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 **xvii**

List of Tables **xix**

List of Algorithms **xxi**

Abbreviations **xxiii**

Symbols **xxv**

Glossary **xxvii**

1 Introduction **1**
1.1 Scientific Contributions . . . 1
1.1.1 Novel Approach Towards Knowledge Acquisition . . . 1
1.1.2 Knowledge Acquisition Platform Implementation as Technical Contribution . . . 2
1.1.3 A Shift From NL Patterns to Logical Knowledge Representation in Conversational Agents . . . 2
1.2 Thesis structure . . . 2

2 Background and Problem Definition **3**
2.1 Knowledge Representation, Engineering and Inference . . . 3
2.2 Context (Information) Extraction . . . 4
2.3 Knowledge Acquisition . . . 4
2.4 Natural Language Processing . . . 4
2.5 Crowdsourcing . . . 4

3 Related Work **5**
3.1 Labour Acquisition . . . 8
3.1.1 Cyc . . . 8
3.1.2 ThoughtTreasure . . . 8
3.1.3 HowNet . . . 8
3.1.4 Open Mind Common Sense (OMCS) . . . 9
3.1.5 GAC/MindPixel . . . 9
3.1.6 Semantic Knowledge Source Integration (SKSI) . . . 9
3.2 Interaction Acquisition . . . 10
3.2.1 Interactive User Interfaces . . . 10
3.2.1.1 KRAKEN . . . 10
3.2.1.2 User Interaction Agenda (UIA) . . . 11
3.2.1.3 Factivore . . . 11
3.2.1.4 Predicate Populator . . . 11

3.2.1.5 Freebase . . . 11
3.2.1.6 OMCommons (Open Mind Commons) . . . 12
3.2.2 Games . . . 12
3.2.2.1 20Q (20 Questions) . . . 12
3.2.2.2 Verbosity . . . 12
3.2.2.3 Rapport . . . 13
3.2.2.4 Virtual Pet . . . 13
3.2.2.5 Goal Oriented Knowledge Collection (GOKC) . . . 13
3.2.2.6 Collabio (Collbaorative Biography) . . . 13
3.2.3 Interactive Natural Language Conversation . . . 14
3.2.4 AIML(Artificial Intelligence Mark-up Language . . . 14
3.2.5 ChatScript . . . 15
3.2.6 CyN . . . 16
3.3 Mining Acquisition . . . 16
3.3.1 Populating Cyc from the Web (PCW) . . . 17
3.3.2 Learning Reader . . . 17
3.3.3 Never Ending Language Learner (NELL) . . . 17
3.3.4 KnowItAll . . . 18
3.3.5 Probase . . . 18
3.3.6 TextRunner . . . 19
3.3.7 ReVerb . . . 19
3.3.8 R2A2 . . . 19
3.3.9 ConceptMiner . . . 20
3.3.10 DBPedia . . . 20
3.3.11 YAGO (Yet Another Great Ontology) . . . 20
3.3.12 KNEXT . . . 21
3.4 Reasoning Acquisition . . . 22
3.4.1 Cyc Predicate Populator + FOIL . . . 22
3.4.2 Plausible Inference Patterns (PIP) . . . 22
3.4.3 AnalogySpace . . . 23
3.4.4 Cyc Wiki . . . 23
3.5 Acquistion of Geospatial Context . . . 24
3.5.1 Extracting Places from Traces of Locations . . . 24
3.5.2 Discovery of Personal Semantic Places based on Trajectory Data Mining . . . 24
3.5.3 Applying Commonsense Reasoning to Place Identification . . . 25

4 Knowledge Acquisition Approach 27
4.1 Architecture . . . 27
4.1.1 Interaction Loop . . . 29
4.1.1.1 Machine to Human Interaction (MHI) . . . 30
4.1.1.2 Human to Machine Interaction (HMI) . . . 31
4.1.1.3 Machine Mediated Human to Human Interaction (MMHHI) 31
4.2 Knowledge Base . . . 31

5 Real World Knowledge Acquisition Implementation 33
5.1 Cyc . . . 33

6 Evaluation 35

7 Conclusions 37

References **39**

Bibliography **45**

Biography **47**

List of Figures

Figure 4.1: General Architecture of the KA system, with an interaction loop presented as arrows. 28
Figure 4.2: Possible interaction types between the user and Curious Cat KA System. 30

List of Tables

Table 3.1: Structured overview of related KA systems 7
Table 3.2: AIML Example 14
Table 3.3: AIML example of saving info to variables 15
Table 3.4: AIML Example of remembering answers on specific questions 15
Table 3.5: Simple ChatScript example 16
Table 3.6: Simple ka (remembering) example 16
Table 3.7: Simple ChatScript question/answer/remember example 16

List of Algorithms

Algorithm 3.1: Staypoint Detection Algorithm 1 (SPD1) 25
Algorithm 3.2: Staypoint Detection Algorithm 2 (SPD2) 26

Abbreviations

AST	...	Abstract Syntax Tree
CC	...	Curious Cat (a name of the knowledge acquisition application and platform that is a side result of this thesis)
CSK	...	Common Sense Knowledge
CYC	...	An AI system (Inference Engine and Ontology), developed by Cycorp Inc.
CycKB	...	Cyc Knowledge Base (Ontology part of Cyc system)
CycL	...	Cyc Lanugage
FOIL	...	First Order Inductive Learner
GAC	...	Generic Artificial Consciousness
GOKC	...	Goal-Oriented Knowledge Collection
GWAP	...	Games With A Purpose
HMI	...	Human to Machine Interaction
JSI	...	Jožef Stefan Institute
KA	...	Knowledge Acquisition
KB	...	Knowledge Base
KDML	...	Knowledge SDatabase Mark-up Language
LSA	...	Latent Semantic Analysis
MHI	...	Machine to Human Interaction
MIT	...	Massachusetts Institute of Technology
MMHHI	...	Machine Mediated Human to Human Interaction
MPI	...	Max Planck Institute
MSR	...	Microsoft Research
NL	...	Natural Language
NLP	...	Natural Language Processing
NP	...	Noun Phrase
NTU	...	National Taiwan University
OIE	...	Open Information Extraction
PMI	...	Pointwise Mutual Information
POI	...	Point of Interest
POS	...	Part of Speech
POS:X	...	Abbreviatios for Part of Speech tags used by POS parsers
POS:CC	...	Coordinating conjunction
POS:CD	...	Cardinal Number
POS:DT	...	Determiner
POS:EX	...	Existential there
POS:FW	...	Foreign Word
POS:IN	...	Preposition or subordinating conjunction
POS:JJ	...	Adjective
POS:JJR	...	Adjective, comparative
POS:JJS	...	Adjective, superlative
POS:LS	...	List item marker

POS:MD ... Modal
POS:NN ... Noun, singular or mass
POS:NNS ... Noun, plural
POS:NNP ... Proper noun, singular
POS:NNPS ... Proper noun, plural
POS:PDT ... Predeterminer
POS:POS ... Possesive ending
POS:PRP ... Personal pronoun
POS:PRP$ ... Possesive pronoun
POS:RB ... Adverb
POS:RBR ... Adverb, comparative
POS:RBS ... Adverb, superlative
POS:RP ... Particle
POS:SYM ... Symbol
POS:TO ... to
POS:UH ... Interjection
POS:VB ... Verb, base form
POS:VBD ... Verb, past tense
POS:VBG ... Verb, gerund or present participle
POS:VBN ... Verb, past participle
POS:VBP ... Verb, non-3rd person singular present
POS:VBZ ... Verb, 3rd person singular present
POS:WDT ... Wh-determniner
POS:WP ... Wh-pronoun
POS:WP$ ... Possesive wh-pronoun
POS:WRB ... Wh-adverb
PTT ... Taiwanese Bulletin Board System
SKSI ... Semantic Knowledge Source Integration
SPD ... Staypoint Detection
TUW ... The University of Waikato, New Zealand
UL ... University of Leipzig
UoM ... University of Manheim
UoR ... University of Rochester
UW ... University of Washington

Symbols

$\wedge$ ... logical conjunction (and). The statement $A \wedge B$ is true if both A and B are true, otherwise it is false.

Glossary

AST (iAbstract Syntax Tree) is an abstract representation of Wikipedia page as parsed from DBPedia parser. Something like DOM tree for Wikipedia instead for pure HTML

OIE (Open Information Extraction is a paradigm introduced by Oren Etizoni in his TextRunner system. The main idea of this paradigm is that the knowledge acquisition system is not pre-determined to extract some specific facts, patterns, etc, but is open-ended, extracting large set of relational tuples without any human input.

PMI (Pointwise Mutual Information is a measure which captures co-occurence relationsip between terms in a big corpus.

SKSI (Semantic Knowledge Source Integration) is a *Cyc* sub-system for external knowledge integration.

Upper Ontology (also top-level, foundation or core ontology) is the part of ontology (or knowledge base), which defines the core objects that serve as a main knowledge building blocks to construct the full knowledge base.

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).

We propose and implement novel approach to automated knowledge acquisition using the user context obtained from a mobile device and knowledge based conversational crowdsourcing. The resulting system named Curious Cat has a multi objective goal, where KA is the primary goal, while having an intelligent assistant and a conversational agent as secondary goals. The aim is to perform KA effortlessly and accurately while having a conversation about concepts which have some connection to the user, allowing the system (or the user) to follow the links in the conversation to other connected topics. We also allow to lead the conversation off topic and to other domains for a while and possibly gather additional, unexpected knowledge. For illustration see the example conversation sketch in Table I, where topic changes from a specific restaurant to a type of dish. In this example case, the conversation is started by the system when user stays at the same location for 5 minutes.

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(Zang et al. 2013). 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)(Lenat 1995), 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 for 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 Problem Definition

This chapter describes the challenges and components that Knowledge Acquisition system such is presented in this work (*Curious Cat*) have to address and bring together into a working workflow in order to be a coherent KA system able to satisfy the goals of general (common) Knowledge Acquisition.

Curious Cat is a KA system making use of existing knowledge, logical inference, crowd-sourcing and mobile context, to trigger natural language questions at user-appropriate moments and then incorporate the answers consistently into the existing KB. To successfully do this, there are many inter connected steps addressing a broad range of as yet not completely solved problems from multiple fields of artificial intelligence, machine learning, natural language processing and human computer interaction. Additionally to that, given that the approach uses crowd-sourcing, there is an additional complexity of technical implementation and scalability.

For more structured explanation of the approach and challenges involved, this section is grouped into sub-sections describing main challenges. Each section then references our approach to it (implementation) and also related works that gives overview of similar approaches.

Curiois Cat is knowledge driven, meaning that knowledge is connecting all of the components, including the user interaction, and storing of the results into the KB (section 2.1). Its user context is obtained through a mobile sensor mining in a real world application that monitors the user's activity and location through mobile GPS and accelerometer sensors. This raw data is then corrected, clustered, classified and enriched before inserted into the KB as knowledge (section 2.2. The newly asserted context can trigger forward chaining operation of the inference engine (section 2.1) which can results in logical representation of a new question (section 2.3) or a statement that the system intends to show to the user. The aforementioned logical formula is then converted to natural language (sections 2.4) and presented to the user through a mobile app. When the user answers, his NL answer is converted back to logic (section 2.4), checked by the inference engine agains the existing knowledge for consistency, and inserted as new peace of knowledge into the KB (section 2.3). After iteraction like that, the system determines whether to continue the conversational path with the user or not. Newly acquired knowledge is then used to check with other users for validity (section 2.5) and is then used by the inference to produce new questions/comments/suggestions (section 2.3).

2.1 Knowledge Representation, Engineering and Inference

dada

2.2 Context (Information) Extraction

dada

2.3 Knowledge Acquisition

dada

2.4 Natural Language Processing

dada

2.5 Crowdsourcing

dada

Chapter 3

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, and since then resulting in a big number of types and implementations of approaches and technologies/algorithms. The difficulty of acquiring and maintaining the knowledge was soon noticed and was Goined as *Knowledge Acquisition Bottleneck* in 1977(Feigenbaum 1977). In more recent survey of KA approaches (Zang et al. 2013), authors categorize all of the KA approaches into four main groups, regarding the source of the data and the way knowledge is acquired:

- *Labour Acquisition.* This approach uses human minds as the knowledge source. This usually involves human (expert) ontologists manually entering and encoding the knowledge.
- *Interaction Acquisition.* As in Labour Acquisition, the source of the knowledge is coming from humans, 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 automaticaGly 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, focusing more on closely related approaches and extending where necessary. According to this classification, our work 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 with various tasks), 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 by sending these generated tasks to many users simultaneously.

