently the results of his/her dissertation.

[1] The PhD thesis can also be divided into Parts — this is not encouraged, but possible and supported by this template.

- *Appendices* (optional)
  - Appendix A Title
  - Appendix B Title
  - ...

  If desired, the candidate can place some of his/her relevant papers in the appendices. In order to do so, he/she must obtain permissions from publishers.

- *Back Matter*
  - References
  - Bibliography
  - Biography
  - Index (optional)

In addition, each author must prepare the cover page and spine using the two `.doc` files provided by IPS.

7.2 Sectioning Example

This is a section. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1 Subsection

This is a subsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.1 Subsubsection

This is a subsubsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.1.1 Paragraph This is a paragraph. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.1.1.1 Subparagraph This is a subparagraph. It is not possible to get deeper than this. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.1.1.2 Subparagraph This is a subparagraph. It is not possible to get deeper than this. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.1.2 Paragraph This is a paragraph. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.2 Subsubsection

This is a subsubsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.2.1 Paragraph This is a paragraph. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.1.2.2 Paragraph This is a paragraph. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.2 Subsection

This is a subsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.2.1 Subsubsection

This is a subsubsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

7.2.2.2 Subsubsection

This is a subsubsection. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text. Some more text.

Chapter 8

Floating Bodies

Floating bodies are figures, tables and algorithms.

8.1 Figures

Captions should be placed below figures as shown in Figure 8.1. If a caption is shorter than the line width, it should be centered.

Figure 8.1: A large IPS logo.

On the other hand, if a caption is very long (see Figure 8.2), only its first (short) part should be put in the List of Figures.

Figure 8.2: A small IPS logo. The IPS has its own logo and a uniform graphic image, which is used on all its documents.

8.2 Tables

Similar rules apply also to captions of tables, with the exception that captions are placed above tables (see Table 8.1).

Table 8.1: A simple table.

A	B	C
12	9834	327
51	2234	97

8.3 Algorithms

Algorithm 8.1 presents an algorithm example.

Algorithm 8.1: An algorithm example.

Data: this text
Result: complete understanding

```
initialization;
while not at end of this document do
    read current section;
    if understood then
        go to next section;
        current section becomes this one;
    else
        go back to the beginning of current section;
    end
end
```

Chapter 9

Equations and Measurement Units

9.1 Equations

Small equations are often written in-line (within the text), for example $j^\star = \sigma T^4$, while larger ones need to be displayed in the following way:

$$\sigma = \frac{2\pi^5 k^4}{15c^2h^3} = 5.6704 \times 10^{-8} Js^{-1}m^{-2}K^{-4} \tag{9.1}$$

All displayed equations need to be numbered so that they can be referenced later in the text (for example, Eq. (9.1) presents the Stefan's (or Stefan-Boltzmann) constant).

9.2 Measurement Units

The candidate can choose a standard for measurement units and has to consistently use it throughout the thesis.

Chapter 10

Definitions and Theorems

10.1 Definitions

See the formal definition of the right triangle in Definition 10.1.

Definition 10.1 (Right triangle). A *right triangle* is a triangle in which one angle is a 90-degree angle.

10.2 Theorems

The Pythagorean theorem is a relation in Euclidean geometry among the three sides of a right triangle. It states that the square of the hypotenuse (the side opposite the right angle) is equal to the sum of the squares of the other two sides (**pythagoras**).

Theorem 10.1 (Pythagorean theorem). *In every right triangle with sides a and b and hypotenuse c, the following holds:*

$$a^2 + b^2 = c^2 \tag{10.1}$$

See Appendix A for the proof of this theorem.

Chapter 11

Reference Formatting

References should be formatted using either the IEEE or APA 6th edition formatting style. This template uses the latter one.

11.1 IEEE

Please see the template using the IEEE style.

11.2 APA 6th Edition

References are cited using the `\parencite` command. For example, see (**mihailovic06**). The alternative `\textcite` is used when the authors are mentioned in text. For example, let us cite the work by **depolli13** References to journal articles that have not yet been published should contain the `doi`, as in (**tusar14**).

Multiple references can be cited at the same time (**kobal04**; **grace10**; **novak12eng**; **zupanc13**). Beside books and journal articles, parts of books (**smodis09**), technical reports (**ivekovic13eng**), PhD theses (**dovgan14eng**) and MSc theses (**tusar07eng**) can be included in the references.

Finally, on-line sources can be referenced, too, see Section 10.2.

Appendix A

Proofs of Theorems

A.1 Proof of the Pythagorean Theorem

Let us prove the Pythagorean Theorem from page 21.

Proof. This proof is based on the proportionality of the sides of two similar triangles, that is, upon the fact that the ratio of any two corresponding sides of similar triangles is the same regardless of the size of the triangles.

Let ABC represent a right triangle, with the right angle located at C, as shown in Figure A.1. We draw the altitude from point C, and call H its intersection with the hypotenuse AB. Point H divides the length of the hypotenuse into two parts.

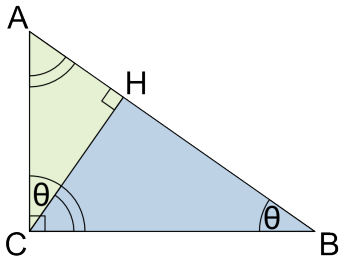

Figure A.1: Similar triangles used in the proof of the Pythagorean theorem.

The new triangle ACH is similar to triangle ABC, because they both have a right angle (by definition of the altitude), and they share the angle at A, meaning that the third angle will be the same in both triangles as well, marked as θ in Figure A.1. By a similar reasoning, the triangle CBH is also similar to ABC.

Similarity of the triangles leads to the equality of ratios of corresponding sides:

$$\frac{BC}{AB} = \frac{BH}{BC} \text{ and } \frac{AC}{AB} = \frac{AH}{AC}. \tag{A.1}$$

The first result equates $\cos\theta$ and the second result equates $\sin\theta$.

These ratios can be written as:

$$BC^2 = AB \times BH \text{ and } AC^2 = AB \times AH. \tag{A.2}$$

Summing these two equalities, we obtain:

$$BC^2 + AC^2 = AB \times BH + AB \times AH = AB \times (AH + BH) = AB^2, \tag{A.3}$$

which, tidying up, is the Pythagorean theorem:

$$BC^2 + AC^2 = AB^2. \tag{A.4}$$

□

Bibliography

Publications Related to the Thesis

All publications related to the thesis should be referenced in the text.

Other Publications (optional)

...

Biography

The author of this thesis ...