HOKKAIDO UNIVERSITY

Title	Expanding Common Sense Knowledge Database using Data Mining on Reference Corpus
Author(s)	KRAWCZYK, Marek
Degree Grantor	北海道大学
Degree Name	博士(情報科学)
Dissertation Number	甲第12640号
Issue Date	2017-03-23
DOI	https://doi.org/10.14943/doctoral.k12640
Doc URL	https://hdl.handle.net/2115/65726
Type	doctoral thesis
File Information	Marek_Krawczyk.pdf

Hokkaido University Collection of Scholarly and Academic Papers : HUSCAP

Doctoral Dissertation

Expanding Common Sense Knowledge Database using Data Mining on Reference Corpus

参照コーパスに対するデータマイニングを用いたコモンセンス知識データベースの拡張

Marek Krawczyk

Language Media Laboratory,
Graduate School of Information Science and Technology,
Hokkaido University

January 2017

Abstract

In the information society of today it becomes increasingly difficult to have a creative and enjoyable conversation without a broad background knowledge. People like to discuss concepts and ideas, and this requires a mixture of commonsense and general knowledge spanning through a broad spectrum of topics. If we want machines to understand such conversations and interact with people using natural language, it is necessary to equip them with the ability to not only analyze linguistic features such as syntax, but also with means to process semantics, that is the knowledge of the underlying concepts represented by the surface text. The conceptual information is gathered in large-scale general knowledge bases such as Cyc, YAGO and ConceptNet. In this thesis I focus on ConceptNet, a knowledge representation project that provides a large semantic graph describing general human knowledge. I have chosen ConceptNet as it captures a wide range of common sense concepts and relations, and its simple semantic network structure makes it easy to use and manipulate. ConceptNet's basic data unit is an assertion, that is two concepts connected with a relation. ConceptNet was designed to contain knowledge collected by the Open Mind Common Sense project's website, as well as knowledge from similar websites and online word games which automatically collect general knowledge in several languages. This open-source knowledge base is used for many applications such as topic-gisting, affect-sensing, dialog systems, daily activities recognition, social media analysis and handwriting recognition. Manual expansion of the knowledge base would be a long and labor-intensive process. For example, nadya.jp, an online project that aims to gather knowledge by using a game with a purpose, since its launch in 2010 has been able to introduce a little over 43,500 entries to ConceptNet. It is therefore evident that we need to develop automatic methods to gather new data. Creating such method and confirming its effectiveness is the goal of my research.

This thesis presents a method for extracting IsA assertions (hyponymy relations), AtLocation

assertions (informing of the location of an object or place), LocatedNear assertions (informing of neighboring locations), CreatedBy assertions (informing of the creator of an object) and MemberOf assertions (informing of group membership) automatically from Japanese Wikipedia XML dump files. I use the Hyponymy extraction tool v1.0, which analyzes definition, category and hierarchy structures of Wikipedia articles to extract IsA assertions and produce an information-rich taxonomy. From this taxonomy I extract additional information, in this case AtLocation, LocatedNear, CreatedBy and MemberOf types of assertions, using the presented original method. The method exploits the qualities of the Japanese writing system to gather new assertions by analyzing linguistic patterns. As a further step an automatic extraction of general common sense knowledge assertions is being performed on the basis of previously generated instance-related IsA, AtLocation, CreatedBy and MemberOf assertions. The acquired assertions would be suitable for introduction to the Japanese part of the ConceptNet common sense knowledge ontology.

The presented experiments prove that the research goal has been achieved on a large scale: the method produced satisfactory results, and helped to acquire 5,866,680 IsA assertions with 96.0% reliability, 131,760 AtLocation assertion pairs with 93.5% reliability, 6,217 LocatedNear assertion pairs with 98.5% reliability, 270,230 CreatedBy assertion pairs with 78.5% reliability and 21,053 MemberOf assertions with 87.0% reliability. The proposed method surpassed the baseline system in terms of both precision and the number of acquired assertions. Further processing of the data produced additional 74,226 AtLocation assertions, 330,418 CreatedBy assertions and 1,355 MemberOf assertions representing general common sense knowledge triplets based on at least 50 instance-related examples. The reliability of top 100 samples was assessed at the level of 98.5%, 91.5% and 71.0% respectively.

The presented method and results prove that it is possible to extract general and common sense knowledge automatically from a large reference corpus. The results could be refined further by applying more sophisticated methods, however this approach already shows a direction which future research may take in order to expand the common sense knowledge databases useful for various AI related applications.

Japanese Abstract

学位論文内容の要旨

今日の情報社会では，幅広い背景知識がなければ，創造的かつ楽しい会話を行うことがますます困難になっている．人々は概念やアイデアについて議論することを好み，またそれを行うためには，幅広い話題にまたがる常識と一般知識の融合が必要である．コンピュータがこのような会話を理解し，自然言語を使用する人々との対話を行うためには，構文などの言語的特徴を分析するだけではなく，意味を処理する手段，すなわちテキストの表層に表現されている概念の知識が必要である．そうした概念情報は，Cyc，YAGO，ConceptNetなどの大規模な一般的な知識ベースに集約されている．

本学位論文では，人間の一般的な知識を記述する大規模なセマンティックグラフを提供する知識表現プロジェクトであるConceptNetに焦点をあてる．ConceptNetを利用した理由は，このデータベースが幅広い常識の概念や関係を捉えており，単純な意味ネットワーク構造が使いやすく操作しやすいという特徴を持つためである．ConceptNetの基本データユニットはアサーションと呼ばれ，これは2つの概念とこれらの概念間の関係を表す．ConceptNetはOpen Mind Common Senseプロジェクトのウェブサイトで収集された知識だけではなく，複数の言語で一般的な知識を自動的に収集する類似のウェブサイトやオンライン単語ゲームの知識を含むように設計されている．このオープンソースの知識ベースは，トピック推定，感情認識，対話システム，日常活動の認識，ソーシャルメディア分析，手書き文字認識などの多くのアプリケーションで使用されている．

一方，人手による知識ベースの拡張は長くて労働集約的なプロセスになる．例えば，2010年に開始されて以来，連想ゲームを使って知識を集めているオンラインプロジェクトのnadya.jpはConceptNetにわずか43,500件のエントリーを投入することしかできていない．したがって，新しいデータを収集するためには自動的な知識獲得の手法を開発する必要があることは明らかである．従って，ConceptNetの知識を自動的に獲得する手法の開発を行うことが本研究の目的である．

本学位論文では日本語ウィキペディアのデータからIsAアサーション（下位関係），AtLocationアサーション（ものまたは場所の位置を示す），LocatedNearアサーション（近隣の場所を示す），CreatedByアサーション（オブジェクトの作成者を示す），またはMemberOfアサーション（グループのメンバーを示す）を自動的に抽出する方法を提案する．提案手法では，IsAアサーションと拡張した階層構造の処理結果を抽出するために，ウィキペディアの記事の定義文，カテゴリ，階層構造を分析することにより上位下位関係を抽出するツールを使用する．このツールによって拡張した階層構造の処理結果から提案手法を用いて，追加の情報，すなわちAtLocation，LocatedNear，CreatedByおよびMemberOfアサーションを抽出する．この手法では，言語学的なパターンを分析することによって新しいアサーションを収集するために，日本語の表記体系の特徴を利用する．また，更なる知識の自動抽出のために，以前に生成されたIsA，AtLocation，CreatedByおよびMemberOfアサーションのインスタンスに基づいて，一般的な常識知識のアサーションを自動的に抽出する．獲得したアサーションはConceptNetの常識知識オントロジーの日本語の部分への導入に適している．

性能評価実験では，本研究の目的が大規模なデータで達成されたことを証明した．96.0%の信頼性を持つ5,866,680個のIsAアサーション，93,5%の信頼性を持つ131,760個のAtLocationアサーション，98.5%の信頼性を持つ6,217個のLocatedNearアサーション，78.5%の信頼性を持つ270,230個のCreatedByアサーション，及び87.0%の信頼性を持つ21,053個のMemberOfアサーションを獲得することができ，提案手法の有効性が確認された．提案手法では精度と取得されたアサーションの数のいずれもベースラインシステムを上回った．また，データの獲得処理では，少なくとも50のインスタンスに基づいて，一般的な常識知識を表す74,226個のAtLocationアサーション，330,418個のCreatedByアサーションおよび1,355個のMemberOfアサーションを生成した．上位100サンプルの信頼性は，それぞれ98.5%，91.5%および71.0%のレベルで評価された．

提案手法および実験結果から大規模な参照コーパスから一般的および常識的知識を自動的に抽出することが可能であることが確認された．一方で，より洗練された手法を適用することによって提案手法ををさらに改善させることが可能であるが，このアプローチは様々な人工知能関連アプリケーションに有用なコモンセンス知識データベースを拡大するために今後の研究が必要とする方向性を示していると考えられる．

Contents

Abstract **i**

List of Tables **vii**

List of Figures **viii**

1 Introduction **1**
1.1 Introduction . . . 1
1.2 Research Background . . . 1
1.3 Contribution . . . 2
1.4 Thesis Structure . . . 3

2 Related research **5**
2.1 Introduction . . . 5
2.2 General and common sense knowledge . . . 5
2.3 Knowledge representations . . . 7
2.3.1 Cyc . . . 7
2.3.2 YAGO . . . 8
2.3.3 Probase . . . 9
2.3.4 ConceptNet . . . 10
2.3.5 Quantitative analysis of ConceptNet 5.3 . . . 13
2.4 Knowledge acquisition methods . . . 16
2.5 Conclusion . . . 18

3 Data mining on reference corpus **20**
3.1 Introduction . . . 20
3.2 Hyponymy relation as IsA relation . . . 20
3.3 Extracting other relations . . . 27
3.3.1 AtLocation assertions acquisition . . . 30
3.3.2 LocatedNear assertions acquisition . . . 34
3.3.3 CreatedBy assertions acquisition . . . 34
3.3.4 MemberOf assertions acquisition . . . 38
3.4 Generalizing over assertions . . . 39
3.5 Performance optimization . . . 41

4 Evaluation **43**
4.1 Introduction . . . 43
4.2 Evaluation methodology . . . 43
4.3 IsA assertions evaluation . . . 45
4.4 AtLocation assertions evaluation . . . 47
4.5 LocatedNear assertions evaluation . . . 47
4.6 CreatedBy assertions evaluation . . . 48
4.7 MemberOf assertions evaluation . . . 51
4.8 General assertions evaluation . . . 52
4.9 Conclusion . . . 53

5 Discussion **55**
5.1 Introduction . . . 55
5.2 Case study of potential application . . . 55

6 Conclusion **57**
6.1 Overall conclusions . . . 57
6.2 Future work . . . 58

Bibliography **59**

Appendices **69**

Research Achievements **93**

Acknowledgments **95**

List of Tables

3.1 Examples of extracted hyponymy pairs presented by the method's authors [87]. . . . 25
3.2 Examples of augmented hyponymy relations generated by Yamada *et al.* [92] method. 29

4.1 Evaluation results for IsA relations. 46
4.2 Examples of generated IsA assertions. 46
4.3 Evaluation results for AtLocation relations in comparison with the nadya.jp baseline. 47
4.4 Examples of generated AtLocation assertions. 48
4.5 Evaluation results for LocatedNear relations 48
4.6 Examples of generated LocatedNear assertions. 49
4.7 Evaluation results for CreatedBy relations. 49
4.8 Examples of generated CreatedBy assertions. 50
4.9 Examples of erroneous CreatedBy assertions. 50
4.10 Evaluation results for MemberOf relations. 51
4.11 Examples of generated MemberOf assertions. 51
4.12 Evaluation results for the acquired relations. 53
4.13 Examples of generated general assertions. 53

List of Figures

2.1 Example of a single edge connecting two nodes. 12
2.2 Graph structure of ConceptNet (from [50]). 12
2.3 Main languages represented in ConceptNet 5.3 and the number of their assertions. . . 13
2.4 Relations of ConceptNet 5.3. 14
2.5 Relations of the Japanese section of ConceptNet 5.3. 15

3.1 Example of Japanese Wikipedia page with marked elements for extraction (upper part) and its source code (lower part). 22
3.2 Example of English Wikipedia page with marked elements for extraction (upper part) and its source code (lower part). 23
3.3 Patterns for finding plausible hypernym X (from [87]). 25
3.4 Procedure of Yamada *et al.* method (from [92]). 28
3.5 Flowchart of the proposed method. 30
3.6 Procedure of AtLocation relation extraction module. 33
3.7 Procedure of LocatedNear relation extraction module. 35
3.8 Procedure of CreatedBy relation extraction module. 37
3.9 Procedure of MemberOf relation extraction module, phase one. 39
3.10 Procedure of MemberOf relation extraction module, phase two. 39
3.11 Outline of the proposed method for general assertion acquisition exemplified with AtLocation assertion type. 41

Chapter 1

Introduction

1.1 Introduction

This thesis summarizes the results of the research into automatic general and common sense knowledge acquisition. This chapter serves to introduce the thesis. More specifically, Section 1.2 introduces the background of the research. Section 1.3 describes the contribution made as a result of the research. Finally, Section 1.4 describes the structure of the remainder of this thesis.

1.2 Research Background

Artificial intelligence is a field of study investigating the creation of intelligent machines. The goal of this study is to create computer systems able to operate on the human level of intelligence, and beyond. The combination of human analytical abilities and the computational speed of machines could lead to one of the biggest leaps the civilization have ever made. One of the approaches towards creating systems which think like humans is to equip them with the same knowledge humans possess, and then teach them how to manipulate it as people do. Therefore, in order to harmoniously cooperate with humans, intelligent systems will sooner or later require common sense knowledge and reasoning.

The reasons why artificial intelligence needs to be equipped with common sense knowledge are twofold. Firstly, various expert systems require it to function properly. Expert systems are designed to deal with solving a strictly defined set of problems and cannot deal with tasks which are out of their scope. Equipping such systems with common sense would enable them to assess whether the task at hand could be solved by them or not. Additionally, common sense could guide them while analyzing new situations [1]. Secondly, common sense knowledge is crucial in the implementation of

truly interactive systems. Those systems do not only need to understand natural language, but also correctly interpret the users' intentions, plans, preferences, feelings, context and so on. Human beings learn all those things by interacting with other humans since the moment they are born. Computers lack both such experience and cognitive skills to use it. Therefore it is necessary to gather such knowledge into a machine readable form so that they can utilize it as the basis for further observations and development.

Common sense knowledge is to some extent language and culture dependent. People from different parts of the planet share some ideas that can be unknown by other groups, and such ideas are most often expressed by language. As artificial intelligence and common sense knowledge related research is conducted mainly in English speaking countries, and because of the special role English has on the international scene, this language is dominating the available knowledge repositories. Japanese language is greatly underrepresented in such data pools. A common sense knowledge base that already contains Japanese language entries and would be good candidate for expansion is called ConceptNet [2]. This knowledge representation project provides a large semantic graph describing general human knowledge. ConceptNet has been chosen as a target of expansion as it captures a wide range of common sense concepts and relations, and its simple semantic network structure makes it easy to use and manipulate. It is already utilized for many applications such as topic-gisting [3], affect-sensing [4], dialog systems [5], daily activities recognition [6], social media analysis [7] and handwriting recognition [8]. Enabling systems using Japanese language to perform such tasks by introducing more Japanese language assertions into the database is the goal of this work.

1.3 Contribution

This thesis proposes an original and novel method for an automatic large-scale Japanese language common sense knowledge extraction from a open-domain reference corpus. The novelty of the proposed method consists of three parts: (a) it uses Japanese language information-rich taxonomy extracted from a large scale reference corpus, in this case Wikipedia, as an input; (b) it acquires data about individuals creating AtLocation, LocatedNear, CreatedBy and MemberOf relations between them by exploiting the qualities of Japanese language; (c) it acquires data about general concepts creating AtLocation, CreatedBy and MemberOf relations between them by generalizing over previously

extracted individuals data. However, the biggest contribution is the output the proposed method. By applying the presented methods I was able to gather 5,866,680 IsA assertions, 131,760 AtLocation assertions, 6,217 LocatedNear assertions, 270,230 CreatedBy assertions and 21,053 MemberOf assertions related to individuals in Japanese language which, as experiments show, represent accuracy estimated at the levels of 96.0%, 93.5%, 98.5%, 78.5% and 87.0% respectively. Additionally, the applied method allowed for the acquisition of general assertions. The goal of this research was to introduce new data into the ConceptNet's database. Considering the fact that the Japanese section of ConceptNet 5.3 has only 6,315 IsA assertions, 10,259 AtLocation assertions and no LocatedNear, CreatedBy or MemberOf assertions, the contribution of the newly acquired data could take this knowledge base to the next level of applicability.

1.4 Thesis Structure

This thesis consists of 6 chapters.

The current Chapter 1 acts as an introduction to the thesis. It presents the research background and the contributions the research made to the field of artificial intelligence. Finally, it provides the reader with an overview of the thesis structure.

Chapter 2 reviews the field of common sense acquisition. First it defines the concept of common sense and presents its qualities. Then it reviews various knowledge representation projects, with the major examples described in detail. The focus was put on ConceptNet, the target database, and its introduction is accompanied by the quantitative analysis of its resources. Finally, the chapter describes the existing approaches towards knowledge acquisition.

Chapter 3 contains the description of the proposed method for Japanese language common sense knowledge acquisition. It starts with the introduction of the base methodology and then presents the developed methodology of extracting particular kinds of assertions, in this case AtLocation, LocatedNear, CreatedBy and MemberOf relations. The chapter continues with the description of a method for general assertions acquisition. Finally, it discusses the performance optimization steps undertaken while implementing the methods.

Chapter 4 presents the evaluation of the proposed method's output. It describes the evaluation methodology first, followed by the presentation of evaluation results concerning every type of the

extracted assertions, that is IsA, AtLocation, LocatedNear, CreatedBy, MemberOf, and additionally general assertions.

Chapter 5 discusses the potential applicability of the acquired data.

Finally, Chapter 6 concludes the thesis and discusses potential future work.

Chapter 2

Related research

2.1 Introduction

This chapter presents the research related to the proposed method for common sense knowledge database expansion. Section 2.2 defines the notion of common sense knowledge. Section 2.3 introduces various knowledge representation projects, with prominent examples described in more detail in Subsections 2.3.1 - 2.3.3. Subsections 2.3.4 - 2.3.5 contain a thorough description of the target of the expansion project, that is ConceptNet, and its quantitative analysis. Section 2.4 describes various approaches towards knowledge acquisition. Finally, Section 2.5 concludes the chapter.

2.2 General and common sense knowledge

Marvin Minsky, one of the pioneers of artificial intelligence research, defines common sense knowledge as the kind of facts and concepts that most of us know [9]. He stated that common sense, besides of the knowledge, also includes common sense reasoning skills which people use to apply the knowledge. It is therefore clear that in order to make computers truly intelligent on a human level, it is necessary to both gather the knowledge and explore the methods of manipulating it.

The above definition common sense knowledge is very broad. It is in fact very difficult to define it clearly, that is to set up definite boundaries between knowledge that is common sense and that is not. One approach to solve this would be not to set any boundaries at all. However, gathering the total of human knowledge seems to be a daunting task. Enumerating certain qualities of common sense knowledge would be a good starting point in this endeavor.

Common sense knowledge is something possessed and shared by a group of people. Considering

the level of social stratification, defining such group and its knowledge would be very complex. The assumption therefore would be to treat the whole of society as such group. Another aspect of common sense knowledge would be its fundamentality - people understand it so well that they take it for granted and assume that every one around possess it as well. This results in the next quality of such knowledge - implicitness. As everyone is aware of it, it it not necessary to talk or write about it. This creates a serious problem for systems trying to extract such knowledge from textual data. Considering the above mentioned issue of defining the boundaries of such knowledge, it would have to be tremendously large scale, both in the amount of gathered pieces of information and in diversity of such information. This implies the open-domain quality - it can not be limited to a specific section of reality and it needs to refer to all possible domains. Finally, a perfect common sense knowledge would have to be default, that is representing assumptions about typical cases of normal life. Therefore such knowledge would be open to revision rather than definitely correct [10]. As the set of default facts about the world is a subject to dynamic change along the timeline of human history, it can be speculated that a finite and complete common sense knowledge is not an attainable goal, but rather a direction to follow.

Common sense knowledge can be divided into three types: factual knowledge, ontological knowledge and rules. Factual knowledge contains information about the surrounding reality. As mentioned before the pool of such knowledge cannot be limited to a particular domain. Its spectrum needs to be as broad as possible, covering all aspect of life. Of course we can imagine that some pieces of information would be known to a broader section of the society than others, and therefore should be treated as more fundamental. This aspect should be taken into consideration while designing systems for creating the knowledge repositories and those conducting common sense reasoning. Ontological knowledge describes terms in some domain represented as statements about concepts and properties. A particular kind of ontological knowledge is taxonomic knowledge, which covers concept and relation hierarchy. Taxonomies create a base of every ontology. Finally, rules include the knowledge about the laws governing the reality and are hardest to acquire.

The focus of this work is the extraction of factual knowledge representing the biggest possible spectrum of domains. This factual knowledge can then act as the basis for inferring general statements that can be used for common sense reasoning. In order to augment the process of gathering the knowledge it is useful to look at the knowledge representation projects and investigate how they

organize the data. Knowing the organization it is then possible to design a method that would extract information fitting to the representation's paradigm.

2.3 Knowledge representations

In order to make computers able to operate on knowledge and decode its meaning, common sense and general knowledge should be coded in a machine readable form. Usually knowledge representations organize the data into three main groups: individuals, concepts and relations. Individuals refer to particular objects or persons existing in the world, such as Michael Jordan, terms such as Chicago Bulls and so on. Concepts refer to collection of individuals, such as NBA player. Relations reflect the relationship between individual or concepts, such as (Michael Jordan, MemberOf, Chicago Bulls). There are many projects aiming at gathering knowledge and organizing it into a consistent digital form. The examples of such projects include ThoughtTreasure [11], HowNet [12], KNEXT [13], Freebase [14], DBPedia [15] and NELL [16]. The most prominent and relevant systems has been described in detail in the following sections.

2.3.1 Cyc

The Cyc project was initiated in 1984 by Douglas B. Lenat who later founded Cycorp Inc., a company which continues the development process. The goal of the initiative is to create a formalized database of English language common sense knowledge processable by computers to enable them to perform reasoning tasks on that knowledge [17]. Cyc consists of a Cyc knowledge base and a collection of Cyc inference engines. In order to codify the knowledge Lenat designed a formal language called CycL, with syntax deriving from predicate calculus and Lisp programming language. The knowledge base consists of terms and assertions connecting those terms. The assertions may refer both to simple facts as well as general rules. The knowledge contained within the database is grouped using so called micro-theories, which act as contexts in which the gathered pieces of information are true. At the current stage the knowledge base contains over 500,000 terms including about 17,000 types of relations, and about 7,000,000 assertions connecting the terms [18]. Cycorp Inc. provides three versions of the knowledge base: freely available OpenCyc, ResearchCyc, which is a full version of the database available to research institution, and the commercial EnterpriseCyc. The project's inference

engine is able to perform general logic deduction and other operations over the knowledge base such as inheritance and automatic classification.