Previous works that can compare with our solution is divided into the systems that exploit existing knowledge (generated anew during acquisition or pre-existing from before

Fix this, refe
to chapters i
to specifici w

in other sources) (**Witbrock2003**; **Kvo2010**; **Mitchel2015**; Singh et al. 2002; Forbus et al. 2007; Sharma et al. 2010), reasoning (**Witbrock2003**; Speer 2007; Speer, Lieberman, et al. 2008; Y.-L. Kuo et al. 2010), crowdsourcing (**Singh2002**; Speer, Krishnamurthy, et al. 2009; Y.-L. Kuo et al. 2010; Saulo D. S. Pedro et al. 2012; S D S Pedro et al. 2013), acquisition through interaction (**Pedro2012**; Speer, Krishnamurthy, et al. 2009; S D S Pedro et al. 2013), acquisition through labour(add, probably rather refer to subsections) () and natural language conversation(**Pedro2012**; **Witbrock2003**; Speer 2007; Speer, Krishnamurthy, et al. 2009; Y.-L. Kuo et al. 2010).

Test referencing table (see Table 3.1).

Table 3.1: Structured overview of related KA systems

System	Parent	Reference	Category	Source	Representation	Prior K.	Crowds.	Context
Cyc project (Cycorp)	/	(Lenat 1995)	Labour	K. Exp.	CycL	/	/	/
ThoughtTrasure(Signiform)	/	(Mueller 2003)	Labour	K. Exp.	LAGS	/	/	/
HowNet (Keen.)	/	(Dong et al. 2010)	Labour	K. Exp.	KDML	/	/	/
OMCS/ConceptNet (MIT)	/	(Singh et al. 2002)	Labour	Public	ConceptNet	/	✓	/
GAC/Mindpixel(McKinstry)	/	(McKinstry et al. 2008)	Labour	Public	MindPixel	/	✓	/
SKSI (Cycorp)	Cyc	(Masters et al. 2007)	Labour/Integration	K. Exp.	Structured	✓	/	/
KRAKEN (Cycorp)	Cyc	(Panton et al. 2002)	Interaction	D. Exp	CycL	✓	/	/
UIA (Cycorp)	Cyc	(Witbrock, Baxter, et al. 2003)	Interaction	D. Exp	CycL	✓	/	/
Factivore (Cycorp)	Cyc	(Witbrock, Matuszek, et al. 2005)	Interaction	D. Exp	CycL	✓	/	/
Predicate Populator (Cycorp)	Cyc	(Witbrock, Matuszek, et al. 2005)	Interaction	D. Exp	CycL	✓	/	/
CURE (Cycorp)	Cyc	(Witbrock 2010)	Interaction	D. Exp	CycL	✓	/	/
OMCommons (MIT)	OMCS	(Speer 2007)	Interaction	Public	ConceptNet	✓	✓	/
Freebase (Metaweb/Google)	/	(Bollacker et al. 2008)	Interaction	Public	RDF	/	/	/
20 Questions (MIT)	OMCS	(Speer, Krishnamurthy, et al. 2009)	Game	Public	ConceptNet	/	/	/
Verbosity (CMU)		(L. V. Ahn et al. 2006)	Game	Public	/	/	✓	/
Rapport (NTU)	ConceptNet	(Y.-l. Kuo et al. 2009)	Game	Public	ConceptNet	/	✓	/
Virtual Pet (NTU)	ConceptNet	(Y.-l. Kuo et al. 2009)	Game	Public	ConceptNet	/	✓	/
GOKC (NTU)	ConceptNet	(Y.-L. Kuo et al. 2010)	Game	Public	ConceptNet	✓	✓	/
Collabio (MS)	/	(Bernstein et al. 2010)	Game	Public	/	/	✓	/
AIML (Alice foundation)	/	(R. S. Wallace 2003)	Chatbot	/	AIML	/	/	/
Chatscript (Brilligunderstanding)	/	(Wilcox 2011)	Chatbot	/	ChatScript	/	/	/
CyN (Daxtron Labs)	Cyc+AIML	(Wilcox 2011)	Chatbot	/	AIML+Cyc	✓	/	/
PCW (Cycorp)	Cyc	(Matuszek, Witbrock, et al. 2004)	Mining	Web Search	AIML+Cyc	✓	/	/
Learning Reader (NU)	Cyc	(Forbus et al. 2007)	Mining	Web	CycL	✓	/	/
NELL (CMU)	/	(Mitchell et al. 2015)	Mining	Web	Predicate l.	✓	✓	/
KnowItAll(UW)	/	(Etzioni, Popescu, et al. 2004)	Mining	Web Search	text	/	/	/
Probase (MSR)	/	(Wu et al. 2012)	Mining	Web	Propriatery	/	/	/
TextRunner (UW)	KnowItAll	(Soderland et al. 2007)	Mining	Web	text	/	/	/
ReVerb (UW)	TextRunner	(Fader et al. 2011)	Mining	Web	text	/	/	/
R2A2 (UW)	ReVerb	(Etzioni, Fader, et al. 2011)	Mining	Web	text	/	/	/
ConceptMiner (MIT)	ConceptNet	(Eslick 2006)	Mining	Web Search	ConceptNet	✓	/	/
DBPedia (UL&UoM)	Wikipedia	(Lehmann et al. 2015)	Mining	Wikipedia	RDF	/	✓	/
YAGO (MPI)	Wikipedia	(Suchanek et al. 2008)	Mining	Wikipedia	RDF	✓	/	/
Cyc+Wiki (TUW)	Cyc/Wikipedia	(Medelyan et al. 2008)	Mining	Wikipedia	CycL	/	✓	/
KNEXT (MPI)		(Schubert 2002)	Mining	Penn Treebank	/	/	/	/
P. Populator+FOIL (Cyc)	Predicate Populator	(Witbrock, Matuszek, et al. 2005)	Reasoning	Induction	CycL	✓	/	/
PIP (NU)	Cyc	(Sharma et al. 2010)	Reasoning	Induction	CycL	✓	/	/
AnalogySpace (MIT)	OMCS	(Speer, Lieberman, et al. 2008)	Reasoning	Analogy	ConceptNet	✓	/	/

3.1 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). Labour acquisition is the most expensive acquisition type, but it gives a high quality knowledge. It is often a crucial initial step in other KA types as well, since it can help to have some pre-existing knowledge to be able to check the consistency of the newly acquired knowledge. Labour Acquisition is often present in other KA types, even if not explicitly mentioned, since it is implicitly done when defining internal workings and structures of other KA processes. While we checked other well known systems that are result of Labour Acquisition, Cyc (mentioned below) is the most comprehensive of them and was picked as a starting point and main background knowledge and implementation base for this work.

3.1.1 Cyc

The most famous and also most comprehensive and expensive knowledge acquired this way, is Cyc KB, which is part of Cyc AI system (Lenat 1995). It started in 1984 as a research project, with a premise that in order to be able to think like humans do, the computer needs to have knowledge about the world and the language like humans do, and there is no other way than to teach them, one concept at a time, by hand. Since 1994, the project continued through Cycorp Inc. company, which is still continuing the effort. Through the years Cyc Inc. employed computer scientists, knowledge engineers, philosophers, ontologists, linguists and domain experts, to codify the knowledge in the formal higher order logic language CycL (Matuszek, Cabral, et al. 2006a). As of 2006 (Matuszek, Cabral, et al. 2006b), the effort of making Cyc was 900 non-crowdsourced human years which resulted in 7 million assertions connecting 500,000 terms and 17,000 predicates/relations (Zang et al. 2013), structured into consistent sub-theories (Microtheories) and connected to the Cyc Inference engine and Natural Language generation. Since the implemtentation of our approach is based on Cyc, we give a more detailed description of the KB and its connected systems in section 5.1 on page 33. Cyc Project is still work in progress and continues to live and expand through various research and commercial projects.

3.1.2 ThoughtTreasure

Approximately at the same time(1994) as Cyc Inc. company was formed, Eric Mueller started to work on a similar system, which was inspired by Cyc and is similar in having a combination of common sense knowledge concepts connected to their natural language presentations. The main differentiator from Cyc is, that it tries to use simpler representation compared to first-order logic as is used in Cyc. Additionally, some parts of *ThoughtTreasure* knowledge can be presented also with finite automata, grids and scripts(Mueller 1999; Mueller 2003). In 2003 the knowledge of this system consisted of 25,000 concepts and 50,000 assertions. ThoughtTreasure was not so successfull as Cyc and ceased all developments in 2000 and was open-sourced on Github in 2015. link as footnote.

3.1.3 HowNet

started in 1999 and is an on-line common-sense knowledge base unveiling inter-conceptual relationships and inter-attribute relationships of concepts as connoting in lexicons of the Chinese and their English equivalents. As of 2010 it had 115,278 concepts annotated with Chinese representation, 121,262 concepts with English representation, and 662,877

knowledge base records including other concepts and attributes (Dong et al. 2010). HowNet knowledge is stored in the form of concept relationships and attribute relationships and is formally structured in KDML (Knowledge Database Mark-up Language), consisting of concepts (called semens in KDML) and their semantic roles.

3.1.4 Open Mind Common Sense (OMCS)

is a crowdsourcing knowledge acquisition project that started in 1999 at the MIT Media Lab(Singh et al. 2002). Together with initial seed and example knowledge, the system was put online with a knowledge entry interface, so the entry was crowd-sourced and anyone interested could enter and codify the knowledge. OMCS supported collecting knowledge in multiple languages. It's main difference from the systems described above (Cyc, HowNet, ThoughtTreasure) is, that it used deliberate crowdsourcing and that it's knowledge base and representation is not strictly formal logic, but rather inter-connected pieces of natural language statements. As of 2013 (Zang et al. 2013), OMCS produced second biggest KB after Cyc, consisting of English (1,040,067 statements), Chinese (356,277), Portuguese (233,514), Korean (14,955), Japanese (14,546), Dutch (5,066), etc. Initial collection was done by specifying 25 human activities, where each activity got it's own user interface for free form natural language entry and also pre-defined patterns like "A hammer is for _____", where participants can enter the knowledge. Although OMCS started to build KB from scratch it shares a similarity to our CC system in a sense that it is using crowd-sourcing and also natural language patterns with empty slots to fill in missing parts. OMCS was later used in many other KA approaches as a prior knowledge, similar way as we use Cyc. After a few versions, OMCS was taken from public access and merged with multiple KBs and KA approaches into an ConceptNet KB[1] (Speer, Chin, et al. 2016), which is now (in 2017) part of Linked Open Data (LOD) and maintained as open-source project.

3.1.5 GAC/MindPixel

Generic Artificial Consciousness was a bold try to make a general AI based on the premise that human thinking can be simulated with answering binary yes/no questions or decisions(**McKinstry20**
where humans have a natural bias toward yes answers. With this in mind, McKinstry founded GAC (later renamed to MindPixel), where internet crowd would answer yes/no questions for shares in the company which would commercially exploit the GAC AI. Initially, one such answer which contributed to GAC knowledge base was called MindPixel, but after a while this became name for the whole system. With this idea, McKinstry's MindPixel was one of the first crowd-sourced efforts for collaborative KB building. While MindPixel shut down after McKinsey stopped working on it, it inspired *OMCS, later OpenMind* and started the crowdsourcing and crowd based cross validation of the facts. Besides the *OMCS*, a similar YES/NO crowdsourcing system lives under the name *Weegy* [2].

3.1.6 Semantic Knowledge Source Integration (SKSI)

No matter how big the underlying knowledge base is, there will always be some missing knowledge, that exists somewhere else. While with knowledge acquisition it is possible to add these missing peaces, sometimes it makes more sense to keep this knowledge externally and only link it properly so it can be used. This is especially true for fast changing data such as stock values, weather forecast information, real-time measurements of traffic/ production line, etc. In such cases it is beneficial if one can link and address this data in the same way

[1] http://conceptnet.io/
[2] http://www.weegy.com

as it would have been included in the original KB. This is the goal of *SKS*(Masters et al. 2007) In order to connect the external source, a "wapper" knowledge needs to be defined in *Cyc* KB, which describes which concepts and instances external source represents, and how to access them (http request, SQL Query, etc.).

