Extracting conceptual structures from multiple sources

Eduard Barbu

Center for Mind/Brain Sciences

Universita degli Studi di Trento
Advisor:prof. Massimo Poesio

A thesis submitted for the degree of

Doctor of Philosophy (PhD)

2009, December

1. Reviewer: Andrea Bozzi

2. Reviewer: Alessandro Lenci

3. Reviewer: Tanya Stivers

Day of the defense: January 18, 2010

Abstract

This thesis extracts conceptual structures from multiple sources: Wordnet, Web Corpora and Wikipedia. The conceptual structures extracted from Wordnet and Web Corpora are inspired by the feature norm effort in cognitive psychology. The conceptual structure extracted from Wikipedia makes the transition between feature norm structures and theory like structures. The main contribution of this thesis can be grouped in two categories:

1. **Novel methods for the extraction of conceptual structures**. More precisely, there are three new methods we developed:
   (a) **Conceptual structure extraction from Wordnet**. We devise a procedure for property extraction from Wordnet using the notion of semantic neighborhood. The procedure exploits the main relations organizing the nouns, the information in glosses and the inheritance of properties principle.
   (b) **Feature Norms like extraction from corpora**. We propose a method to acquire feature norm like structures from corpora using weakly supervised methods.
   (c) **Conceptual Structure from Wikipedia**. A novel unsupervised method for the extraction of conceptual structures from Wikipedia entries of similar concepts is put forward. The main idea we follow is that similar concepts (i.e. those classified under the same node in a taxonomy) are described in a comparable way in Wikipedia. Moreover, to understand the kind of information extracted from Wikipedia we annotate this knowledge with a set of property types.

2. **Evaluation**. Specifically, we evaluate Wordnet as a model of semantic memory and suggest the addition of new semantic relations. We also assess the properties extracted from all sources for a unified test set, in a clustering experiment.

Bunicilor, Marin si Lina, care mi-au luminat copilaria.

To my grandparents, Marin and Lina, who brought light into my childhood.

Ai miei nonni, Marin e Lina, che hanno illuminato la mia infanzia.

Acknowledgements

First I would like to thank to my adviser Massimo Poesio for the guidance and feedback during all stages of this thesis. The seminar he organized in the first year of my Ph.D. program entitled **Concepts in Mind and Language** was one of the best seminars I attended and easily the best seminar to date organized in CIMEC. Researchers in various fields ranging from Psychology to Artificial Intelligence, from Formal Ontology and Logic to Linguistics exchanged ideas about the nature and structure of concepts. This seminar had a great influence in choosing the subject of the thesis. It also allowed me to understand and appreciate different perspectives on conceptual representation from the one I was familiar with.

The second person who deserves my gratitude is my friend Verginica Barbu Mititelu. She gave feedback on the whole work and helped with data collections and evaluation. Besides my adviser she is the only person who has a detailed knowledge about this thesis.

Thirdly, aside from the discussions with my adviser, I benefited from conversations[1] with other members of the young CLIC research group: Marco Baroni, Brian Murphy and Federico Boschetti. The second part of the chapter 5 is the result of the collaboration with Claudio Giuliano and Lorenza Romano.

I also want to thank Gianluca Lebani who patiently annotated and rated the data in the chapter 6.

Last but not least, I like to thank my family who warmly supported and encouraged me during my PH.D. studies: my father Aron Barbu,

[1]Mainly during lunch hours.

my mother Constanta Barbu, my brother Ovidiu Barbu and his wife Maria Barbu and my grandmother Lina Barbu.

Contents

List of Figures

LIST OF FIGURES

List of Tables

1

Preliminaries

The work presented in this thesis except the effort in chapter 7 was published in the following articles: Barbu and Poesio (2008), Barbu (2009), Barbu (2008), Poesio *et al.* (2008) and Barbu and Poesio (2009).

1.1 A brief note on terminology

Terminology is always a problem. This thesis operates with concepts drawn from psychology and Natural Language Processing. Unfortunately, in Cognitive Psychology the properties are called features, which potentially clashes with the term feature as used in Machine Learning. Therefore a more accurate terminological choice for the psychologist would have been to baptize Feature Norms as Property Norms but, unfortunately, we cannot change this. Throughout this work we keep the term feature norm but call the psychologist feature: property. The term feature will be used in the Machine Learning sense. In this thesis the concepts will be typeface bold and properties typeface italic[1]. To make things clear consider the statement : A woman has an arm. We can say that there is a binary relation between the entities **woman** and **arm**, or that all the instances of the **woman** have the property *has arm.* Further the type of the property *has arm* is part. This is the reason why we call Wu and Barsalou taxonomy a taxonomy

[1]Of course bold and italic will be used for general term or definition emphasis as it is customary.

of property types. In a clustering task the property *has arm* can be a clustering feature.

1.2 The Road map of the thesis

For the best experience I advise the reader to read the whole thesis. However this brief chapter presentation is meant to help you choose the road map which best suits your interests:

1. Chapter 1: **Preliminaries** is this chapter. To avoid further misunderstandings I strongly advise the reader to read at least 1.1.

2. Chapter 2: **Introduction** introduces the psychological perspective on concepts and very briefly presents some work related with our own.

3. Chapter 3: **Feature Norms** shows in what way the feature norms are models of semantic memory and compare at the level of properties two norms.

4. Chapter 4: **Feature Norms and Wordnet** evaluates Wordnet as a model of semantic memory by comparing it with the two feature norms previously introduced. The output of this chapter is a new set of relations which the Wordnet builders can implement.

5. Chapter 5: **Extracting Feature Norm like Structures From Corpora** develops a method for extracting Feature Norm like structures from very large corpora. It also offers the preliminary results of a kernel based method for feature norm like property extraction.

6. Chapter 6: **Extracting Richer Knowledge Structures from Wikipedia** presents a new unsupervised method for extracting conceptual structures from Wikipedia.

7. Chapter 7: **The Knowledge Test Set** defines a test set for which we extract conceptual structures from Wordnet, Corpora and Wikipedia. The resulting properties are used in a clustering experiment.

8. Chapter : **Summary, Conclusions and Further Work** summarizes the thesis and emphasis the final conclusions.

9. Annexes. There are three annexes containing the properties extracted for **The Knowledge Test Set** from each source: Wordnet (annex 9.1), corpora(annex 9.2) and Wikipedia(annex 9.3).

The reader interested only in the computational aspect of the thesis can safely skip over the chapter 1. In particular, the chapter 6 can be read independently from the rest of the thesis.

The psychologist interested in the way linguistic resources can be used for extracting properties found in the norms should browse the main chapters (4, 5, 6) for practical examples.

1.3 Software used in the thesis

The software resources used in each chapter are the following:

1. **TreeTagger** it is a language independent POS tagger. Throughout this thesis we perform POS tagging and lemmatization using the English parameter file trained on the PENN treebank.

2. **WordNet::QueryData** it is a Perl API to wordnet databases. It is used in chapter 4.

3. **CWB**. According to the Web description: "The IMS Open Corpus Workbench (CWB) is a collection of tools for managing and querying large text corpora (100 M words and more) with linguistic annotations. Its central component is the flexible and efficient query processor CQP." CQP has many facilities, the mostly used in this thesis being: regular expressions over attribute values of individual corpus positions and regular expressions over sequences of corpus positions. These tools are used in the chapter 5.

4. **UCS**. It is a toolkit implementing many association measures and it is mainly used in collocation extraction. We make use of this tool in the same chapter 5.

5. **jSRE** is an open source Java tool for Relation Extraction. It is used in the second part of the chapter 5.

6. **WWW:Wikipedia** It is Perl module which allows downloading of up to date Wikipedia articles. It is used in chapter 6.

7. **CLUTO** is a software implementing various clustering algorithms. It is described and used in chapter 7.

2

Introduction

As the thesis is about acquiring conceptual structures from different resources, we will briefly review some theories of concepts. In this introduction we will center on the theory of conceptual structure in psychology. There are two main reasons for choosing the psychological perspective on concepts. The first one is that during this thesis the focus will be on the properties generated in psychological experiments. The second reason is that the psychological theories of concepts always consider the work in philosophy and other areas. In the second part of this chapter we will present two methods for extracting conceptual structures from corpora. This subsection is not in any way a survey of the work in conceptual structure acquisition but the choices we made are among those close to our own perspective.

From our point of view no theory of concepts answers the question: What concepts are? Most psychologists and philosophers subscribe to the idea that concepts are mental representations. But this does not solve the ontological status of concepts. It just moves the question from what concepts are to what the representations are[1]. All the so called theory of concepts are properly speaking theories of conceptual structures. They assume that concepts are mental representations and based on this belief they build models of the concept structure.With this essential idea in mind we introduce the following useful terminology:

[1]It is notorious that cognitive science does not have a theory of representation.

Lexical concept. A concept is lexical if it corresponds to lexical units of a particular language. For example, in English the concept **cow** is a lexical concept, whereas the concept **cow with white spots eating grass** is not a lexical concept. Of course there are concepts which are lexical concept in a language and not lexical concepts in a different language. The concepts for which we acquire properties in this thesis are all lexical concepts in English.