At the early stage of development the project mainly depended on labor-intensive manual introduction of knowledge supported by knowledge authorizing tools. These included systems designed to aid trained ontologists [19], experts on a specific field of study [20] [21] and volunteers [22]. Cyc also takes an advantage of textual resources in two approaches: either the facts or rules are extracted from the source, transformed to fit the CycL specifications and introduced into the knowledge base [23], or external repository is treated as an extension of the database [24] [25].

Cyc has been used in several different applications. For example the Research Analyst Assistant provides a multi-domain platform providing analysts with answers to complex questions posed in natural language [26]. The system interprets the question, searches for the information necessary to form the answer and integrates domain-specific and general knowledge with open-source and proprietary databases to build a response either in natural language or by other means suitable to convey information in a particular domain, such as maps, time-lines or charts. The reasoning path and the source data can be presented if necessary. The system may be applied to various domains such as medical record analysis, counter-terrorism analysis or financial analysis. Other applications of Cyc include alternative question answering system [27], word sense disambiguation [28], semantic web [29] [30], and integration of heterogeneous data sources [24].

2.3.2 YAGO

YAGO is an ontology created by Suchanek *et al.* [31]. It is based on WordNet [32], providing a large amount of entities, and Wikipedia, providing a structured taxonomy. Entities covered by the ontology include individuals, classes, relations and fact identifiers. The database contains over 1,000,000 entities and 5,000,000 facts consisting of two entities connected with a relation. Each fact is described by a confidence score, which has a value between 0 and 1. YAGO is extendable as new sources can be added to the ontology.

To extract various kinds of relations YAGO employs a combination of rule-based and heuristic methods. The relations include TYPE relation (for example (Alber Einstein, type Physicist)), the MEANS relation (like (“urban center”, means, city)) and other relations referring to data extractable

from Wikipedia's infoboxes, such as bornInYear. The evaluation of the YAGO resources revealed that the accuracy of the extracted data oscillates around the level of 95%. The number of assertions, which could be counted in millions, were much larger than other ontologies such as OpenCyc or KnowItAll [33].

The second generation of the ontology, YAGO2, employs a paradigm where entities and facts are placed in time and space dimensions [34]. It contains 9,800,000 entities and 447,000,000 facts automatically extracted from Wikipedia, WordNet and GeoNames [35]. The project also employs rule-based extraction method. In the previous version of the system these rules were hardwired into its source code. YAGO2 made the extension of the extraction rules easier by storing them in text files.

The third version is an extension of YAGO knowledge base that combines the information from the Wikipedias in multiple languages [36]. It contains over 4,500,000 entities, 8,900,000 facts, 15,600,000 type facts and 1,300,000 labels automatically extracted from English, German, French, Dutch, Italian, Spanish, Russian, Polish, Arabic and Persian Wikipedias. Japanese language is not represented in the database. The base is publicly accessible through a Web interface and available for download from the creators' website [37].

There are many project that apply YAGO as a resource. The examples of such applications include named entities extraction [38], web search results categorization [39], and recommendation system [40].

2.3.3 Probase

Probase is a project developed by Microsoft Research aiming at building a probabilistic taxonomy of concepts to enable computer systems to conceptualize similarly to human beings [41]. Such conceptualization covers both instantiating concepts, that is providing a typical instance of a given general concept (for example from "largest company" to "China mobile") and abstracting from one or multiple instances to a general concept describing them (for example from "China, India, Brazil" to "emerging markets"). Probase is reported to be unique in two aspects: it has a much larger concepts base (covering 2,700,000 concepts) than other comparable bases, and it measures the plausibility and typicality of the data using probabilities, and this is used to execute probabilistic reasoning with the taxonomy.

The procedure of the project iteratively extracts IsA relations using a set of patterns applied to 1.68

billion Web pages. With each iteration it extract new IsA relation pairs and then uses them to increase the precision and recall of the extraction in the next iteration. The extraction is performed in three steps. First, the procedure generates a list of candidate super-concepts and a list of sub-concepts using extraction patterns. Then the system computes the likelihood ratio between two given candidate super-concepts, and the one representing the highest likelihood is selected as the super-concept. Finally, following an observation that sub-concepts closest to the super-concept are more likely to be valid, the system finds a sub-concept which likelihood given the super-concept is above a threshold, and then assumes all candidate sub-concepts with a higher likelihood to be valid sub-concepts of a given super-concept.

The core version of the IsA data is available for academic use and can be downloaded from the Microsoft Concept Graph website [42]. The applications of Probase include semantic web search [43], short text understanding [44] and open question answering [45].

2.3.4 ConceptNet

The development of ConceptNet started as a part of the Open Mind Common Sense project initiated at Massachusetts Institute of Technology's Media Lab, which aimed at collecting information necessary for various computer applications to understand discourse between human beings. With time the project grew into an international Common Sense Computing Initiative. The first version of ConceptNet provided to the public was ConceptNet 2 [2], which was distributed as a Python data structure together with software to read and operate it. This version of the project contained only English entries, but soon became multilingual with the introduction of a sister project OMCS no Brasil [46] gathering knowledge in Brazilian Portuguese, and GlobalMind [47], gathering knowledge in English, Chinese, Japanese and Korean. ConceptNet in its third version [48] was moved to a SQL database which could be updated more easily, also by users interacting with a Web site. The resources in English and Brazilian Portuguese were kept in two separate databases. The integration of different language resources into one database came with the introduction of ConceptNet 4, which had a normalized database structure and contained the intormation from English OMCS, OMCS no Brasil, GlobalMind and additional OMCS in Dutch [49]. Contributions from other projects, such as online games collecting knowledge in English, Chinese and Japanese were also incorporated into the

database. To facilitate the use of the database within other projects, a Web API was added, which allowed users to access and query the accumulated resources. The motivation for developing a next version of the database was the difficulty of incorporating data from other projects. The data had to be aligned and analyzed in search for duplicates already appearing in the SQL database, which was time and labor intensive. Because of this the alignment was not performed regularly, which resulted in many out-of-sync versions of the database being maintained by separate projects. Restructuring the database for easier integration is especially important to realize the new goals of ConceptNet 5 [50], which is inclusion of knowledge from other crowd-sourced providers, particularly those mining data from Wiktionary [51] and Wikipedia [52]. The goals also include adding links to other sources such as DBPedia [53], Freebase [14] and WordNet [32], as well as supporting machine-reading tools extracting relations from Web pages, such as ReVerb [54].

ConceptNet offers a set of features unique from other knowledge representation projects:

- concepts are represented in many natural languages in form of words and phrases
- it covers not only relationships, but also common-sense associations between concepts usually made by people
- its sources have a wide range of formality and granularity
- contains specific facts as well as common sense knowledge

The sources of ConceptNet 5 include data gathered by OMCS project in English [55], Portuguese [46] and Dutch [49], multilingual data, which includes translations between assertions, from GlobalMind [47], games with a purpose [56] collecting common knowledge in English - Verbocity [57], Japanese - nadya.jp [58] and Chinese [59], a process scanning English Wiktionary [51], WordNet 3.0 [60], and selected fragments of data gathered by DBPedia [61] and ReVerb [54]. With version 5.3 of the database the pool of sources was supplemented with one that rapidly increased the amount of Japanese language assertions: JMdict - a Japanese Multilingual Dictionary [62]. The project, aiming at the compilation of a multilingual lexical database with Japanese as the pivot language, started in 1999 as an offshoot of the EDICT Japanese-English Electronic Dictionary project [63]. The entirety of ConceptNet 5 is available for free download from the project's website [64].

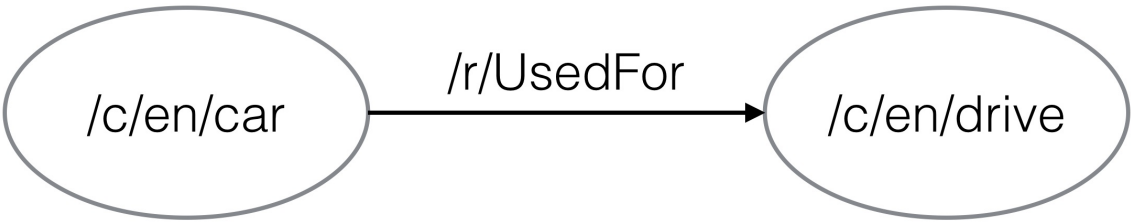


Figure 2.1: Example of a single edge connecting two nodes.

ConceptNet 5 structure consists of a network of nodes and the edges that connect them [50]. Each node is a concept described by a singe word, a word sense or a short phrase written in a natural language. Edges, as mentioned before, are the connections established between the nodes. Figure 2.1 shows an example of an edge. Proper namespaces are applied to describe every edge and create its Universal Resource Identifier. The namespaces include /c/ for concept, /r/ for relation and /s/ for source. The fundamental element of an edge is a relation: a codified description of a relationship between the two connected nodes. A few main examples of relations present in ConceptNet include a general RelatedTo relation, hierarchical IsA relation, PartOf, UsedFor, AtLocation, LocatedNear, HasProperty, CreatedBy, TranslationOf, etc. In total there are 52 kinds of relations. Each edge also contains information about sources of the underlying relation, surface text describing this relation and other additional features.

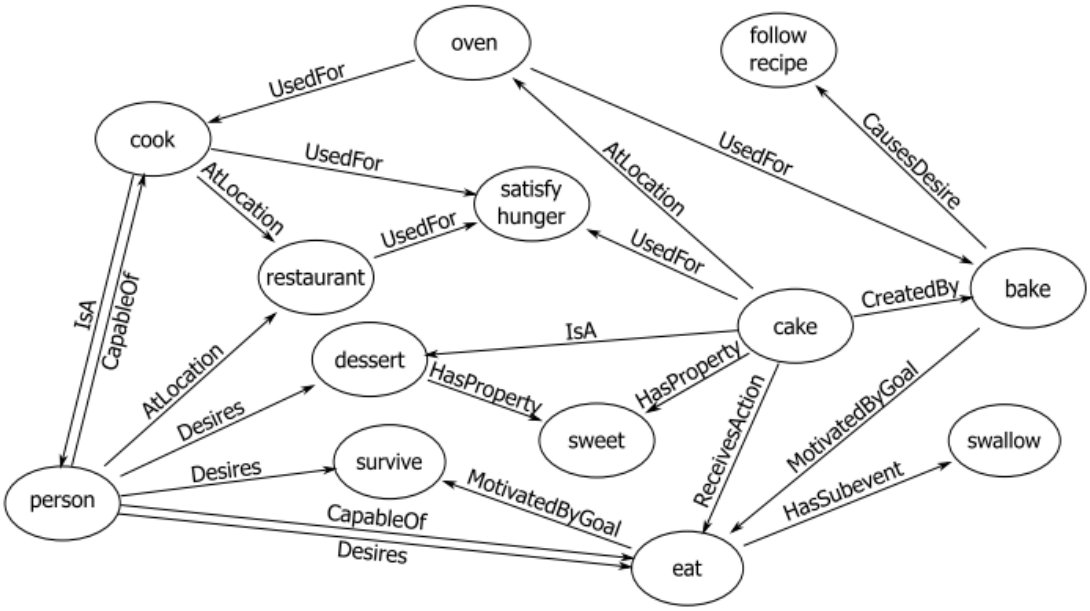


Figure 2.2: Graph structure of ConceptNet (from [50]).

One or more edges create an assertion - the proposition expressed by a relation between two concepts. The goal of providing the ontology with additional sources is to introduce data to create new edges for the graph, which would lead to the establishment of new, meaningful assertions about the surrounding reality. An example of graph describing a certain fragment of reality is shown on Figure 2.2.

2.3.5 Quantitative analysis of ConceptNet 5.3

In order to identify whether there exists a need for the expansion of the Japanese section of ConceptNet, and if so, which types of assertions should be populated, I performed a qualitative analysis of the ontology's version 5.3.

The first question was how strongly the Japanese language is represented in the database in comparison with other languages. In order to confirm this, a script has been prepared to scan through every assertion and check the language code attached to both concepts of the assertion. If both concepts referred to the same language then the assertion was assigned to that language. In cases when one concept belonged to one language and the other to a different one, which is the case for example with TranslationOf assertions, then the assertion was assigned to both languages. Figure 2.3 shows the results of the analysis concerning top 11 languages.

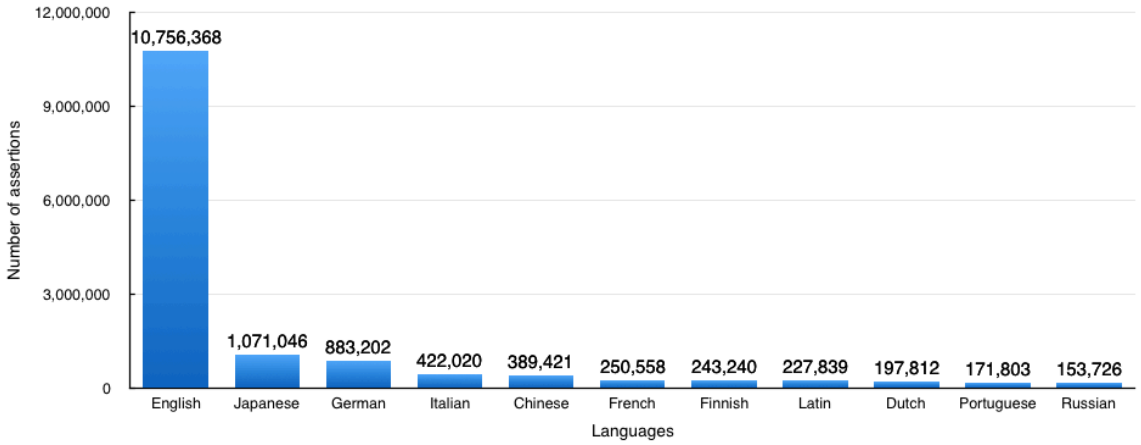


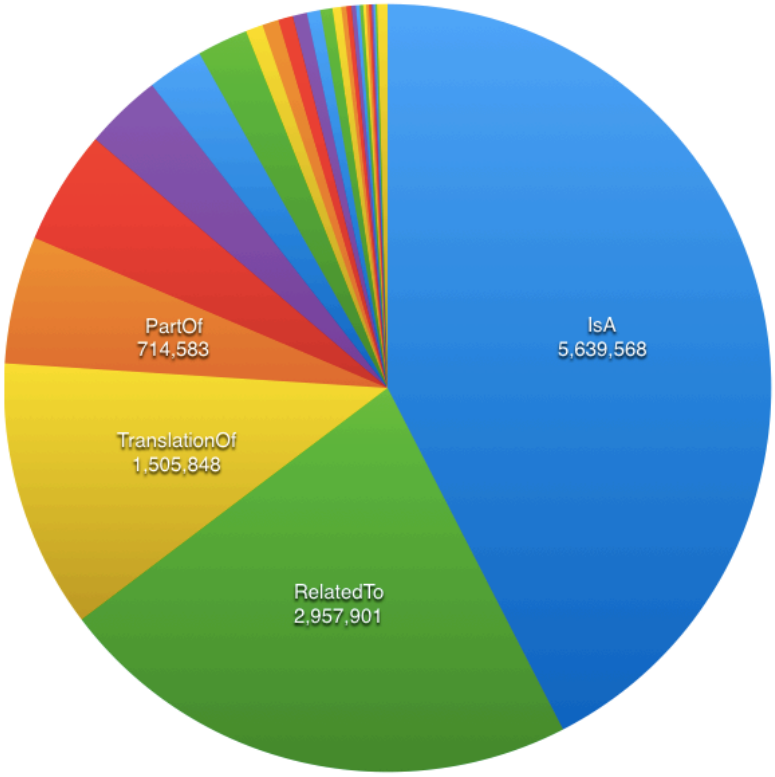
Figure 2.3: Main languages represented in ConceptNet 5.3 and the number of their assertions.

The calculation process revealed that the language with the strongest representation in ConceptNet is English, with over 10,700,000 assertions. This is not surprising taking into consideration the origin of the ontology and general abundance of linguistic sources referring to that language.

Second place belongs to Japanese, however the disproportion in comparison with English is remarkable - with over 1,070,000 assertions, the Japanese section is ten times smaller than the English language part. This without any doubt has a serious impact on the applicability of the database to various applications. The broader the data set gets, the more versatile it becomes in use. It is therefore vital to introduce a large amount of new assertions to the Japanese ConceptNet.

Other languages have proportionally smaller representation in the database. With over 880,000 assertions German reached the third position, over 420,000 assertions included Italian concepts and almost 390,000 assertions related to Chinese.

Another important quantitative aspect of the ontology is the number and proportions of various types of assertions. Figure 2.4 presents a chart showing the distribution of represented relations and a table listing specific numbers of assertions.



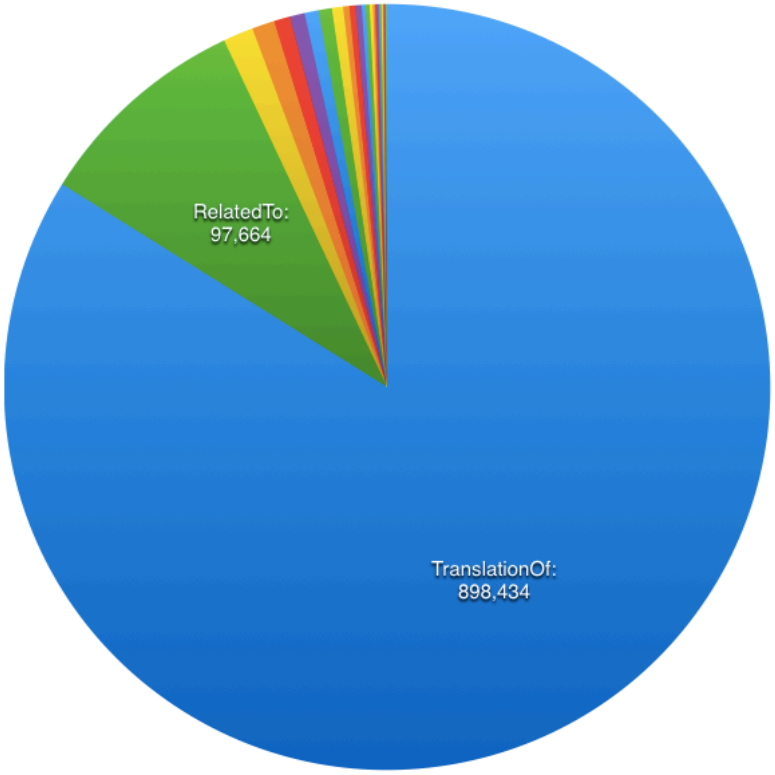
Relation:	Number of assertions:
IsA	5,639,568
RelatedTo	2,957,901
TranslationOf	1,505,848
PartOf	714,583
EtymologicallyDerivedFrom	639,794
AtLocation	432,076
Synonym	314,898
DerivedFrom	284,978
Antonym	96,722
Causes	92,097
UsedFor	81,793
MotivatedByGoal	80,000
HasSubevent	75,164
CapableOf	67,970
Desires	48,716
HasProperty	29,236
HasPrerequisite	27,105
CausesDesire	25,602
CompoundDerivedFrom	21,430
MadeOf	18,466
HasFirstSubevent	16,815
DefinedAs	16,158
MemberOf	12,290
SimilarTo	11,561
ReceivesAction	10,739
HasA	9,863
Others	58,149

Figure 2.4: Relations of ConceptNet 5.3.

The total number of assertions present in ConceptNet 5.3 is 13,289,522. It is clear that the IsA type of relation has the strongest representation in the database. With over 5,600,000 assertions of this kind it makes up for 42.4% of the total number of assertions. With almost 3,000,000 assertions

RelatedTo is the second most common relation present in the database, taking 22.3% of the total. Next is the TranslationOf relation with over 1,500,000 assertions (11.3%) and PartOf with over 714,000 assertions (5.4%).

These proportions will not be very informative unless we confront them with those representing the Japanese portion of the ontology. By detecting any disproportions with the general structure of the database, it will be possible to identify the kinds of relations that need to be populated with additional assertions in the first place. It was therefore necessary to perform a separate analysis of the proportions of relations present in the Japanese section. The results are presented on Figure 2.5.



Relation:	Number of assertions:
TranslationOf:	898,434
RelatedTo:	97,664
CompoundDerivedFrom:	13,655
AtLocation:	10,259
DerivedFrom:	7,225
EtymologicallyDerivedFrom:	6,749
IsA:	6,315
Synonym:	5,985
PartOf:	5,076
MotivatedByGoal:	2,828
HasSubevent:	2,786
CapableOf:	2,679
UsedFor:	1,779
Causes:	1,757
HasProperty:	1,485
Antonym:	971
MadeOf:	765
CausesDesire:	762
NotDesires:	732
InstanceOf:	556
HasPrerequisite:	554
Desires:	469
SymbolOf:	452
HasLastSubevent:	366
HasFirstSubevent:	364
DesireOf:	277
DefinedAs:	102

Figure 2.5: Relations of the Japanese section of ConceptNet 5.3.

The first look on the chart reveals a large disproportion in comparison with the general structure of ConceptNet. The most dominant part of the Japanese section is taken by TranslationOf relations. Almost 900,000 assertions of this type make up for a staggering 83.9% of the total number of Japanese assertions. The reason for the dominance of this type of assertion lies in the introduction of JMdict [62], which provided a source to 873,838 pairs of concepts connected to each other with TranslationOf relation. The second largest relation is the RelatedTo, however with over 97,000 assertions it covers

only 9.1% of the whole. Further relations make up for less than 2% each. This clearly shows that it is necessary to bring the balance between different types of assertions closer to the proportions representing the whole of the ontology. Especially IsA type of assertions is strongly underrepresented. With a little over 6,300 assertions it can not compare with the 42.4% proportion present in the overall distribution. Also many of the relations present in other languages have zero assertions in the Japanese part. The examples of such relations are LocatedNear, CreatedBy and MemberOf. This information would serve as an important lead when designing a strategy for expanding the Japanese section of the ontology.

2.4 Knowledge acquisition methods

In order to choose the path of expanding ConceptNet with Japanese language data in the most effective way considering the existing constraints, it is necessary to review the existing methods of knowledge acquisition. Such methods can be divided into four groups: labor acquisition, interaction acquisition, mining acquisition and reasoning acquisition.

In labor acquisition schema the knowledge is provided directly by trained knowledge engineers or untrained volunteers. Knowledge engineers are usually required to codify the data in a certain formal language to create a resource readable by computers. Volunteers could use some kind of interface to enter new information using natural language. This approach ensures high quality of the acquired data, however it is extremely time and labor intensive. The Cyc Knowledge Base is an example of knowledge representation built this way [17]. To facilitate the process of data entry by domain experts tool such as KRAKEN [20] and a dialogue system based on user interaction agenda [21] were developed. Volunteers can contribute knowledge through interfaces such as Factivore and Predicate Populator [22].