In *SKS* system, this "wrapper" knowledge is asserted as normal *CycL* assertions using special predicates and concepts. The descriptions are structured into three layers, where first (access layer) describes how to access the data (ie. how to connect, send queries and retrieve the content). Then the second (physical schema layer), describes how the data is structured inside the original source. The third layer (logical schema) describes in *CycL*, how the data connects to *CycKB*. Example of the logical layer is semantics on how specific columns translate into the KB. For example, table *Securities*, column *Name*, translates into instance of Cyc concept *#$Equity-Security* with the *#$nameString* linked to the value of the table cell. With this approach, it is possible to seamlessly link the existing KB with external sources, and use *Cyc* inference engine and querying mechanisms on externally connected data without even noticing it's external.

3.2 Interaction Acquisition

Similarly as with Labour KA, interaction Acquisition gets the knowledge from human minds, but in this case the acquisition is an intended side effect, while users are interacting with the software as part of some other activity/task, or as part of a motivation scheme, such as knowledge acquisition games. Besides games, the interaction could be some other user interface for solving specific tasks, or a Natural Language Conversation. This type of acquisition is most strongly correlated with the approach described in this thesis, since Curious Cat uses points (gaming), to motivate users and it interacts with user in NL, while discussing various topics (concepts). It uses the conversation to set up the context and acquire (remember) user's responses and places them properly in to the KB. Sometimes the acquired knowledge is paraphrased and presented back to user to show the 'understanding', which was first tried in OSMC (section 3.1, (Singh 2002)), but there only in non-conversational way as part of the input forms.

3.2.1 Interactive User Interfaces

Interactive user interfaces are the most common representation of interaction acquisition, where the user interface is constructed in a way to help user enter the data and thus make the acquisition much faster and cheaper. Historically, these systems were developed to help the labour acquisition systems, or on top of them, after parent systems reached some sort of maturity and initial knowledge stability. This is the reason why all of these systems rely or are build on top of labour acquisition (section 3.1) or mining acquisition (section 3.3) systems.

3.2.1.1 KRAKEN

system was a knowledge entry tool which allows domain experts to make meaningful additions to CYC knowledge base, without the training in the areas of artificial intelligence, ontology development, or knowledge representation(Panton et al. 2002). It was developed as part of DARPA's Rapid Knowledge Formation (RKF) project in 2000. As its goal was to allow knowledge entry to non-trained experts, it started to use natural language entry and is as this, a first pre-cursor to Curious Cat system and a seed idea for it. It consists of creators, selectors, modifiers of Cyc KB building blocks, tools for consistency checks and tools for using existing knowledge to infer new things to ask. This tool, together with it's

derived solutions was later re-written and integrated into Cyc as CURE system (see below). While KRAKEN and later CURE already used Natural Language generation and parsing, and started with the idea of natural language dialogue for doing the KA, the interaction, it was missing user context (user's had to select or search the concept of interest), and also crowdsourcing aspects. Kraken was also missing rules for explicit question asking. The questions were all related to the selected concept and given as a list of natural language forms.

3.2.1.2 User Interaction Agenda (UIA)

was a web based user interface for KRAKEN KA tool(Panton et al. 2002; Witbrock, Baxter, et al. 2003). It worked inside a browser and it worked as responsive web-app (in 2001) by automatically triggering refresh functionality of the browser. It consisted of a menu of tools that is organized according to the recommended steps of the KE process, text entry box (query, answer, statement), center screen for the main interaction with the current tool, and a summary with a set of colored steps needed to complete current interaction. Similarly as KRAKEN itself, this interface was later improved and integrated into main Cyc system as part of CURE tool.

3.2.1.3 Factivore

This application was a Java Applet user interface for an extended KRAKEN system, meant for quick facts entering (Witbrock, Matuszek, et al. 2005). On the back-end it used the same mechanisms and logical templates, while in the front-end it only allowed facts entering, as opposed to UIA, which also allowed rules (which ended up as not being useful).

3.2.1.4 Predicate Populator

Predicete Populator is a similar tool as *Factivore*, which instead of only collecting instances, allows to add general knowledge about classes. For example, instead of describing facts for a specific restaurant, it can collect general knowledge that is true for all restaurants (Witbrock, Matuszek, et al. 2005). The context of the KA in this case, is given by class concept, a predicate and a web-site which is parsed into CycL concepts. These are then filtered out if they do not match argument constraints of the predicate and then shown to user for selection. As part of the validation, this tool had some problems with correctly acquired knowledge. One of the proposed solutions (never implemented), was to start using volunteers to vote about the correctness. This is already a pre-cursor idea for crowd-sourced voting mechanisms that we used in Curious Cat.

3.2.1.5 Freebase

Freebase started in 2007(Bollacker et al. 2008) and was a large (mostly instance based) crowd-sourced graph database for structured general human knowledge. Initially it was acquired from multiple public sources, mostly Wikipedia. The initial seed was then constantly updated and corrected by the community. On the user interface side, Freebase provides an AJAX/Web based UI for humans and an HTTP/JSON based API for software access. For finding knowledge and also software based editing, it uses Metaweb Query Language (MQL). A company behind freebase was bought by Google in 2010 and incorporated into a Google Knowledge Graph. In 2016 Freebase was incorporated into the Wikidata platform and shut down by Google and is no longer maintained.

3.2.1.6 OMCommons (Open Mind Commons)

This system is an interactive interface to OMCS which can respond with a feedback to user answers and maintain dialogue (Speer 2007). This is similar approach as we do with Curious Cat and shows understanding of the knowledge users enter. The mechanisms behind is by using inference engine to make analogical inferences based on the existing knowledge and new entry. Then it generates some relevant questions and asks user to confirm them. For example, as given from the original paper, *OMCommons* asks: "A bicycle would be found on the street. Is this common sense?". This is then displayed to the user with the justification for the question: "A bicycle is similar to a car. I have been told that a car would be found on the street". Users then click on "Yes/No" buttons to confirm or reject the inferred statement. The interactive interface also allows its users to refine the knowledge entered by other users and see the ratings. Users can also explore what new inferences are result of their new contributions.

3.2.2 Games

Games are a specific sub-section of interaction acquisition, where the actual acquisition is hidden or transformed into much more enjoyable process, maximizing the entertainment of the users. This type of KA was first officially introduced by Luis von Ahn in 2006 (L. von Ahn 2006; Luis von Ahn et al. 2008) under the name 'Games with Purpose' paradigm.

3.2.2.1 20Q (20 Questions)

This is a game with intentional knowledge acquisition task which focuses to the most salient properties of concepts. The game itself is a standard 20 questions game which aims to make one player figure out the concept of discussion by asking yes/no questions and then infer from the answers what the concept could be. The only difference is that the player which is asking is a computer based on OMCS knowledge base. It generates questions in NL, and according to what a player answers, it attempts to guess the concept. To decide what questions to ask, it uses statistical classification methods (Speer, Krishnamurthy, et al. 2009), to discover the most informative attributes of concepts in OMCS KB. After the user answers all the questions, including whether the detected concept was right or not, the concept and the answers will be assigned to proper cluster and thus the characteristics of the object are learned.

3.2.2.2 Verbosity

. Similarly as Q20 above, Verbosity is a spoken game for two persons randomly selected online. It was inspired by Taboo board game(Hasbro n.d.) which required players to state common sense facts without mentioning the secret concept. While having similar game-play as aforementioned board game, Verbositywas developed with the intent to collect common sense knowledge (L. V. Ahn et al. 2006). One player (narrator), gets a secret word concept and needs to give hints about the word to the other player (guesser), who must figure out the word that is described the hints. The hints take the form of sentence templates with blanks to be filled in. For example, if the word is "CAR", the narrator could say "it has wheels." In the experiments, a total of 267 people played the game and collected 7,871 facts. While these facts were mostly a good quality and it was proven that the game can be used successfully, these facts were natural language snippets and were not incorporated into any kind of structure or formal KB.

3.2.2.3 Rapport

This is a KA game based on Chinese OMCS questions, but implemented as a Facebook game to make use of the social connections inside social network. The Game helps users to make new friends or enhance connections with their existing social network by asking and answering questions and matching the answers to other users(Y.-l. Kuo et al. 2009). This game aims to enhance the experience and community engagement and thus functionality of aforementioned *Verbosity* game, by employing simultaneous interaction between all the players versus only 1 to 1 interaction between 2 community members. For evaluation, the answers where multiple users answered the same were considered valid. This game had a similarity with Curious Cat in a sense that it employed the voting mechanism for the same answers, and the repetitive questioning of the same question to multiple users. Authors found out that the agreement between same answers of the repetitive question and voting is 80% or more after at least 2 repetitions of the same question. In 6 months, *Rapport* collected 14,001 unique statements from 1,700 users. Normalized, this is 8.2 answers per user.

3.2.2.4 Virtual Pet

This is a similar game as *Rapport* in a sense that it uses *OMCS* patterns, is in Chinese and is developed by the same authors (Y.-l. Kuo et al. 2009). Instead of Facebook platform, *Virtual Pet* uses PTT (Taiwanese bulleting board system in Chinese language). Instead of direct interaction between the users themselves, users interact with virtual pet and can ask it questions and answer it's questions. In the back-end, the questions the pet asks, are actually questions from other users. This game in 6 months collected 511,734 unique pieces of knowledge from 6,899 users. Normalized this is 74,1 answers per user. While this game attracted much more answers than *Rapport*, the quality of the answers was slightly lower. Authors argue that the reasons behind both is, that users didn't interact directly, but through the virtual pet, so they were less careful whether answers are correct or not.

3.2.2.5 Goal Oriented Knowledge Collection (GOKC)

. This game builds on the findings and approach of *Virtual Pet* KA game. The main improvement is to try and actually make use of the new knowledge inside a given domain (picked by the initial seed questions), to infer new questions. With this the authors tried to fix a drawback of *Virtual Pet*, that through time, the questions and answers become saturated, and the number of new questions and answers falls exponentially through time, with respect to the number of already collected knowledge peaces. This approach is also aligned with the CC approach, which uses existing+ context and new knowledge, to drive the questions. First part of the *GOKC* paper describes analysis of the knowledge collected by *Virutal Pet* game. The second part is a description and evaluation of GOKC KA approach, where authors did 1 week experiment to show that the approach works. During that week the system inferred created 755 new questions, out of which, 12 were reported as bad. Out of these questions 10,572 answers were collected where 9,734 were voted as good. This results in the 92,07% precision. Compared to the game without question expansion (*Virtual Pet*), which has precision of 80.58%, this is an improvement.

3.2.2.6 Collabio (Collbaorative Biography)

. This is also a Facebook based game, with the intention to collect user's tags. While the gathered knowledge is more a set of person's tags than knowledge, it served as an inspiration to *Rapport* and *Virtual Pet*. During the experiment, *Collabio* users tagged 3,800 persons

with accurate tags with information that cannot be found otherwise(Bernstein et al. 2009; Bernstein et al. 2010).

3.2.3 Interactive Natural Language Conversation

Natural Language Knowledge Acquisition methods are special case of Interaction Acquisition systems. While almost all of the approaches already described above (under Interactive User Interfaces and Games subsections) use natural language to some extent, the language processing used is based on relatively small amount of textual patterns, or statements which are not necessary connected into a conversation. Common denominator of these systems is that they intentionally try to acquire knowledge and then use natural language statements to do this. As a side effect and as motivation for users, sometimes consequent questions and answers give a feeling of conversation. On the other side chat-bots, start with the intention to maintain an interesting conversation with the users, and have to do knowledge acquisition only to remember facts and parts of the past conversations to be able to be smart enough, so users do not lose interest. Starting with Eliza(Weizenbaum 1966), these systems evolved, mostly directed by Turning Tests(**Turing?**), implemented as Loebner competitions, trying to pass it[3]. Through the measure of these tests(Bradeško et al. 2012), among a few propriatery chat-bots, two technologies evolved (*AIML, ChatScript*) to be general enough and can be used for conversational engine (chat-bot) construction and also NL knowledge acquisition.