Complex Concept. A complex concept is a concept which has a structure. Complex concepts are composed of simpler concepts which can be either complex concepts or primitive concepts.

Primitive Concept. Unlike complex concepts, the primitive concepts cannot be further decomposed. It was hoped that the primitive concepts can be sensory or perceptual in nature. Thus the conceptual space can be reduced by a series of decomposing operations on perceptual terms. Unfortunately, this hope was never realized and we think it will never be realized[1].

Containment Model. The Containment Model gives an account of complex concepts. According to this model when we tokenize complex concepts; we also tokenize the simpler concepts in their composition.

Inferential Model. Different from the above model, the Inferential Model predicts that the subjects tokenizing the complex concept are only "disposed to" infer the simpler concepts in their composition. The simpler concepts are not necessarily parts of the complex concepts as in the Containment Model. To see the difference between the two models let's consider the concept **sparrow**. The containment model says that when we tokenize the concept **sparrow** we also tokenize the component concept **bird**. The inferential model says that when tokenizing the concept **sparrow** we have the disposition to infer that **sparrow** *is a bird*.

In psychology there are three main classes of theories of conceptual structure known as: the classical theory of concepts, the prototype theory and the theory-theory.

[1]For example, how can one reduce concepts like **beauty** and **truth** to perceptual terms?

2.1 The classical theory of concepts

The classical theory of the concepts is, strictly speaking, a label for a bunch of views that share a common core but vary in detail. The core states that the concepts have a structure of necessary and sufficient conditions. By definition a **necessary condition** of a stament must be satisfied for the statement to be true. A **sufficient condition** is a condition that, if satisfied, guarantees the statement's truth.

For example, in this model the concept **bachelor** is defined by two necessary and suficient properties: *unmarried* and *man*. That is if we encounter a **bachelor** we know that *is a man* and *unmarried* and we also know that if an individual *is a man* and *unmarried* he is a **bachelor**. There are many domains of science (e.g. various branches of mathematics) where almost all concepts are defined in the classical way. Bellow we give some examples of concepts having necessary and sufficient conditions:

1. **Prime Number**. By definition a prime number is a natural number which has exactly two distinct divisors: 1 and itself. Therefore, to be a prime number a mathematical entity should have three necessary and sufficient conditions: *natural number*, *divided by 1*, *divided by itself*

2. **Equilateral triangle**. In geometry an object is an Equilateral triangle if and only if: *it is a triangle* and *all three sides are equal*

3. **Wining the Lottery**. The event **winning the lottery** is equivalent with the events: *buying a lottery ticket* and *the variant on the ticket should be extracted.*

At the representation level the concepts are equivalent to the list of their necessary and sufficient properties. Thus, the classical theory is a representative of the Containment Model. In Natural Language Processing and Artificial Intelligence examples of the Classical Theory are ontologies, though the term ontology is flexible enough to accommodate a correct formulation of theories. In the figure

2.1 the concepts which have the equivalence sign over them are concepts defined in the classical way[1].

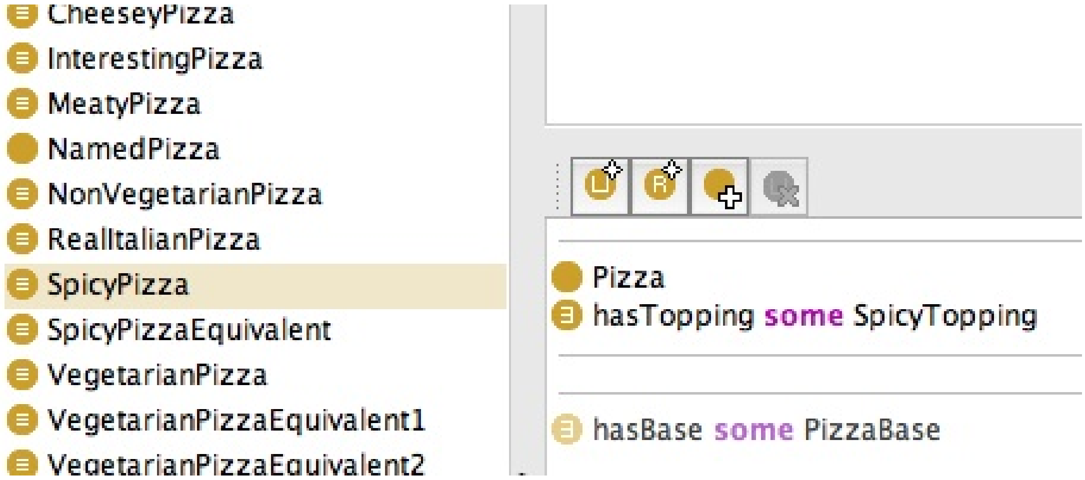

Figure 2.1: A fragment of an Ontology built in Protege illustrating defined concepts -

The ontology presented in the figure above was built in Protege. The necessary and sufficient conditions for the concept **SpicyPizza** detailed in the right panel are: *is a Pizza* and *has Spicy Topping*. In this thesis the reader will find examples of defined concepts in chapter 6.

The Classical theory has a series of drawbacks among which we mention the following ones:

1. **Definition Failure**. The first critique of the theory is that despite many years of research there are still very few defined concepts. The trouble is that not only the vast majority of common sense concepts lack definitions but also concepts fundamental to the science like DNA.

2. **Psychological implausibility**. The affirmation that concepts have definitional structure can be tested. If concepts have definitions then it should be the case that more time would be needed to process complex concepts

[1]Concepts for which the necessary and sufficient conditions are specified.

than simpler or primitive component concepts. It was shown that this is false (Kintsch (1974)) hence the concepts do not have definitional structure.

However the main cause of the abandon in psychology of this theory are the discovery by Rosch of the prototypical effects, the subject of the next section.

2.2 The Prototype Theory

The main evidence the prototype theory is based on is a series of psychological experiments performed by Eleonor Rosch in the 1970s (Rosch and Mervis (1975)). She asked the subjects to order, according to the typicality, items belonging to a category. The results of the experiment show a high inter-subject agreement. For example, in a consistent way, people rate the item **sparrow** as better example of the category **bird** than the item **chicken**. By the same token they say that **trout** is a typical fish and **eel** is an atypical one. To explain the typicality phenomenon the prototype theory was put forward. In the modern formulation the theory states that the conceptual structures encode a statistical analysis of properties[1]. Differently from the classical theory the prototype theory states that the concept properties are neither necessary nor sufficient. From a representational perspective the concepts are weighted lists of properties. Consider a hypothetical list of properties describing the concept **bird**:

1. *has wings* 1
2. *flies* 0.9
3. *lays eggs* 0.8
4. *sings* 0.6 ...

All the properties in the conceptual structure are weighted with a number between 0 and 1. The number represents the probability that a certain property is present in the conceptual structure. The property *has wings* is a highly rated

[1]Rosch would not agree with the modern formulation of the prototype theory (see Rosch and Mervis (1975) for details).

property for the category bird, presumably having the maximum score. The property *sings* instead having a lower score will be a less typical property for birds.

According to the prototype theory the classification of an instance of the concept **bird** will proceeds as follows. The perceptual system extracts a conceptual structure for the instance. Then a simultaneous comparison between extracted conceptual structure for the instance and all stored conceptual structures is performed. The instance will be classified as bird if the two conceptual structures will have sufficient properties in common[1]. Considering the way the classification is performed there is no surprise that the prototype theory generally has better results in the classification task. This is because for the classical theory an instance either belongs to a concept or no: tertium non datur. The prototype theory instead allows for the degree of membership.

Despite its early success there are some problems for the prototype theory, the most relevant being the following ones:

1. In the first place there is a problem with the interpretation of the data that gave rise to prototypically theory. The refutation of the theory goes like this: the fact that, the subjects order the members of categories, does not constitute evidence for typicality effects. In an experiment performed by Sharon Armstrong, Lila Gleitman, and Henry Gleitman(Armstrong *et al.* (1983)) it was shown that, the subjects order according to typicality, even the members of categories with clear cut boundaries. They rank the prime numbers for example. From my point of view it is truly odd that the subjects are willing to say that 5 is a better example of prime number than 61. A number is either prime or not[2].

2. A related problem is that there are a whole bunch of concepts mostly not natural kinds which lack prototype structure. There is no prototype for

[1]There are different algorithms which implement the categorization process. The threshold for classification varies with the algorithm.

[2]In an article surveying the theories of concepts (Rosch (1999)) Rosch very weakly replies to this argument. She only states that "In fact the phenomenon of judging different degrees of membership is so universal that, by a strange twist of logic, it has been used as a refutation of graded structure..."

the following concepts, for example: "things painted in white", "objects weighting more than 3 kg" etc...

3. The compositionality is still a problem for the prototype theory of concepts despite the efforts (see for example Kamp and Partee (1995)). The prototype of complex concepts are not functions of the prototypes of component concepts (Fodor (1981)) [1]. Consider the classical example of the complex concept **pet fish**. The prototype of this concept could be the **goldfish**. The prototype for the first component concept **fish** could be the **trout**. The prototype for the second component concept **pet** could be **dog**. There is no conceivable way of combining the most salient properties of pet prototype like *furry, affectionate etc.* and the properties of fish prototype like *gray, medium-sized ...* to get the properties of **pet fish** prototype: *small, brightly colored.*

Please notice that the classification process for both classical and theory-theory is a very simple process: the extraction of properties for the item to be classified and a matching process with the long term memory stored categorical representations. However, it is clear that this is just part of the story. In reality the classification process heavily relies on the individual background knowledge. The theory in the next section brings the world knowledge to the table.