Interaction acquisition represents an approach where human minds are treated as data sources, and the knowledge is acquired in a process of interaction with human contributors. In this setup the interactive property acts as a motive for the contributors to provide the knowledge in an enjoyable and productive way. There are usually two forms of interaction employed: interactive user interfaces and games. Interactive interfaces provide feedback to the users when they enter new knowledge, which encourages them to contribute even more information [65]. When using a game, a tedious

action of entering knowledge is transformed into more enjoyable process of playing the game. The knowledge is accumulated in the background, possibly without the users not really realizing it while playing the game. The advantage of such solution is a considerable reduction of cost in comparison with the labor acquisition approach. The game or an interactive user interface has to be programmed once and then the players contribute knowledge for free. However, in practice the approach reveals its drawbacks. The quality of the acquired knowledge may be low, as often users enter meaningless pieces of information to test how the system would react or when they try to be original and funny. It has also been observed that systems of that kind tend to have a limit to gathering new data: users tend to enter similar answers to general questions, and the amount of original answers is limited [66]. Another problem is retention - it is challenging to keep the users interacting with an interface or a game for a long period of time. The design has to be entertaining enough to keep people engaged and willing to produce large amounts of data. Examples of systems providing data to knowledge bases through means of interaction include OMSC [55], Verbocity [67], nadya.jp [58], Common Consensus [68], 20 Questions [69] and Virtual Pet game [59]. An interesting recent development in the area of interactive approach towards Japanese language common sense knowledge acquisition has been presented by Otani *et al.* [70]. The system acquires knowledge with the help of a quiz game introduced as a module of a popular smartphone spoken dialogue system Yahoo! Voice Assist [71]. The approach is similar to the word game used by nadya.jp, however the implementation of the game into already popular dialogue system resulted in the acquisition of a large amount of Japanese common sense knowledge. The gathered information is intended to be incorporated into ConceptNet.

Using reasoning acquisition approach, potential knowledge can be inferred automatically from already possessed knowledge. Such reasoning techniques include analogical reasoning, inductive reasoning and plausible reasoning. Analogical reasoning helps to discover facts or properties relating to concepts on the basis of information held about similar concepts. Inductive or plausible reasoning may be exploited to produce rules from basic facts that would be true in certain contexts. This approach may be very useful in overcoming the explicitization problem: human beings tend to ignore the existence of commonsense knowledge. As it is so obvious to them, they assume that everyone around them already possesses it and so there is no point in explicitly stating the obvious. Reasoning acquisition may help to solve this problem by assisting in creating a formal description of those im-

plicit facts. However, in order for such methods to work, they require a large amount of previously formalized knowledge to operate with. Examples of such systems include FOIL, an inductive logic programming system applied to Cyc [72], Learner [73] and AnalogySpace [74].

Finally, knowledge may be acquired automatically from a text corpus using mining acquisition approach. To execute such operation it is necessary to employ natural language processing techniques. Once a researcher develops a mining system, the system can process either domain-specific corpora (such as books or newspapers) or open-domain corpora (such as the Web). Web-oriented systems, due to the larger scale of the available input, can gather much more knowledge than the ones analyzing domain-specific input. Using an open-domain source would also help to solve the diversity problem: as common sense knowledge is domain and type independent, it is important to retrieve it from a wide spectrum of domains. Designing a system able of analyzing various Web pages can be very challenging due to the heterogeneous nature of the Web. An alternative approach would be to create a method for analyzing a large open-domain corpus that has a formalized structure. Applying such system would be very beneficial in terms of efficiency, that is the speed of gathering knowledge, and automation, that is the reduction of manual labor necessary for acquiring data. Examples of projects following the mining acquisition approach include Cyc, to be more precise its subsystem for querying the Web using Google search engine [23], KnowItAll [33], TextRunner [75], R2A2 [76], or knowledge extraction system from Chinese online encyclopedias [77].

2.5 Conclusion

The analysis of the structure of ConceptNet 5.3, as well as the review of existing approaches towards knowledge acquisition lead to the following conclusions. Considering the current state of the Japanese part of the database and the challenges the various ways of expansion must face, it would be most effective to apply the mining acquisition approach on a structured, domain-independent corpus. This would not only solve the problem of efficiency-automation balance, which is crucial considering the limitations of the available human and time resources. Such choice of the mining target would help to overcome the diversity problem appearing while processing sources referring to a particular domain. An obvious candidate for such source would be Wikipedia. As it is created by general public, and not by a finite group of specialists, it is safe to assume that it represents a pool of collective knowledge

shared by the whole population of Japanese language speakers. The domain-independency condition is met as well, as Wikipedia articles cover a virtually limitless variety of topics. The above conclusions became the guidelines for the development of a method of expanding the Japanese section of ConceptNet.

The reader may wonder how the knowledge about various entities contained within Wikipedia articles could help to extract common sense knowledge. As I demonstrate in the latter part of the thesis, explicitly expressed facts may lead us to more general conclusions. By looking at a large number of tangible examples it is possible to learn some unwritten truths, which are obvious to humans, but still unknown to computers.

Chapter 3

Data mining on reference corpus

3.1 Introduction

This chapter presents the proposed method for automatic acquisition of assertions suitable for the introduction to the Japanese section of the ConceptNet. Section 3.2 describes the approach of using hyponymy relations as the ontology's IsA assertions and a method of acquiring them. Section 3.3 shows the process of gathering other types of assertions by first presenting the underlying method for generating information-rich taxonomy, and then referring to the methodology of mining AtLocation assertions in Subsection 3.3.1, LocatedNear assertions in Subsection 3.3.2, CreatedBy assertion in Subsection 3.3.3 and MemberOf assertions in Subsection 3.3.4. Section 3.4 presents a method for gathering general assertions generated on the basis of the previously extracted data. Finally, Section 3.5 describes the solution to performance issues met during the implementation of the proposed method.

3.2 Hyponymy relation as IsA relation

As it has been shown in Section 2.3.5, IsA relation is represented by the highest number of assertions in ConceptNet. However, only around 6,400 assertions of that type could be found in the Japanese section of the ontology. It became clear that populating this relation should be the first goal in the path of increasing the number of Japanese language entries.

IsA relation of ConceptNet 5.3 covers both hyponymy relation, as well as concept-instance relation as defined by linguistic literature [78]. For example, assertions built on the basis of sentences "National university is a kind of university" and "Tokyo University is an instance of university" would

be incorporated into this relation. Therefore including hyponymy pairs into the base would be a valid task. There is a number of methods that can be used to automatically acquire sets of two lexical terms connected by a hyponymy relation [31] [79] [80] [81] [82] [83] [84]. However, a method that would best serve the purpose of acquiring assertions valid for introduction to the Japanese part of ontology was created by Sumida *et al.* [85].

The key idea of the method is to focus on the analysis of a consistently structured reference corpus. This would allow for a much more effective and cost-efficient extraction of hyponymy pairs in comparison with the attempt to perform such operation on unstructured documents. In this case Wikipedia is considered as such structured corpus. Wikipedia is created with the use of MediaWiki software package [86], which interprets the source code written in the Wiki markup syntax to generate web pages accessible by the public. What is important, the Wiki markup syntax is stricter than the HTML syntax, and its use is regulated by the editorial policy. As a result Wikipedia, having a consistent structure of its articles, becomes a much more suitable target of information extraction in comparison with generic HTML documents (see Figure 3.1 for an example of Japanese Wikipedia page and its source code which is being analyzed by the method; figure 3.2 presents English language equivalent of the page) The elements of Wikipedia articles targeted by the analysis include headings, bulleted lists, ordered lists and definition lists. These elements, when marked with a given number of special symbols, such as "=" or "*", create a hierarchical structure. The method's goal is to extract hyponymy relations from such structures.

An improved version of the method [87] retrieves hyponymy relations in two steps: first it extracts the hyponymy relation candidates from hierarchical layouts, and then selects valid hyponymy relations using a classifier based on Support Vector Machines [88]. The candidates are obtained by considering the title of each marked-up item as a hypernym candidate, and titles of all subordinate marked-up items as its hyponym candidates. The SVM-based classifier uses the following features to select proper hyponymy relations from the candidates:

- Part-of-speech tag - a unique dimension in the feature space is assigned for each POS tag; when the candidate hypernym/hyponym consists of a morpheme with a given POS tag, then the value of corresponding element of the feature vector is set to 1; in case when the candidate consists of more than one morpheme, then the feature vectors of all morphemes are summed; the POS

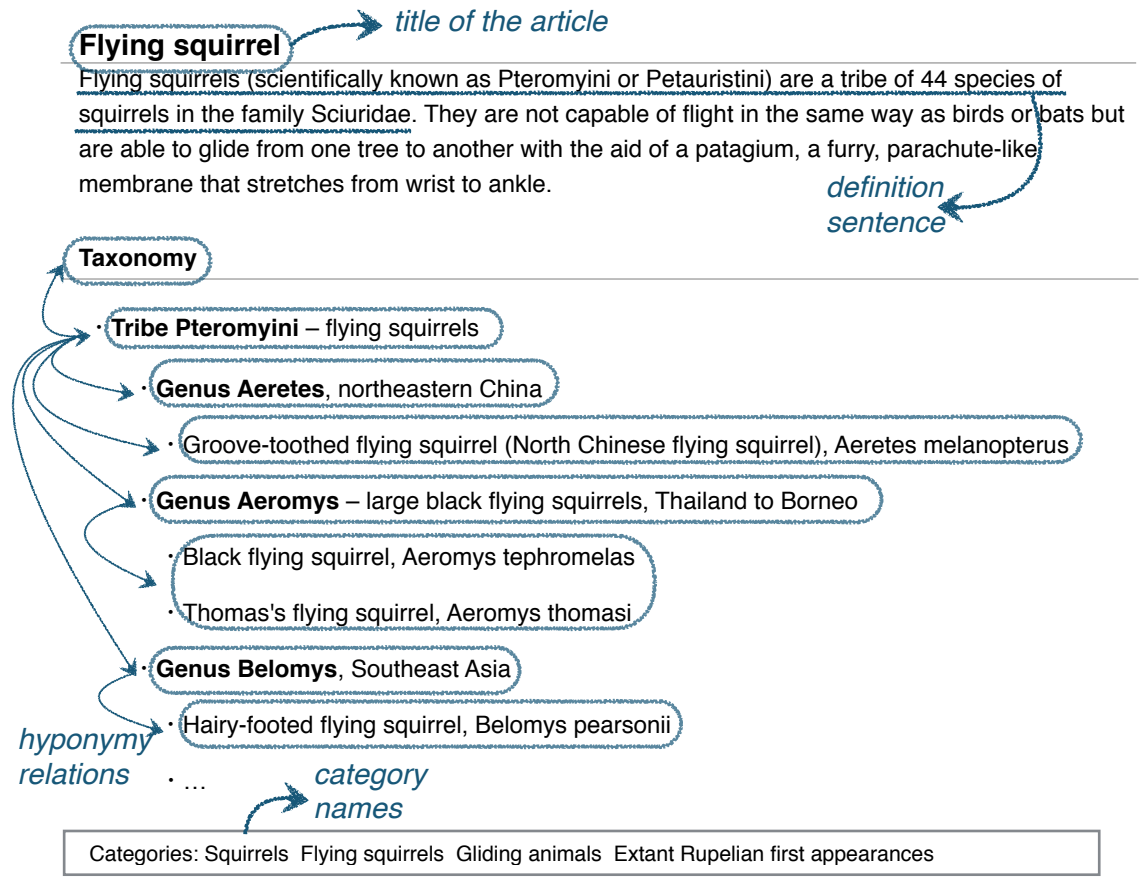
```
'''モモンガ''' （摸摸具和）は、[[ネズミ目]]（齧歯目）[[リス科]][[リス亜科]][[モモンガ族]]に属する小型[[哺乳類]]の総称。滑空によって飛翔する性質を持つ[[リス]]の仲間。また、狭義には特に、'''[[ニホンモモンガ]]''' ({{SnameIlPteromys momonga}}) を指す。漢族語では鼯鼠とよぶ。

== 分類 ==

*[[モモンガ族]] {{SnameIlPteromyini}} - 15属、約45種
**ミゾバムササビ属 {{SnameIlAeretes}}
***ミゾバムササビ {{SnameIlAeretes melanopterus}}
**クロムササビ属 {{SnameIlAeromys}}
***クロムササビ {{SnameIlAeromys tephromelas}}
***トマスクロムササビ {{SnameIlAeromys thomasi}}
**ケアシモモンガ属 {{SnameIlBelomys}}
***ケアシモモンガ {{SnameIlBelomys pearsoni}}

[[Category:リス科|ももんか]]
```

Figure 3.1: Example of Japanese Wikipedia page with marked elements for extraction (upper part) and its source code (lower part).




```
'''Flying squirrels''' (scientifically known as '''Pteromyini''' or '''Petauristini''') are a [[tribe (biology)|tribe]] of 44 [[species]]
of [[squirrel]]s in the [[family (biology)|family]] [[Squirrel|Sciuridae]]. They are not capable of flight in the same way as
[[bird]]s or [[bat]]s but are able to glide from one tree to another with the aid of a [[patagium]], a furry, parachute-like
membrane that stretches from wrist to ankle.

==Taxonomy==

'''Tribe Pteromyini''' – flying squirrels
**Genus ''[[Aeretes]]'', northeastern [[China]]
***[[Groove-toothed flying squirrel|Groove-toothed flying squirrel (North Chinese flying squirrel)]], ''Aeretes
melanopterus''
**Genus ''[[Aeromys]]'' – large black flying squirrels, [[Thailand]] to [[Borneo]]
***[[Black flying squirrel]], ''Aeromys tephromelas''
***[[Thomas's flying squirrel]], ''Aeromys thomasi''
**Genus ''[[Belomys]]'', Southeast Asia
***[[Hairy-footed flying squirrel]], ''Belomys pearsonii''

[[Category:Squirrels]]
[[Category:Flying squirrels| ]]
[[Category:Gliding animals]]
[[Category:Extant Rupelian first appearances]]
```

Figure 3.2: Example of English Wikipedia page with marked elements for extraction (upper part) and its source code (lower part).

tag of the last morpheme is mapped to a different dimension;

- Morpheme - the candidate's morphemes are mapped to the dimensions of feature vectors, as some morphemes, such as "genus", appearing at the end of a candidate hypernym increase the probability of the relation being valid; the last morpheme is mapped to a separate dimension;

- Expression - mapping each expression of the candidates to an element in a feature vector may help to detect strings that may appear among the marked-up item's title, but which can not form a correct hypernym or hyponym, such as "Background" or "Note";

- Attribute - an observation has been made that if a relation candidate includes an attribute, for example "Anatomy" as attribute of different creatures, such relation is invalid as it can not create a proper hyponymy; the authors of the method generated a set of 40,733 attributes to be mapped to the dimensions of feature space;

- Layer - an observation has been made that if a hyponymy relation is extracted from the bottom of the hierarchy structure, such relation has a higher probability of being valid; to include this fact during the classification, each type of marking items (that is headings, bulleted lists, ordered lists or definition lists) from which the candidates are extracted is mapped to an element of a feature vector;

- Distance - it has been observed that when a distance d between candidates for hypernym and hyponym on a hierarchy structure equals 1, the probability of the relation being valid increases; to note this phenomenon during the classification, the distance d is mapped to two elements of the feature vector: when it equals 1, and separately when it is above that value;

- Pattern - this feature is based on an observation made during the preparation of the first version of the method [85] which suggested that when the candidate hypernym is obtained from a hypernym that can be matched to one of the patterns shown in Figure 3.3, then it is more likely to be correct;

- Last character - if the last character of both hypernym and hyponym candidates is the same, then such relation is more likely to be correct, as the last characters are likely to convey major

X の一覧 (list of X), X 一覧 (list of X), X 詳細 (details of X), X リスト (X list), 代表的な X(typical X), 代表 X (typical X), 主要な X (popular or typical X), 主な X (popular or typical X), 主要 X (popular or typical X), 基本的な X (basic X), 基本 X (basic X), 著名な X (notable X), 大きな X (large X), 他の X (other X), 一部 X (partial list of X), *X の詳細 (details of X), *代表的 X (typical X), *基本的 X (basic X), *著名 X (notable X), *一部の X (partial list of X)

Figure 3.3: Patterns for finding plausible hypernym X (from [87]).

semantic content of Japanese compound nouns; to pass this information to the classifier, this feature is set to one if the repetition of the last character occurs;

By applying all of the above mentioned features to a SVM-based classifier, the authors of the method were able to achieve 89.7% accuracy of the extracted hyponymy pairs (refer to the original paper for the evaluation methodology and additional measurements) [87]. Table 4.2 presents examples of the acquired hyponymy pairs.

Table 3.1: Examples of extracted hyponymy pairs presented by the method's authors [87].

Hypernym	Hyponym
kōen (park)	*Motomiya-kōen* (Motomiya Park)
kōkyō-shisetsu (public institution)	*rōjin-fukushi-sentā* (welfare center for the elderly)
kōgu (tool)	*baisu* (vice)
saiji (festival)	*unagi-matsuri* (eel festival)
wakusei (planet)	*Tennōsei* (Uranus)
mizuumi (lake)	*Tanzawa-ko* (Lake Tanzawa)
kōkū kaisha (airline company)	*Tai Kokusai Kōkū* (Thai Airways International)
chōkokuka (sculptor)	*Isamu Noguchi*

Two additional methods of extracting hyponymy pairs from Wikipedia have been proposed as well - extraction from definition sentences, and extraction from category pages. Wikipedia definition sentences were previously used to mine hyponymy relations for named entity recognition [89]. This method has been developed for English language Wikipedia content analysis and adapting it to the use with Japanese language articles required some modifications inspired by Tsurumaru *et al.* [90]. The method exploits sentential patterns appearing in dictionary definitions. As Wikipedia contains such sentence at the beginning of each article, it is easy to recognize and analyze them in search of hyponymy pairs. The authors of the method manually prepared 1,334 patterns to facilitate the process. The method of extracting hyponymy relations from category pages has been inspired by the research conducted by Suchanek *et al.* [31]. In that case the process could be augmented by using WordNet information [32] to achieve better results. However, in case of Japanese language hyponymy extraction such support was not used although Japanese version of WordNet has been developed [91]. The method therefore is very simple and uses only Wikipedia resources: it regards a category name given on the top of a category page as a hypernym and every position listed on the page as its hyponyms. The authors of the method admit that this approach introduces a considerable amount of noise to the output. The precision of extraction of hyponymy relations from definition sentences and from category pages have been assessed at the level of 77.5% and 70.5% respectively. Considering the low number of IsA assertions already present in the ConceptNet 5.3, a decision has been made to apply the described methods to a more recent version of Wikipedia and conduct experiments aiming at assessing the quality of the data acquired by all three methods. If the accuracy of the methods would achieve satisfactory results then the output could be introduced to the ConceptNet's data pool. Although Yamada *et al.* suggests that these hyponymy pairs are not informative enough to be useful for NLP tasks such as question answering [92], however they do fall into the scope of ConceptNet's domain of common sense and general knowledge. They are simple enough not to interfere with the ConceptNet's usage flexibility, yet informative enough to introduce new and valuable input to the knowledge base.

To implement and test the described methods I used the Hyponymy extraction tool v1.0 [93], an open-source program which takes Wikipedia's XML dump files as input. It consists of four modules, three of which deal with extraction of hyponymy pairs from different parts of Wikipedia content:

hierarchy structures, definition and category, as described earlier. The program utilizes the Pecco library [94] (SVM-like machine learning tool) to perform the classification of hyponymy pairs candidates. The results of the implementation of the tool are described in Chapter 4.

3.3 Extracting other relations

The fourth module of the Hyponymy extraction tool v1.0 executes a method for information-rich taxonomy creation developed by Yamada *et al.* [92], which utilizes the results of the previously described methods [85] [87]. The underlying idea behind the development of the method is as follows. Hyponymy relations extracted from the hierarchical structures of Wikipedia articles, and especially hypernyms of such pairs, may be considered as too abstract or vague to be directly used for different applications, such as question answering. An example of such pair would be a hypernym "*sakuhin*" (work) and a hyponym "*Abatā*" (Avatar). To make such pair more useful, the hypernym would have to be enriched with some additional information to make it more specific. This information may be provided by the title of the Wikipedia article from which such hyponymy pair was extracted. Using the previous example, if the hyponymy pair was extracted from a hierarchy structure included in a Wikipedia article entitled "*Jēmuzu Kyameron*" (James Cameron), then adding the title to the hypernym with a proper postposition would make it more specific.

Figure 3.4 presents the procedure of the method using the previous example. First the hyponymy relations are extracted from the hierarchical structures of a given Wikipedia article. These relations are treated as basic hyponymy relations by the method (left part of the diagram). Next, each hypernym is augmented with a modifier, which is the title of the Wikipedia article from which the hyponymy relation was extracted. The modifier is added to the hypernym with the Japanese particle "*no*", a genitive case marker expressing English words such as [of], [by], [in] or ['s]. Such augmented, intermediate hyponymy relation is later referred to as T-INTER (central part of the diagram; the "T" in the name of the relation stands for "title" and "INTER" stands for "intermediate"). Finally, yet another intermediate concepts pair is introduced to the taxonomy. It is created by generalizing the hypernym of T-INTER, and to be more specific, by generalizing the title of the Wikipedia article from which the basic hyponymy relation was extracted. To do this the method requires a hypernym of the title. Such hypernym is generated from the definition sentence or category name following the previous

methods [85] [87]. The title is replaced with its hypernym, and the newly createated generalized intermediate concept pair is referred to as G-INTER (see the right part of the diagram; the "G" in the name of the relation stands for "general" and "INTER" stands for "intermediate").

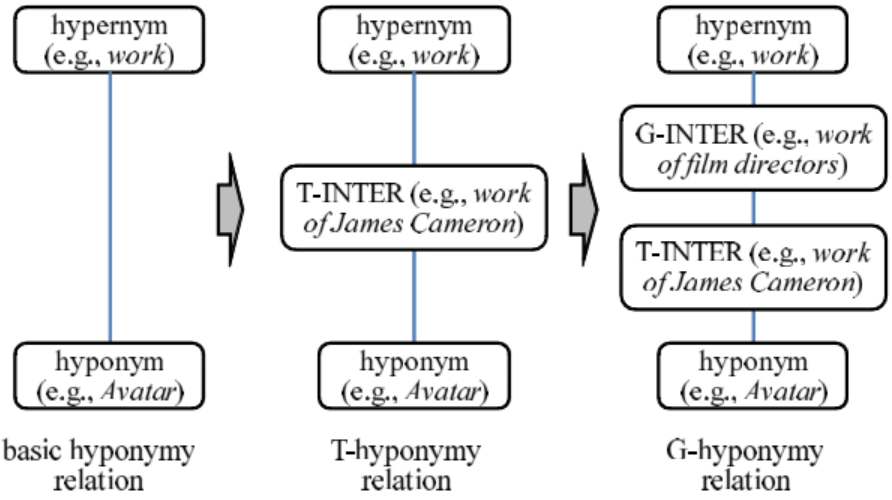


Figure 3.4: Procedure of Yamada *et al.* method (from [92]).

Examples of augmented hyponymy relations are shown in Table 3.2. As we can see the generated augmented hypernyms are too specific to be incorporated into ConceptNet directly. However, a closer inspection of the data reveals that some additional information about their corresponding hyponyms may be extracted from them, such as information concerning location, neighboring locations, creator and so on. Knowledge about location and creator may be directly transferred into ConceptNet as already built-in AtLocation, LocatedNear and CreatedBy relations. It should be noted that according to the ConceptNet documentation [95] the CreatedBy relation relates to processes, however inspection of the existing CreatedBy assertions show that they include creations and their authors as well. The remaining part of the acquired information related to the hyponyms may be represented by a more general RelatedTo relation.