3.2.4 AIML(Artificial Intelligence Mark-up Language

is an XML based scripting language. It allows developers of chat-bots, to construct a pre- defined natural language patterns and their responses. These definitions are then fed into an AIML engine, which can match user inputs with the patterns and figure out what response to write. AIMLs syntax consists mostly of input rules (categories) with appropriate output. The pattern must cover the entire input and is case insensitive. It is possible to use a wildcard (*) which binds to one or more words. The simplest example of AIML pattern with appropriate response is presented in Table 3.2. This pattern detects user's questions like "Do you have something on the menu?" and responds with "We have everything on the menu."

Table 3.2: AIML Example

```
<Category>
   <pattern> Do you have * on the menu </pattern>
   <template>
      We have everything on the menu.
   </template>
</Category>
```

AIML allows recursive calls to its own patterns, which allows for some really complicated and powerful patterns, covering many examples of input. Regarding the knowledge acquisition, AIML has an option to store parts of the textual patterns as variables and thus store information for later.

The AIML example on Table 3.3 can remember keywords following "I just ate" pattern, like "I just ate pizza". If user at some point later says "I am hungry", the bot is able to

[3] http://www.loebner.net

Table 3.3: AIML example of saving info to variables

```
<category>
    <pattern>I just ate *</pattern>
    <template>
        Nice choice! <set name = "food"> <star/></set>
    </template>
</category>
<category>
    <pattern>I am hungry</pattern>
    <template>
        Eat another <get name = "food"/>?
    </template>
</category>
```

respond with "Eat another pizza". In a combination with "<that>" tag, which matches previous computer's response, AIML can be used to construct specific knowledge acquisition questions (Table 3.4). The given example is using AIML 1.0, which was later improved with AIML2.0(R. Wallace 2013) which introduced the *<Learn>* tag, but mechanism stayed mostly the same.

Table 3.4: AIML Example of remembering answers on specific questions

```
<category>
    <pattern>*</pattern>
    <that>What did you order</that>
    <template>
        Was it good? <set name = "menuItem"><star/></set>
    </template>
</category>
```

While AIML language with appropriate engine can remember specific facts, the mechanism is purely keyword based and cannot really count as structured knowledge. Additionally, since it only remembers direct facts, it would be really hard to construct an acquisition of all types of food for example. AIML based chatbots were winning Loebner's competitions in the years from 2000 to 2004, but were later outcompeted by Chatscript based bots and propriatery solutions.

3.2.5 ChatScript

is an NLP expert system consisting of textual patterns rules. It was designed by Bruce Wilcox (Wilcox 2011) and besides patterns it has mechanisms for defining concepts, triple store for facts, own inference engine POS tagger and parser. From the measure of how close the system is to pass the Turing Test as measured by Loebner's competitions, *ChatScript* surpassed *AIML*, and is its successor, since both systems are open sourced. It was designed purposely to be simpler to use and have more powerful tools for NLP and knowledge acquisition which is integral part chatbot systems. A simple example from AIML (Table 3.2) can be re-written in much shorter form as ChatScript rule (Table 3.5).

Similarly the example from Table 3.3 can be writen as:

Table 3.5: Simple ChatScript example

```
?: (do you have * on the menu) We have everything on the menu.
```

Table 3.6: Simple ka (remembering) example

```
s: (I just ate _*) Nice choice!
s: (I am hungry) East another _0?
```

Similarly, example from Table 3.4 in ChatScript looks like:

Table 3.7: Simple ChatScript question/answer/remember example

```
t: What did you order?
        a: (_*) \$menuItem=_0 Was it good?
```

While the above examples repeats the functionality of AIML, ChatScript is more powerful and can remember facts in the shape of (subject verb object) and act on them. This is done with using *ĉreatefact* and *f̂indfact* functions.

3.2.6 CyN

CyN is an AIML interpreter implementation with additional functionality to be able to access Cyc inference engine and KB for both, storing the knowledge and also for querying(Coursey 2004). This was done by introduction of new AIML tags:

- <*cycterm*> Translates an English word/phrase into a Cyc symbol.
- <*cycsystem*> Executes a CycL statement and returns the result.
- <*cycrandom*> Executes a CycL query and returns one response at random.
- <*cycassert*> Asserts a CycL statement.
- <*cycretract*> Retracts a CycL statement.
- <*cyccondition*> Controls the flow execution in a category template.
- <*guard*> Processes a template only if the CycL expression is true

3.3 Mining Acquisition

This category of KA systems try to make use of big text corpus-es available online or otherwise on some digital media. Because the core idea of writing is to share information, there is a lot of knowledge in the texts that can be extracted and converted into a structured knowledge that can later be used by computers. Due to wast size and availability of the data on the internet, mining is most often done on the web resources. Since most of the data format from these corpuses is text, these techniques are particularly strong in the using various NLP techniques, which are often combined with existing knowledge to correct mistakes and check consistency.

3.3.1 Populating Cyc from the Web (PCW)

Since whole idea of Cyc system is to gain enough knowledge through manual work, to be able to learn on itself after some point, the Cyc team is looking int into other means of knowledge acquisition which can automatize or speed-up the KA process. One of the approaches is by mining facts from the Internet by issuing appropriate search engine queries(Matuszek, Witbrock, et al. 2004).

Because Cyc KB is really big, the first step of this approach is to select appropriate part of the kb (concepts and related queries) which are in the interest of the system. For initial experiments a set of 134 binary predicates was selected. These predicates were then used to scan the KB and find the missing knowledge, which was converted to CycL queries. These queries were then converted to NL and queried on a web search engine. The results are then converted back to CycL through NL to logic engine of Cyc. After the conversion these are converted back to NL and re-searched on the web, to check whether the results still hold, and then as the last step, the results are checked for consistency (whether they can be asserted into Cyc). From the initial 134 predicates, the system generated 348 queries, 4290 searches. It found 1016 facts, out of which 4 were rejectec due to inconsistency, 566 rejected by search engine (not same results), 384 were already known to Cyc, and finally 61 new consistent facts were detected as valid. After human review, the findings were that only 32 facts were actually correct.

3.3.2 Learning Reader

Learning Reader(Forbus et al. 2007) is a prototype knowledge mining system that combines NLP, large KB (CycKB) and analogical inference into an automated knowledge extraction system that works on simplified language texts. The system uses Direct Memory Access Parsing (DMAP(Martin et al. 1986)) to parse text and convert it into CycL concepts which are then checked by the inference engine whether they can form correct CycL statements. The prototype consisted of 30,000 NDAP patterns (a quick approximation would be to imagine CyN patterns- chapter 3.2.6). After parsing and syntax checks, found CycL sentences were checked by the inference within CycKB, whether knowledge is new or not. Only new logical statements were then asserted. Additional feature of this prototype system is that it includes question answering mechanism, which can be used by evaluators to check what the system learned. This same mechanism is also used to try to generate new facts (elaborate) and also questions based on newly acquired knowledge. The experiment on 62 written stories improved the recall from 10% to 37% and kept the accuracy as 99.7% compared to original 100%. By using additional inference and conjecture based inference, the recall raised to 60% while accuracy dropped to 90.8%.

3.3.3 Never Ending Language Learner (NELL)

NELL(Mitchell et al. 2015) is a text mining KA system running 24/7 with the goal to extract knowledge, use this knowledge to improve itself and extract more knowledge. NELL was started in January 2010 and as of 2015 acuired 80 million confidence weighted new beliefs. NELL consists of many different learning tasks (for different types of knowledge), where each task also consist of the performance metrics, so the system can asses itself and check if the learning task itself is also improving through time. In 2015 it consisted of 2500 learning tasks. Some of learning task examples:

- Category Classification
- Relation Classification

- Entity Resolution
- Inference Rules among belief triples

After learning tasks, there is a *Coupling Constraings* component, which combines results of learning tasks. The potentially useful knowledge gets asserted into the KB as candidates, where the assertions are checked by knowledge integrator module which integrates the assertions into the KB, or rejects them.

After some initail initial KB had been gathered, the CMU text-mining knowledge NELL also started to apply a crowdsourcing approach(Saulo D. S. Pedro et al. 2012), using natural language questions to validate its KB. In a similar fashion as Curious Cat, NELL can use newly acquired knowledge, to formulate new representations and learning tasks. There is, however, a distinct difference between the approaches of NELL and Curious Cat. NELL uses information extraction to populate its KB from the web, then sends the acquired knowledge to Yahoo Answers, or some other Q/A site, where the knowledge can be confirmed or rejected. By contrast, Curious Cat formulates its questions directly to users (and these questions can have many forms, not just facts to validate), and only then sends the new knowledge to other users for validation. Additionally, Curious Cat is able to use context to target specific users who have a very high chance of being able to answer a question.

3.3.4 KnowItAll

KnowItAll(Etzioni, Popescu, et al. 2004) is a domain independent web fact extraction system that s specific search engine queries to find new instances of specific classes. It starts it extraction with a small seed of class names and NL patterns like "NP1 such as NP2". The classes and patterns are then used to find instances, new classes and also new extraction phrases by analyzing the results of the web search engines. For example, Googling: "Cities such as *", will return a lot of statements with instances of cities. After these cities are extracted, the names can be used for furhter Googling and by analyzing the phrases in which these cities appear, new patterns can be found, and so on. As part of the experiment, KnowItAll ran for 5 days and extracted over 50,000 instances of cities, states, countries, actors and films. To asses the correctnes of the extractions, the system can fill-in the instances into the various patterns and check the hit-count returned by the search engine./ This then compares by the hit-count of the instance itself and uses this to asses the probability of the instance really belonging to the detected class. For example: comparing the hit-count of "Ljubljana", "Cities such as Ljubljana" and "Planets such us Ljubljana", the system can figure out that Ljubljana is most likely indeed an instance of the class city.

KnowItAll was the first of the systems that inspired an *Open Information Extraction (Open IE)* paradigm(Etzioni, Fader, et al. 2011) which resulted in many other IE systems such us TextRunner, ReVerb and R2A2. The main ide of this paradigm is to avoid hand labeled examples and domain specific verbs and nouns when approaching textual patterns which can lead to open (without specifying the targets) knowledge extraction on a web scale.

3.3.5 Probase

Probase is a probabilistic taxonomy of concepts and instances consisting of 2.7 million of concepts extracted from 1.68 billion of web pages (Wu et al. 2012). The main difference between Probase and other KBs is that Probase is probabilistic as opposed of "black and white" KB. On the other hand, even if it has much more concepts, it is sparse in the

knowledge, since it only uses isA relation (taxonomy). Probase was compared to other taxonomies such as WordNet, YAGO and Freebase in the sense of recall and precition of isA relations. Probase was found to be most comprehensive (biggest recall), while losing at precision measure against YAGO (92.8% vs 95%).

Probase was later renamed as *Microsoft Concept Graph*, and has accessible API[4] which in 2017 consists of 5,401,933 concepts, 12,551,613 instances and 87,603,947 isA relations.

3.3.6 TextRunner

TextRunner is a successor of KnowItAll system (Soderland et al. 2007) and is the first to introduce Open Information Extraction (OIE) paradigm, which's main idea is that it is open-ended and can extract information autonomously without any human intervention which would fix the system to some specific domain or set of concepts/relations. *TextRunner* was ran through over 9 million of web pages, and compared to *KnowItAll* reduced the error rate for 33% on comparable set of extractions. Throughout the experiments, *TextRunner* collected 11mio of high probability tuples and 1 mio concrete facts. *TextRunner* consists of three components.

Self-Supervised Learner is a component started first which takes a small corpus of documents as an input and then outputs a classifier that can detect candidate extractions and classify them as trustworthy or not.

Single-Pass Extractor is a component that makes a single pass through the full corpus and extracts tuples for all possible relations. These tuples are then sent to the classifier trained before, which then marks them as trustworthy or not. Only trustworthy tuples are retained.

Redundancy-Based Assesor is the last step which assigns a probability to each retained tuple, based on the probabilistic model or redundancy (Downey et al. 2005).

3.3.7 ReVerb

With the experiments done with *TextRunner* and *WAE*, it became obvious that OIE systems have a lot of noise and inconistencies in the results. For this reason two syntactical and lexical constraints were introduced in *ReVerb* OIE system(Fader et al. 2011). This helps with removing the incoherent extractions such as "recalled began" which was extracted from sentence "They recalled that Nungesser began his career as precint leader", or uninformative extractions like "Faust, made a deal" extracted from "Faust made a deal with the devil".(Fader et al. 2011). When started, *ReVerb* first identifies relation phrases that match the constraints, then if finds approriate pair of appropriate noun phrase rguments for each identified phrase. The resulting extractions are then given a confidence score using logistic regression classifier.