2.3 The Theory-Theory

The name theory-theory is misleading. The nomenclature lead many people believe that the concepts are theories[2], which is not true. Basically the agreed formulation of the theory-theory says that concepts are part of larger structures called theories. To understand a concept is to understand its role in a particular theory. The inspiration for the theory-theory comes from the philosophy of science. We cannot grasp in isolation the concepts **atom** or **energy** but to get

[1] The component concepts should also have prototypes. If they do not have the problem is even worse.

[2] The unfortunate terminology has a basis. There are researchers who believe that concepts are theories. For details the reader should consult the book (Margolis and Laurence (1999)).

their meaning we should learn physics. This observation sounds very well but the trouble begins when we want to make precise what a theory is. And maybe some efforts in Artificial Intelligence can help us answer this question.

There is a similarity between the theory-theory and the CYC Lenat (1995) effort in Artificial Intelligence. CYC is a project started in 1984 that attempts to formalize the common-sense knowledge. The concepts in CYC have meaning only in a context[1]. The context comprises a set of assertions which encode IF-THEN rules, background assumptions and facts relevant to a domain. Nevertheless, following the theory of context of John McCarthy the set of assertions do not exhaust the context but they are just evaluated as true or false in the context. Therefore in CYC the contexts are first order citizens. Unfortunately, despite the massive effort invested in CYC[2] the results are much less impressive than expected. The choosing of contexts in CYC is somehow arbitrary from a theoretical perspective.

Unlike CYC theories which are guided by practical applications the psychological theories should have cognitive import. The main problem of the theory-theory is that nobody was able to give a satisfying account of the notion of theory. To our knowledge nobody was able to say which constructs count as theories and which do not count. Likewise nobody was able to say where to cut the boundaries between the theories. The failure of CYC project should warn us not to be over-optimistic about this theory of concepts.

Once specified, the theories can be easily casted in a logical language or in the form of association rules or in other appropriate form. In the chapter 6 we extract richer conceptual structures than simple list of properties. They are not theories but I think they represent a step in the right direction. Of course the three theories presented above summarize the main lines of discussion about the concepts. Based on the extensive criticism coming from many quarters, the initial theories were changed to accommodate this criticism[3]. It is not our purpose to describe all the objections and responses. For a detailed treatment the interested reader can consult (Margolis and Laurence (1999)) or (Murphy (2002)).

[1]The context is also called a micro-theory.

[2]People encoded by hand the rules relevant to a domain.

[3]The modified classical theory of concepts is called the neo-classical theory, for example.

2.4 The extraction of conceptual structures from corpora

2.4.1 The Extraction of Qualia Structures

The Qualia structures have origin in Aristotle modes of explanation (Kronlid (2003)). The main idea is that the conceptual structure[1] is revealed along four dimensions called qualia:

1. **Constitutive**. It reflects the relation between an object and its constituent parts. The relation is well studied in Natural Language Processing and it is encoded in Wordnet (see chapter 4 for details)

2. **Formal**. This is another name for the ubiquitous IS-A relation. Again this relation is also encoded in Wordnet. Unfortunately, like Wordnet builders, Pustejovsky does not distinguish the true taxonomic relations from false taxonomic relations (see chapters 4 and/or 6 for details).

3. **Telic**. It represents the purpose or the function of the object.

4. **Agentive**. The definition of the agentive role is a little bit vague: according to Pustejovsky it denotes the factors involved in bringing about the object.

Cimiano and Wenderoth (Cimiano and Wenderoth (2005)) acquire qualia structures using web as a corpora. To this end following the seminal work of Hearst (Hearst (1998)) they built by hand a set of lexico-syntactic patterns meant to extract the four property types which make the Qualia Structure. The process of acquiring Qualia structures from the web has the next steps:

1. **Clue generation**. The clues are queries which indicate the qualia to be extracted. For example, if we consider the Hearst-like pattern: NP_0 such as NP_1, $NP_2 \ldots NP_n$ and you want to extract the formal qualia for the concept **cow** you first made the query "such as cow" and sent it to a searching engine.

[1]This structure disclose the meaning of the nouns.

2. **Document Downloading and Preprocessing**. The first n^1 most relevant documents are downloaded and preprocessed: POS tagging and lemmatization.

3. **Query Matching**. The query made at step 1 is matched on the downloaded documents and the corresponding qualia is extracted. For the previous pattern presumably the following properties will be extracted : *animal, herbivore, meat.*

4. **Qualia Weighting**. The extracted qualia are weighted according to the Jaccard coefficient that takes into consideration the searching engine hits.

The results extracted for the constitutive and the formal roles are readily interpretable. It is relatively easy to judge if a noun extracted represents a part or a superordinate of the concept we started with. The agentive role give the least information possible. As a rule the role is represented by very general verbs like: *make*, *generate*, *create*[2].

2.4.2 Extraction of other property schema

Another attempt to extract conceptual structures from the Web is that of Abdulrahman and Poesio(Almuhareb and Poesio (2005), Poesio and Almuhareb (2005)). Based on the work of Guarino (Guarino (1992)) and Pustejovsky Pustejovsky (1995) they define the following conceptual structure [3]:

1. **Part and Related Objects** . Guarino argued for the separation between components of the objects and arbitrary two place relations. For technical reasons the authors preferred to mix these two distinct relations.

2. **Quality**. These are mostly perceptual qualities like: size, color, but also include some other economic and social properties (e.g. *price*).

3. **Related-Agent**. These are binary relational roles whose subject is an animated agent.

[1] In the original work n is equal to 10.

[2] It would be interesting to say how something is created not just that it is created.

[3] The next property types are called by the authors attributes

4. **Activity**. These property types map onto Pustejovsky Telic and Agentive Role.

In the first phase the authors generate candidate words using a very general pattern. Afterwards two classifiers are trained: a binary classifier which distinguish an attribute from a non attribute and 5-way classifier[1]. The classification process relies on 4 types of information: morphological information, an attribute model, a question model and an attribute usage model. The morphological information says if a noun is derived from an adjective or verb. The attribute model uses the properties of the candidate attributes. The question model make use of the information extracted from the web with the questions: What is the ATTR of CONCEPT or When is the ATTR of CONCEPT. Finally the usage model quantifies the usage of the candidate nouns as attributes. The binary classier achieved an accuracy of 81.82% as evaluated through cross-validation, which corresponds to an F value of .892 at recognizing attributes and .417 at recognizing non-attributes. The 5-way classier achieved an accuracy of around 80% at cross validation, corresponding to an F value over .8 for quality, activity, and part / related object, of .95 for related agent, and .538 for non-attribute.

[1] This classifier should tell if a candidate is either: Part and Related Objects, Quality, Related-Agent or Activity

3

Feature Norms

3.1 Contribution

This chapter makes the transition between the introduction and the main part of the thesis. Here we present three essential resources which will be used throughout the thesis: two feature norms and a taxonomy classifying the properties in the norms. The main contribution of this chapter is a comparison between two norms at the level of generated properties. This comparison assesses the stability of the semantic memory content represented by feature norms.

3.2 Introduction

Feature norms are ways of making explicit the content of semantic memory. In a task called feature generation[1] the subjects list the most salient properties for a set of concepts. The concepts used in most feature generation tasks are basic level concepts representing concrete objects. In a celebrated series of experiments in the 70s Rosch and Mervis (Rosch and Mervis (1975)) asked their subjects to produce features for twenty members of six basic level categories. Subsequently they demanded the subjects to rank the respective members according to how good examples they are for the respective categories. For example, the subjects

[1]The task is also known as: feature production or feature listing. All the terms will be used in this thesis.

were asked to rank the concepts chair, piano and clock in function of how representative examples they are for the category furniture. One major finding of their study was that typicality of a concept is highly correlated with the total cue validity for the same concept. That is, the most typical items are those that have many properties in common with other members of the category and few properties in common with members outside the category. Subsequent research replicated the results of Rosch and Mervis, but nowadays it is acknowledged that besides cue validity there are other factors that determine the typicality (Barsalou (1985)). Following Rosch, other researchers (Ashcraft (1978), Moss *et al.* (2002)) built feature norms and used them for investigations of the semantic memory. The norms became the empirical material for constructing computational theories about information encoding, storage and retrieval from semantic memory. Following the line of research that started with Rosch and Mervis, the norms are also used to examine the relation between semantic representations and prototypicality. From a formal point of view feature norms are databases storing concepts and their properties. The feature production task has four particularities:

1. **Empirical Soundness**. For excluding short term memory effects the concepts in a session should not be similar. In case the concepts are presented in the same session the experimenter should make sure that there does not exist a linear sequence of concepts belonging to the same category (e.g. there are not presented three animals in sequence for example).