The proposed procedure of acquiring additional information slightly differs between the versions responsible for extracting particular relation, but its general structure is presented in Figure 3.5 First it scans the G-INTER using a handcrafted primary rule set in search of tags referring to items fitting to one of the designated categories: location, neighboring location, creator or member. Next it filters

Table 3.2: Examples of augmented hyponymy relations generated by Yamada *et al.* [92] method.

Original Hypernym	G-INTER	T-INTER	Hyponym
tōjō-jinbutsu (character)	*SF eiga no tōjō-jinbutsu* (character of SF movie)	*WALL-E no tōjō-jinbutsu* (character of WALL-E)	M.O
seihin (product)	*kigyō no seihin* (product of a company)	*Silicon Graphics no seihin* (product of Silicon Graphics, Inc.)	IRIS Crimson
sakuhin (work)	*America no shōsestu-ka no sakuhin* (work of American novelist)	*J.D. Salinger no sakuhin* (work of J.D. Salinger)	A boy in France
machi (town)	*England no shu no machi* (town in a county in England)	*East Sussex no machi* (town in East Sussex)	Uckfield
kantoku (director)	*musical eiga no kantoku* director of a musical)	*Ame ni Utaeba no kantoku* (director of Singin' in the Rain)	Stanley Donen
ibento (event)	*Hōsō-kyoku no ibento* (event of a broadcasting station)	*Fuji Television no ibento* (event of Fuji Television Co., Ltd.)	*Odaiba dotto komu* (Odaiba dot com)

the basic hypernym through a secondary rule set to exclude items that would introduce noise. If the basic hypernym is positively assessed by the secondary rule set, the procedure assumes that the phrase acquired by deleting the descriptor from the G-INTER is a valid tag corresponding to a particular relation. In the next stage the method compares the validated tag with the content of the T-INTER to extract a concept suitable for becoming an element of a new assertion. Finally, the newly acquired concept is joined to the base hyponym with a proper relationship tag to extract a new relation. In post-processing phase the procedure removes duplicates of assertions to output a list of unique assertions of a given type following a format "A relation B", where A and B refer to the extracted concepts.

The effectiveness of the method mainly depends on the number and nature of introduced rules to both the primary and secondary rule sets. The method at this stage uses 55 primary rules and 14 secondary rules, which led to extraction of assertions concerning location, neighboring locations, creators and members. I have created a manually crafted set of rules written in Python regular expressions format, using heuristics after analysis of the input data. The reason why I chose this kind

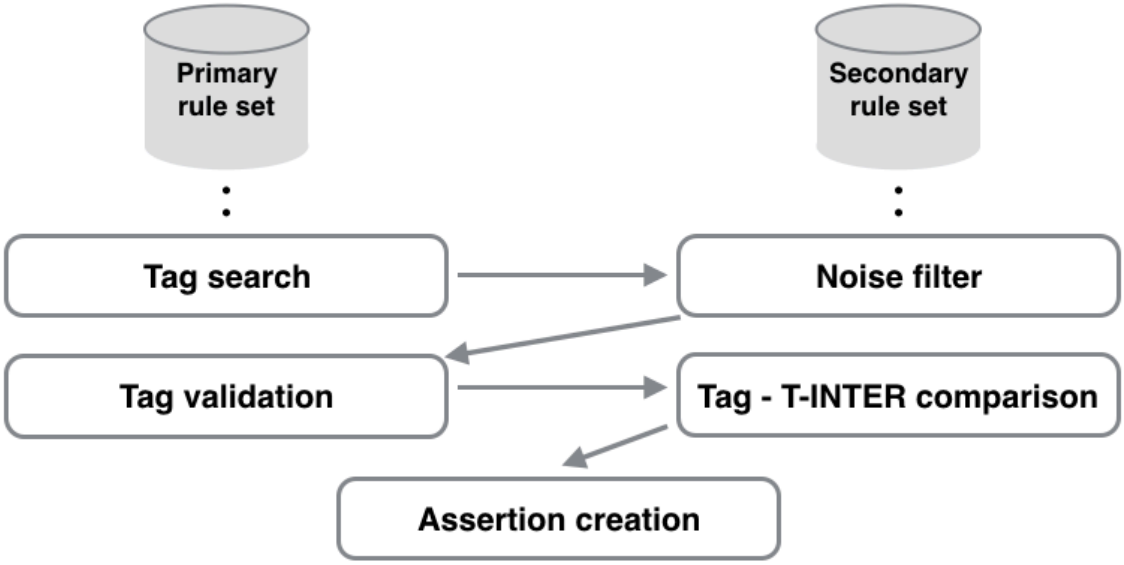


Figure 3.5: Flowchart of the proposed method.

of approach is the fact that the information units contain Chinese or Japanese characters or strings of characters indicating a type of location, a city, province, school, creator or member. I use the rules to detect these characters or strings, and this way the method is able to obtain the named entities referring to locations, creators, members and other concepts. Due to the qualities of the Japanese language's writing system these rules are often very simple, containing a single character, but are still effective for detecting the language units suitable for extraction. For example, the secondary rules used for detecting people include the suffix "~sha", which describes different professions. For English such a shortcut would be harder to apply, and therefore person detection would require a much larger rule set covering a long list of names of professions and appropriate suffixes (like "~er", "~or" or "~ist"). The following subsections describe in detail the procedure of acquiring each of the four kinds of instance-related assertions.

3.3.1 AtLocation assertions acquisition

In order to acquire a large number of assertions representing AtLocation relation the procedure uses a primary rule set consisting of 21 items. The secondary rule set, consisting of 12 items, ensures the high level of reliability of the output data. I utilize primary rule set to scan through the G-INTER element of the list in search of suitable candidates for extraction. The rules present in the primary set

of the AtLocation extraction module are as follows:

1. (.*)市町村\](.*)
2. (.*)市町村,.*\](.*)
3. (.*)特別区\](.*)
4. (.*)特別区,.*\](.*)
5. (.*)町\](.*)
6. (.*)県\](.*)
7. (.*)県,.*\](.*)
8. (.*)\[郡,.*\](.*)
9. (.*)の都市\](.*)
10. (.*)の都市,.*\](.*)
11. (.*)\[都市,.*\](.*)
12. (.*)所在地\](.*)
13. (.*)\[共和国.*\](.*)
14. (.*)国\](.*)
15. (.*)国,.*\](.*)
16. (.*)国家\](.*)
17. (.*)国家,.*\](.*)
18. (.*)省都,.*\](.*)
19. (.*)の省都\](.*)
20. (.*)のタウン\](.*)

21. (.*)のタウン,.*\](.*)

Rules number 1-2 refer to municipalities, 3-4 to special districts, 5, 20 and 21 to towns, 6-7 to prefectures, 8-11 to cities, 12 to locations, 13-17 to countries, and finally 18-19 to metropolitan areas. Most of the rules are created in two forms, taking into consideration the fact that the G-INTER may include multiple tags. The rules utilize the fact that the tags are confined by square brackets within the G-INTER. For example, G-INTER tags describing *Hiji* town include “town”, “castle town”, and “municipality in *Oita* prefecture”. Therefore it is in most cases necessary to create rules that will detect tags which are followed both by a coma and by a square bracket at the end of tags list.

The secondary rule set of the AtLocation relation extraction module consists of the following items:

1. (.*)人
2. (.*)者
3. (.*)人物
4. (.*)性
5. (.*)キャスト
6. (.*)タレント
7. (.*)アーティスト
8. (.*)テナント
9. (.*)パーソナリティ
10. (.*)優
11. (.*)姉妹都市
12. 隣接(.*)

Rules number 1-9 refer to people and other names indicating human beings, such as “cast”, “talent”, “artist”, “tenant” and “personality”. The reason why I introduced these rules is the fact that it is not possible to build an AtLocation kind of relationship between a place and a person, as people, even though born in a specific location, are most often not limited to existing in that particular location. Rules 10-12 help to filter out other noise inducing items such as sister cities or adjacent locations. The rules were prepared manually after the extensive analysis of the input, that is the augmented hyponymy relations list [92].

Figure 3.6 presents the procedure of extracting AtLocation relations. The script goes through every line of the input list and divides it to four parts, that is the original hypernym, G-INTER, T-INTER and hyponym. When any of the elements enumerated in the primary rule set is found in the G-INTER (Step 1), then the procedure utilizes the secondary rule set to scan through the original hypernym. Next the script checks that the original hypernym does not include the elements from the secondary rule set (Step 2). If such condition is met and all noise inducing items have been filtered out, then it verifies that the content of G-INTER confined by square brackets is a valid location tag (Step 3). As shown in the example from Figure 3.6, it checks that “county in England” is a valid tag to describe a location. Further step of the method is to read the descriptor part of the G-INTER, that is the part outside the square brackets, and to remove it from the G-INTER, leaving only the concept needed for data extraction (Step 4). This way, as shown in the previous example, it can extract the knowledge that the county we can refer to in that particular case is East Sussex. Finally the extracted concept is being connected to the hyponym with AtLocation relation to create a new assertion, for example Uckfield-AtLocation-East Sussex.

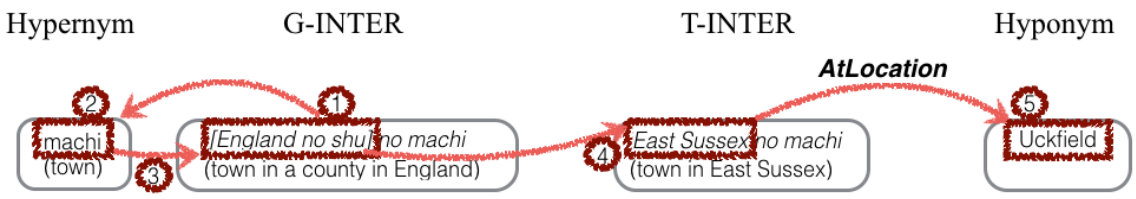


Figure 3.6: Procedure of AtLocation relation extraction module.

3.3.2 LocatedNear assertions acquisition

LocatedNear assertions acquisition module is a good example of applying very simple rules which exploit the qualities of Japanese writing system to gather high-accuracy data. Both primary and secondary rule sets of this module consist of only one rule each. The primary rule is:

1. 隣(.*)

The character in this rule, either standing on its own or as a part of a characters compound, indicates a state of being close, next door or neighboring. By confirming whether the searched string has this character at its beginning, the system is able to detect a wide variety of expressions indicating close physical proximity. The secondary rule set is also minimalistic and consists of one item:

1. 郡

The character in this rule signifies counties. I have introduced it to the secondary rule set after preliminary experiments conducted on the data acquired by applying the primary rule. One of the annotators evaluating the output indicated that assertions with American counties as one or both of the concepts are ambiguous, as there are counties with the same name in different states. In order to prevent the ambiguous items from lowering the quality of the output data I have made a decision to create a secondary rule detecting counties to filter them out.

Figure 3.7 presents the procedure of extracting LocatedNear relations. In this module the procedure is in most parts analogous with the path of AtLocation extraction, however there is one major difference: the primary and secondary rule sets are both applied to the basic hypernym. First (Step 1) the script checks whether the hypernym includes the character indicating physical proximity and then (Step 2) confirms that it does not refer to counties. Further steps correspond with the ones described in the previous section.

3.3.3 CreatedBy assertions acquisition

CreatedBy assertions acquisition module represents an alternative approach: minimalistic rules can not be applied here if the script is to generate high accuracy data. Preliminary experiments revealed

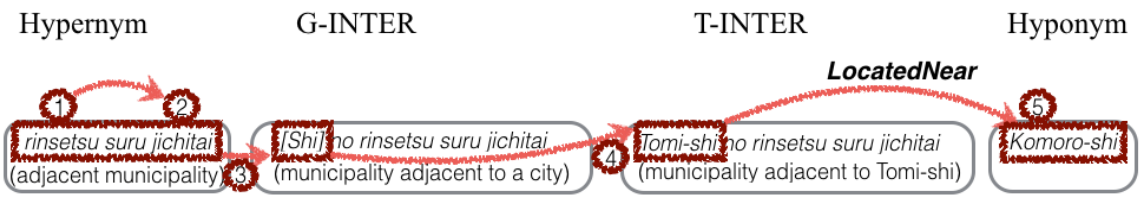


Figure 3.7: Procedure of LocatedNear relation extraction module.

that extracting creator information is more complex and creates some challenges. While extracting location-related information, the introduced rules may be simple and straightforward. In the case of creators, the rules not only have to cover the qualities of the writing system, but also take into consideration the importance of particular roles while creating a given piece of work. For example, human annotators indicated that a number of professionals taking part in the creation of films may not be considered as the creators of these films. Actors, actresses and voice actors, even if they make a great contribution to the work, should not be labeled as its creators. Further experiments showed that similarly animators, animation directors, sound directors, and storyboard creators do not qualify to be included in the common sense CreatedBy assertions. After taking all the above mentioned aspects into consideration, the preliminary rule set consists of the following items:

1. (.*)漫画家\](.*)
2. (.*)漫画家,.*\](.*)
3. (.*)イラストレーター\](.*)
4. (.*)イラストレーター,.*\](.*)
5. (.*)作詞家\](.*)
6. (.*)作詞家,.*\](.*)
7. (.*)作曲者\](.*)
8. (.*)作曲者,.*\](.*)
9. (.*)監督\](.*)
10. (.*)監督,.*\](.*)

11. (.*)脚本家\](.*)

12. (.*)脚本家,.*\](.*)

13. (.*)小説家\](.*)

14. (.*)小説家,.*\](.*)

15. (.*)演出家\](.*)

16. (.*)演出家,.*\](.*)

17. (.*)デザイナー\](.*)

18. (.*)デザイナー,.*\](.*)

19. (.*)画家\](.*)

20. (.*)画家,.*\](.*)

21. (.*)クリエイター\](.*)

22. (.*)クリエイター,.*\](.*)

23. (.*)作家\](.*)

24. (.*)作家,.*\](.*)

25. (.*)写真家\](.*)

26. (.*)写真家,.*\](.*)

27. (.*)アーティスト\](.*)

28. (.*)アーティスト,.*\](.*)

29. (.*)音楽家\](.*)

30. (.*)音楽家,.*\](.*)

31. (.*)ミュージシャン\](.*)

32. (.*)ミュージシャン,.*\](.*)

Rules number 1-2 refer to manga artists, 3-4 to illustrators, 5-6 to lyricists, 7-8 to composers, 9-10 to directors, 11-12 to scriptwriters, 13-14 to novelists, 15-16 to stage directors, 17-18 to designers, 19-20 to painters, 21-22 to creators, 23-24 to writers, 25-26 to photographers, 27-28 to artists, and 29-32 to musicians. Again the rules are prepared in two forms to cover cases of the searched tag being in the middle or at the end of the tags list confined by square brackets.

The secondary rule set of the CreatedBy relation extraction module consists of only one following item:

1. (.*)声優

The role of this element is to filter out entries referring to voice actors. As the human annotators indicated during the preliminary experiments, this role excludes people from being considered as creators of a particular film or animation.

Figure 3.8 presents the procedure of extracting CreatedBy relations. As it was the case with LocatedNear relation extraction module, I apply both primary and secondary rule sets to investigate the same element of the enriched taxonomy, but in this case the procedure scans through the tags enclosed by square brackets in the G-INTER. After confirming that the G-INTER tag or tags include one of the elements listed in the primary rule set (Step 1), the scripts filters out noise inducing items (Step 2), in this case elements that have a voice actor as a tag in G-INTER. Further steps correspond with the ones described in section 3.3.1.

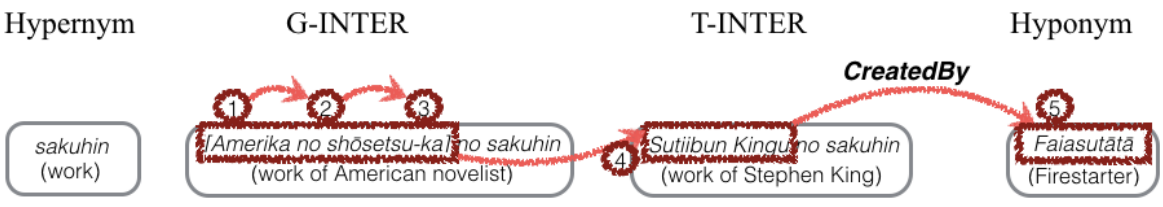


Figure 3.8: Procedure of CreatedBy relation extraction module.

3.3.4 MemberOf assertions acquisition

The last module of the proposed method is responsible for extracting assertions consisting of two concepts connected by a MemberOf relation. This module uses the elements of the primary rule set selectively in order to achieve higher coverage. It also has a built-in procedure for normalizing phrases referring to members: it checks whether phrases present in the information-rich taxonomy contain "current member" expression and replaces it with "member" for further processing. The rules present in the primary set of the MemberOf extraction module are as follows:

1. (.*)のメンバー
2. .*,(.*)のメンバー,.*\](.*)
3. .*,(.*)のメンバー\](.*)
4. \[(.*)のメンバー,.*\](.*)

All of the above rules refer to expressions denoting members. The form of the rules is determined by the element they will target, as well as the position of the element in case of tags confined within square brackets.

The secondary rule set of this module consists of only two elements:

1. .*の過去のメンバー.*
2. .*の以前のメンバー.*

Both rules serve to filter out items that refer to past members, as it has been pointed out by human annotators that people who are no longer members of particular groups or organizations should not be included in the MemberOf assertions pool.

The procedure of MemberOf assertions extraction module has two phases. Phase one is presented in Figure 3.9 and executes the following program: first the rule number one from the primary rule set is used to scan the T-INTER of the taxonomy (Step 1). If the searched phrase is found the script assumes that the part of the T-INTER that precedes the searched phrase is a valid organization or group tag (Step 2), and then connects it with the hyponym to form a new assertion (Step 3).

Figure 3.9: Procedure of MemberOf relation extraction module, phase one.

The analysis of the input data revealed that in many cases the G-INTER's tags contain membership information referring to the object named in the T-INTER. Phase two of MemberOf assertion extraction method, presented on Figure 3.10, was designed to exploit this fact. The script analyzes the G-INTER's tags list one by one with primary set's rules 2-4 to search for phrases indicating membership (Step 1). If such phrase is found, then the secondary rule set is is used to filter out noise inducing items (Step 2). If the analyzed phrase passes the check, then the program assumes that the phrase preceding the one expressed by the primary rule in the G-INTER it is a valid group or organization name (Step 3). Next the script deletes the G-INTER tag descriptor from the T-INTER to retrieve a member (Step 4) and finally links is established between the member and the organization or group (Step 5).

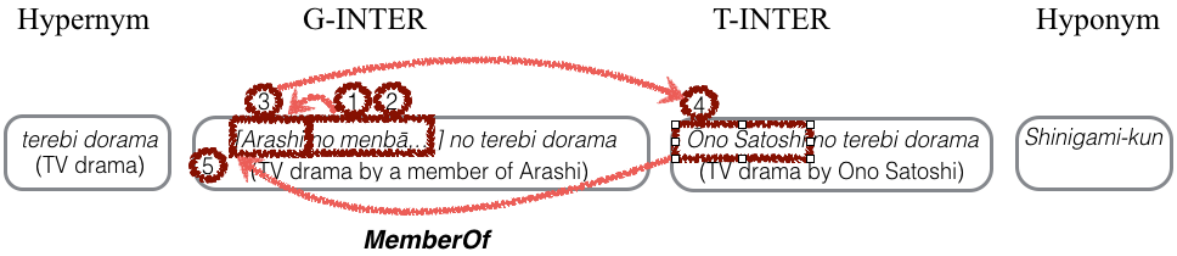


Figure 3.10: Procedure of MemberOf relation extraction module, phase two.

3.4 Generalizing over assertions

Wikipedia contains a lot of information related to instances of certain concepts, such as Salvador Dali as an instance of an artist. Filling up ConceptNet with instances is a valid task, as it is very hard to establish the boundaries of common sense knowledge – facts that are obvious to one group of people overlap to a large proportion with the knowledge of another group, but there is always a discrepancy. This issue raises a question: would it be possible to come to more general conclusions on the basis of

the numerous instances?

This approach would be consistent with analogical reasoning [96], a process which is inseparable with human cognition. Analogies are useful for explaining new concepts, for example very abstract ideas, such as “electricity”, can be depicted with more concrete, tangible examples, such as “water flow”. Analogy is also useful for communication and persuasion. To explain very complex global environmental phenomena one can use a smaller scale example to depict an impact of human intervention on the natural balance, for example comparing earth to Easter Island. By analyzing the effects of overpopulation and exploitation of the island’s ecology which in consequence led to a rapid loss of species, famine and collapse of social structures, it is possible to convey a much more convincing picture about a global state of the environment. In this case however, gathering data about instance related concepts would support making predictions within a given domain: if a large number of representatives of a certain group hold the same quality, or has a relation with another general concept, then it is safe to assume that a given representative of this group also holds such quality. Gathering instance-related data, perceived from a point of view of example-based learning theory [97], can be treated as a preparatory phase leading to learning from worked examples. In such scenario learners study problems with solutions already given (in this case it would be concepts belonging to a particular group having a given relation with another given concept) before being confronted with a problem-solving task. As the learners need a basis of useful examples from which they can draw analogies and conclusions, an analogy-making mechanism would require a set of instance-related pieces of information to draw a general conclusion.

In order confirm whether it would be possible to create such mechanism and make general conclusions on the basis of the data acquired with the previously presented procedure, I have prepared and tested the following method. Figure 3.11 presents the proposed method’s schema. First the script imports the additional information lists representing AtLocation, CreatedBy and MemberOf relations. LocatedNear relations have been excluded from this analysis, as the preliminary experiments revealed that a low number of assertions representing a given relation yields unsatisfactory results. Next each assertion is analyzed one by one: for both concepts in the assertion the script finds their hypernyms in the generated IsA relations list. When the hypernyms are found, assertions representing all possible combinations between concept A’s hypernyms and concept B’s hypernyms are generated. The pro-

cess is repeated for all assertions in the additional information list. Finally, the script produces a list of the generated hypernym assertions together with their respective occurrence frequency. This way the method indicates which assertions are created on the basis of the highest number of examples. The hypothesis is that the assertions with the highest occurrence frequency represent general, common sense observations. The number and reliability level of the data acquired with the proposed method is presented in the Section 4.8 and the solution to the performance issues which appeared while implemented the method is presented in section 3.5.

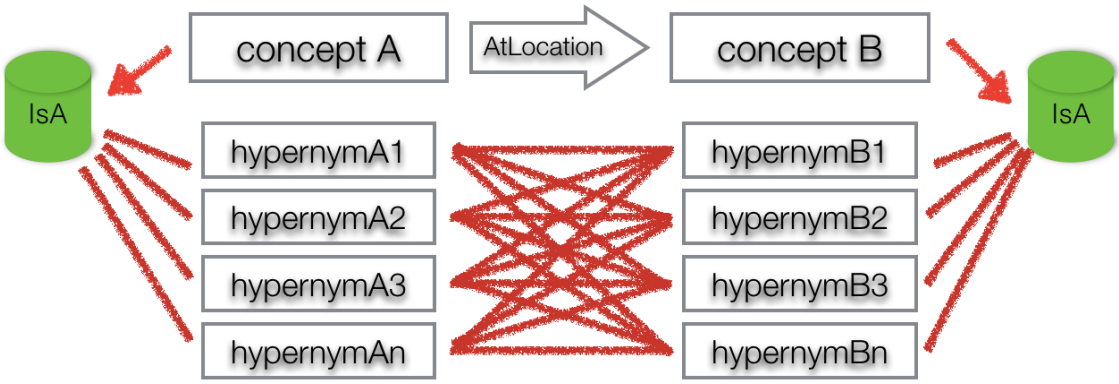


Figure 3.11: Outline of the proposed method for general assertion acquisition exemplified with AtLocation assertion type.