3.3.8 R2A2

R2A2 is another improvement in OIE paradigm, since previous systems assumed that relation arguments are only simple noun phrases. Analysis of *ReVerb* errors showed that 65% of errors is on the arguments side (the relation was ok). To fix this, *R2A2* system goes somehow into the direction of kb based KA sytems like Curious Cat with argument constraings.(Etzioni, Fader, et al. 2011). The difference is that *R2A2* is not using hard logic and inference, but rather statistical classifier to detects class constraints (bounds) of the arguments. Compared to *ReVerb*, *R2A2* has much higher precision and recall.

[4] https://concept.research.microsoft.com

3.3.9 ConceptMiner

ConceptMiner is a KA system built by Ian Scott Eslick as part of his master thesis(Eslick 2006), with the main hipothesis that the seed knowledge collected from vounteers can be then used to bootstrap automatic knowledge acquisition. *ConceptMiner* specifically focuses on binary semantic relationships such as cause, effect, intent and time. The system relies on the prior volunteer knowledge from *ConceptNet* and tests its hipothesis with experimental extractions of knowledge aroun three semantic relations: desire, effect and capability.

As a first step, the systemn uses knowledge around predicates *DesireOf*, *EffectOf* and *CapableOf* from *ConceptNet*, to construct web-search queries. The results of these are then used to derive general patterns for aforementioned relations. For example an existing knowledge *(DesireOf "dog" "attention")*, when converted to search engine query: "dog * bark", results in patterns like:

- "My/PRP$ dog/NN loves/VBZ attention/NN ./."
- "Horseback/NN riding /VBG dog /NN attracts/VBZ attention/NN."

While not all of the patterns are of the same quality, with the sheer number of repetitions, it is possible to extract more probable ones. This then results in general patterns such as *<X>/NN loves/VBZ <Y>/NN*. These can be then used to issue a lot of search queries with various combinations of wors, to extract instances of 'who desires what'. These potential instances then go into the last step (filtering). As part of this step, *ConceptMiner* removes badly formed statements, concepts not included in *ConceptNet* KB, and concepts with low PMI score (see abbreviations and glossary).

3.3.10 DBPedia

DBPedia is crowd-sourced RDF KB, extracted from Wikipedia pages and made publicly available(Lehmann et al. 2015). As of 2017 the English DBPedia contains 4.58 million knowledge pieces, out of which 4.22 million in a consistent ontology, including 1,445,000 persons, 735,000 places, 411,000 creative works (123,000 music albums, 87,000 films and 19,000 video games), 241,000 organizations (58,000 companies, 49,000 educational institutions), 251,000 species and 6,000 diseases[5]. *DBPedia* is also localized into 125 languages, so all-together it consists of 38.3 million knowledge pieces. It is also linked to YAGO categories.

Acquisition mechanism is automatic and consists of the following steps:

- Wikipedia pages are downloaded from dumps or through API and parsed into an Abstract Syntax Tree (AST)
- AST is forwarded to various extractor modules. For example, extractor module can find labels, coordinates, etc. Each of the extractor modules can convert it's part of AST into RDF triples.
- The collection of RDF statements as returned from the extractors is written into an RDF sink, supporting various format such as NTriples, etc.

3.3.11 YAGO (Yet Another Great Ontology)

YAGO is an ontology built automatically from *WordNet* and *Wikipedia*(Suchanek et al. 2008). Latest version *YAGO3* is built from multiple languages and as of 2015 consist of

[5] http://wiki.dbpedia.org/about

4,595,906 entities, 8,936,324 facts, 15,611,709 taxonimy facts and 1,398,837 labels (Demner-Fushman et al. 2015). The facts were extracted from Wikipedia category system and info boxes, using a combination of rule-based and heuristic methods, and then enriched with hierarchy (taxonomic) relations taken from WordNet. Since building YAGO is automatized, each next run of the script can use existing knowledge for type and consistency checking. This kind of type checking helps YAGO to maintain its precision at 95%(Suchanek et al. 2008).

3.3.12 KNEXT

With the premise that there is a lot of general knowledge available in texts, which lays beneath explicit assertional content, Schubert build *KNEXT* KA system(Schubert 2002) which extracts *general "possibilistic" propositions* from text. The main difference towards other KA mining systems is that before combining meanings from a phrase, the meanings are abstracted (generalized) and simplified. For example, abstraction of "a long, dak corridor" yields "a corridor". Or "a small office at the end of a long dark corridor" yields "an office". This kind of abstraction, together with weakening of relations into a possibilistic form, starts to represent presumptios about the world. The extraction follows five steps:

- Pre-process and POS tag the input
- Apply a set of ordered patterns to the POS tree recursively
- For each successfully matched subtree, abstract the interpretations using semantic rule patterns
- Collect the phrases expected to hold general "possibilistic" propositions
- Formulate the propositiions and output these together with simple English representations.

From an example input statement "Blanche knew something must be causing Stanley's new, strange behavior but she never once connected it with Kitti Walker.", the output looks like this:

```
A female-individual may know a preposition
(:Q DET FEMALE-INDIVIDUAL) KNOW[V] (:Q DET PROPOS)
something may cause a behavior
(:F K SOMETHING[N]) CAUSE[V] (:Q THE BEHAVIOR[N])
a male-individual may have a behavior
(:Q DET MALE-INDIVIDUAL) HAVE[V] (:Q DET BEHAVIOR[N])
a behavior can be new
(:Q DET BEHAVIOR[N]) NEW[A])
a behavior can be strange
(:Q DET BEHAVIOR[N]) STRANGE[A])
a female-individual may connect a thing-reffered-to with a female
   -individual
(:Q DET FEMALE-INDIVIDUAL) CONNECT[V]
(:Q DET THING-REFERRED-TO)
(:P WITH[P] (:Q DET FEMALE-INDIVIDUAL)))
```

The authors position their system as an addition to systems like *Cyc*, and conducted their KA experiments on Treebank corpora resulting in around 60% of propositions marked as "reasonable general claims"(Schubert and Tong 2003).

3.4 Reasoning Acquisition

Compared to other types of KA, *Reasoning Acquisition* in its essence, doesn't need any external data-sources, but it uses existing knowledge and machine inference to automatically infer additional facts from the existing knowledge. While deductive reasoning can come with new facts from the premises and rules, the reasoning that has a chance to produce higher value (non obvious) findings is inductive and analogical resasoning. Analogical reasoning can find new facts of some concept based on properties of similar concepts. On the other side, infuctive reasoning can find find new probable rules based on current observations of the KB.

3.4.1 Cyc Predicate Populator + FOIL

With the initial experiments conducted from Labour Rule Acquisition done as part of *Factivore* and *Predicate Populator*(Witbrock, Matuszek, et al. 2005), it was found that getting inference rules from crowdsourcing and untrained human labour is ineffective and slow. For this reason, inductive logic inference mechanism based on FOIL (First Order Inductive Learner) approach (J.R. et al. 1997) was added to the systen. Experiments were conducted on a set of 10 predicates from the KB, which generated 300 new rules. Of these rules, 7.5% were found to be correct and 35% correct with minor editing to make them well formed (assertible to Cyc). This way, rule acquisition was speed up for quite a lot, since previous experiments showed that human experts produce rules with the rate around three per hour, while with FOIL, they can review and double check for correctnes around twenty rules per hour.

3.4.2 Plausible Inference Patterns (PIP)

The main idea of *PIP* system is to learn new plausible inference rules (patterns of plausible reasoning) by combining existing knowledge and reinforcement learning and thus improve the system's question answering abilities(Sharma et al. 2010). It is based on *Cyc* knowledge base, especially its predicate hierarchy represented by *#$genlPred* predicate, and *PredicateType* which represents second order collection of predicates that can be grouped together by some common features. For example *(#$genlPreds #$holds #$touches)*, means that each time something is holding something else, it is also touching it. The system scans the KB for rules containing predicates and tries to generate *PIPs* out of them, meaning that it tries to replace the predicates in the rules with the appropriate *PredicateType* which is linked to prior predicate. If there are more than 5 ways to generate the same PIP (from various predicate instances), then this new rule is accepted as "valid". Example of such PIP is:

$$(\#\$familyRelationsSlot(?x, ?y) \wedge \#\$familyRelationSlot(?y, ?z)) \Longrightarrow \#\$personalAssociationPredicate(?y, ?z)$$

where *#$familyRelationsSlot* and *#$personalAssociationPredicate* are instances of *#$PredicateType* and thus collections of other predicates, and *?x,?y,?z* are variables representing other concepts (**sharma2010**). This PIP or inference rule, tells the special inference algorithm that two predicates of type *#$familyRelationsSlot* cand imply *#$personalAssociationPredicate*, whent a proper combination of antecedent bindings on *?x,?y, ?z* is can be proved. Part of the *PIP* system is also an inference algorithm (FPEQ), which can make use of the above rule and come up with the possible solutions. As the first step,

it constructs a graph connected to the concepts used in the Q/A query. Then, the algorithm searches for the assertions in the graph matching a set of existing PIP inference rules. As the last steps, it returns the maches (i.e. filled-in PIPs with specific predicates and concepts) which can answer the query. As an additional step, authors introduuced reinforcement learning, where a small amount of user feedback (+1 or -1 voting) for final answers, can improve the PIP selection and also the level of generalizations when generating PIPs. Overall, the experiments showed that the approach increased Q/A recall for quite a lot (120% improvement), while minimally reduced the precision (94% of the baseline)(Zang et al. 2013).

3.4.3 AnalogySpace

This system is meant to work over big, but unexact knowledge bases (such as OMCS), where the need it to be able to make rough conclusions based on similarities, as opposed on hard and absolute logical facts. *AnalogySpace* does just that, by performing analogical closure by dimensionality reduction of semantic network (KB graph)(Speer, Lieberman, et al. 2008). This system tried to represent a new synthesis between a standard symbolic reasoning and statistical methods. For this, a similar technique to Latent Semantic Analysis (LSA) is being used, where strong assertions are used as opposed to weak semantics of word co-occurences in the document. In the LSA matrix, on one axis are concepts from *OMCS*, and on the other axis a features of these concepts, which yelds a sparse matrix of very high dimension. Then Singular Value Decomposition (SVD) is used on the matrix to reduce the dimensionality. This results in *principal components*, which represent the most important aspects of the knowledge. Then semantic similarity can be determined using linear operations over the resulting vectors. In a sense, this dimensionality reduction is acting as a generalization process for the kb. This way it is easier to calculate similarities between resulting concept vectors and can thus make generalizations based on these similarities, even if original concept didn't have some of the exact assertions that would enable inference engine to use it in the inference process and thus come with good answer. Results of experminents have shown that more than 70% of the resulting assertions were marked as true by human validators.

3.4.4 Cyc Wiki

Given that Wikipedia contains a lot of knowledge which is not structured in a way to be useful for inference engines directly, it makes sense to either try to structure it (as was done with *YAGO*), or link it with some existing ontology such as *CycKB*, which was done as part of this research task (Medelyan et al. 2008). Authors first filtered out of pootential pool of Cyc concepts all non-common sense concepts, which ended with 83,897 of them. Then these concepts were string matched with Wikipedia concept names based on their name directly, or based on *#$nameString* predicate. For example, Cyc concept *#$Virgo-Constellation* would be mathced to Wikipeadia page *Virgo (Constellation)*. After this step, the mappings that have 1 to 1 relationship with Wikipedia pages and do not point to Wiki disambiguation pages, are considered properly aligned. But the mappings that have more than 1 Wikipedia result then go into the disambiguation phase. With this approach, authors were able to get the precision of 96.2% and recall 64.0% when no disambiguation was needed, and precision 93% and recall 86.3% with the disambiguation part.

3.5 Acquistion of Geospatial Context

This sub section covers the works and approaches that can relate to our context mining/acquisition implementation. While the approaches themselves are more from the data-mining domain, as opposed to knowledge acquisition, in *Curious Cat*, the results are converted in the knowledge and asserted directly as a contextual knowledge about the users. For this reason, this related work is in its own subsection. In the Table 3.1, the approaches here are not listed, since they don't count as a knowledge acquisition, but more a custom data-mining solutions which support our proposed KA approach.