2. **Property normalization**. The property normalization means finding classes of equivalence among the properties listed by subjects. The operation is necessary because the same property can be listed in different ways by different subjects (e.g. *used for transportation*, *transport*, etc.). The final properties stored in the norms will thus be distinct.

3. **Property weighting**. The properties in the norms have a weight representing the number of subjects who listed them. Moreover, not all properties are stored in the final norm, but only those having a weight above a defined threshold.

4. **Schema Annotation**. Most feature norms have a schema meant to annotate the properties according to their types.

The first three characteristics above, distinguish the feature production from similar efforts in AI (e.g. Concept Net H.Liu and Singh (2004))[1].

3.3 Concepts and the Brain

The psychological theories of concepts presented in the previous chapter regard mainly the conceptual structure. All the theories assume that concepts are mental representations which correspond to categories of objects in the world. We saw that the concepts are mentally represented as list of properties in the case of classical theory of concepts and prototype theories or richer schema in the case of theory-theory. However, the above theories do not say anything about the brain representation of concepts.

The orthodoxy in cognitive science is that the concepts are represented in the brain by amodal symbols. The perceptual system is responsible for transducing the modal specific states into amodal symbols. Further the orthodox theory assumes that the brain is a modular system with unique principles.

A newer class of theories about the representation of concepts in the brain are the modality specific theories. Opposed to the orthodox view, the modality specific theories reject the argument that concepts are represented by transduced amodal symbols. The main reason of rejection being that the orthodox view has no experimental bases. It is not capable of making prediction and it cannot be falsified because it is powerful enough to "explain" any experimental result post-factum. The modality specific theories claim that concepts are represented in hierarchical organized association areas. According to Barsalou the modality specific theories have the advantage of bridging the gap between the perception and cognition. For example, the extraction of the properties of the concept **car** by the visual system proceeds as follows (Barsalou (2003)):

[1]There remains to be seen how the final representation stored in the norms differ from Concept Net like representations but this is not an objective of this thesis.

"During visual processing of a car, for example, populations of neurons fire for edges, vertices and planar surfaces, whereas others fire for orientation, color and movement. The total pattern of activation over this hierarchically organized distributed system represents the entity in vision" .

By the same token distinct populations of neurons will represent other perceptual qualities: the taste, the color, the introspective properties, etc. The final concept representation is derived by integrating all modality specific information. One of the main construct of this theory of brain representation of concepts is that of simulator. A simulator is a generator of contextual conceptual representations. In this model a category is not represented in a static way by a definite representation. Instead the simulator implements dynamicity by allowing the integration of new information at different processing stages. Therefore the same concept is represented by different properties in different contexts(Barsalou (2003)):

"On one occasion, the CAR simulator might produce a simulation of travelling in car, whereas on others it might produce simulations of repairing a car, seeing a car park and so forth."

3.4 The structure of McRae and Garrard feature norms

In this section we introduce the two norms used in this thesis which will be called from now on: Garrard (Garrard *et al.* (2001)) and McRae(McRae *et al.* (2005)) norms. The methodology for building the norms differs in some details from one researcher to the other. For example, unlike Rosch and Mervis neither Garrard nor McRae impose their subjects time limits for property listing task. The purpose of McRae's experiment is the obtaining of information about the mental representation of concepts. The stimuli in the experiment are basic level concepts representing concrete objects. Each concept is shown on an empty page and the subject is expected to list its most salient properties. In the task description session experimenters hinted the subjects about the nature of the knowledge they were expected to provide. First they named the kind of properties presumed to be listed: properties related to the shape and physical appearance of objects represented by the target concepts, part properties, functional properties, behaviour related properties, etc. Moreover, the subjects were presented with examples of concepts and properties. To get a feeling of what kind of properties are recorded in the McRae feature norm we list the properties for 3 concepts (**airplane**, **apple** and **eagle**) in the table 3.1

Concept	Properties listed by subjects
Airplane	*flies*, *found in airports*, *made of metal* *has a propeller*, *has engines* *has wings*, *crashes*, *used for travel* *is fast*, *is large*, *used for transportation* *requires pilots*, *used for passengers*
Apple	*a fruit*, *eaten in pies*, *e.g. granny smith* *grows on trees*, *has a core* *is delicious*, *is green*, *is crunchy* *is round*, *is worm infested*, *is yellow* *tastes sour*, *tastes sweet*, *used for cider*
Eagle	*a bird*, *a carnivore*, *a predator* *builds nests*, *eats*, *has a beak* *flies*, *lays eggs*, *has claws*, *symbol of U.S.* *lives in mountains*, *symbol of freedom*

Table 3.1: Concept Description Examples For McRae feature norm

In Garrard and colleagues's experiment each concept was presented on a piece of paper having four fields. The fields help the subjects to formulate the property and they were intended to supply the following property types:

- Classification properties (under the field CATEGORY)
- Descriptive properties (under IS field)
- Parts (under HAS field)

- Abilities (under CAN field)

In the table 3.2 we list some of the properties produced by the subjects in Garrard Experiment for the same three concepts above.

Concept	Properties listed by subjects
Airplane	*is a aircraft*, *is a vehicle*, *is large* *is made of metal*, *is fast* *can make a noise*, *has cockpit*, *has fuselage* *has propellor*, *has seat*, *has tailplane*, *can carry passengers* *has rudder*, *has controls* , *has flap*, *can fly*
Apple	*is a fruit*, *has pips*, *has skin*, *can rot*, *has maggots* *is round*, *has stalk*, *has flesh* , *can be bought* *is sweet*, *is coloured*, *is hard*, *can be sold* *is juicy*, *is sour*, *has white flesh*, *can grow* *is small*, *is edible*, *can be cooked*, *is found on trees* *can fall*, *can be picked*, *can ripen*, *can be preserved*
Eagle	*is a bird*, *has claws*, *has wings*, *is carnivorous* *has legs*, *has large wings*, *has feet*, *is predator* *has two legs*, *has eyes*, *has hooked beak* *is golden*, *is grey*, *can fly*, *can swoop*, *can walk* *can lay eggs*, *can hover*, *can reproduce*, *can carry* *has good eyesight*, *is dangerous*, *has nest*, *is wild* *is found in mountains*, *is rare*, *is protected*, *is strong*

Table 3.2: Concept Description Examples For Garrard feature norm

In addition to the properties listed for concepts, the feature norms contain a wealth of interesting information. We mention only three fields that are particularly important:

1. **Dominance** is a field indicating the number of subjects that listed a certain property. It reflects the weight of the property in the mental representation

of a concept: the higher the dominance, the more salient the property.

2. **Distinctiveness** reflects the percent of members of a concept for which a specific property is listed. It is a measure of how good the individual properties are in distinguishing the concepts. For example, *has trunk* is a highly distinctive property for the concept **elephant** because it helps distinguishing the members of this class from the other animals not members. Instead *has tail* is a lowly distinctive property because elephants share this property with other animals.

3. **Classification**. The third field, the most significant from our point of view, gives a classification of properties in the feature norms.

Unfortunately, the two feature norms contain different property classifications schemas. Garrard feature norm has a relatively simple but, nevertheless, controversial property classification. The properties have one of the four types: categorizing, sensory, functional or encyclopedic. The categorizing properties represent taxonomic classification of concepts (e.g. **lion** *is an animal*). The sensory properties are defined by reference to the sensory modality used to perceived the objects (e.g. **the bus** *is coloured* or **the apple** *is sour*). The functional properties describe an activity or the use someone makes of an item (**monkeys** *can run*, **a brush** *can apply paint*). Finally the encyclopedic are defined negatively meaning that all properties which are not superordinate, sensory or functional are encyclopedic.

In our opinion the sensory properties are not well defined. For example, in the Garrard feature norm some properties denoting parts of the sophisticated modern apparatus are typed with the label sensory (e.g. *the rotor* or *the control* of an **helicopter**). Even if some can argue that we "see" these parts, their identification as parts is largely based on the knowledge of the structure and the functions of a modern vehicle and therefore the label sensory is misleading. The property types used to categorize the properties in McRae feature norm are based on Wu and Barsalou schema. In next sections we discuss this schema and the theory behind it.

3.5 Wu and Barsalou taxonomy

Much more interesting is the classification employed by McRae and colleagues. They use a taxonomic classification, a slightly modified version of Wu and Barsalou taxonomy (Wu and Barsalou (2006)). The original Wu and Barsalou taxonomy is an application of Barsalou amodal theory for the interpretation of properties generated in feature production task. Because the original theory sees the brain symbols as perceptual in nature, we expect that the classification schema property types to be linked with perceptual simulations. According to Wu and Barsalou there are two lines of evidence which suggest that the subjects use simulations in feature production task:

1. **The occlusion phenomenon**. The argument runs like this. If the subjects use simulations then we expect that the occluded properties[1] to be generated less frequently than properties directly visible. The reason is that for "finding" the occluded properties an extra computational effort is necessary. For example, the occluded properties: *dirt* and *root* for the concept **lawn** should be less generated than the directed perceived color property *green*. According to Wu and Barsalou this phenomenon is pervasive in the experiments.

2. **The imagery task**. The fact that the distribution of the properties in feature listing task is similar to that in a task when the participants are instructed to describe images, constitutes further evidence for the amodal representation system.