3.5 Performance optimization

The operation of scanning through every line of AtLocation, CreatedBy and MemberOf assertions lists and searching for hypernyms of both concepts in the assertion greatly increases the time complexity of the algorithm. The complexity reaches to the level of $2n \times m$ where n is the length of the list of assertions of a given type and m is the length of the list of IsA assertions. In order to optimize the performance of the algorithm and minimize the required time for finishing the operation several different approaches have been made. The first step was to minimize the program's hard drive operations. Instead of reading the files containing the processed assertions list and the IsA list line by line, the program loaded both lists to the operational memory and performed the search there. This alone had a positive impact on the performance, but still the time to finish the analysis of a single relation file, for example CreatedBy, could be counted in days, in this case 6 days. A radical reduction of the

necessary calculation time came with the introduction of multiprocessing. The multiprocessing functionality has been implemented in the following way. First the main program loads the assertions list to be processed and the IsA assertions list to the operational memory. Then a dedicated procedure is used to check the number of cores available on the machine the program is executed on. The program then uses a for loop to launch a number of subprocesses equal to the number of cores detected, and run them each using a separate core. Each subprocess takes a single position from the analyzed assertion list as input. The position consists of two elements, concept A and concept B. The subprocess searches for each of these elements in the IsA list's hyponym position. If a match is found, the corresponding hypernym is added to a temporary list of concept A's hypernyms or concept B's hypernyms. Next a list of all possible pairs between concept A's hypernyms and concept B's hypernyms is created and returned as the output of the subprocess. The main program adds the returned list to a global list and loads the subprocess with another position from analyzed assertion list. The loop of adding next position from the analyzed assertion list to a subprocess and receiving the subprocess' output list continues until the list of assertions finishes. At that moment the main program closes all subprocesses one by one after it receives their outputs. The global list is then transformed into a dictionary which key is a tuple consisting of concept A's hypernym and concept B's hypernym pair, and the value is the number of times such pair appeared in the global list. Finally the dictionary is sorted from the most frequent to the least frequent pair and saved to a file (sorting is necessary to find pairs created on the basis of the highest number of examples). As a result of implementing multiprocessing in the described manner, the process of analyzing the mentioned before CreatedBy relation file has been reduced from 6 days to 11,5 hours while running on a 32 core machine.

Chapter 4

Evaluation

4.1 Introduction

The purpose of this this chapter is to present the evaluation of the data acquired by the introduced and proposed methods. Section 4.2 describes the evaluation methodology applied to assess the discussed datasets. Further sections show the evaluation referring to the particular datasets. Section 4.3 presents the results of applying method for gathering IsA assertions. Section 4.4 presents the assessment of the generated AtLocation assertion dataset. Section 4.5 contains the evaluation of LocatedNear Assertions. Section 4.6 shows the assessment of CreatedBy relations sample. Following Section 4.7 refers to the MemberOf assertions evaluation results. Section 4.8 contains the results of general assertions assessment. Finally, Section 4.9 concludes the chapter.

4.2 Evaluation methodology

To verify the reliability level declared by Sumida *et al.* [87] and evaluate the proposed method for obtaining additional relations I used the 2014-11-04 version of the Japanese Wikipedia dump data as the input to the definition, category and hierarchy modules of the Hyponymy extraction tool v1.0 running at 93% precision rate using SVM-based classifier trained with the biggest available training set. Next I obtained 2,738,211 basic hypernym–G-INTER–T-INTER–basic hyponym sets by running the fourth 'extended' module of the Hyponymy extraction tool v1.0 on the same Wikipedia dump data.

The 93% reliability level declared by the authors of the Sumida *et al.* [87] method has been verified by three human annotators, whose task was to evaluate a sample of the data and decide whether the

extracted pairs a) represent a correct hyponymy relation, b) represent related concepts but not in a hyponymy relation, or c) represent unrelated concepts. The annotators assigned 1, 0.5 and 0 points respectively to 300 randomly selected assertions. If two or more annotators assessed an item as belonging to one category, their decision was regarded as the evaluation output. In cases where their decisions varied (which happened 10 times), the author of this thesis decided the score. Appendix A contains a full list of the evaluated IsA assertions. The procedure follows a modified Sumida *et al.* [87] evaluation method.

The decision to assign 0.5 points to related concepts has been made after the analysis of common sense knowledge evaluation methods applied in related research. For example, to evaluate knowledge gathered in ConceptNet Speer *et al.* [50] proposed a five grade classification of the evaluated assertions. In order to evaluate the knowledge the annotators had to decide whether assertions were true, sometimes true, vague, false/nonsense, or indicate that they do not know the answer. For the use of the current study such evaluation would be impractical as it would be difficult to assign numerical accuracy values to the assessed data sets. That is why in this case there are only three categories: true, related and false. Whether the related concepts are useful or not depends on a particular application of a knowledge base. As it has been demonstrated in the potential application case study (Chapter 5) sometimes related concepts are as helpful as concepts connected by a particular relation. As the destination of the generated data is a knowledge base for universal use, related concepts should not be excluded from the analysis and therefore 0.5 points were assigned to them. In case of a stricter evaluation approach, only concepts related by the particular relation could be considered.

By applying the proposed method for extracting additional information it was possible to produce pairs representing AtLocation, LocatedNear, CreatedBy MemberOf relation. For comparison, nadya.jp, the baseline system, has provided only a set of AtLocation relations and no LocatedNear, CreatedBy or MemberOf relations during four years of its operation. In the case of AtLocation pairs, the evaluation covered 100 pairs randomly selected from the proposed method's output and 100 pairs randomly selected from nadya.jp's AtLocation assertions [58] (the number of evaluated pairs was adjusted to balance the proportion between the total number of pairs and the test sample). While evaluating LocatedNear, CreatedBy and MemberOf relations, a comparison with the baseline was not

possible, as ConceptNet 5.3 does not yet contain any LocatedNear, CreatedBy or MemberOf pairs in its Japanese language section. These assertions were therefore evaluated independently. The evaluation procedure follows the previously applied one: 1 point being applied to correct AtLocation, LocatedNear, CreatedBy or MemberOf assertions, 0.5 point to related concepts but not in the evaluated relation, and 0 points to unrelated concepts. In 15 cases the annotators' evaluation was inconsistent, and therefore the first author decided the score. Appendices B-E contain full lists of the evaluated assertions.

In order to assess the effectiveness of the generalization method, a full evaluation of the top 100 samples from each category was performed. A broader evaluation was impossible due to human resources and time constraints. Top 100 samples are defined as those which received the highest occurrence frequency in the process of cross-referencing IsA relations list with each of the AtLocation, CreatedBy and MemberOf lists. Additional initial assessment was performed on 100 random samples from LocatedAt, CreatedBy and MemberOf general assertion sets consisting of items created on the basis of at least 50 examples from the instance-related data. The purpose was to check whether the 50 examples threshold is suitable for obtaining high accuracy data. The evaluation procedure follows the previously applied one: both top and random assertion datasets have been verified by three human annotators, whose task was to evaluate the data and decide whether the extracted pairs a) represent a correct relation of a given type, b) represent related concepts but not in a given relation, or c) represent unrelated concepts. The annotators assigned 1, 0.5 and 0 points respectively to the tested assertions, following the previously presented rationale. If two or more annotators assessed an item as belonging to one category, their decision was regarded as the evaluation output. In cases where their decisions varied (25 cases out of 300 for top samples and 7 cases out of 300 for random samples), the first author decided the score.

4.3 IsA assertions evaluation

Applying Sumida *et al.* [87] method to the 2014-11-04 version of the Japanese Wikipedia dump data resulted in obtaining 6,014,194 hypernym-hyponym pairs. The number of unique hyponymy pairs was 5,866,680, which indicates that 147,514 pairs have been extracted by more than one module.

Table 4.1 presents the evaluation results. 283 pairs were assessed as representing a correct hyponymy relation, 10 pairs as related concepts but not in a hyponymy relation and 7 as unrelated concepts. This results in 96.0% accuracy value of the tested sample, which surpasses the 93% declared by Sumida *et al.* To measure the agreement level between judges, Randolph's free marginal multirater Kappa was used instead of Fleiss' fixed-marginal multirater Kappa, due to high agreement low Kappa paradox [98]. The level of overall agreement between annotators was 86.9%, and the Kappa value was 0.80, which indicates that the annotation judgement was in substantial agreement. Examples of the extracted IsA assertions that have been positively verified by the annotators are presented in Table 4.2.

Table 4.1: Evaluation results for IsA relations.

Correct hyponymy	Related concepts	Unrelated concepts	Accuracy	Total number of pairs
0.943 (283/300)	0.033 (10/300)	0.023 (7/300)	0.960	5,866,680

Table 4.2: Examples of generated IsA assertions.

Kurausu Arofusu (Klaus Allofs)	IsA	*Werudā Burēmen no senshu* (Werder Bremen player)
dai ni-ji Shōwa kitte (second Showa stamp)	IsA	*Nihon no futsū kitte* (Japanese definitive stamp)
Makai Suikoden (Hell's Water Margin)	IsA	*Nihon no SF shōsetsu* (Japanese SF novel)
Sakurai Ikuko	IsA	*josei* (female)
Jon Windamu (John Wyndham)	IsA	*SF sakka* (SF writer)
Rūsu taifū (Ruth typhoon)	IsA	*taifū* (typhoon)
chōritsu toshokan (town library)	IsA	*kyōiku shisetsu* (educational facility)
Tamura Hitoshi	IsA	*puro yakyū senshu* (professional baseball player)

4.4 AtLocation assertions evaluation

The proposed method produced 131,760 pairs representing AtLocation relation. For comparison, nadya.jp, the baseline system, has provided only 8,706 AtLocation relations. Table 4.3 shows the evaluation results of the proposed AtLocation pairs generation method in comparison with the baseline system. 88 pairs generated by the method were evaluated as representing a correct AtLocation relation, 11 pairs as related concepts but not in an AtLocation relation, and 1 as unrelated concepts. This results in a 93.5% accuracy value. In the case of the baseline system, 64 pairs were evaluated as correct AtLocation assertions, 20 as related concepts but not in an AtLocation relation, and 16 as unrelated concepts. The accuracy value for the baseline system is 74.0%. The level of overall agreement between annotators was 73.6% and the Kappa value was 0.60, which indicates that the annotation judgment was in moderate agreement. Examples of the extracted AtLocation assertions that have been positively verified by the annotators are presented in Table 4.4.

Table 4.3: Evaluation results for AtLocation relations in comparison with the nadya.jp baseline.

	Correct AtLocation	Related concepts	Unrelated concepts	Accuracy	Total number of pairs
Proposed	0.880 (88/100)	0.110 (11/100)	0.010 (1/100)	0.935	131,760
Baseline	0.640 (64/100)	0.200 (20/100)	0.160 (16/100)	0.740	8,706

$p < 0.001$, t-score = 4.6291

4.5 LocatedNear assertions evaluation

The number of pairs representing LocatedNear assertion generated by the proposed system reached the value of 6,217. Table 4.5 contains the evaluation result of the accumulated relations. 97 pairs were evaluated as correct LocatedNear pairs, 3 as related concepts and none as unrelated concepts, which results in 98.5% accuracy. The level of overall agreement between annotators was 86.6% and the Kappa value was 0.80, which indicates that the annotation judgment was in substantial agreement. Examples of the extracted LocatedNear assertions that have been positively verified by the annotators are presented in Table 4.6.

Table 4.4: Examples of generated AtLocation assertions.

Tomato Ginkō (Tomato Bank)	AtLocation	*Okayama-shi* (Okayama city)
Mariina Ōdōri (Marina Boulevard)	AtLocation	*A Korūnya* (A Coruna)
Wōren Shinrin-kyoku Kūkō (Warren USFS Airport)	AtLocation	*Aidaho-gun* (Idaho County)
Hoshinomiya Jinja (Hoshinomiya Temple)	AtLocation	*Minami-mura* (Minami village)
Ōtao hoikuen (Otao nursery)	AtLocation	*Sakai-shi* (Sakai city)
Shinzutsumi Shizen Kōen (Shinzutsumi nature park)	AtLocation	*Kurihara-shi* (Kurihara city)
Sandi Fukku (Sandy Hook)	AtLocation	*Eriotto-gun* (Elliott County)
Hoteru Kadoya (Kadoya Hotel)	AtLocation	*Tochigi-shi* (Tochigi city)

Table 4.5: Evaluation results for LocatedNear relations

Correct LocatedNear	Related concepts	Unrelated concepts	Accuracy	Total number of pairs
0.970 (97/100)	0.030 (3/100)	0.000 (0/100)	0.985	6,217

4.6 CreatedBy assertions evaluation

In case of pairs representing CreatedBy relation, the method was able to produce 270,230 assertions. Table 4.7 contains the evaluation result of the generated CreatedBy relations. 60 pairs were evaluated as correct CreatedBy pairs, 37 as related concepts and 3 as unrelated concepts, which results in 78.5% accuracy. The level of overall agreement between annotators was 71.6% and the Kappa value was 0.57, which indicates that the annotation judgment was in moderate agreement. Examples of the extracted CreatedBy assertions that have been positively verified by the annotators are presented in Table 4.8.

The analysis of the relatively low accuracy score of the assessed CreatedBy assertions revealed the

Table 4.6: Examples of generated LocatedNear assertions.

Ōgoe-machi (Ogoe city)	LocatedNear	*Ono-machi* (Ono city)
Iseri-gawa (Iseri river)	LocatedNear	*Konoha-gawa* Konoha river
Shin Edo-gawa Kōen (New Edo River Park)	LocatedNear	*Kōdansha Noma Kinenkan* (Kodansha Noma Memorial Museum)
Daitingu (Daiting)	LocatedNear	*Monhaimu* (Monheim)
Sahoro Yūsu Hosteru (Sahoro Youth Hostel)	LocatedNear	*Obihiro Yachiyo Yūsu Hosteru* (Obihiro Yachiyo Youth Hostel)
Kumotori-yama (Mount Kumotori)	LocatedNear	*Karamatsuo-yama* (Mount Karamatsuo)
Goshogawara-shi (Goshogawara city)	LocatedNear	*Sotogahama-machi* (Sotogahama town)
Gujō Keisatsujo (Gujo Police Station)	LocatedNear	*Ōno Keisatsusho* (Ono Police Station)

Table 4.7: Evaluation results for CreatedBy relations.

Correct CreatedBy	Related concepts	Unrelated concepts	Accuracy	Total number of pairs
0.600 (60/100)	0.370 (37/100)	0.030 (3/100)	0.785	270,230

following: in 24 cases it was the annotators' opinion that actors, voice actors, animators, storyboard creators or sound directors cannot be considered as creators of works they contribute to. Although it would be valid to include such persons in the RelatedTo kind of relationship with the work they helped to create, defining them as creators would go against common sense. This is a valid observation and it will be taken into consideration when re-designing and expanding the rule set for the next version of the algorithm. The reason why such items went through the applied rules are as follows: there is a group of people covered by Wikipedia articles who perform more than one role. If a person is a director, actor, voice actor and writer, all those attributes will be enumerated in the G-INTER's tags enclosed by square brackets. In case when a person is only an actor in one film, and a director of another film, the method will indicate such person as a creator in both cases. This issue will

Table 4.8: Examples of generated CreatedBy assertions.

Dāku Hōsu (Dark Horse)	CreatedBy	*Jōji Harison* (George Harrison)
Kaze (Wind)	CreatedBy	*Kubota Kōtarō*
Manuke-na Ōkami (Sheep Wrecked)	CreatedBy	*Maikeru Rā* (Michael Lah)
The Point of View	CreatedBy	*Aran Kurosurando* (Alan Crosland)
Bun Bun Bun Bun!! (Boom, Boom, Boom, Boom!!)	CreatedBy	*Benga Bōizu* (Vengaboys)
Genki-na Burōkun Hāto (Healthy Broken Heart)	CreatedBy	*Matsumoto Takashi*
Haru no Hi (Spring Day)	CreatedBy	*Watanabe Takuya*
When the Birds Fly South	CreatedBy	*Sutanton A Koburentsu* (Stanton A. Coblentz)

have to be resolved in the future to increase the accuracy of the output. There were also cases of assertions assessed as invalid due to errors passed from the output of the Hyponymy extraction tool to the proposed method. Table 4.9 contains examples of assertions that were assessed as erroneous by the annotators.

Table 4.9: Examples of erroneous CreatedBy assertions.

Shishi no ketsumyaku (Lion bloodline)	CreatedBy	*Ozawa Hitoshi* (actor)
Rōdo 88 (Road 88)	CreatedBy	*Tomita Yasuko* (actress)
Tsurupika Hagemaru (Little Baldy Hagemaru)	CreatedBy	*Zen Sōichirō* (storyboard creator)
Kaiketsu Zorori (Incredible Zorori)	CreatedBy	*Yamada Etsuji* (sound director)
Kishin Dōji Zenki (Zenki)	CreatedBy	*Hayashi Akemi* (animator)
Human (incomplete name error)	CreatedBy	*Nikoruson Beikā* (Nicholson Baker)

4.7 MemberOf assertions evaluation

Applying the proposed method resulted in obtaining 21,053 pairs representing MemberOf relation. Table 4.10 contains the evaluation result of the generated MemberOf assertions. 76 pairs were evaluated as correct MemberOf pairs, 22 as related concepts and 2 as unrelated concepts, which results in 87.0% accuracy. The level of overall agreement between annotators was 80.6% and the Kappa value was 0.71, which indicates that the annotation judgment was in substantial agreement. Examples of the extracted MemberOf assertions that have been positively verified by the annotators are presented in Table 4.11.

Table 4.10: Evaluation results for MemberOf relations.

Correct MemberOf	Related concepts	Unrelated concepts	Accuracy	Total number of pairs
0.760 (76/100)	0.220 (22/100)	0.020 (2/100)	0.870	21,053

Table 4.11: Examples of generated MemberOf assertions.

Henning Schmitz	MemberOf	*Kurafutowāku* (Kraftwerk)
Dir.F	MemberOf	*Suiyōbi no Kanpanera* (Wednesday Canpanella)
Ōno Satoshi	MemberOf	*Arashi*
Nishimura Akihiro	MemberOf	*Nikkan Giin Renmei* (Japan-Korea Parliamentarians' Union)
Nils Lindenhayn	MemberOf	*Ji Ōshan* (The Ocean)
Murata Megumi	MemberOf	*Melon Kinenbi*
Richādo Ōkusu (Richard Oakes)	MemberOf	*Sūēdo* (Suede)
Suzuki Daisuke	MemberOf	Day After Tomorrow

In the 13 cases the annotators decided that the generated MemberOf assertion refer to the former

member of relative group, and therefore assigned it as the related concepts. The question whether these pairs should be considered as representing concepts in MemberOf relation is currently under discussion. If we would consider that the status of a member, once granted, is not temporary, then the accuracy rate of the tested sample would be higher, reaching 93.5%.

4.8 General assertions evaluation

When the threshold of 50 examples has been applied to the instance-related datasets, it resulted in obtaining 74,226 AtLocation relation, 330,418 CreatedBy relation and 1,355 MemberOf relation general assertions. These quantities are referred to as "Number of 50+ examples pairs" in Table 4.12, which presents the evaluation results. In case of AtLocation general assertions 98 pairs were assessed as representing correct AtLocation relation, 1 pair as related concepts but not in AtLocation relation and 1 as unrelated concepts. This results in 98.5% accuracy value of the tested sample. The level of overall agreement between annotators was 74.0%, and the Kappa value was 0.61, which indicates that the judgement was in substantial agreement. In case of CreatedBy general assertions 83 pairs were assessed as representing correct CreatedBy relation, 17 pairs as related concepts but not in CreatedBy relation and none as unrelated concepts. This results in 91.5% accuracy value of the tested sample. The level of overall agreement between annotators was 42.6%, and the Kappa value was 0.14, which indicates that the judgement was in slight agreement. In case of MemberOf general assertions 68 pairs were assessed as representing correct MemberOf relation, 6 pairs as related concepts but not in MemberOf relation and 26 as unrelated concepts. This results in 71.0% accuracy value of the tested sample. The level of overall agreement between annotators was 65.6%, and the Kappa value was 0.48, which indicates that the judgement was in moderate agreement. Table 4.13 presents examples of generated general assertions that have been positively verified by the annotators.

Initial assessment of 100 random samples from AtLocation, CreatedBy and MemberOf general assertion sets taken from items created on the basis of at least 50 examples from the instance-related data revealed the following results: in case of AtLocation assertions the annotators assessed the sample as representing 7.0% accuracy, CreatedBy samples received 66.0% accuracy score and MemberOf assertions were evaluated at the level of 52.0% accuracy. These results clearly indicate that estab-

Table 4.12: Evaluation results for the acquired relations.

	Correct relations	Related concepts	Unrelated concepts	Accuracy of top 100 pairs	Number of 50+ examples pairs	Accuracy of random 100 pairs
AtLocation	0.980 (98/100)	0.010 (1/100)	0.010 (1/100)	0.985	74,226	0.070
CreatedBy	0.830 (83/100)	0.170 (17/100)	0.000 (0/100)	0.915	330,418	0.660
MemberOf	0.680 (68/100)	0.060 (6/100)	0.260 (26/100)	0.710	1,355	0.520

Table 4.13: Examples of generated general assertions.

toshi oyobi machi (city and town)	AtLocation	*gun* (province)
shōgakkō (elementary school)	AtLocation	*machi* (city)
sakuhin (work)	CreatedBy	*zonmei jinbutsu* (living person)
shutsuen sakuhin (performance art)	CreatedBy	*bunkajin* (cultural figure)
zonmei jinbutsu (living person)	MemberOf	*Nihon no kashu gurūpu* (Japanese singer group)
owarai geinin (comedian)	MemberOf	*Nihon no owarai konbi* (Japanese comic duo)

lishing a rigid threshold level for all types of assertions leads to inconsistent results: 66.0% could be considered as an acceptable accuracy, while 7.0% is much below desirable performance. It is therefore evident that there is a need for further development of a method for assigning the acceptability threshold. Presenting reasons for the discrepancy in the above-mentioned initial results would be a target of a separate study.

4.9 Conclusion

The presented results show that IsA relation pairs generated by the definition, category and hierarchy of the Hyponymy extraction tool v1.0, as well as AtLocation, LocatedNear and MemberOf relation

pairs extracted by the proposed method may be incorporated into ConceptNet as a part of general factual knowledge. Considering the number of the newly acquired assertions as well as reliability of the data in comparison with the resources already present in the knowledge base, such operation would be beneficial for ConceptNet. CreatedBy relation pairs could also be added after the revision of introduced rules and a substantial increase of the accuracy rate.