3.5.1 Extracting Places from Traces of Locations

This algorithm (we refer to it as SPD1 - Staypoint Detection), is one of the first papers/approaches that started to mine the raw GPS data into more user friendly notion of location - place.(Kang et al. 2005) The main idea of the algorithm is to be able to cluster the raw GPS locations into the particular places (home, work, specific restaurant, etc.) that user visited. This is achieved with a combination of radius(D_t) and time-based (T_t) clustering tresholds, where a cluster is a set of GPS coordinates that are within radius ($dist(loc_1) - dist(loc_n) < D_t$) during a time which is longer than a threshold ($time(loc_n) - time(loc_1) > T_t$). With this approach, it is possible to cluster locations based on how long one stays within a region, wihtout knowing the number of clusters in front (as is necessary with more standard clustering approaches). The core of the algorithm has a benefit to be really simple and can be easily ran on the phone.

The algorithm pseudocode is shown below (algorithm 3.1), where CL represents currently detecting cluster (staypoint), VP is detected cluster (staypoint or Visit Point) and *plocs* is temporal locations to be discarded or added to cluster later. Regarding the thresholds, T_t is Time threshold, T_d is distance treshold.

While the above alogorithm robustly detects stay-points, it completely ignores paths between. Additional weak-point is that it doesn't hadndle well when more than one GPS coordinate wrongly jumps due to GPS accuracy. This was fixed by the improved algorithm we developed as part of *Curious Cat* system, and simultanously, the GPS accuracy part as well by the algorithm described below (subsection 3.5.2).

3.5.2 Discovery of Personal Semantic Places based on Trajectory Data Mining

This work(Lv et al. 2016)(SPD2), builds on top of the first *SPD* algoritm described in subsection 3.5.1. It improves the stay-point detection and incorporates it into the broader system which is able to map particular visits into the exact places (points of interest - POIs) and then additionally able to predict next locations of tracked users. With this functionality (especially mapping to exact points of interest), this approach shares a lot of similarities with the context mining approach that is employed within the *Curious Cat* system. The main improvement of this algorithm, is that it takes into the account the discontinuous characteristics of phone GPS signal (it gets lost inside buildings, its accuracy drops under the trees). This is done by introducing a second T_{dt} threshold, which is must be higher than T_d, and sets a tolerated distance between two consequent stay-points to be merged on the fly. Additionally, these thresholds are not fixed, but dynamically calculated based on the distribution of distances between raw GPS points.

The simplified algorithm (removed unnecessary if/else statements) is shown below (algorithm 3.2), where CL represents currently detecting cluster (staypoint), PC is previous cluster and VP is detected cluster (staypoint or Visit Point). Regarding the thresholds, T_t

Algorithm 3.1: Staypoint Detection Algorithm 1 (SPD1)

```
Data: raw GPS coordinates
Result: cluster representing one staypoint
if distance(CL, loc) < D_t then
    add loc to CL;
    clear plocs;
else
    if plocs.length > 1 then
        if duration(cl) > T_t then
            add CL to VP;
        end
        clear CL;
        add plocs.end to CL;
        clear plocs;
        if distance(CL,loc) < D_t then
            add loc to CL;
            clear plocs;
        else
            add loc to plocs;
        end
    else
        add loc to plocs;
    end
end
```

is Time threshold, T_d is distance treshold and T_{dt} is tolerated distance threshold for later merging.

Similarly as *SPD1*, also *SPD2* has a prblem of ignoring the GPS coordinates that are part of paths (or moves) between stay-points.

3.5.3 Applying Commonsense Reasoning to Place Identification

Clustering raw sensor data into locations, paths, activities and travel modes is often not enough. Besides knowing the location, time of arrival and duration of stay, contextual knowledge benefits from the information about the type of place and name, or even the exact ID from some database, where additional information can be looked up. This is part of the *CuriousCat* system, which was inspired by and is based on the related work of Marco Mamei(Mamei 2010). In this work, the author shows how *Cyc* knowledge base can be used to improve automatic place identification. The approach is using probabilistic ranking the list of candidates (retrieved from *Foursquare* in consideration of the common sense likelihood, which is based on the user profile, features of the day (exact time, noon, afternoon, morning, midday,...), previous visits and type of probable place. The approach was validated and given that the GPS accuracy of N95 phones (used in the experiment) is worse than that of phones in 2017, it achieved the accuracy of 75% for business type of places.

Algorithm 3.2: Staypoint Detection Algorithm 2 (SPD2)

Data: raw GPS coordinates
Result: cluster representing one staypoint

if *distance(CL, loc)* $< T_d$ **then**
  add loc to CL;
else if *duration(CL)*$>T_t$ **then**
  append CL to VP;
  CL = 0, PC = 0;
else if *interval(CL, PC)* $> T_t$ **and** *distance(CL,PC)*$< T_dt$ **then**
  CL = combine (CL, PC);
  append CL to VP;
  PC = 0;
else
  PC=CL;
  CL=0;
end

Chapter 4

Knowledge Acquisition Approach

This chapter defines the terms, formal structure and steps that form our proposed KA approach. First it introduces the general architecture and interaction loop that defines the sequence of interactions and steps involved in the process(section 4.1). In the second part, it formalizes the upper ontology and logical constructs required for the KA approach (ref to chapter). After that, each of the crucial steps is described in more detail through examples and additions to the core logical structure defined earlier.

4.1 Architecture

The proposed KA system consists of multiple interconnected technologies and functionalities which we grouped into logical modules according to the problems they are solving (as also defined in chapter 2). This was done in order to minimize the complexity, improve the maintenance costs and allowing switching the implementations of separate sub-modules. Additionally such logical grouping increases the explainability and general understanding of the system.

On Figure 4.1 these modules are represented with the boxes, and their functionality groups are presented with the colors (see the figure legend). Arrows represent the interaction and workflow order and initiation (the interaction is initiated/triggered from the origin of the arrow).

We can see that the central core of the system is the knowledge base (modules marked in purple and letter A). The knowledge base consists of *Upper Ontology* gluing everything together, *Common Sense Knowledge* to be able to "understand" user's world and check the answers for consistency, *Meta Knowledge* for enabling inference about its internal structures, *User Context KB* to hold current user context and *Knowledge Acquisition Rules* to drive the KA process from within the KB, using logical inference.

Next to the KB, is an *Inference Engine* that performs inference over the knowledge from the KB. Its modules are represented with the red color and letter B. The inference engine needs to be general enough to be able to perform over full KB, and should be capable of meta-reasoning (over the meta-knowledge and KA knowledge in the KB) about the KB's internal knowledge structures. In cases when the inference engine have some missing functionalities, some of these tasks can be supplemented by the *Procdeural Support* module. In the proposed system, inference engine handles almost all of the core KA operations, which can be separated into the following modules:

- *Consistency Checking* module which can asses the user's answers and check whether they fit within the current KB knowledge.

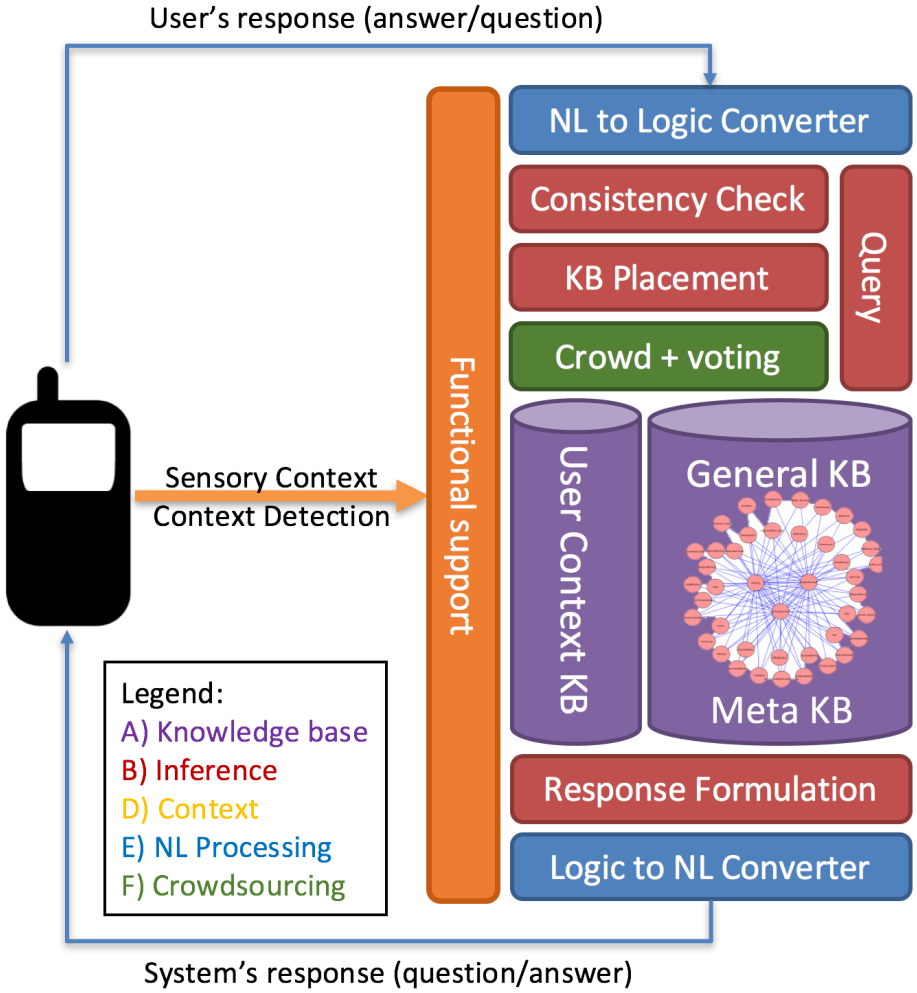

Figure 4.1: General Architecture of the KA system, with an interaction loop presented as arrows.

- *KB Placement* module which decides where into the subtree of the KB the answer should be placed.
- *Querying* module, which employs the inference engine to answer questions that are coming from the user through NL to Logic converter.
- *Response Formulation* module, which employs the KA meta-knowledge and do inference about what to say/ask next. Results of this module are then forwarded to the Logic to NL converter and then to the user.

Tightly integrated with the knowledge base and inference engine is a *Crowdsourcing Module*, which monitors crowd (multipl user's) answers and is able to remove (or move to different contexts) the knowledge from the KB, based on its consistency among multiple users. If some piece of knowledge inside the KB is questionable, the module marks it as such and then *Response Formulation* module checks with other users whether it's true or not and should maybe be removed or only kept in the one user's part of the KB. This module is represented in Green color and letter F.

At the entry and exit point of the system workflow, there are NLP processing modules which can convert logic into the natural language and vice versa. These modules are used

for natural language communication with the users. Tese two modules are represented in Blue and letter E.

On the side of the Figure, there is a procedural module (depicted with Orange color and letter D), which is a normal software module (in our implementations written in procedural programming language), which glues everything together. It contains a web-server, authentication functionalities, machine learning capabilities, connections to external services and context mining and other functions that are hard to implement using just logic and inference. This module is taking care of the interactions between submodules.

All of the modules are triggered either through the contextual triggers (also internal, like when timer detects the specific hour or time of day), or by the users. When the context changes, it causes the system to use inference engine to figure out what to do. Usually, as a consequence it results in a multiple options like questions or comments. Then it picks one and sends a request to the user. This triggering is represented with the arrows, where the blue arrows represent natural language interaction, and the orange one represents structured or procedural interaction, when the procedural module classifies or detects any useful change in the sensor data sent into the system py the part running on the mobile phone.

4.1.1 Interaction Loop

As briefly already mentioned above, besides architecture, Figure 4.1 also indicates a system/user interaction loop represented by arrows. Orange arrow (pointing directly from the phone towards the system) represents the automatic interaction or triggers that the phone (client) is sending to the system all the time. This provides one part of the user context. After the precedural part analyses the data (as described in ref to SPD) and enter findings into the KB as context, this often triggers the system to come up with a new question, or context related info. Example of such a trigger is, when user changes a location and the system figures out the name and type of the new place. On the other side, Blue arrows represent the Natural Language interaction which can happen as a result of automatic context (Orange arrow), or some other reason causing new knowledge appearing in the KB. New knowledge can appear as a consequence of answering a question from the same user, or some other user. This shows, how the actual knowledge (even if entered automatically through procedural component) is controlling the interaction, and explains how the system is initialized and how its main pro-activity driver is implemented. Examples of such initialization of the interaction is presented in Script 1 and Script2. Additionally, the user can trigger a conversation at any point in time either by continuing the previous conversation or simply starting a new one.