Just like the Wu and Barsalou taxonomy, the modified version used by McRae and collegues has two levels[2]; at the coarsest level the properties are classified as taxonomic, entity, situational or introspective. At this level the modified taxonomy is identical with the original Wu and Barsalou taxonomy. At the next level each mentioned category of properties is again further divided. The modified Wu and Barsalou taxonomy used by McRae has at the second level 27 categories. In

[1]The occluded properties are those properties not visible.

[2]From now on we will refer to the slightly modified version used to annotate McRae feature norm also as Wu and Barsalou taxonomy when there is no possibility of confusion.

general the modifications are not substantial and the curious reader can consult McRae paper for details(McRae *et al.* (2005)). We will not discuss here either the original taxonomy or the modifications operated to the original taxonomy by McRae and colleagues. However, throughout this thesis when needed we will clarify the meaning of some property types. By way of example we show a partial description of the concept **accordion**:

- *a musical instrument* (Taxonomic $->$ Superordinate),
- *has keys* (Entity $->$ External Component),
- *produces music* (Entity$-$ $>$Entity Behaviour).

In the above description three properties of the concept **accordion** are listed. The first property states that the accordion is a musical instrument, and according to Wu and Barsalou category it is classified at the first level as taxonomic property and at the second level as a superordinate property (Taxonomic $-$ $>$Superordinate). The other two properties *has keys* and *produces music* are classified as being Entity$-$ $>$External Component and Entity$-$ $>$External Behavior property type respectively. The Entity$-$ $>$Component property type comprise those that are external components of the object to be described whereas Entity$-$ $>$Behaviour property type encompasses properties denoting the behavior of the object under description.

3.6 A comparison between Garrard and McRae feature norms

Table 3.3 gives a quantitative assessment of the two feature norms. The first row of the table records the number of concepts in each feature norm; the second row gives the number of concept-property pairs in the each of the two feature norms and the last row lists the average number of properties per concept for each feature norm.

It can be noted that the average number of properties per concept is twice as bigger in Garrard feature norm than in McRae feature norm. This fact is

	McRae feature norm	Garrard feature norm
Number of Concepts	541	62
Concept-property Pairs	7275	1657
Average No. property/Concept	13.4	26.7

Table 3.3: A quantitative evaluation of McRae and Garrard feature norms

partially explained by the different strategies used for property generation. As shown in a previous section the subjects in Garrard experiment have to fill in the fields already present on the page. For example, they see on the page the concept **Elephant** and the beginning of a property (*has* ...) that expresses in most cases part properties. The subjects should retrieve from the memory only part of a property (*legs* for example). In McRae experiment the subjects see only the concepts and should retrieve the whole property (e.g.*has legs*) from the memory.

In this section we investigate the similarity of McRae and Garrard feature norms. To achieve this, we performed a semi-automatic mapping between the concept-property pairs in the two norms following the next steps:

- Identification of the common concepts in the two feature norms. The process is trivial because the concepts in most cases have the same notation in both norms. However in some cases the notation differs (e.g. **airplane** in McRae feature norm is **aeroplane** in Garrard feature norm). In these cases we check if the different notations are part of the same synset in Wordnet[1].

- Once the set of common concepts is identified an automatic mapping between the properties is performed under the assumption that the properties which have the same last word are equivalent (e.g. *tastes sweet* versus *is sweet*)

- Finally we perform a manual verification of the accuracy of the automatic mapping together with the manual mapping of the properties which could not be mapped automatically.

[1]For example, **airplane** and **aeroplane** belong to the same Wordnet synset.

The mapping links between the two feature norms are in most cases one to one, but there are cases in which the mapping is either one to many or many to one. An example of the one to one mapping: the concept-property pair (**orange**-*grows on trees*) in McRae feature norm is mapped onto the concept-property pair (**orange**-*is found on trees*) in Garrard feature norm. An example of one to many mapping: the unique concept property pair (**key**-*used for locking or unlocking doors*) in McRae feature norm is mapped on two concept-property pairs (**key**-*can open doors* + (**key**-*can lock doors*) in Garrard feature norm.

Mapped Concepts	50
McRae Concept-property pairs	765
Garrard Concept-property pairs	1326
Mapped Pairs	430

Table 3.4: The mapping between McRae and Garrard feature norms

The mapping results are presented in table 3.4. We found a set of 50 concepts common to both feature norms (Mapped Concepts line in the table). For this common set of concepts we list the number of concept-property pairs present in each norm ("McRae Concept-property pairs" and "Garrard Concept-property pairs" respectively). Finally, "Mapped Pairs" represents the number of concept-property pairs the feature norms have in common. As one can see looking at the same table, 56 % of concept-property pairs listed in McRae feature norm are present in Garrard feature norm, but only 32 % of the concept-property pairs in Garrard feature norm are also present in McRae feature norm. The problem is how to make sense of these differences. The second finding namely that 68 % of the concept property pairs in Garrard feature norm are not in McRae feature norm can be explained by the different number of property/concept in the norms.

More problematic is how to interpret the first finding, 43 % of the concept-property pairs in the McRae feature norm are not in the Garrard feature norm. This fact poses serious problems for the computational theories of the semantic memory based on feature norms but we will not address the problem in this thesis. Suffices to say that the great variation between the two property-norms should warn us to consider the norms gold-standards.

One can object against the comparison we made and argue that the similarity between the norms is to be found at the level of property types. Even if the subjects do not list exactly the same properties, they tend to list the same kind of properties (e.g. parts, superordinates, etc.). Notwithstanding this fact, the property types are too general to be an accurate description of the semantic memory. What really matters is not the fact that the subjects produce parts or superordinates for the concepts in the test set but the distinct content of their semantic representation. Specifically what parts, superordinates, etc. they produce and in what way they are consistent across different feature norms.

We performed a second comparison between the norms using the categories in Wu and Barsalou taxonomy. In particular we were interested what percent of properties under a property types (e.g. Made Of properties) present in McRae feature norm are also present in Garrard feature norm.

property Type	CFPM	CFPG	CFPG / CFPM
Made Of	32	27	0.84
Superordinate	67	54	0.80
External Component	32	27	0.84
Entity Behaviour	171	129	0.75
External Surface Property	102	64	0.62
Internal Component	21	13	0.61
Internal Surface Property	18	11	0.61

Table 3.5: Per property type comparison between the McRae and Garrard feature norms

Table 3.5 gives the results of this comparison. All the figures refer to the Mapped Concept set (see table 3.4 for more details). The meaning of the columns of the table is the following:

- Property Type: the column lists the categories in Wu and Barsalou taxonomy omitting the first level of classification

- CFPM: represents the number of concept-property pairs classified in McRae feature norm under a specific property type. For example, there are 32 properties in the McRae feature norm classified as instances of Made Of property type, 67 exemplifying the Superordinate property type and so on.

- CFPG: has the same signification as the above column just that it refers to the Garrard feature norm. Thus from the 32 properties classified as instances of Made Of property-type in the McRae feature norm, 27 have been mapped on Garrard feature norm.

- The last column represents the ratio between the CFPG and CFPM columns. From the table 3.5 those property types classifying less than 11 properties or having a score less than 0.51 were eliminated.

The property types successfully mapped from McRae feature norm to Garrard feature norm are parts (Made Of, External Component, Internal Component), taxonomic properties (Superordinate), the properties classified under Entity $->$ Behaviour and the properties denoting external and internal surface properties. Other properties not listed in the table, those classified under Systemic Property or Origin, for example, are peculiar to McRae feature norm.

4

Feature Norms and Wordnet

4.1 Contributions

The contributions of this chapter are the following:

1. **Property extraction**. We devise a procedure for property extraction from Wordnet using the notion of semantic neighborhood. The procedure exploits the main relations organizing the nouns, the information in glosses and the inheritance of properties.

2. **Cognitive assessment of Wordnet**. We evaluate the Wordnet as a cognitive resource by comparing its conceptual descriptions with the descriptions in Garrard and McRae feature norms. Based on this comparison we suggest the addition of new relations in the future versions of Wordnet.

4.2 What is Wordnet

Wordnet is the most used lexical resource in Natural Language Processing. Wordnet was used for a variety of tasks:Word Sense Disambiguation(Agirre and Martinez (2000),Agirre and Rigau (1996)), machine translation (K.Knight (1993),Dorr and Katsova (1998)), information retrieval (Langer and Hickey (1997),Magnini and Strapparava (2001),Mandala *et al.* (1998)), Name Entity Recognition (Magnini *et al.* (2002)) etc. However, at the beginning it was not designed to be used in

Natural Language Processing tasks but to be a model of the human lexicon(Miller *et al.* (1990)). The main source of inspiration for Wordnet organization were various psycholinguistic studies(Caramazza and Berndt (1978), Charles and Miller (1989)). The basic unit of meaning in Wordnet is the synset which represents a set of synonyms. Two words are synonymous if the substitution of one for the other in a certain linguistic context does not change the meaning of the proposition they are part of. Obviously for satisfying the substitution criterion the two words should have the same part of speech.