In case of the generalization method, the top 100 assertions represent a satisfactory accuracy level and could be introduced to ConceptNet as representing common sense knowledge. These results clearly indicate that establishing a rigid threshold level for all types of assertions leads to inconsistent results. It is therefore evident that there is a need for further development of a method for assigning the acceptability threshold.

Chapter 5

Discussion

5.1 Introduction

The purpose of this chapter is to present the case study of a system that can profit from the data acquired by the proposed method.

5.2 Case study of potential application

In order to verify the potential applicability of the acquired data to a working system, a book recommendation system scenario was taken into consideration. The reason for choosing such an approach is that recommendation systems are usually knowledge-based and, especially at the beginning of the operation, suffer from an insufficient amount of available data vectors [99]. The investigated Japanese book recommendation system is currently being created at Hokkaido University. The system is being designed to consist of five modules, each performing book recommendation based on a different set of data: attributes (title, author, publisher, sales date, genre, price), content description, users' reviews, Amazon sales-based suggestions, and attributes plus reviews. A preliminary survey performed among the system's test users revealed that the attribute-based module represents the lowest reliability: the test users' opinions suggested that recommendations made on the basis of the authors' name and title similarity were very often misleading. However, to improve the effectiveness of attribute-based recommendation, the system could be provided with more input for building additional vectors. Therefore it would be useful to verify whether the data extracted by the proposed method could potentially be applied for this purpose. The analysis covered the system's working data, consisting of 106,415 book titles accompanied with authors' names. The data was gathered from the Amazon Japan

website [100]. In order to test the data acquired with the proposed method against books that are popular in Japanese society, texts which had less than 30 reviews at a Japanese book review sharing site, Dokusho Meter [101] were filtered out. Such operation resulted in a list of 14,055 book titles accompanied by their 18,988 authors' names. A test script has been created to search the title and author data using the IsA and CreatedBy relation pairs. As a result, additional information about the author or authors of 13,007 books (92.5% of the studied sample) has been found. To be more precise, the script found information concerning 15,685 authors' names (82.6%). The additional information includes other works created by the authors, the authors' place of birth, occupations and other characteristics included in the IsA and CreatedBy relation bases. These clues may be used to create more detailed profile of each author, which could be utilized when comparing them with other authors to make book recommendations. Further information concerning the title of 538 volumes (3.8%) has been extracted as well. In total the data produced by the proposed method was able to provide the system with additional, potentially useful information concerning 13,038 positions, which is 92.7% of the analyzed sample. Each book found in the data received an average of 28 additional information vectors. On the basis of these findings, it is possible to put forward a hypothesis that the data acquired by the proposed method have a strong potential for application to a practical use. As the approach of the creators of the discussed book recommendation system is to move away from conventional collaborative filtering to more complex and innovative semantic feature analysis-based recommendation, the data produced by the proposed method would provide the fundamental element necessary for realizing that approach. Proving the aforementioned hypothesis, however, would have to be the object of a separate, extensive study performed upon the completion of the current system.

Chapter 6

Conclusion

6.1 Overall conclusions

This thesis presented a method for automatic acquisition of common sense knowledge triplets from the Japanese Wikipedia. It resulted in the acquisition of instance related IsA, AtLocation, LocatedNear, CreatedBy and MemberOf assertions with accuracy estimated at the levels of 96.0%, 93.5%, 98.5%, 78.5% and 87.0% respectively. Additional processing of the acquired data resulted in a set of AtLocation, CreatedBy and MemberOf assertions representing general common sense knowledge with accuracy estimated at the levels of 98.5%, 91.5% and 71.0% respectively for the 100 samples which received the highest occurrence frequency in the process of cross-referencing IsA relations list with each of the AtLocation, CreatedBy and MemberOf lists. The accuracy of randomly selected 100 samples was lower, which revealed the need for further investigation regarding the acceptability threshold. As the Japanese part of ConceptNet 5.3 consists of 1,071,046 assertions, a contribution of the newly acquired assertions would be significant. It would mean an almost sixfold increase and could potentially make ConceptNet applicable to many Japanese language analysis problems. Moreover, as Wikipedia is a constantly expanding source, it would be possible to acquire more assertions simply by applying the proposed method to the updated Wikipedia XML dump files.

The applicability of ConceptNet is not limited to any particular branch of data analysis. Therefore it could be speculated that the results of the proposed method may not only augment the effectiveness and scope of already created tools, but also may contribute to the development of new directions and approaches, as depicted by the presented book recommendation system example.

6.2 Future work

In order to extend the functionality of the proposed method, an update to the primary and secondary rules could be performed, which would allow the system to increase its accuracy and the scope of extracted information. It would be interesting to explore the possibility of using a machine learning algorithm for automatic rule generation combined with the already present heuristics. Such a combination could potentially be more effective in increasing accuracy, as well as finding new rules to extract even more relations.

As it has been demonstrated in the latter part of the Evaluation section, there is a need for further development of a method for dynamically assigning the acceptability threshold to balance the accuracy level and the number of extracted assertions. Applying machine learning methods to train a model for classifying assertions as belonging to a particular category based on a set of features should be also taken into consideration.

In order to effectively utilize the pairs representing related concepts it would be useful to create an interface for the evaluation of the method's output by Japanese native speakers. By applying methods similar to those observer in games with purpose it would be possible to acquire new and original assertions as well.

Bibliography

[1] J. McCarthy, "Some expert systems need common sense," *Annals of the New York Academy of Sciences*, vol. 426, no. 1, pp. 129–137, 1984.

[2] H. Liu and P. Singh, "ConceptNet - a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, no. 4, pp. 211–226, 2004.

[3] R. H. Speer, C. Havasi, K. N. Treadway, and H. Lieberman, "Finding your way in a multi-dimensional semantic space with Luminoso," in *Proceedings of the 15th international conference on Intelligent user interfaces*. ACM, 2010, pp. 385–388.

[4] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "SenticSpace: visualizing opinions and sentiments in a multi-dimensional vector space," in *Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2010, pp. 385–393.

[5] S. J. Korner and T. Brumm, "RESI - a natural language specification improver," in *Proceedings of IEEE International Conference on Semantic Computing (ICSC'09)*. IEEE, 2009, pp. 1–8.

[6] J. Ullberg, S. Coradeschi, and F. Pecora, "On-line ADL recognition with prior knowledge," in *Proceedings of the 2010 conference on STAIRS 2010: Proceedings of the Fifth Starting AI Researchers' Symposium*. IOS press, 2010, pp. 354–366.

[7] E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Sentic Computing: exploitation of common sense for the development of emotion-sensitive systems," in *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer, 2010, pp. 148–156.

[8] Q.-F. Wang, E. Cambria, C.-L. Liu, and A. Hussain, "Common sense knowledge for handwritten Chinese text recognition," *Cognitive Computation*, vol. 5, no. 2, pp. 234–242, 2013.

[9] M. Minsky, *The emotion machine*. Pantheon New York, 2006.

[10] L.-J. Zang, C. Cao, Y.-N. Cao, Y.-M. Wu, and C. Cun-Gen, "A survey of commonsense knowledge acquisition," *Journal of Computer Science and Technology*, vol. 28, no. 4, pp. 689–719, 2013.

[11] E. T. Mueller, *Natural language processing with Thought Treasure*. Signiform New York, 1998.

[12] Z. Dong and Q. Dong, *HowNet and the Computation of Meaning*. World Scientific, 2006.

[13] L. Schubert, "Can we derive general world knowledge from texts?" in *Proceedings of the second international conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 2002, pp. 94–97.

[14] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. ACM, 2008, pp. 1247–1250.

[15] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer *et al.*, "DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia," *Semantic Web*, vol. 6, no. 2, pp. 167–195, 2015.

[16] A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. Hruschka Jr, and T. M. Mitchell, "Toward an architecture for never-ending language learning." in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI-10)*, vol. 5, 2010, p. 3.

[17] D. B. Lenat, "CYC: A large-scale investment in knowledge infrastructure," *Communications of the ACM*, vol. 38, no. 11, pp. 33–38, 1995.

[18] Knowledge base - Cycorp. Website. Accessed: 15.12.2016. [Online]. Available: http://www.cyc.com/kb/

[19] J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubézy, H. Eriksson, N. F. Noy, and S. W. Tu, "The evolution of Protégé: an environment for knowledge-based systems

development," *International Journal of Human-computer studies*, vol. 58, no. 1, pp. 89–123, 2003.

[20] K. Panton, P. Miraglia, N. Salay, R. C. Kahlert, D. Baxter, and R. Reagan, "Knowledge formation and dialogue using the KRAKEN toolset," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence*, 2002, pp. 900–905.

[21] M. Witbrock, D. Baxter, J. Curtis, D. Schneider, R. Kahlert, P. Miraglia, P. Wagner, K. Panton, G. Matthews, and A. Vizedom, "An interactive dialogue system for knowledge acquisition in Cyc." in *Proceedings of the 18th International Joint Conference on Artificial Intelligence*. Citeseer, 2003, pp. 138–145.

[22] M. J. Witbrock, C. Matuszek, A. Brusseau, R. C. Kahlert, C. B. Fraser, and D. B. Lenat, "Knowledge begets knowledge: Steps towards assisted knowledge acquisition in Cyc." in *AAAI Spring Symposium: Knowledge Collection from Volunteer Contributors*, 2005, pp. 99–105.

[23] C. Matuszek, M. Witbrock, R. C. Kahlert, J. Cabral, D. Schneider, P. Shah, and D. Lenat, "Searching for common sense: populating Cyc from the Web," in *Proceedings of the 20th National Conference on Artificial Intelligence*, 2005, pp. 1430–1435.

[24] J. Masters, C. Matuszek, and M. Witbrock, "Ontology-based integration of knowledge from semi-structured Web pages," Technical Report, Cycorp, Tech. Rep., 2006.

[25] O. Medelyan and C. Legg, "Integrating Cyc and Wikipedia: Folksonomy meets rigorously defined common-sense," in *Wikipedia and Artificial Intelligence: An Evolving Synergy, Papers from the 2008 AAAI Workshop*, 2008, p. 65.

[26] Research analyst assistant - Cycorp. Website. Accessed: 15.12.2016. [Online]. Available: http://www.cyc.com/enterprise-solutions/solutions/research-analyst-assistant/

[27] J. Curtis, G. Matthews, and D. Baxter, "On the effective use of Cyc in a question answering system," in *Proceedings of the IJCAI05 Workshop Knowledge and Reasoning for Answering Questions (KRAQ05)*, 2005, pp. 61–70.

[28] J. Curtis, J. Cabral, and D. Baxter, "On the application of the Cyc ontology to word sense disambiguation." in *FLAIRS Conference*, 2006, pp. 652–657.

[29] C. D Pierce, D. Booth, C. Ogbuji, C. Deaton, E. Blackstone, and D. Lenat, "SemanticDB: a semantic Web infrastructure for clinical research and quality reporting," *Current Bioinformatics*, vol. 7, no. 3, pp. 267–277, 2012.

[30] M. Witbrock, K. Panton, S. L. Reed, D. Schneider, B. Aldag, M. Reimers, and S. Bertolo, "Automated OWL annotation assisted by a large knowledge base," in *Workshop Notes of the 2004 Workshop on Knowledge Markup and Semantic Annotation at the 3rd International Semantic Web Conference ISWC2004*, 2004, pp. 71–80.

[31] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[32] G. A. Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[33] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the Web: An experimental study," *Artificial intelligence*, vol. 165, no. 1, pp. 91–134, 2005.

[34] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Artificial Intelligence*, vol. 194, pp. 28–61, 2013.

[35] GeoNames. Website. Accessed: 15.12.2016. [Online]. Available: http://www.geonames.org

[36] F. Mahdisoltani, J. Biega, and F. Suchanek, "Yago3: A knowledge base from multilingual Wikipedias," in *Proceedings of 7th Biennial Conference on Innovative Data Systems Research*. CIDR Conference, 2014.

[37] Max-Planck-Institut fur Informatik: YAGO. Website. Accessed: 15.12.2016. [Online]. Available: http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/

[38] L. Abouenour, K. Bouzoubaa, and P. Rosso, "Using the Yago ontology as a resource for the enrichment of named entities in Arabic WordNet," in *Proceedings of The Seventh International Conference on Language Resources and Evaluation (LREC 2010) Workshop on Language Resources and Human Language Technology for Semitic Languages*, 2010, pp. 27–31.

[39] A. Ren, X. Du, and P. Wang, "Ontology-based categorization of Web search results using YAGO," in *Proceedings of International Joint Conference on Computational Sciences and Optimization (CSO)*, vol. 1. IEEE, 2009, pp. 800–804.

[40] Y. Hu, Z. Wang, W. Wu, J. Guo, and M. Zhang, "Recommendation for movies and stars using YAGO and IMDB," in *Proceedings of 12th International Asia-Pacific Web Conference (APWEB)*. IEEE, 2010, pp. 123–129.

[41] W. Wu, H. Li, H. Wang, and K. Q. Zhu, "Probase: a probabilistic taxonomy for text understanding," in *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*. ACM, 2012, pp. 481–492.

[42] Microsoft Concept Graph and concept tagging release. Website. Accessed: 15.12.2016. [Online]. Available: https://concept.research.microsoft.com/

[43] Y. Wang, H. Li, H. Wang, and K. Q. Zhu, "Toward topic search on the Web," Citeseer, Tech. Rep., 2010.

[44] Y. Song, H. Wang, Z. Wang, H. Li, and W. Chen, "Short text conceptualization using a probabilistic knowledgebase," in *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Three*. AAAI Press, 2011, pp. 2330–2336.

[45] A. Fader, L. Zettlemoyer, and O. Etzioni, "Open question answering over curated and extracted knowledge bases," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 1156–1165.

[46] J. Anacleto, H. Lieberman, M. Tsutsumi, V. Neris, A. Carvalho, J. Espinosa, M. Godoi, and S. Zem-Mascarenhas, "Can common sense uncover cultural differences in computer applica-

tions?" in *Proceedings of IFIP International Conference on Artificial Intelligence in Theory and Practice*. Springer, 2006, pp. 1–10.

[47] H. Chung, "GlobalMind-bridging the gap between different cultures and languages with common-sense computing," Ph.D. dissertation, Massachusetts Institute of Technology, 2006.

[48] C. Havasi, R. Speer, and J. Alonso, "ConceptNet 3: a flexible, multilingual semantic network for common sense knowledge," in *Recent advances in natural language processing*. Citeseer, 2007, pp. 27–29.

[49] N. Eckhardt, "A kids open mind common sense," Ph.D. dissertation, Tilburg University, 2008.

[50] R. Speer and C. Havasi, "Representing general relational knowledge in ConceptNet 5." in *Proceedings of International conference on language resources and evaluation (LREC)*, 2012, pp. 3679–3686.

[51] Wiktionary. Website. Accessed: 13.11.2015. [Online]. Available: https://www.wiktionary.org/

[52] Wikipedia. Website. Accessed: 13.11.2015. [Online]. Available: https://www.wikipedia.org/

[53] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "DBpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[54] O. Etzioni, M. Banko, S. Soderland, and D. S. Weld, "Open information extraction from the Web," *Communications of the ACM*, vol. 51, no. 12, pp. 68–74, 2008.

[55] P. Singh, T. Lin, E. T. Mueller, G. Lim, T. Perkins, and W. L. Zhu, "Open mind common sense: Knowledge acquisition from the general public," in *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*. Springer, 2002, pp. 1223–1237.

[56] L. Von Ahn, "Games with a purpose," *Computer*, vol. 39, no. 6, pp. 92–94, 2006.

[57] L. Von Ahn, M. Kedia, and M. Blum, "Verbosity: a game for collecting common-sense facts," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 2006, pp. 75–78.

[58] K. Nakahara and S. Yamada, "Development and evaluation of a Web-based game for commonsense knowledge acquisition in Japan," in *Unisys Technology Review no. 107*, 2011, pp. 295–305.

[59] Y.-l. Kuo, J.-C. Lee, K.-y. Chiang, R. Wang, E. Shen, C.-w. Chan, and J. Y.-j. Hsu, "Community-based game design: experiments on social games for commonsense data collection," in *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2009, pp. 15–22.

[60] C. Fellbaum *et al.*, "WordNet: An electronic lexical database MIT press," *Cambridge MA*, 1998.

[61] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer, "DBpedia spotlight: shedding light on the web of documents," in *Proceedings of the 7th International Conference on Semantic Systems*. ACM, 2011, pp. 1–8.

[62] J. Breen, "JMDict: a Japanese-multilingual dictionary," in *Proceedings of the Workshop on Multilingual Linguistic Resources*. Association for Computational Linguistics, 2004, pp. 71–79.

[63] J. W. Breen, "Building an electronic Japanese-English dictionary," in *Japanese Studies Association of Australia Conference*. Citeseer, 1995.

[64] ConceptNet 5. Website. Accessed: 15.12.2016. [Online]. Available: http://conceptnet5.media.mit.edu/

[65] P. Singh *et al.*, "The public acquisition of commonsense knowledge," in *Proceedings of AAAI Spring Symposium: Acquiring (and Using) Linguistic (and World) Knowledge for Information Access*, 2002.

[66] Y.-L. Kuo and J. Y.-j. Hsu, "Goal-oriented knowledge collection." in *AAAI Fall Symposium: Commonsense Knowledge*, 2010.

[67] R. Speer, C. Havasi, and H. Surana, "Using Verbosity: Common sense data from games with a purpose." in *FLAIRS Conference*, 2010.

[68] H. Lieberman, D. Smith, and A. Teeters, "Common Consensus: a Web-based game for collecting commonsense goals," in *ACM Workshop on Common Sense for Intelligent Interfaces*, 2007.

[69] R. Speer, J. Krishnamurthy, C. Havasi, D. Smith, H. Lieberman, and K. Arnold, "An interface for targeted collection of common sense knowledge using a mixture model," in *Proceedings of the 14th international conference on intelligent user interfaces*. ACM, 2009, pp. 137–146.

[70] N. Otani, D. Kawahara, S. Kurohashi, N. Kaji, and M. Sassano, "Large-scale acquisition of commonsense knowledge via a quiz game on a dialogue system," in *Proceedings of Open Knowledge Base and Question Answering Workshop of the 26th International Conference on Computational Linguistics (COLING 2016)*, 2016, pp. 11–20.

[71] Onsei Asisuto - Yahoo! Japan. Website. Accessed: 15.12.2016. [Online]. Available: http://v-assist.yahoo.co.jp

[72] J. R. Quinlan and R. M. Cameron-Jones, "Induction of logic programs: FOIL and related systems," *New Generation Computing*, vol. 13, no. 3-4, pp. 287–312, 1995.

[73] T. Chklovski, "Learner: a system for acquiring commonsense knowledge by analogy," in *Proceedings of the 2nd international conference on Knowledge capture*. ACM, 2003, pp. 4–12.

[74] R. Speer, C. Havasi, and H. Lieberman, "AnalogySpace: Reducing the dimensionality of common sense knowledge." in *Proceedings of the 23rd national conference on Articial intelligence*, vol. 8, 2008, pp. 548–553.

[75] M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni, "Open information extraction from the Web." in *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, vol. 7, 2007, pp. 2670–2676.

[76] A. Fader, S. Soderland, and O. Etzioni, "Identifying relations for open information extraction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 1535–1545.

[77] Z.-c. Wang, Z.-g. Wang, J.-z. Li, and J. Z. Pan, "Knowledge extraction from Chinese wiki encyclopedias," *Journal of Zhejiang University SCIENCE C*, vol. 13, no. 4, pp. 268–280, 2012.

[78] D. A. Cruse, *Lexical semantics. Cambridge textbooks in linguistics*. Cambridge University Press, Cambridge, UK., 1986, vol. 12.

[79] M. A. Hearst, "Automatic acquisition of hyponyms from large text corpora," in *Proceedings of the 14th conference on Computational linguistics - Volume 2*. Association for Computational Linguistics, 1992, pp. 539–545.

[80] E. Hovy, Z. Kozareva, and E. Riloff, "Toward completeness in concept extraction and classification," in *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2*. Association for Computational Linguistics, 2009, pp. 948–957.

[81] J.-H. Oh, K. Uchimoto, and K. Torisawa, "Bilingual co-training for monolingual hyponymy-relation acquisition," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1*. Association for Computational Linguistics, 2009, pp. 432–440.

[82] S. P. Ponzetto and M. Strube, "Deriving a large scale taxonomy from Wikipedia," in *Proceedings of the 22st National Conference on Artificial Intelligence (AAAI07)*, vol. 7, 2007, pp. 1440–1445.

[83] V. Nastase and M. Strube, "Decoding Wikipedia categories for knowledge acquisition." in *Proceedings of the 23rd Conference on the Advancement of Artificial Intelligence*, vol. 8, 2008, pp. 1219–1224.

[84] R. Snow, D. Jurafsky, and A. Y. Ng, "Learning syntactic patterns for automatic hypernym discovery," *Advances in Neural Information Processing Systems 17*, 2004.

[85] A. Sumida and K. Torisawa, "Hacking Wikipedia for hyponymy relation acquisition." in *Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP)*, vol. 8, 2008, pp. 883–888.

[86] MediaWiki. Website. Accessed: 8.12.2016. [Online]. Available: https://www.mediawiki.org/wiki/MediaWiki

[87] A. Sumida, N. Yoshinaga, and K. Torisawa, "Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia." in *Proceedings of the 6th International Language Resources and Evaluation (LREC08)*, 2008.

[88] V. N. Vapnik, *Statistical learning theory*. Wiley New York, 1998, vol. 1.

[89] J. Kazama and K. Torisawa, "Exploiting Wikipedia as external knowledge for named entity recognition," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007, pp. 698–707.

[90] H. Tsurumaru, T. Hitaka, and S. Yoshida, "An attempt to automatic thesaurus construction from an ordinary Japanese language dictionary," in *Proceedings of the 11th coference on Computational linguistics*. Association for Computational Linguistics, 1986, pp. 445–447.

[91] H. Isahara, F. Bond, K. Uchimoto, M. Utiyama, and K. Kanzaki, "Development of the japanese wordnet." in *Proceedings of 6th International conference on Language Resources and Evaluation (LREC)*, 2008.

[92] I. Yamada, C. Hashimoto, J.-H. Oh, K. Torisawa, K. Kuroda, S. De Saeger, M. Tsuchida, and J. Kazama, "Generating information-rich taxonomy from Wikipedia," in *Proceedings of 4th International Universal Communication Symposium (IUCS)*. IEEE, 2010, pp. 97–104.

[93] Version1.0 : Hyponymy extraction tool. Website. Accessed: 13.11.2015. [Online]. Available: https://alaginrc.nict.go.jp/hyponymy/

[94] N. Yoshinaga. pecco - c++ library for efficient classification with conjunctive features. Website. Accessed: 13.11.2015. [Online]. Available: http://www.tkl.iis.u-tokyo.ac.jp/~ynaga/pecco/

[95] R. Speer. Relations . commonsense/conceptnet5 wiki . github. Website. Accessed: 13.11.2015. [Online]. Available: https://github.com/commonsense/conceptnet5/wiki/Relations

[96] D. Gentner, "Analogical reasoning, psychology of," *Encyclopedia of cognitive science*, 2003.