According to the interactions described above, the proposed KA system have two options for the interaction. Human to machine (HMI), when users initiate interaction, and machine to human (MHI), when the system initiates the interaction. The specifics of both, which cannot fit in Figure 4.1 are explained in the following sub-sections 4.2.1 and 4.2.2. On top of this, the design of the system allows a novel type of interaction which combines multiple users and machine into one conversation, while presenting this to the users as a single conversation track with the machine. This becomes useful when the system doesn't have enough knowledge to be able to answer user's questions, but it has just enough to know which other users to ask (i.e. when someone is asking a question about specigic place and there is no answer in the KB, *Curious Cat* can ask other users that it knows had been there). This type of the interaction can be called Machine Mediated Human to Human interaction (MMHHI). This allows the system to answer questions also when it doesn't know them, while simultaneously also store and remember the answers, either parsing them and assert them directly to KB, or leave them in NL for later Knowledge

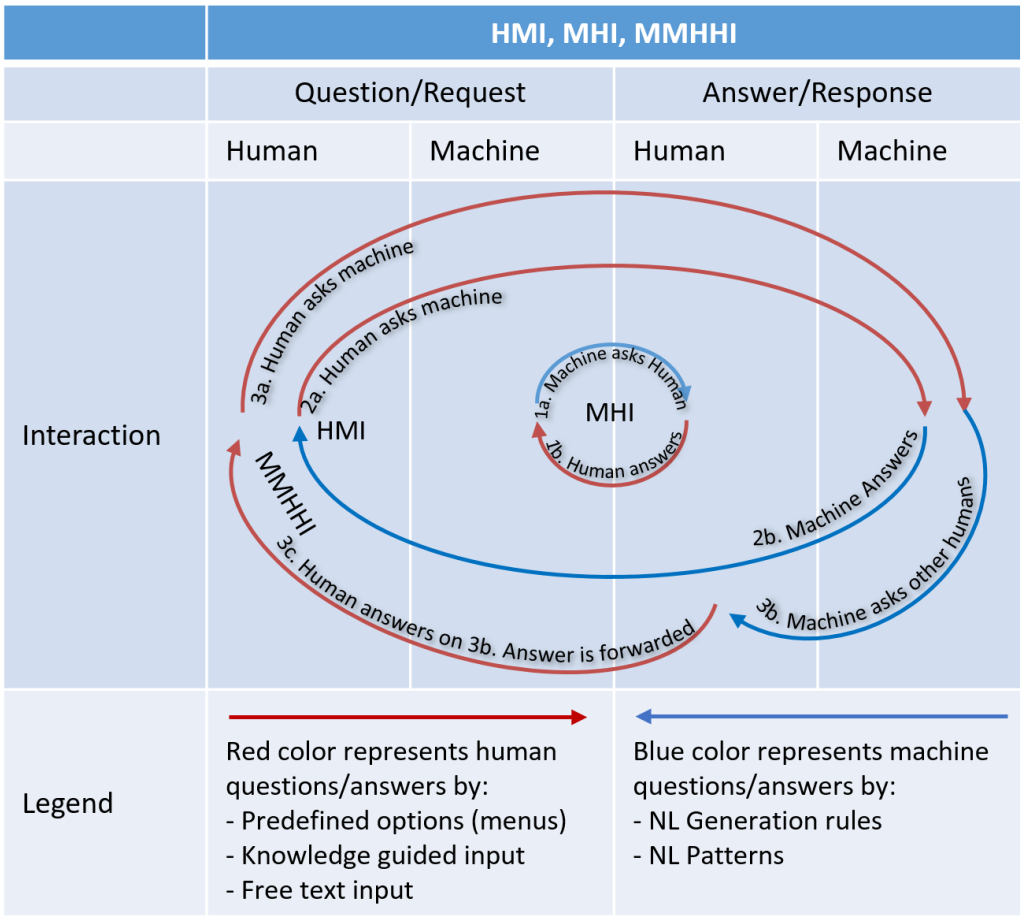

Figure 4.2: Possible interaction types between the user and Curious Cat KA System.

Mining analysis. The possible interaction types are also presented on Figure 4.2.

4.1.1.1 Machine to Human Interaction (MHI)

The most basic form of interaction between the CC system and the user, which we also use the most, is when something triggers a change in the KB and CC decides it's the time to ask or tell something. On the Figure 4.2, this is represented by the most inner loop (MHI). Example of this is when the context part of *Procedural* module classifies a new location and then asserts it into the KB. This then triggers Inference Engine which results in a new user query (same as defined in ref to CCWantsToAskLocation).

$$ccWantsToAsk(CCUser1, (userLocation(CCUser1, CCLoc1)))$$

This query then goes through logic to NL ref conversion, which is then presented to the user in NL like "Where are we? Are we at the restaurant X?". This is represented on Figure 4.2 with a blue arrow on the inner circle, marked with 1a. The presentation is handled by the client and can be in a written form, or through the text-to-speech interface. User can then answer this question and thus close the interaction loop (blue arrow marked with 1b), possibly causing a new one with her answer. The intference triggering, language rules and mechanisms for context detection are described in more detail in sections ref, ref, ref respectively. This type of the interaction is where the users answer questions and is thus part of the main research topic of this thesis.

4.1.1.2 Human to Machine Interaction (HMI)

dada

4.1.1.3 Machine Mediated Human to Human Interaction (MMHHI)

dada

4.2 Knowledge Base

Internally KB has three components. The main part, which should in any real implementation of the system also be the biggest, is the common-sense knowledge and its upper ontology over which we operate. This part of the system contributes the most to the ability to check the answers for consistency. The more knowledge already exists, the easier becomes to assess the answers. The second part is the user Context KB, which stores the contextual knowledge about the user. This covers the knowledge that the user has provided about himself (section 4.4.2) and the knowledge obtained by mining raw mobile sensors (section 4.4.1). This is represented as the orange arrow, pointing into the context part of the KB. The sensor based context allows the system to proactively target the right users at the right time and thus improve the efficiency and accuracy and also stickiness of the KA process. The third KB part, is the meta-knowledge and KA rules that drive the dialog and knowledge acquisition process (section 4.3.3). Although in our implementation we used Cyc KB and tested Umko KB, the approach is not fixed to any particular knowledge base. But it needs to be expressive enough to be able to cover the intended knowledge acquisition tasks and meta-knowledge needed for the system?s internal workings. After the KB, the second most important part of the architecture is an inference engine (in Fig. 2 marked in red and letter B), which is tightly connected to the knowledge base. The inference engine needs to be able to operate with the concepts, assertions and rules from the KB and should also be capable of meta-reasoning about the knowledge base?s internal knowledge structures. As the individual components (indicated with red color in Fig. 2) suggest, the inference engine is used for: ? Checking the consistency of the users? answers (e.g., can you order a car in a restaurant if it?s not food?). ? Placement of new knowledge inside the KB. ? Querying the KB to answer possible questions. ? Using knowledge and meta-rules to produce responses based on the user and her/his context input (similar in function to the scripts in script-based conversational agents).

Fig. 2. General Architecture of the KA system, with a simple interaction loop At both ends of the stacked chain in Fig. 2, there are natural language processing components (marked in blue and with letter E), which are responsible for logic-to-language and language-to-logic conversion (sections 2.4 and 4.5). These are crucial if we want to interact with users in a natural way and thus avoid the need for users to be experts in first order logic. This module and its components are described in more detail in section 4.5. Besides the main interaction loop, which implicitly uses crowdsourcing while it interacts with the users, there is an additional component (marked in green and with letter F). This ?crowdsourcing and voting? component handles and decides, which elements of knowledge (logical assertions) can be safely asserted and made ?visible? to all the users and which are questionable and should stay visible only to the authors of the knowledge. If the piece of knowledge is questionable, the system marks it as such and then the question formulation process will check with other users whether it?s true or not. This is described in more detail in section 4.7. In addition to logic-based components presented above, there is a functional driver system (marked in orange), which glues everything together, forwards

the results of inference to the NL converters, accepts and asserts the context into the KB, handles the synchronization between the instances of the systems, etc.

Chapter 5

Real World Knowledge Acquisition Implementation

5.1 Cyc

TBW

Chapter 6

Evaluation

TBW
chapters that

Chapter 7

Conclusions

We came to the following conclusions ...

References

Ahn, L. von (2006). "Games with a Purpose." In: *Computer* 39.6, pp. 92–94. ISSN: 0018-9162. DOI: 10.1109/MC.2006.196. URL: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=1642623.

Ahn, Luis Von, Mihir Kedia, and Manuel Blum (2006). "Verbosity : A Game for Collecting Common-Sense Facts." In: *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pp. 75–78. ISBN: 1595931783.

Ahn, Luis von and Laura Dabbish (2008). "Designing games with a purpose." In: *Communications of the ACM* 51.8, p. 57. ISSN: 00010782. DOI: 10.1145/1378704.1378719.

Bernstein, Michael et al. (2009). "Collabio: a game for annotating people within social networks." In: *Proceedings of the 22nd annual ACM symposium on User interface software and technology (UIST '09)*, pp. 97–100. ISSN: 00325910. DOI: 10.1145/1622176.1622195. URL: http://dl.acm.org/citation.cfm?id=1622195.

— (2010). "Personalization via friendsourcing." In: *ACM Transactions on Computer-Human Interaction* 17.2, pp. 1–28. ISSN: 10730516. DOI: 10.1145/1746259.1746260.

Bollacker, Kurt et al. (2008). "Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge." In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*. SIGMOD '08. New York, NY, USA: ACM, pp. 1247–1250. ISBN: 978-1-60558-102-6. DOI: 10.1145/1376616.1376746. URL: http://doi.acm.org/10.1145/1376616.1376746.

Bradeško, Luka and Dunja Mladenić (2012). "A Survey of Chabot Systems through a Loebner Prize Competition." In: *Proceedings of Slovenian Language Technologies Society Eighth Conference of Language Technologies*, pp. 34–37. ISBN: ISBN 978-961-264-048-4.

Coursey, Kino (2004). "LIVING IN CYN : MATING AIML AND CYC TOGETHER WITH PROGRAM N." In:

Demner-Fushman, Dina et al. (2015). "YAGO3: A Knowledge Base from Multilingual Wikipedias." In: *Conference on Innovative Data Systems Research (CIDR)*. URL: http://scholar.google.com/scholar?hl=en%7B%5C&%7DbtnG=Search%7B%5C&%7Dq=intitle:Automatic+Event+and+Relation+Detection+with+Seeds+of+Varying+Complexity%7B%5C#%7D0%7B%5C%%7D5Cnhttp://www.aclweb.org/anthology/C12-1129%7B%5C%%7D5Cnhttp://dx.doi.org/10.1016/j.jbi.2013.08.010%7B%5C%%7D5Cnhttp://www.informatik.uni.

Dong, Zhendong, Qiang Dong, and Changling Hao (2010). "HowNet and Its Computation of Meaning." In: *Coling 2010* August, pp. 53–56. DOI: 10.1142/9789812774675.

Downey, Doug, Oren Etzioni, and Stephen Soderland (2005). "A probabilistic model of redundancy in information extraction." In: *IJCAI International Joint Conference on Artificial Intelligence*, pp. 1034–1041. ISSN: 10450823. DOI: 10.1016/j.artint.2010.04.024.

Eslick, Ian Scott (2006). "Searching for Commonsense." Doctoral dissertation.

Etzioni, Oren, Anthony Fader, et al. (2011). "Open Information Extraction: The Second Generation." In: *Proc. Int. Joint Conf. Artificial Intell.*

Etzioni, Oren, Ana-maria Popescu, et al. (2004). "Web-Scale Information Extraction in KnowItAll (Preliminary Results)." In: *WWW 2004*.