Wordnet has a different organization for nouns, verbs, adjectives and adverbs. The nouns are organized in inheritance hierarchies. The verbs are mainly structured by various entailment relations and adjectives are principally organized in clusters around two antonyms. The adverbs are linked with the adjectives they are derived from. Interesting for our work is the organization of nouns in Wordnet. Inspired chiefly by the work of Quillian on semantic networks the noun meaning is structured by hyponymy relation. Hyponymy[1] is a transitive and hierarchical relation. There is an economical advantage in organizing the noun lexicon around this relation. The properties are inherited from superordinates to subordinates, therefore it is necessary to explicitly store only the distinctive properties for each concept. For example, in an hierarchy **canary**$->$**bird**$->$**animal** the property *has skin* is stored in the entry of the concept **animal**, the property *has wings* in the entry of **bird** and the property *sing* in the entry of **canary**. By the inheritance principle it can be inferred that the concept **canary** has also the properties *has skin* and *has wings*. There is some psychological evidence which justifies the hierarchical organization of nouns in semantic memory (Miller *et al.* (1990)); however, there is no evidence for the property inheritance. The noun hierarchy of the last version of Wordnet has as root the highly abstract concept **entity**. At the second level the hierarchy is partitioned in: **physical entity**, **abstract entity** [2]. The other semantic relation structuring the nouns in Wordnet is meronymy,

[1]Because of a critique (see next section) in the actual version of Wordnet it is made a distinction between the hyponymy relation holding between concepts and instance_of relation which holds between a concept and an instance.

[2]On the second level there is a third concept called **thing**. We believe that this is a mistake and we hope that in the future versions of Wordnet it will be corrected.

which induces a part hierarchy. There are three kinds of meronymic relations in Wordnet:

1. **Constituent Part**. In general it links the instances of concepts with the their constituent parts. For animals the constituent are visible and distinctive parts of their bodies (e.g. **lion** has **mane**). The constituent of machines are either the visible parts (**car** has part **car door**)or the components related with the machine functioning (**car** has part **engine**).

2. **Substance**. The Substance relations specify the stuff the objects are made of. For example: **feather** has substance **keratin**.

3. **Member** . The member relation can be formalized as the element operation for sets $\in$ (e.g. **human race** has member **people**).

The meronyms of the higher concepts in the taxonomic hierarchy are inherited by the subsumed concepts.

4.3 An ontological assessment of Wordnet Relations

In the past other researchers suggested improvements of Wordnet (e.g. Priss (1998)) and proposed methods for populating the lexical ontology with instances or adding to it new relations(Amaro *et al.* (2006)). A Wordnet reorganization based on a formal ontology analysis of relations was proposed by Gangemi, Guarino and others (Gangemi *et al.* (2003)). They suggested a thorough redesign of the upper level of Wordnet ontology and the redefinition of some relations. Among the problems met was a conflation of concepts and instances. In a logical language concepts are interpreted as sets of instances. The versions of Wordnet anterior to 2.1 did not operate this distinction. For example in Princeton Wordnet 1.6 the instance Fall with the gloss "the lapse of mankind into sinfulness because of the sin of Adam and Eve" was considered a hyponym of the concept **event** whereas it should have been an instance of the same concept. Because of this critique starting with version 2.1 of Wordnet the initial Hyponym relation was split into two relations: a proper Hyponym relation holding between concepts and the Instance_Of relation holding between an instance and a concept. Other drawbacks of Wordnet from the point of view of the above analysis were: heterogeneous generality for the concepts found at the same level in the Wordnet noun hierarchy or the violation of subsumption constraints for some concepts.

The formal analysis starts from the assumption that Wordnet should be a formal ontology and not a lexical ontology. But Wordnet builders never claimed that the resource is a formal ontology. Wordnet construction started from psycholinguistic principles and intended to be a model of human semantic memory. For example even the distinction between instances and concepts and the splitting of relation of hyperonymy in two sub-relations is not justified from a psycholinguistic perspective. The original intention of the Wordnet team was not to provide a formal definition of a relation. To test if two concepts stay or not in the hyperonym relation is the task of the native English speaker:

"A concept represented by the synset x, x',... is said to be a hyponym of the concept represented by the synset y, y', ... if native speakers of English accept sentences constructed from such frames as An x is a (kind of) y." However because Wordnet is mainly used in Natural Languages Processing tasks at least the distinction between concepts and instances is useful from a practical point of view. Moreover, for some applications and even for wordnet builders a precise definition of a relation is better than a poorly specified one.

Unlike the formal methodology of Gangemi and Guarino we evaluate the Wordnet structure from psychological and cognitive perspective. To the basic question (which was also the chief question that Miller's team posed): What is the mental representation of concepts? we think that an answer is given by the feature norms. The advantage of feature norms over other ways of collecting common sense knowledge like CYC or Concepts Net is the normalization and ordering of properties by salience. We do not think that the norms are the definitive question to the problem of common-sense but they are a step in the right direction. Therefore we will compare Wordnet representations with the two feature norm introduced before.

4.4 Acquisition of Wordnet properties

To compare the concept descriptions in the two feature norms with the concept descriptions in Wordnet we first mapped the concepts in the two feature norms onto the Wordnet synsets. The mapping procedure has two steps: the first one is fully automatic and in the second one the manual intervention is necessary.

1. We guess the most likely assignment between the concepts in the two feature norms and the corresponding Wordnet synsets. To achieve this the hyperonym chains for all the synsets containing the concepts in the two feature norms are generated. For the concepts in the norms the taxonomic classification is given either by the Category field in (Garrard feature norm) or for Superordinate property type (McRae feature norm). We perform the intersection between the taxonomic classification of the concepts in the feature norm and the hyperonyms in the hyperonymic chains. As shown in figure 4.1 there are two senses of the word apple in Wordnet, the first one (apple[1]) refers to the fruit and the second one (apple[2]) refers to the tree. One of the hyperonyms of the concept apple[1] (fruit) is identical with the classification of the concept apple in the feature norms. Therefore, we find that the apple should be mapped on the first sense of apple in Wordnet (apple[1]).

2. There are cases when the automatic mapping cannot be performed either because the concepts in the feature norm lack a taxonomic classification or because the intersection has more than one element (this happen because the sense distinction in Wordnet is sometimes fined grained). In these cases these concepts are manually mapped onto the corresponding synsets.

Before presenting the algorithm for Wordnet property extraction it is useful to define some terms we will use in this section: **Projection Set**, **Semantic Neighborhood** and **Wordnet property**:

Projection Set: The set of synsets that represent the mappings of the concepts in the feature norms onto Wordnet is called the projection set.

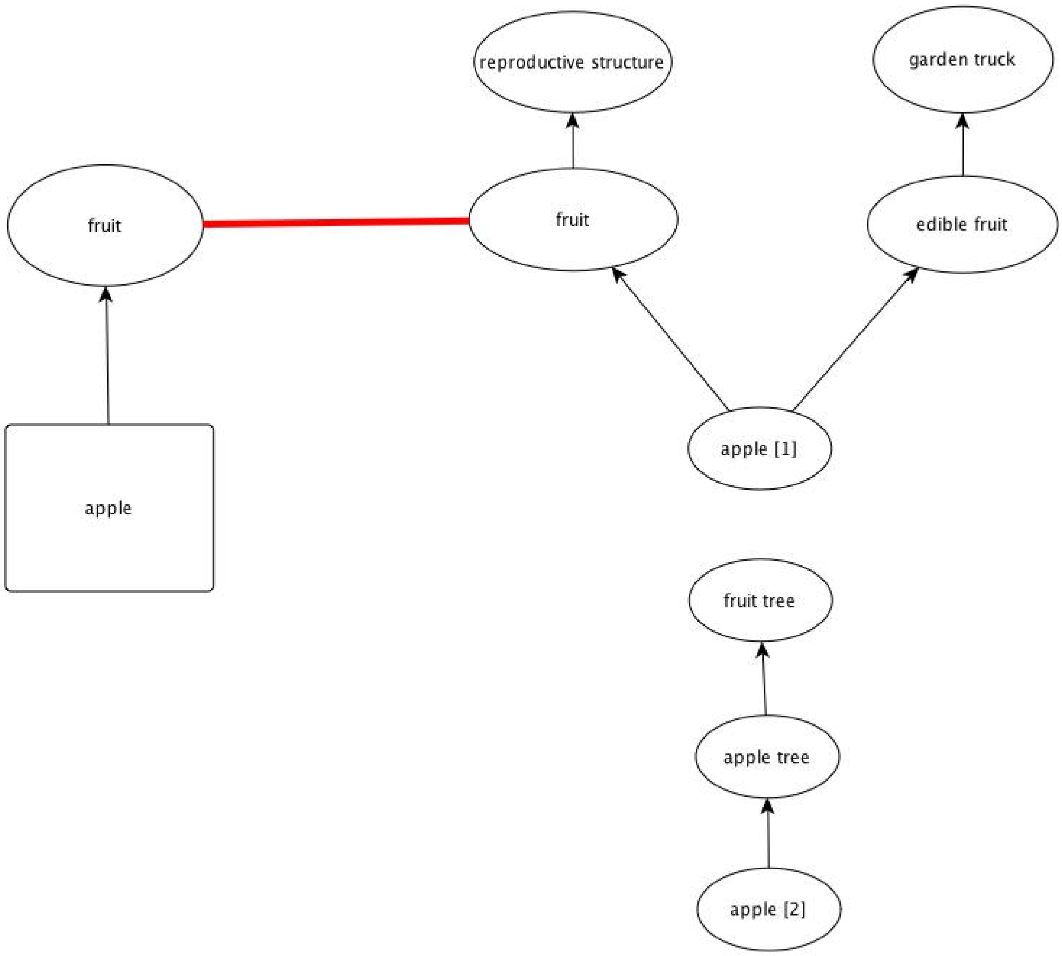

Figure 4.1: The mapping between the concepts in the feature norms and Wordnet -

Semantic Neighborhood: The semantic neighborhood of a synset s is a graph <N, R> where N is a finite set of nodes representing Wordnet synsets and R is a set of relations linking the nodes. To each synset in N we attach its corresponding gloss.