[97] A. Renkl, "Toward an instructionally oriented theory of example-based learning," *Cognitive Science*, vol. 38, no. 1, pp. 1–37, 2014.

[98] J. J. Randolph, "Free-marginal multirater kappa (multirater k [free]): An alternative to Fleiss' fixed-marginal multirater kappa." *Online Submission*, 2005.

[99] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-start recommendations," in *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2002, pp. 253–260.

[100] Amazon.co.jp. Website. Accessed: 13.11.2015. [Online]. Available: http://www.amazon.co.jp/

[101] Dokusho Meter. Website. Accessed: 13.11.2015. [Online]. Available: http://bookmeter.com/

Appendices

Appendix A - IsA assertions evaluation set

市川笑三	IsA	歌舞伎役者
マーティ・コードバ	IsA	ミネソタ・ツインズの選手
アランの妹	IsA	キャスト
富山駅中央駅	IsA	富山地方鉄道の鉄道駅
大木英夫	IsA	聖学院大学の教員
秋葉啓太	IsA	出身者
ダニエル・シュタイナー	IsA	オーストリアの俳優
鍋田恭孝	IsA	愛知県出身の人物
グレン・L・マーティン	IsA	かつて存在したアメリカ合衆国の航空機メーカー
FMちゅーピー	IsA	コミュニティFM放送局
フソバクテリア門	IsA	真正細菌
原夫次郎	IsA	出身者
たかじんのそこまで言って委員会	IsA	バラエティー・教養番組
プレイン・フォー・ア・ミラクル	IsA	収録曲
防府市立向島小学校	IsA	公立小学校
ロシアのスノーボード選手	IsA	ロシアのスポーツ選手
城東市民体育館	IsA	日本のハンドボール競技施設
石屋友美	IsA	所属タレント
トヨタ・タコマ	IsA	トヨタの車種
中ノ口川	IsA	新潟県の河川
安川悦子	IsA	ゲスト
IMB医療事務スクール	IsA	スクール
せんず	IsA	アイテム
中居ちはる	IsA	ワープエンタテインメント女優
石田淳也	IsA	育成選手
田中寛人	IsA	アナウンサー・キャスター
安国寺	IsA	臨済宗妙心寺派の寺院
モグラング	IsA	キャラクター
馬越幸子	IsA	日本の歌手

井沢まさみ	IsA	日本のイラストレーター
カール・ラデック	IsA	大粛清犠牲者
西田川郡西郷村	IsA	山形県の廃止市町村
フリッツ・マハループ	IsA	オーストリア系アメリカ人
久下	IsA	登場人物
壬生忠岑/源重之	IsA	歌人
ARIEL	IsA	作品
西尾理弘	IsA	東京外国語大学の人物
仙台シティエフエム	IsA	この番組が聴取できたコミュニティFM局
池内敏	IsA	存命人物
ジェニー・ローズ	IsA	覆面レスラーとして活動していたプロレスラー
デイヴィッド・レクタ・ルディシャ	IsA	ケニアのオリンピック金メダ
亀山佳明	IsA	日本の社会学者
後肢太陰脾経	IsA	経絡
A・J・エリス	IsA	過去の主な所属選手
アレックス・カブレラ	IsA	野球選手
緑丘B遺跡出土物	IsA	町の文化財
Tapponia_superba	IsA	ササグモ科の属種
ジローラモ・ディルータ	IsA	作曲家
天文図書室	IsA	理学系研究科・理学部図書室
ヨハン・ローゼンミュラー	IsA	ドイツの作曲家
張富士夫	IsA	日本の実業家
ファイターズスタジアム	IsA	日本プロ野球の本拠地野球場
日別朝夕大御饌祭	IsA	祭事
松村明	IsA	辞典編纂者
今夜もシャララ_ぽっぷるfeeling	IsA	ラジオ番組
駒澤大学	IsA	新制大学
富樫あずさ	IsA	芸能人
白井吉治	IsA	本作品に登場する著名なゴルファー
ジョン・モウ	IsA	イギリスの地質学者
河内淳一	IsA	日本の作曲家
梁山駅	IsA	慶尚南道の鉄道駅
ザ・ワン	IsA	ドキュメンタリー・ビデオ
二宮清純	IsA	ボクシング解説者
放浪記	IsA	成瀬巳喜男の監督映画
SF_Short_Films_〜仲良き事は良きことかな〜	IsA	映画
ヒイロ・ユイ	IsA	架空のテロ
大宮氏	IsA	大須賀党構成員

村松清美	IsA	登場人物
古馬主要戦	IsA	競走
天城山心中	IsA	自殺事件
ヒロト・アマギワ	IsA	登場人物
豊田通商ファイティングイーグルスの選手	IsA	NBDLの選手
ポーギー	IsA	スタージョン級原子力潜水艦
富谷銑太郎	IsA	法政大学の人物
政王	IsA	登場人物
美樹工業	IsA	ジャスダック上場企業
ベルナール・ギー	IsA	登場人物
大久保忠世	IsA	キャスト
カンヌ映画祭	IsA	公認映画祭
歌うボキャブラ天国	IsA	フジテレビ系番組
トコトン!サタデー	IsA	放送終了した番組
ハードオフ	IsA	出店しているテナント
石田敦子	IsA	準レギュラー
長野雅弘	IsA	存命人物
田原伸吾	IsA	野球指導者
各国の漫画	IsA	漫画
国家情報センター	IsA	中華人民共和国の企業
法相宗	IsA	門跡寺院
管野秀夫	IsA	日本の写真家
ファロットの休日	IsA	作品
洗谷	IsA	ジャジメントグループ関係者
眠る前のテレパシー	IsA	歌
京セラコミュニケーションシステム	IsA	入居企業
犬神サーカス団	IsA	出場バンド・アーティスト
松法港南防波堤灯台	IsA	北海道の灯台
ニライカナイからの手紙	IsA	作品
殉国七士廟	IsA	A級戦犯
ア゛ーくん	IsA	登場人物
新里門駅	IsA	駅
サラトフ州	IsA	採用しているロシアの地域
赤ずきんチャチャ	IsA	過去の主な提供番組
Cubase_VST_5.0_Score	IsA	過去に販売されたグレード
ヤマハミュージックメディア	IsA	配信出版社
VOICE_CREW	IsA	過去の主なネット番組
大沢次郎左衛門	IsA	キャスト
ハイテク選書ワイド	IsA	レーベル

埼玉県防災航空隊	IsA	組織
ガブリエラ・パルッツィ	IsA	出身者
阿島陣屋	IsA	現存しない長野県の建築物
宇宙戦艦ヤマト	IsA	アニメ映画
シャコタンの唄	IsA	シングル
英雄広場駅	IsA	駅
埼玉県立大宮工業高等学校	IsA	埼玉県高等学校
屋島の戦い	IsA	戦い
ケシカスくん	IsA	擬人化キャラクターを主人公にした物語
韓国正教会	IsA	コンスタンティノープル総主教座との一致にある教会
石絵未季	IsA	日本のAV女優
The_John_Doe_Associates	IsA	著書
奇術	IsA	登場人物
ノーブルズ郡	IsA	郡
サンマー・ホリデイ_Summer_Holiday	IsA	監督作品
Endless_Rain	IsA	収録曲
山崎真	IsA	日本のサッカー選手
ニューヨーク	IsA	アメリカ合衆国のドラマ映画
エル・アンヘル	IsA	存命人物
ふじかわ農業協同組合	IsA	山梨県の農業協同組合
工藤晶	IsA	選手として登場する人物
非情のライセンス	IsA	腸捻転解消でMBSから移行したネット番組
ユリアン・ロイス	IsA	世界陸上選手権ドイツ代表選手
スティーブ・ガッド	IsA	フュージョンの主なアーティスト
島村俊治	IsA	NHKの元職員アナウンサー
権田二毛作	IsA	キャスト
山口県版	IsA	かつて存在していた地域版
ポール・ホーガン	IsA	オーストラリアの俳優
アクトベ	IsA	カザフスタンの都市
夜のおもちゃ・ぜんじろげ!	IsA	ラジオ番組
Attacked_kuma3	IsA	作品
成美那	IsA	対戦型格闘ゲームのキャラクター
伍子胥	IsA	作品
カルヴァート	IsA	兵員輸送艦
台北市を舞台とした映画作品	IsA	台北市を舞台とした作品
ラングレイ	IsA	人間
武市英雄	IsA	出身者

寝取られファイター_ヤリっちんぐ!_ROUND	IsA	ゲーム
水谷氏	IsA	日本の氏族
相模原市立広田小学校	IsA	神奈川県小学校
本庄家系譜	IsA	家系譜
ボーク	IsA	メインベルトの小惑星
コドモ警察	IsA	テレビドラマ
振聴	IsA	麻雀用語
NINETY	IsA	コナミのゲームソフト
リチャード・フロリダ	IsA	社会学者
神代薬品	IsA	かつて存在した日本の医薬品卸売企業
週末の部屋シリーズ_2_真夜中の部屋で	IsA	出演作品
湯河原万葉公園	IsA	神奈川県の公園
渡竹	IsA	ハイスクール!奇面組の登場人物
江波戸さんちのにぎやかな正月	IsA	出演作品
老老介護	IsA	介護
真崎猟	IsA	登場人物
クリストフ・ガンズ	IsA	存命人物
スレイマン・ディアワラ	IsA	OGCニースの選手
あなたも時給___万円	IsA	法律とは関係ないコーナー
ミカド	IsA	住宅機器メーカー
エクトル・スカローネ	IsA	モンテビデオ出身の人物
佐々十郎	IsA	東京都出身の人物
清水紘治	IsA	京都市出身の人物
新医療創造支援本部	IsA	附属機関
クンワル・インドラジット・シン	IsA	ネパールの首相
土曜ワイド劇場_/_火災調査官・紅蓮次郎_完全密室無人出火のトリック	IsA	出演作品
紅樹林駅	IsA	台北捷運の鉄道駅
ラプター	IsA	登場人物
全日本テニス選手権	IsA	NHK教育テレビ番組
荒川亮	IsA	東京都出身の人物
とぎ汁	IsA	米加工品
九段会館	IsA	東京都区部中部の観光地
オーケストラの姉妹	IsA	作品
貨幣小博物館	IsA	名所
やっぱりヤンチャー	IsA	道徳・生活指導をテーマとした番組
能代清	IsA	数学者
吉田有希	IsA	過去のワンエイトプロモーション所属者

High_School_Girl	IsA	作品
カール・ベーム	IsA	人物
スコア	IsA	モントリオールを舞台とした作品
稲沢市立山崎小学校	IsA	小学校
小杉町ケーブルテレビ	IsA	ケーブルテレビ局
マメロのオキテ	IsA	過去のコーナー
小嶋由紀代	IsA	存命人物
成田自動車教習所	IsA	日本の自動車教習所
藤原頼通	IsA	平安時代の歌人
ビリー・ミルズ	IsA	オリンピック陸上競技アメリカ合衆国代表選手
美園上八雲神社	IsA	神社
奥野谷浜駅	IsA	日本の鉄道駅
西品治団地	IsA	中国地方の住宅団地
玉川聖学院中等部・高等部	IsA	日本のプロテスタント系高等学校
吉沢春水	IsA	東京都出身の人物
ギルツ・イェーカブソンス	IsA	存命人物
佐伯田公	IsA	キャスト
遼寧省鞍山市	IsA	行政区画
WAITING_FOR_YOUR_LOVE	IsA	収録曲
常磐大学幼稚園	IsA	私立幼稚園
萩原信二	IsA	ナレーター
クラウス・アロフス	IsA	ヴェルダー・ブレーメンの選手
北旭川駅	IsA	境界駅
チェロとオーケストラのためのレガティシモ	IsA	オーケストラ曲
国分盛重	IsA	独眼竜政宗の登場人物
琴ノ浦温山荘庭園	IsA	名勝
祭文	IsA	演芸
岩井俊二	IsA	日本の作詞家
第二次昭和切手	IsA	日本の普通切手
四角い宇宙で待ってるよ	IsA	曲名
地域医療機能推進機構大阪病院	IsA	大阪府の医療機関
魔界水滸伝	IsA	日本のSF小説
よゐこ	IsA	交流のある人物
森岡貞香	IsA	島根県出身の人物
上杉純雄	IsA	日本の実業家
征矢千鶴	IsA	日本のタレント
桜井郁子	IsA	女性
佐久市立中佐都小学校	IsA	小学校

ルース台風	IsA	日本の台風
鹿川菖蒲	IsA	登場人物
宮崎充登	IsA	出身者
横須賀弾薬整備補給所	IsA	配置部隊・機関
大垣市立赤坂中学校	IsA	岐阜県中学校
三浦亜沙妃	IsA	MOODYZ女優
町立図書館	IsA	教育施設
フェリ・カフラック	IsA	ウィーン出身の人物
トーントン寝台車火災事故	IsA	イギリスの鉄道事故
飛田穂洲	IsA	ゆかりの人物
劇場版	IsA	タイムトラベルを題材としたアニメ映画
プリンセサ・サンディー	IsA	来日外国人選手
ゴットン	IsA	登場キャラクター
安江良介	IsA	日本の雑誌編集者
東神奈川車掌区	IsA	東日本旅客鉄道の車掌区
ライブSE	IsA	作品
セルジオ・キャロリーノ	IsA	チューバ奏者
建みさと	IsA	テレビCM出演者
高良鉄夫	IsA	沖縄県出身の人物
アンディ・リード	IsA	登場人物
唐人町商店街	IsA	日本の商店街
西川勉	IsA	広島大学附属中学校・高等学校の人物
ブラックバーン・ファイアブランド	IsA	登場する航空機
池田草庵	IsA	幕末の人物
キャサリン・ハーディ	IsA	ジョージア州の人物
稲敷市立高田小学校	IsA	茨城県小学校
松岡みゆき	IsA	神奈川県出身の人物
多村仁志	IsA	プロ野球選手
キャット・モンロー	IsA	登場キャラクター
山原和敏	IsA	台湾職業棒球大聯盟の選手
埼玉県立八潮南高等学校	IsA	高等学校
le_joli_monde_/_ジョリー・モンド	IsA	ファッションブランド
ふと気付けば	IsA	収録曲
夜光運河	IsA	川崎区の運河
クレアリデル・ヴァナント・シャウラ	IsA	登場人物
トルコの歌手	IsA	各国の歌手
アトリエ	IsA	オリジナルアルバム
千部山真教寺	IsA	名所・旧跡・景勝地
L'altro_giardino	IsA	作品

ポリメタクリレート樹脂溶液系接着剤	IsA	有機系接着剤
青春ミステリ	IsA	日本独自のサブジャンル
竹田光訓	IsA	サムスン・ライオンズの選手
アーロン・ブルックス	IsA	アメリカ合衆国の野球選手
茶臼山	IsA	山岳
イヴァン・フィッシャー	IsA	ウィーン国立音楽大学出身の人物
スーパーどどん波	IsA	ドラゴンボールの技
共和政ローマ	IsA	ローマの歴史
大蛇飛駆寒虫	IsA	登場人物
岩嶋雅奈未	IsA	かつて所属していたタレント
ぱるエンタープライズ	IsA	芸能プロダクション
広東レオパーズ	IsA	中国のプロ野球チーム
天野修	IsA	明海大学の教員
趙公明	IsA	神
熊本県立牛深高等学校	IsA	熊本県高等学校
小黒八七郎	IsA	長岡市出身の著名人
転送	IsA	琉球諸島及び大東諸島に関する日本国とアメリカ合衆国との間の協定
NACK5	IsA	出演作品
アルトサクソフォン協奏曲	IsA	作品
平田敬作	IsA	キャスト
川本進	IsA	人
藤沢周平傑作選	IsA	過去に放送した番組
真田アサミ	IsA	過去のTABプロダクション所属者
グリーリー・エステイツ	IsA	ポスト・ハードコア・バンド
ジョン・ウィンダム	IsA	SF作家
美雪沙織のお熱いのがお好き!	IsA	出演作品
龍谷中学校・高等学校	IsA	佐賀県の私立高等学校
林剛史	IsA	神戸市出身の人物
白河市立釜子小学校	IsA	福島県小学校
蔵治光一郎	IsA	武蔵中学校・高等学校の人物
D.C.S.S._〜ダ・カーポ_セカンドシーズン〜_外伝ドラマ_Vol.3_魔法使いの足跡	IsA	ドラマCD
浅田彰	IsA	ゆかりのある人物
ズリーナ・ムニョス	IsA	存命人物
軍事学者	IsA	分野別の学者
菅原一樹	IsA	弟子
星乃けい	IsA	北九州市出身の人物

灰とダイヤモンド	IsA	エクスタシーレコードのアルバム
スプルーアンス級駆逐艦	IsA	水上戦闘艦
七飯郵便局	IsA	北海道の郵便局
恩納バイパス	IsA	バイパス
アークエとガッチンポーてんこもり	IsA	広報部時代に宣伝を出掛けていた番組
ヤスハラケミカル	IsA	当地域に本社を置く主な企業

Appendix B - AtLocation assertions evaluation set

Proposed method evaluation set

網走市	AtLocation	常呂町
大久保熨斗吉商店	AtLocation	つくばみらい市
下益城中学校	AtLocation	隈庄町
高千穂峰	AtLocation	霧島町
静岡弁	AtLocation	静岡県
八幡小学校	AtLocation	山梨市
美山支所	AtLocation	南丹市
黒川明神山	AtLocation	世羅町
大津市立富士見小学校	AtLocation	大津市
ちょうちん岩	AtLocation	湯沢町
愛知県立美和高等学校	AtLocation	美和町
サウスワース	AtLocation	キトサップ郡
内船郵便局	AtLocation	南部町
帖佐小学校	AtLocation	姶良市
長崎県立佐世保東翔高等学校	AtLocation	佐世保市
下毛郡	AtLocation	豊前国
日本映画学校	AtLocation	川崎市
土佐山田町立平山小学校	AtLocation	土佐山田町
玉城町立田丸小学校	AtLocation	玉城町
売木村立売木中学校	AtLocation	売木村
舌鼓	AtLocation	山口市
鎮国守国神社	AtLocation	桑名市
アタック!!	AtLocation	掛川市
北消防署筑波分署	AtLocation	つくば市
藍葉	AtLocation	久喜市
長野県商工新聞	AtLocation	長野県
ラドラ川	AtLocation	ルーゴ県
鹿児島県立大島高等学校	AtLocation	奄美市
トマト銀行	AtLocation	岡山市
瑞穂市図書館	AtLocation	瑞穂市
牛久市立牛久第一中学校	AtLocation	牛久市
朱円簡易郵便局	AtLocation	斜里町
野田市役所	AtLocation	野田市
太田尾保育園	AtLocation	西海市
キャンディ国立博物館	AtLocation	キャンディ
ハニムーン島州立公園	AtLocation	ピネラス郡

彦根市立稲枝東小学校	AtLocation	彦根市
東京線	AtLocation	富山市
マリーナ大通り	AtLocation	ア・コルーニャ
星宮神社	AtLocation	美並村
ウォーレン森林局空港	AtLocation	アイダホ郡
半熟革命	AtLocation	浦安市
鳥取県済生会境港総合病院	AtLocation	境港市
美馬市立岩倉小学校	AtLocation	美馬市
桜川村立浮島小学校	AtLocation	桜川村
遠賀郡	AtLocation	遠賀町
キミとボクとエデンの林檎	AtLocation	神戸市
川西町立千手小学校	AtLocation	川西町
長沼健太郎	AtLocation	高岡市
山交バス	AtLocation	朝日町
みらさか竹工房はなかご	AtLocation	三良坂町
紋散透鐔_金象嵌銘_林又七	AtLocation	八代市
石川豊人	AtLocation	越中国
西部児童館	AtLocation	御津町
芦北海浜総合公園	AtLocation	芦北町
梅美台公園	AtLocation	木津川市
大島新田多目的グラウンド	AtLocation	杉戸町
新堤自然公園	AtLocation	栗原市
上川郡	AtLocation	和寒町
沼津市立第二中学校	AtLocation	沼津市
Emerson	AtLocation	コロンビア郡
宮山古墳群	AtLocation	浜島町
胆振地方男女平等参画センター	AtLocation	室蘭市
川北神社の赤松	AtLocation	標津町
脇本海水浴場	AtLocation	阿久根市
水木	AtLocation	多賀町
あさひ保育園	AtLocation	和泉市
一般県道	AtLocation	海南町
徳島市立富田保育所	AtLocation	徳島市
薬師堂	AtLocation	大戸村
佐々木敏夫	AtLocation	豊後高田市
岡山県立倉敷琴浦高等支援学校	AtLocation	倉敷市
ホテル加登屋	AtLocation	栃木市
大浜海水浴場	AtLocation	壱岐市
館林市立長良保育園	AtLocation	館林市
布袋町立布袋中学校	AtLocation	布袋町

広戸川	AtLocation	広戸村
山之口小学校	AtLocation	山之口町
サンディフック	AtLocation	エリオット郡
柳原河川敷運動場	AtLocation	栃木市
青森市立筒井南小学校	AtLocation	青森市
北海信用金庫倶知安支店	AtLocation	倶知安町
会津若松市立門田小学校	AtLocation	会津若松市
豊後大野市役所	AtLocation	豊後大野市
第一石鹸株式会社	AtLocation	板倉町
小山市立間々田東小学校	AtLocation	小山市
花石郵便局	AtLocation	今金町
南房総市国保富山病院	AtLocation	南房総市
一般国道	AtLocation	総社市
こもれびに揺れる魂のこえ	AtLocation	神戸市
池月公園	AtLocation	美馬町
赤丸城跡	AtLocation	高岡市
ねむのき幼稚園	AtLocation	青梅市
Spruce_Hill_Township	AtLocation	ダグラス郡
大神幼稚園	AtLocation	日出町
一宮町立染河内小学校	AtLocation	一宮町
湯の里雪祭り・百八灯	AtLocation	魚沼市
早良町立早良中学校	AtLocation	早良町
舟山島	AtLocation	定海区
牡蛎	AtLocation	湖西市

Baseline method evaluation set

火	AtLocation	お参り
清水寺	AtLocation	京都
シャワー	AtLocation	海
木	AtLocation	温泉
バスタオル	AtLocation	お風呂場
トマト	AtLocation	食料品店
搭乗券	AtLocation	空港
弓	AtLocation	射的
ぬいぐるみ	AtLocation	ベッド
社会人	AtLocation	会場
砂利	AtLocation	公園
太陽	AtLocation	砂浜

長椅子	AtLocation	停留所
太陽	AtLocation	海辺
ろうそく	AtLocation	墓場
影	AtLocation	外
防波堤	AtLocation	岸
ミルク	AtLocation	ポット
工具	AtLocation	学校
骨	AtLocation	牛
ライト	AtLocation	コンサート
テーブル	AtLocation	フランス料理店
海溝	AtLocation	水中
おつまみ	AtLocation	食品店
火	AtLocation	葬式
園芸用品	AtLocation	ショップ
ドライヤー	AtLocation	ふろ場
車	AtLocation	銀行
救命胴衣	AtLocation	海
木材	AtLocation	カンセキ
テープ	AtLocation	美術館
オルガン	AtLocation	音楽
イス	AtLocation	レストラン
記念ボール	AtLocation	野球場
花	AtLocation	庭園
俳優	AtLocation	歌舞伎座
子犬	AtLocation	車の中
テーブル	AtLocation	食堂
扉	AtLocation	玄関
肝臓	AtLocation	胃
衣類	AtLocation	納戸
レシピ	AtLocation	戸棚の中
非常階段	AtLocation	玄関
搭乗券	AtLocation	飛行場
中落ち	AtLocation	肉
地震	AtLocation	樹海
園芸用品	AtLocation	東急ハンズ
バスタオル	AtLocation	バス
住む家のない人	AtLocation	河原
一本締め	AtLocation	パーティー
フォークリフト	AtLocation	瓶
叫び	AtLocation	オペラ