Fader, Anthony, Stephen Soderland, and Oren Etzioni (2011). "Identifying relations for open information extraction." In: *Proceedings of the Conference on ...* pp. 1535–1545. ISSN: 1937284115. DOI: 10.1234/12345678. arXiv: arXiv:1411.4166v4. URL: http://dl.acm.org/citation.cfm?id=2145596%7B%5C%%7D5Cnhttp://dl.acm.org/citation.cfm?id=2145596%7B%5C%%7D5Cnhttp://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.1089%7B%5C&%7Drep=rep1%7B%5C&%7Dtype=pdf%7B%5C%%7D5Cnhttp://www.cs.washington.edu/research/projects/aiweb/media/papers/etzioni-ijca.

Feigenbaum, E. A. (1977). "The Art of Artificial Intelligence: Themses and Case Studies of Knowledge Engineering." In: *Proceedings of the 5th International Joint Conference of Artificial itelligence*, pp. 1014–1029.

Forbus, Kenneth D et al. (2007). "Integrating Natural Language , Knowledge Representation and Reasoning , and Analogical Processing to Learn by Reading Learning Reader : The System." In: *Proceedings of AAAI-07: Twenty-Second Conference on Artificial Intelligence*. Vancouver,BC.

Hasbro (n.d.). *Taboo board game*. URL: https://www.hasbro.com/common/documents/dad288731c4311ddbd0b0800200c9a66/2BF862075056900B1021F6D7061EDCC7.pdf.

J.R., Quinlan and Cameron-Jones R.M. (1997). "Induction of Logic Programs: FOIL and related systems." In: *New Generation Computing* 13.3, pp. 287–312. DOI: 10.1007/BF03037228. URL: https://doi.org/10.1007/BF03037228.

Kang, Jong Hee et al. (2005). "Extracting places from traces of locations." In: *ACM SIGMOBILE Mobile Computing and Communications Review* 9.3, p. 58. ISSN: 15591662. DOI: 10.1145/1094549.1094558.

Kuo, Yen-Ling and Jane Yung-jen Hsu (2010). "Goal-Oriented Knowledge Collection." In: *AAAI Fall Symposium: Commonsense Knowledge*, pp. 64–69. URL: http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/viewPDFInterstitial/2278/2605.

Kuo, Yen-ling et al. (2009). "Community-Based Game Design: Experiments on Social Games for Commonsense Data Collection." In: *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pp. 15–22. ISSN: 978-1-60558-193-4. DOI: 10.1145/1600150.1600154. URL: http://dl.acm.org/citation.cfm?id=1600150.1600154.

Lehmann, Jens et al. (2015). "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia." In: *Semantic Web* 6.2, pp. 167–195. ISSN: 22104968. DOI: 10.3233/SW-140134.

Lenat, Douglas Bruce (1995). "Cyc: A Large-Scale Investment in Knowledge Infrastructure." In: *Communications of the ACM* 38.22.

Lv, Mingqi et al. (2016). "The discovery of personally semantic places based on trajectory data mining." In: *Neurocomputing* 173, pp. 1142–1153. ISSN: 18728286. DOI: 10.1016/j.neucom.2015.08.071. URL: http://dx.doi.org/10.1016/j.neucom.2015.08.071.

Mamei, Marco (2010). "Applying Commonsense Reasoning to Place Identification." In: *International Journal of Handheld Computing Research* 1.2, pp. 36–53. DOI: 10.4018/jhcr.2010040103. URL: http://dx.doi.org/10.4018/jhcr.2010040103.

Martin, E and I Riesbeck (1986). "Uniform Parsing and Inferencing for Learning." In: *Proceedings of AAAI-86*, pp. 257–261.

Masters, James, Cynthia Matuszek, and Michael Witbrock (2007). "Ontology-Based Integration of Knowledge from Semi-Structured Web Pages." In: *Cycorp*.

Matuszek, Cynthia, John Cabral, et al. (2006a). "An Introduction to the Syntax and Content of Cyc." In: *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. AAAI Press.

— (2006b). "An Introduction to the Syntax and Content of Cyc." In: *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*. AAAI Press.

Matuszek, Cynthia, Michael Witbrock, et al. (2004). "Searching for Common Sense : Populating Cyc ™ from the Web." In: *Search*.

McKinstry, Chris, Rick Dale, and Michael J. Spivey (2008). "Action dynamics reveal parallel competition in decision making." In: *Psychological Science* 19.1, pp. 22–24. ISSN: 09567976. DOI: 10.1111/j.1467-9280.2008.02041.x.

Medelyan, Olena and Catherine Legg (2008). "Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense." In: *Proceedings of the WIKIAI Wikipedia and AI Workshop at the AAAI* 8, pp. 13–18. URL: http://www.aaai.org/Papers/Workshops/2008/WS-08-15/WS08-15-003.pdf.

Mitchell, T et al. (2015). "Never-Ending Learning." In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence (AAAI-15)*.

Mueller, Erik T (1999). *A database and lexicon of scripts for ThoughtTreasure*. Vol. 1999. CogPrints ID cog00000555 http://cogprints.soton.ac.uk, Article No. 0003004.

— (2003). *ThoughtTreasure: A natural language/commonsense platform*. URL: http://alumni.media.mit.edu/%7B~%7Dmueller/papers/tt.html (visited on 01/01/2017).

Panton, Kathy et al. (2002). "Knowledge Formation and Dialogue Using the KRAKEN Toolset." In: *Proceedings of the Fourteenth National Conference on Innovative Applications of Artificial Intelligence*, pp. 900–905.

Pedro, S D S, A P Appel, and E R Hruschka Jr (2013). "Autonomously reviewing and validating the knowledge base of a never-ending learning system." In: *Proceedings of the 22nd ...* pp. 1195–1203. ISBN: 9781450320382. URL: http://dl.acm.org/citation.cfm?id=2488149.

Pedro, Saulo D. S. and Estevam R. Hruschka (2012). "Collective intelligence as a source for machine learning self-supervision." In: *Proceedings of the 4th International Workshop on Web Intelligence & Communities - WI&C '12*. 3, p. 1. ISBN: 9781450311892. DOI: 10.1145/2189736.2189744. URL: http://dl.acm.org/citation.cfm?id=2189736.2189744.

Schubert, Lenhart (2002). "Can we derive general world knowledge from texts?" In: *Proceedings of the second international conference on Human Language Technology Research*, p. 94. DOI: 10.3115/1289189.1289263. URL: http://portal.acm.org/citation.cfm?doid=1289189.1289263.

Schubert, Lenhart and Matthew Tong (2003). "Extracting and evaluating general world knowledge from the Brown corpus." In: *Proceedings of the HLT-NAACL 2003 workshop on Text meaning - Volume 9*, pp. 7–13. DOI: 10.3115/1119239.1119241. URL: http://dx.doi.org/10.3115/1119239.1119241.

Sharma, Abhishek and Kenneth D Forbus (2010). "Graph-Based Reasoning and Reinforcement Learning for Improving Q/A Performance in Large Knowledge-Based Systems." In: *2010 AAAI Fall Symposium Series*, pp. 96–101. ISBN: 9781577354840. URL: http://www.aaai.org/ocs/index.php/FSS/FSS10/paper/download/2246/2596.

Singh, Push (2002). "The Public Acquisition of Commonsense Knowledge Push Singh The Diversity of Commonsense Knowledge." In: *AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, pp. 47–53. URL:

http://www.aaai.org/Papers/Symposia/Spring/2002/SS-02-09/SS02-09-011.pdf.

Singh, Push et al. (2002). "Open Mind Common Sense: Knowledge acquisition from the general public." In: *Cooperative Information Systems Oct. 30-Nov. 1 2002*, pp. 1223–1237. ISSN: 03029743. DOI: 10.1007/3-540-36124-3_77. URL: http://portal.acm.org/citation.cfm?id=646748.701499.

Soderland, Stephen et al. (2007). "Open information extraction from the web." In: *International Joint Conference On Artificial Intelligence*, pp. 2670–2676. ISSN: 00010782. DOI: 10.1145/1409360.1409378. URL: http://portal.acm.org/citation.cfm?id=1625705.

Speer, Robert (2007). "Open mind commons: An inquisitive approach to learning common sense." In: *Proceedings of the Workshop on Common Sense and Interactive Applications*. URL: http://www.fatih.edu.tr/%7B~%7Dhugur/inquisitive/Open%20Mind%20Commons%20An%20Inquisitive%20Approach%20to.PDF.

Speer, Robert, Joshua Chin, and Catherine Havasi (2016). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge." In: Singh 2002. arXiv: 1612.03975. URL: http://arxiv.org/abs/1612.03975.

Speer, Robert, Jayant Krishnamurthy, et al. (2009). "An interface for targeted collection of common sense knowledge using a mixture model." In: *Proceedings of the 14th International Conference on Intelligent User Interfaces*, pp. 137–146. DOI: 10.1145/1502650.1502672.

Speer, Robert, Henry Lieberman, and Catherine Havasi (2008). "AnalogySpace : Reducing the Dimensionality of Common Sense Knowledge." In: *AAAI'08 Proceedings of the 23rd national conference on Artificial intelligence*, pp. 548–553. ISBN: 9781577353683.

Suchanek, Fabian M., Gjergji Kasneci, and Gerhard Weikum (2008). "YAGO: A Large Ontology from Wikipedia and WordNet." In: *Web Semantics* 6.3, pp. 203–217. ISSN: 15708268. DOI: 10.1016/j.websem.2008.06.001. arXiv: arXiv:1203.5073v1.

Wallace, Richard (2013). "AIML 2.0 Draft Specification." URL: http://www.alicebot.org/style.pdf.

Wallace, Richard S. (2003). *The Elements of AIML Style*. Tech. rep. Alice AI Foundation. DOI: 10.1.1.693.3664. URL: http://www.alicebot.org/style.pdf.

Weizenbaum, Joseph (1966). "ELIZA–A Computer Program For the Study of Natural Language Communication Between Man and Machine." In: *Communication of the ACM* 9.1, pp. 36–45. ISSN: 00010782. DOI: 10.1145/365153.365168.

Wilcox, Bruce (2011). "Beyond Façade: Pattern Matching for Natural Language Applications." In: *Gamasutra*, pp. 1–5. URL: http://www.gamasutra.com/view/feature/6305/beyond%7B%5C_%7Dfa%EF%BF%BDade%7B%5C_%7Dpattern%7B%5C_%7Dmatching%7B%5C_%7D.php?page=1.

Witbrock, Michael (2010). "Acquiring and Using Large Scale Knowledge Knowledge Capture : Mixed Initiative." In: *Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces*. Cavtat/Dubrovnik, pp. 37–42. URL: http://ieeexplore.ieee.org/document/5546360/?reload=true%7B%5C&%7Dtp=%7B%5C&%7Darnumber=5546360.

Witbrock, Michael, David Baxter, et al. (2003). "An Interactive Dialogue System for Knowledge Acquisition in Cyc." In: *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. Acapulco, Mexico.

Witbrock, Michael, Cynthia Matuszek, et al. (2005). "Knowledge Begets Knowledge: Steps towards Assisted Knowledge Acquisition in Cyc." In: *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, pp. 99–105. URL: http://www.aaai.org/Papers/Symposia/Spring/2005/SS-05-03/SS05-03-015.pdf.

Wu, Wentao et al. (2012). "Probase: A probabilistic taxonomy for text understanding." In: *Proceedings of the 2012 ACM SIGMOD ...* pp. 481–492. ISSN: 00043702. DOI: 10.1016/j.artint.2011.01.003. URL: http://dl.acm.org/citation.cfm?id=2213891.

Zang, Liang-Jun et al. (2013). "A Survey of Commonsense Knowledge Acquisition." In: *Journal of Computer Science and Technology* 28.4, pp. 689–719. ISSN: 1000-9000. DOI: 10.1007/s11390-013-1369-6. URL: http://link.springer.com/10.1007/s11390-013-1369-6.

Bibliography

Publications Related to the Thesis

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

Journal Articles

Bradesko, Luka et al. (2017). "Curious Cat-Mobile, Context-Aware Conversational Crowdsourcing Knowledge Acquisition." In: *ACM Trans. Inf. Syst.* 35.4, 33:1–33:46. DOI: 10.1145/3086686. URL: http://doi.acm.org/10.1145/3086686.

Conference Paper

Luka, Bradesko et al. (2016). "Conversational Crowd based and Context Aware Knowledge Acquisition Chat Bot." In: *8th IEEE International Conference on Intelligent Systems IS'16*, pp. 239–252. ISBN: 9781509013548.

Other Publications (optional)

...

Biography

The author of this thesis ...