Wordnet property: A Wordnet property of a concept is any word in the synsets of its semantic neighborhood and any noun together with its modifying adjectives or any verb in the glosses of the synsets of its semantic neighborhood.

There will be two projection sets, one for each feature norm: **McRae Projection Set** and **Garrard Projection Set** respectively. In the rest of this section when we use the term Projection Set without qualification we refer to both projection sets. The semantic relations considered for the generation of the semantic neighbourhood are : hyperonymy and meronymy. Both are transitive and inheritance relations, therefore the generation of concept properties can profit from the hierarchical organization of Wordnet.

The property extraction for the concepts represented by the synsets in the projection set is performed from the semantic neighborhood of each synset. The algorithm for the extraction of Wordnet properties has three steps:

1. The semantic neighborhoods of each synset in the projection set is generated. The relations in R are hyperonymy and meronymy. The semantic neighbourhood of a synset includes the most significant hyperonyms and all meronyms. A hyperonym is significant if it is not one of the top hyperonyms in the following list :

   (a) (whole, unit): an assemblage of parts that is regarded as a single entity

   (b) (object, physical object): a tangible and visible entity; an entity that can cast a shadow

   (c) (physical entity): an entity that has physical existence

   (d) (entity): that which is perceived or known or inferred to have its own distinct existence (living or nonliving)

The hyperonyms in the above list are too general to be useful. The ideal properties that can be extracted from the gloss of the synset (whole, unit) are: *an assemblage of parts*, *regarded as a single entity.* These properties can be both interesting and disputable from an ontological point of view, but they are too abstract and never produced by subjects. The information in glosses is crucial in the process of property extraction. It is straightforward that the information in glosses of the superordinate concepts also applies to superordinate concepts. Interestingly, in many cases the properties present in the glosses of the parts can be transferred to the whole. By way of example, consider the gloss of the concept **fuselage**, one part of the **airplanes**: "the central body of an airplane that is designed to accommodate the crew and passengers". The properties: *accommodate the crew* and *accommodate passengers* can be propagate to the whole (**airplane**).

2. Second, all the glosses of the synsets from the semantic neighborhood are part of speech tagged and lemmatized. The part of speech tagging and the lemmatization is performed with TreeTagger. The part of speech tagger uses an English parameter file trained on Penn Treebank.

3. Thirdly we extract all the Wordnet properties for the concepts representing the synsets in the projection set. The properties for the concepts represented by the synsets in the semantic neighborhood are propagated to the concepts represented by the synsets in the projection set. Moreover, the duplicate properties are eliminated because it is possible that the same property to be generated more than once. For example, one of the hyperonyms of the concept **apple** is **fruit**, but the word **fruit** is also present in the gloss of the concept **apple**.

Figures 4.2 and 4.3 show a part of the semantic neighborhood for the concepts **apple** and **airplane**. The main nodes in the graph are represented by the synsets and glosses of the concepts **apple** and **airplane**, respectively. The edges of the graph are labeled with the hyperonym or meronym relations [1]. The properties extracted by our algorithm for the two examples above are the following:

[1]The hyperonym relations are colored red and the meronym relations are colored green.

- **apple**={*fruit, red, yellow, green skin, sweet, tart crisp whitish flesh, edible fruit, grow, edible reproductive body, seed plant, sweet flesh, fresh fruit, vegetable, market, rind, peel, skin, produce, green goods, green groceries*}
- **airplane**= {*airplane, aeroplane, plane, aircraft, fuselage, windshield, windscreen, vehicle, fly, fixed wing, powered, propeller, jet, central body, accommodate crew, passengers , transparent screen, protect, occupants* }

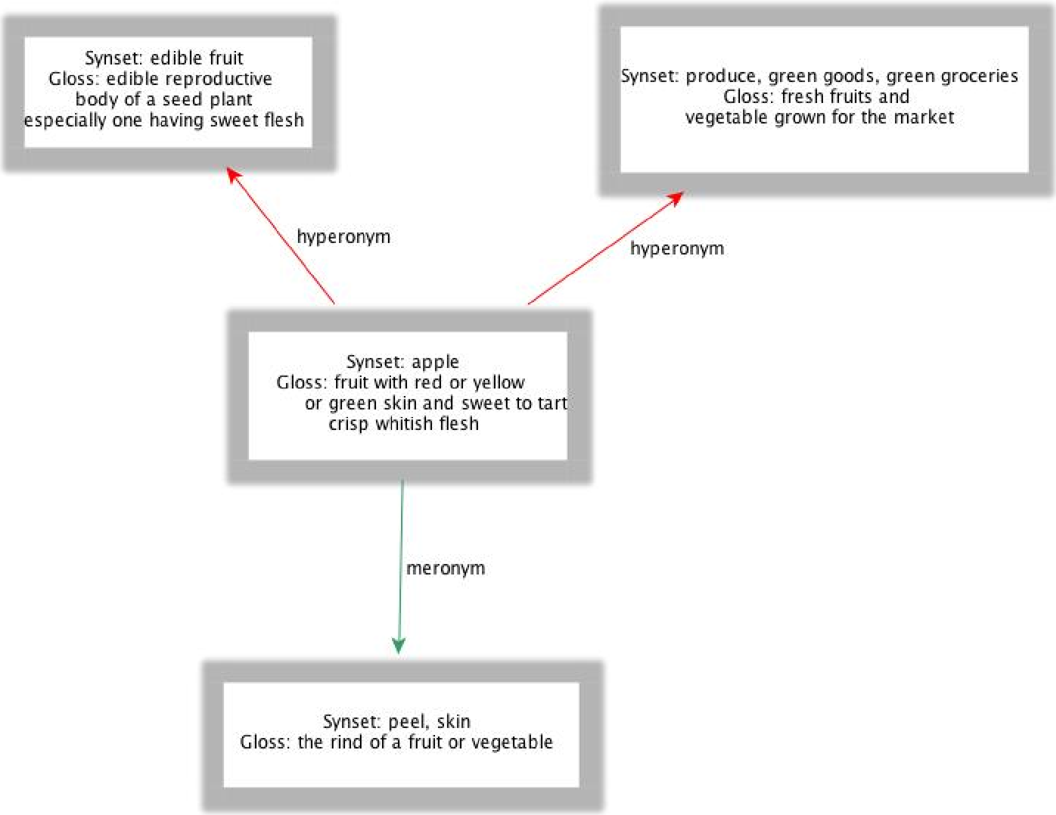

Figure 4.2: The semantic neighborhood for the concept apple -

Please observe that the algorithm extracts many good properties like *fruit, red, yellow, skin* for the concept **apple** or *vehicle, fixed winged, passengers* for the concept **airplane**, but it also extracts dubious properties like *protect* in the case of the concept **airplane**. A reasonable number of false properties will not affect