どら焼き	AtLocation	戸棚の中
牛乳瓶	AtLocation	道の駅
かみそり	AtLocation	はげ
肝臓	AtLocation	体内
コピーバンド	AtLocation	花見会場
灰皿	AtLocation	テーブルの上
救急車	AtLocation	病院
蜘蛛	AtLocation	ディズニーランド
デジタルカメラ	AtLocation	ビックカメラ
法被	AtLocation	お祭り
歯模型	AtLocation	クリニック
フトン	AtLocation	寝室
半纏	AtLocation	おみこし
虎	AtLocation	動物園
おたま	AtLocation	台所
出雲大社	AtLocation	山陰
柵	AtLocation	動物園
ほうちょう	AtLocation	台所
電車の駅	AtLocation	ヨーロッパ
使い捨てカメラ	AtLocation	コンビニ
信仰	AtLocation	それぞれの心の中
座敷	AtLocation	和室
和服	AtLocation	押し入れ
テント	AtLocation	キャンプ場
りんごの木	AtLocation	庭園
ホッチキス	AtLocation	アルミニウム
浮浪者	AtLocation	河川敷
ポプラ並木	AtLocation	旭川
花火	AtLocation	浜辺
マイク	AtLocation	スタンド
灰皿	AtLocation	ラジオ局
お墓	AtLocation	寺院
マグマ	AtLocation	火山の近く
救命胴衣	AtLocation	ボート
食卓	AtLocation	サンタ
照明	AtLocation	体育館
ブラックバス	AtLocation	湖
エアクリーナー	AtLocation	航空機
カルテ	AtLocation	診療所
クジラ	AtLocation	海

ガードマン	AtLocation	通り
オックスフォード	AtLocation	ボストン
お線香	AtLocation	仏壇
ベット	AtLocation	寝室
美術館	AtLocation	ニューヨーク
首相	AtLocation	内閣
ジュース	AtLocation	商店街
太鼓	AtLocation	盆踊り

Appendix C - LocatedNear assertions evaluation set

美瑛町	LocatedNear	石狩国上川郡
大越町	LocatedNear	小野町
井芹川	LocatedNear	木葉川
八千代市立村上東中学校	LocatedNear	八千代松陰高等学校
高千穂町	LocatedNear	豊後大野市
鉾田市	LocatedNear	小美玉市
喜多方市	LocatedNear	会津若松市
イオンモール成田	LocatedNear	サイクルベースあさひ
上平村	LocatedNear	桶川町
うきは市	LocatedNear	八女市
池上運河	LocatedNear	南渡田運河
高森工業団地	LocatedNear	平工業団地
大木町	LocatedNear	久留米市
八郷町	LocatedNear	石岡市
三和町	LocatedNear	境町
サロマ湖畔ユースホステル	LocatedNear	小清水はなことりの宿ユースホステル
小美玉市	LocatedNear	笠間市
彦根市	LocatedNear	東近江市
軽井沢町	LocatedNear	佐久市
海上町	LocatedNear	干潟町
サホロユースホステル	LocatedNear	帯広八千代ユースホステル
城里町	LocatedNear	笠間市
岡崎市	LocatedNear	額田郡幸田町
小諸市	LocatedNear	東御市
白井市	LocatedNear	船橋市
山方町	LocatedNear	金砂郷町
岡山南警察署	LocatedNear	倉敷警察署
鬼石町	LocatedNear	藤岡市
天龍村	LocatedNear	浜松市
日高川町	LocatedNear	有田郡
板垣峠	LocatedNear	巡見峠
見附市	LocatedNear	三条市
新江戸川公園	LocatedNear	講談社野間記念館
大黒運河	LocatedNear	恵比須運河
三和区	LocatedNear	津有区
佐敷町	LocatedNear	玉城村
麦草峠	LocatedNear	丸山展望台
ハーバーランド駅	LocatedNear	JR神戸線
志布志市	LocatedNear	曽於郡大崎町

横浜紅葉坂 LocatedNear 神奈川県立青少年センター
筑紫野市 LocatedNear 筑紫郡那珂川町
黒部市 LocatedNear 下新川郡
奥州市 LocatedNear 和賀郡西和賀町
田村市 LocatedNear 葛尾村
豊富町 LocatedNear 稚内市
つくばみらい市 LocatedNear 守谷市
国立病院機構旭川医療センター LocatedNear 旭川市総合体育館
雲南市 LocatedNear 安来市
ダイティング LocatedNear モンハイム
奥千丈岳 LocatedNear 国師岳
雲取山 LocatedNear 唐松尾山
佐野プレミアム・アウトレット LocatedNear イオンモール佐野新都市
滑川市 LocatedNear 富山市
埼玉県立所沢西高等学校 LocatedNear 国立病院機構西埼玉中央病院
釧路郡 LocatedNear 標茶町
原市町 LocatedNear 春岡村
武雄市 LocatedNear 嬉野市
枝幸郡 LocatedNear 中頓別町
東白川郡棚倉町 LocatedNear 大子町
勝浦郡生比奈村 LocatedNear 多家良村
名寄市 LocatedNear 美深町
南天山 LocatedNear 両神山
東白川郡棚倉町 LocatedNear 矢祭町
日田おおやまユースホステル LocatedNear 湯布院カントリーロードユースホステル
酉谷山 LocatedNear 熊倉山
佐賀市 LocatedNear 小城市
読谷村 LocatedNear 恩納村
金沢中警察署 LocatedNear 南砺警察署
近江希望が丘ユースホステル LocatedNear ユースホステル和迩浜青年会館
伊達家霊廟 LocatedNear 有備館
鹿島町 LocatedNear 大野村
大田区立東蒲小学校 LocatedNear 聖跡蒲田梅屋敷公園
東京ファッションタウン LocatedNear パレットタウン
道中庵ユースホステル LocatedNear パイラ松島・奥松島ユースホステル
一関市 LocatedNear 栗原市
海部郡 LocatedNear 海陽町
大品山 LocatedNear 鍬崎山
大島町 LocatedNear 小杉町
足利市 LocatedNear 桐生市

対馬西山寺ユースホステル LocatedNear 武雄温泉ユースホステル
雲取山 LocatedNear 白岩山
ケーブルネットワーク淡路 LocatedNear テレビ鳴門
帯広八千代ユースホステル LocatedNear 池田北のコタンユースホステル
甲斐駒ヶ岳 LocatedNear アサヨ峰
石狩振興局 LocatedNear 長沼町
館山市 LocatedNear 千倉町
五所川原市 LocatedNear 外ヶ浜町
東庄町 LocatedNear 干潟町
郡上警察署 LocatedNear 大野警察署
網走郡 LocatedNear 弟子屈町
鷺洲上公園 LocatedNear シティタワー大阪福島
三次ユースホステル LocatedNear 岩国ユースホステル
東白川郡鮫川村 LocatedNear 浅川町
留萌自動車学校 LocatedNear 留萌市立病院
牧区 LocatedNear 三和区
大田区 LocatedNear 川崎区
東京国際空港 LocatedNear 木更津市
小野町 LocatedNear 田村市
塩谷郡高根沢町 LocatedNear 芳賀町
和泊町 LocatedNear 知名町

Appendix D - CreatedBy assertions evaluation set

スターライト	CreatedBy	加藤正人
I_am_GHOST	CreatedBy	松井寛
ブン・ブン・ブン・ブン!!	CreatedBy	ベンガボーイズ
とある魔術の禁書目録	CreatedBy	柳沢テツヤ
獅子の血脈	CreatedBy	小沢仁志
月刊少女野崎くん	CreatedBy	中嶋敦子
冠二郎冠Revolution	CreatedBy	蜂須賀健太郎
NEVER_GIVE_UP!_〜TO_WORLD〜	CreatedBy	真理絵
The_Point_of_View	CreatedBy	アラン・クロスランド
エノケンの魔術師	CreatedBy	木村荘十二
ダーク・ホース	CreatedBy	ジョージ・ハリスン
Father_Fucker_＃3	CreatedBy	冴樹高雄
ファイアスターター	CreatedBy	スティーヴン・キング
雁_滑稽堂_明治後期	CreatedBy	小原古邨
耳助漫遊記	CreatedBy	田中正雄
魔大陸の鷹	CreatedBy	赤城毅
ラヴ・パレード	CreatedBy	ジェフ・スコット・ソート
太陽の勇者ファイバード	CreatedBy	直井正博
ロード88	CreatedBy	富田靖子
恋にいのちを	CreatedBy	神山繁
ハイスクール!奇面組	CreatedBy	中村隆太郎
Dr.とウェディングベル♥	CreatedBy	南原兼
なつみん	CreatedBy	マツリセイシロウ
まぬけなオオカミ	CreatedBy	マイケル・ラー
マジック・タッチ	CreatedBy	マイケル・ホイ
青い山脈	CreatedBy	吉永小百合
沈める雪	CreatedBy	安部慎一
In_the_Tracks_of_Maurice_Jarre	CreatedBy	ジャン＝ピエール・モッキー
ナオミに捧ぐ――愛も汚辱のうちに	CreatedBy	佐藤友哉
つるピカハゲ丸くん	CreatedBy	善聡一郎
ドラえもん_新・のび太と鉄人兵団_〜はばたけ_天使たち〜	CreatedBy	伊東伸高
スナオになりたいね／種ともこ	CreatedBy	草野マサムネ
絵そらごと	CreatedBy	相見とし子
必殺仕事人V・激闘編	CreatedBy	京本政樹
まくら泥棒	CreatedBy	ねむようこ
風	CreatedBy	久保田光太郎
アグネス・チャンのビューティーフード	CreatedBy	アグネス・チャン

中居正広の金曜日のスマたちへ	CreatedBy	角田陽一郎
就職戦線異状なし	CreatedBy	坂元裕二
アイドル天使ようこそようこ	CreatedBy	首藤剛志
火うち箱	CreatedBy	ハンス・クリスチャン・アンデルセン
隠忍術SHINOBI	CreatedBy	小沢和義
わずかいっちょまえ	CreatedBy	星里もちる
マーティン・GPCPA1	CreatedBy	高橋優
ネフィリムセシルの目	CreatedBy	伊藤結花理
地獄歌占	CreatedBy	みとせのりこ
うえきの法則	CreatedBy	湖山禎崇
ロト6_スペシャルライブ_This_is_秋の音楽祭!	CreatedBy	中西忠司
かいけつゾロリ	CreatedBy	山田悦司
めらんこりい白書	CreatedBy	阿久悠
愛媛県立新居浜西高等学校校歌	CreatedBy	岡本敏明
メダロット	CreatedBy	橋本昌和
ど根性ガエル	CreatedBy	福冨博
浅田弘幸画集	CreatedBy	浅田弘幸
私は泣かない	CreatedBy	神佑輔
元気なブロークン・ハート	CreatedBy	松本隆
鬼神童子ZENKI	CreatedBy	林明美
素敵な気分De!	CreatedBy	榊原郁恵
題名のない映画_Film_ohne_Titel	CreatedBy	ヒルデガルト・クネフ
When_the_Birds_Fly_South	CreatedBy	スタントン・A・コブレンツ
世界一受けたい授業	CreatedBy	ローラ
スパイダーマン	CreatedBy	西沢利明
Jean	CreatedBy	インコグニート
一私小説書きの日乗	CreatedBy	西村賢太
ねずみ物語	CreatedBy	犬山イヌコ
レーシング小僧_嵐	CreatedBy	池沢早人師
旅立ちの日に_詞	CreatedBy	松井孝夫
龍ヶ嬢七々々の埋蔵金	CreatedBy	花澤香菜
春の日	CreatedBy	渡辺拓也
おねがいマイメロディ	CreatedBy	金崎貴臣
Ellington	CreatedBy	デューク・エリントン
青蛾館２５周年記念	CreatedBy	長田育恵
ハミングバード情報局	CreatedBy	椎名へきる
スイングゴルフ_パンヤ_2ndショット!	CreatedBy	釘宮理恵
A_Song_for_Shelter_/_Ya_Mama	CreatedBy	ファットボーイ・スリム

浮気人間絵図	CreatedBy	山口洋子
少年Gメン3_秘密情報	CreatedBy	藤田茂
雨にぬれても	CreatedBy	村松真理
フルメタル・パニック!	CreatedBy	山田尚子
The_Second_Raid		
グラゼニ	CreatedBy	アダチケイジ
さらば宇宙戦艦ヤマト	CreatedBy	白土武
すくぅ〜る・らぶっ!3〜未来へのアレグレット〜	CreatedBy	小倉結衣
Human	CreatedBy	ニコルソン・ベイカー
同級生	CreatedBy	鹿目けい子
あたしンち	CreatedBy	折笠富美子
ボスコニアン	CreatedBy	岩谷徹
IMALU	CreatedBy	村山☆潤
復讐の掟_ROGUE_COP	CreatedBy	大野雄二
いい旅・夢気分	CreatedBy	わぐりたかし
がんばれわんにゃんお助け隊!!	CreatedBy	なりゆきわかこ
ジャングルDEいこう!	CreatedBy	坂本佳栄子
Maceo_Parker/Roots_Revisited	CreatedBy	ドン・プーレン
がんばれ!!タブチくん!!	CreatedBy	小林治
ボン・ヴォヤージュ_Bon_Voyage	CreatedBy	ティエリー・アルボガスト
Kyoto_Jazz_Massive	CreatedBy	菱山正太
となグラ!	CreatedBy	明田川仁
ショート寸前！	CreatedBy	桜井雪
KING_OR_CURSE	CreatedBy	坂本裕次郎
色即ぜねれいしょん	CreatedBy	山本浩司
風立ちぬ	CreatedBy	ヴェルナー・ヘルツォーク

Appendix E - MemberOf assertions evaluation set

本間昭光	MemberOf	ポルノグラフィティ
二之湯智	MemberOf	自民党国際人材議員連盟
舟岡昭浩	MemberOf	ザ・やなせふなおか
AMIYA	MemberOf	SUPER_DROP_BABIES
平原四郎	MemberOf	地域・生活者起点で日本を洗濯
HIRO	MemberOf	トウスケプロジェクト
レイ・クレスポ_Rey_Crespo	MemberOf	スカ・クバーノ
テリー・チャイムズ	MemberOf	白い暴動
UMEKEN	MemberOf	BAGDAD_CAFE_THE_trench_town
みうらじゅん	MemberOf	大島渚
ヴィクトル・ヘムグレン	MemberOf	コンストラクデッド
スコット・メルカド	MemberOf	キャンドルボックス
Samoth	MemberOf	エンペラー
レブ・ビーチ	MemberOf	ウィンガー
ヤコブ・ロゴルト_Jacob_Krogholt	MemberOf	ウィザリング・サーフェイスの解散時
向井慧	MemberOf	ボーイフレンド
エール橋本	MemberOf	東京ペールワン
デイヴィッド・カヴァデール	MemberOf	ラスト・コンサート・イン・ジャパン
MOUSE_THE_PEACE_MC	MemberOf	Trio_The_Clock
マリリン・マックー	MemberOf	フィフス・ディメンション
ボビー・ロンデイネリ	MemberOf	闇からの一撃
ダーシャ・シャシナ	MemberOf	セレブロ
川上啓之	MemberOf	突然段ボール
遠峯ありさ	MemberOf	黒BUTAオールスターズ
越川弘志	MemberOf	ザ・カーナビーツ
Henning_Schmitz	MemberOf	クラフトワーク
金成公信	MemberOf	ギンナナ
タカシ	MemberOf	ミスター・カイト
ミーヤ	MemberOf	レ・ロマネスク
高嶺格	MemberOf	ダムタイプ
杉山晋太郎	MemberOf	ザ・スターリン
スチュアート・キャシディ	MemberOf	Kバレエカンパニー
木津みずき	MemberOf	大笑点
シリル・ネヴィル_Cyril_Neville	MemberOf	ネヴィル・ブラザーズ
金澤直樹	MemberOf	かげぼうし
安西卓丸	MemberOf	ふくろうず
リエ	MemberOf	東京ブラススタイル
マーク・グリーンウェル	MemberOf	バルサゴス
Nils_Lindenhayn	MemberOf	ジ・オーシャン

マイケル・マドックス	MemberOf	キル・ハンナ
_Michael_Maddox		
Dir.F	MemberOf	水曜日のカンパネラ
R.Yanagihara	MemberOf	スピアメン
菊池篤	MemberOf	日高央
ブルーノ・バーベイ	MemberOf	マグナム・フォト
村田めぐみ	MemberOf	メロン記念日
アヤノ	MemberOf	ヴィドール
伊東裕扶子	MemberOf	川崎純情小町☆
花房里枝	MemberOf	さんみゅ〜
BAGI	MemberOf	ONE_TRACK_MIND
ロン・ストライカート	MemberOf	メン・アット・ワーク
Nat	MemberOf	クロマティックス
マリア・ガウフィン	MemberOf	ザイアフィング
渡部和正	MemberOf	ビビリ劇団
ロバート・ディレオ	MemberOf	ストーン・テンプル・パイロッツ
田中利幸	MemberOf	ステーキハウス
デヴィッド・ヒューズ	MemberOf	トラッシュキャン・シナトラズ
竹安堅一	MemberOf	フラワーカンパニーズ
松井正道	MemberOf	三叉路
花男	MemberOf	太陽族
研次郎	MemberOf	君が咲く山
ジョン・スワン	MemberOf	フラタニティ
Graham	MemberOf	カリフォルニア・ワイブス
Nakko	MemberOf	Osaka翔Gangs
西村明宏	MemberOf	日韓議員連盟
ヘンリー	MemberOf	スーパージュニア
ルドルフ・アドルファス・ジョーダン	MemberOf	サンディニスタ!
Loïc_Rossetti	MemberOf	ジ・オーシャン
トニ・ニュメリン	MemberOf	モルス・プリンシピアム・エスト
Risa	MemberOf	NKO☆Lovers
野津友那乃	MemberOf	さくら学院
三尾ケイジ	MemberOf	キャットフラメンコダンサーズ
Ryan	MemberOf	グループラヴ
長谷川閑史	MemberOf	TPP交渉への早期参加を求める国民会議
Dave_Quackenbush	MemberOf	ヴァンダルズ
山瀬功治	MemberOf	オシムジャパン
デイヴィッド・ドレイマン	MemberOf	デヴァイス
リチャード・オークス	MemberOf	スウェード

逢沢一郎	MemberOf	日韓議員連盟
七条好	MemberOf	あひる艦隊
エド	MemberOf	札幌スーパーギャグメッセンジャーズ
コーヘイ	MemberOf	トリッパー
永井路夫	MemberOf	すっとんトリオ
樋口豊	MemberOf	SWEET_STRANGE_LIVE
グッチ裕三	MemberOf	ビジーフォー
J.J.	MemberOf	テデスキ・トラックス・バンド
FUJIWARA	MemberOf	大笑点
John_Butler	MemberOf	ジョン・バトラー・トリオ
Simen	MemberOf	アークチュラス
筑波礼子	MemberOf	ハナ肇とクレージーキャッツ
ダン・スミス	MemberOf	ノイゼッツ
小沼達也	MemberOf	レベッカ
幸助	MemberOf	幸助・福助
トミー・ドリーマー	MemberOf	アライアンス
高橋栄徳	MemberOf	ラニアルズ
ジョージ・バビット	MemberOf	ザ・ベンチャーズ
MATCHLESS_DC	MemberOf	チャットモンチー
クラウス・クリーガー	MemberOf	タンジェリン・ドリーム
伊藤ツヨシ	MemberOf	THE_STREET_BEATS
鈴々木保香	MemberOf	バニーガール向上委員会
鈴木大輔	MemberOf	Day_after_tomorrow

Research Achievements

Journals

1. Marek Krawczyk, Rafal Rzepka and Kenji Araki: ”Extracting location and creator-related information from Wikipedia-based information-rich taxonomy for ConceptNet expansion”, *Knowledge- Based Systems*, 2016, Volume 108, pp. 125-131.

International Conferences

1. Marek Krawczyk, Rafal Rzepka and Kenji Araki: ”Populating ConceptNet knowledge base with Information Acquired from Japanese Wikipedia”, in *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC)* 2015, pp. 2985-2989.

2. Marek Krawczyk, Rafal Rzepka and Kenji Araki: ”Exploiting Wikipedia-based Information-rich Taxonomy for Extracting Location and Creator Related Information for ConceptNet Expansion”, in *Proceedings of the 7th Language and Technology Conference (LTC)*, 2015, pp. 383-387.

3. Marek Krawczyk, Rafal Rzepka, and Kenji Araki: ”Rule-based Approach to Extracting Location, Creator and Membership-related Information from Wikipedia-based Information-rich Taxonomy for ConceptNet Expansion”, in *Proceedings of Language on Computers IJCAI 2016 Workshop*, pp. 29-35.

National Conferences

1. Marek Krawczyk, Rafal Rzepka and Kenji Araki: ”Extracting ConceptNet Knowledge Triplets from Japanese Wikipedia”, in *Proceedings of The Twenty First Annual Meeting of The Association for Natural Language Processing (NLP 2015)*, 2015, pp. 1052-1055.

2. Marek Krawczyk, Yuki Urabe, Rafal Rzepka and Kenji Araki: ”A-dur: Action Duration Calculation System”, in *Technical Report of Language Engineering Community Meeting*, SIG-LSE-B301-7, 2013, pp.47-54.

Awards

1. BEST PAPER AWARD for:

   Marek Krawczyk, Rafal Rzepka and Kenji Araki: ”Exploiting Wikipedia-based Information-rich Taxonomy for Extracting Location and Creator Related Information for ConceptNet Expansion”, in *Proceedings of the 7th Language and Technology Conference (LTC)*, 2015, pp. 383-387.

Acknowledgments

I would like to sincerely thank my supervisor, Prof. Kenji Araki, from the Graduate School of Information Science and Technology, Hokkaido University, for the invaluable guidance in writing this thesis.

I would like to express my sincere gratitude to the members of the review committee, Prof. Tsuyoshi Yamamoto, Prof. Miki Haseyama and Prof. Yūji Sakamoto, from the Graduate School of Information Science and Technology, Hokkaido University, for their valuable insights and help in refining this thesis.

I would also like to sincerely thank Assistant Prof. Rafal Rzepka, from the Graduate School of Information Science and Technology, Hokkaido University, for the countless hours of assistance and fruitful discussion over the course of performing the work described in this thesis.

Furthermore, I would like to sincerely thank everyone at the Language Media Laboratory, Graduate School of Information Science and Technology, Hokkaido University for their invaluable support and assistance.

Finally, I would like to thank the Ministry of Education, Culture, Sports Science and Technology, Japan, for the opportunity to study in Japan on a government scholarship.