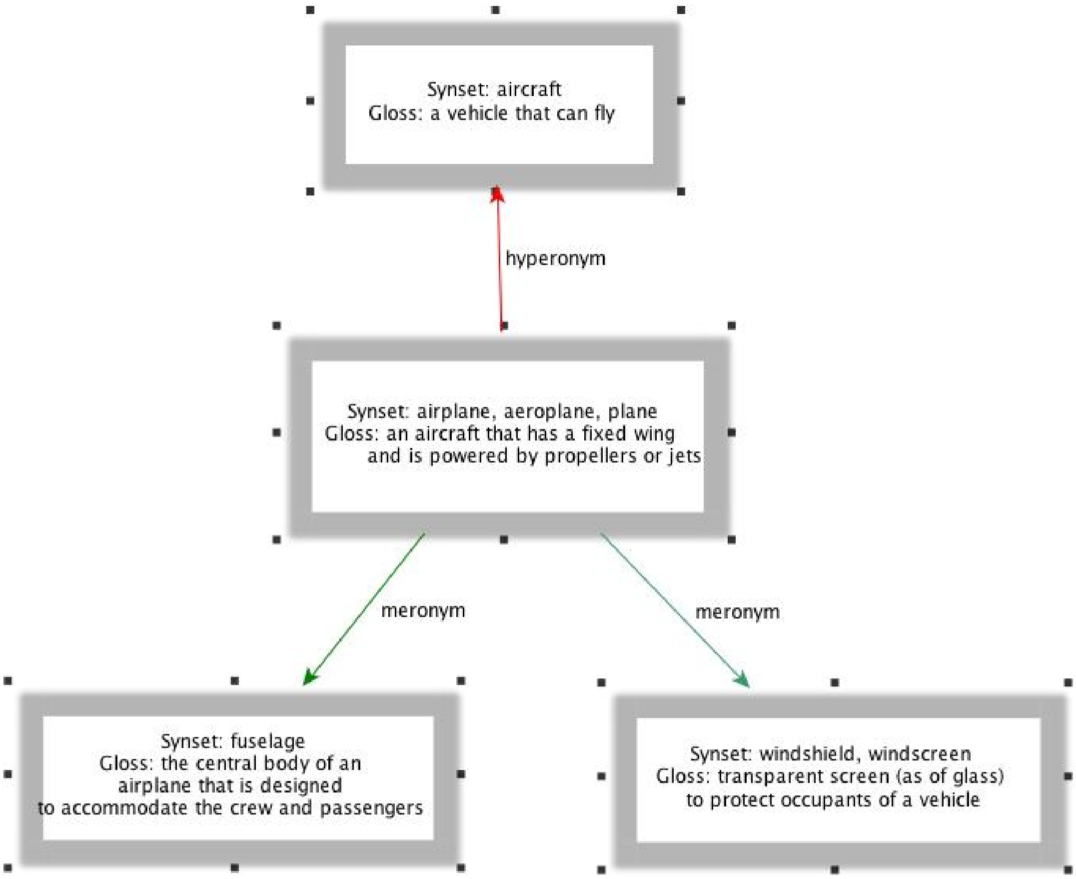

Figure 4.3: The semantic neighborhood for the concept airplane -

the comparison process between Wordnet and feature norm. The precision will be of concern only when someone wants to use Wordnet for property collection.

4.5 The comparison between Feature norms and Wordnet

Feature norms, as we showed before, are built having in mind the computational modeling of semantic memory. Therefore one would expect to find in Wordnet many properties produced by subjects in feature generation task. In this section we compare the properties extracted from Wordnet with the properties recorded in the McRae and Garrard norms. The purpose of this comparison is twofold: first we want to benchmark the Wordnet as a cognitive resource and second we want to see if and how Wordnet can be used in the process of automatic property collection. The results of the Wordnet assessment as a cognitive resource would allow Wordnet builders to improve the resource in the future by adding new kinds of relations. The automatic acquisition property technique will enable the collection of feature norm like properties from Wordnet.

4.5.1 Global Comparison

In performing the automatic comparison between feature norm and Wordnet we had to make two simplifying assumptions. In both Garrard and McRae feature norms *has legs* and *has four legs*, for example, are considered to be distinct properties. We neglect the cardinality and collapse these properties into one: *has legs*. We also consider that in case a property expresses a two-place relation and the relation is not explicitly defined in Wordnet (e.g. meronym or hyperonym), the presence of the arguments of the relation in Wordnet is sufficient for deciding that the relation linking the arguments in Wordnet is the same relation expressed by the feature norm property. For example, if we want to decide if the property *used for cooking* for the concept **pot** exists in Wordnet and we find the word *cooking* in the semantic neighborhood of the concept **pot** then we assume that the relation that holds between **pot** and *cooking* is the functional relation *used* *for*. For most properties in the feature norm this is true, but there are some cases when our second assumption is false.

Table 4.1 shows the proportion of the concept-property pairs in each feature norm found in Wordnet.

Feature Norm	CP pairs in feature norm	CP pairs in Wordnet	Percent in Wordnet
McRae	6925	2108	30%
Garrard	1537	342	22%

Table 4.1: A global comparison between feature norms and Wordnet

The meaning of the columns of the table 4.1 is the following:

- The **CP pairs in feature norm** column lists the number of concept-property pairs in each feature norm.
- The **CP pairs in Wordnet** column gives the number of concept-property pairs in the intersection between each feature norm and Wordnet.
- The last column shows the percent of the properties in the feature norm estimated to be in Wordnet.

It can be seen that the percent of concept-property pairs in the intersection between McRae feature norm and Wordnet is higher than the percent of concept-properties in the intersection Garrard feature norm and Wordnet (30 % vs. 22 %).

4.5.2 Per Property Type Comparison

This comparison is assessing which property types are better represented in Wordnet, which are weakly represented or lacking. Figure 4.4 shows the mapping between the properties extracted for the concept **apple** from Wordnet and the properties annotated with the corresponding property types in both feature norms. The property *skin* in Wordnet maps onto the property *has skin* which is a sensory property in Garrard feature norm and an External Component in McRae feature norm. The results of this comparison are given in the next two tables. The signification of the tables columns (4.2 and 4.3) is the following:

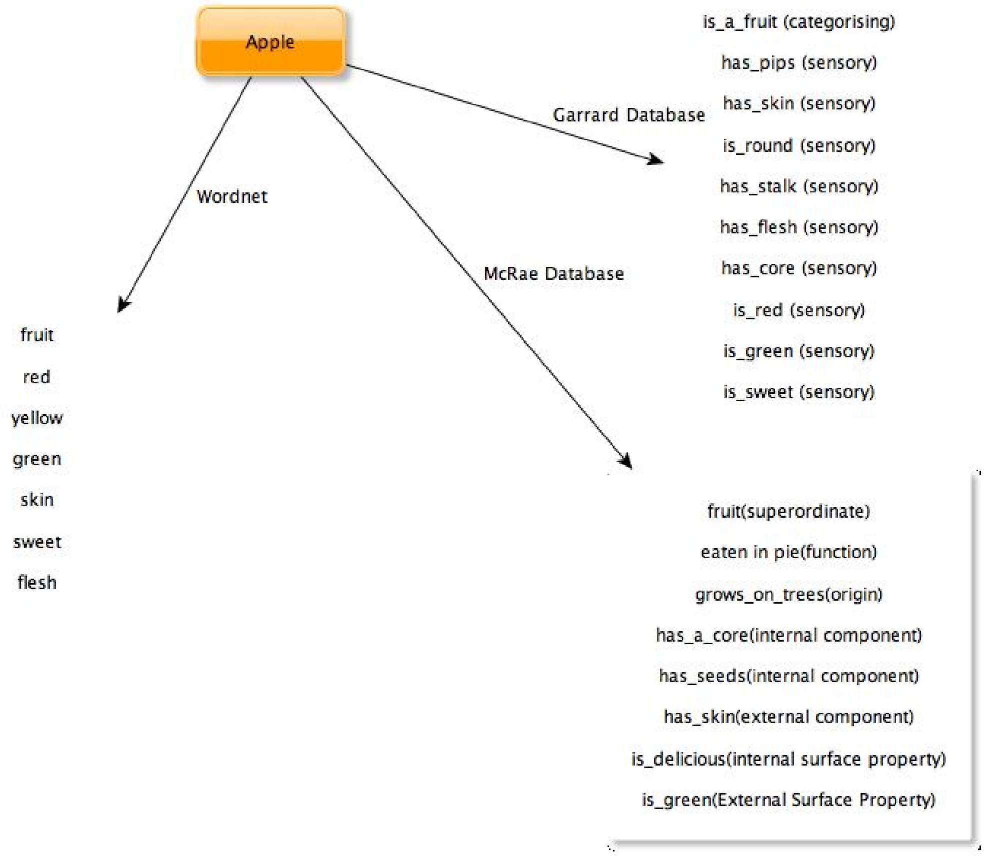

Figure 4.4: Mapping the property types for the concept apple -

- Property Type - represents the classification of the properties in each feature norm. As discussed earlier the two feature norms have distinct classification schemas.
- CP Pairs in feature norm - for each property type the number of properties in each feature norm is listed.
- CP Pairs in Wordnet - it numbers the concept property pairs in the intersection between the feature norms and Wordnet for each property type.

- Percent in Wordnet - the estimation of the percent of properties in the intersection between the feature norm and Wordnet for a certain property type. The lines of the table are ordered after the values of this column.

property Type	**CP pairs in feature norm**	**CP pairs in Wordnet**	**Percent in Wordnet**
Categorizing	115	83	72%
Sensory	737	190	25%
Encyclopedic	241	26	11%
Functional	444	43	10%

Table 4.2: Per property type comparison between Garrard feature norm and Wordnet

Table 4.2 gives the comparison for Garrard feature norm. The properties are classified using Garrard classification schema. As one expected the property type better covered by Wordnet is the classification type, 72 % of the classification properties produced by the subjects in Garrard experiment are found in Wordnet. All other property types are not so well represented, the second place is taken by the sensory properties with 25 %. Garrard classification schema is very crude and do not give us much information. Consider for example all the sensory properties of the concept **airplane**.

airplane={*is large, is made of metal, is fast, has cockpit, has fuselage, has propellor, has seat, has tailplane, has rudder, has control, has flap*}.

Under the label sensory there are lumped different kinds of properties: substance (*made of metal*), components (*has cockpit*), qualities across multiple dimensions: speed (*is fast*), size (*is large*).

Table 4.3 giving the comparison for McRae norm lists only the property types classifying more than 50 properties in the feature norms. Meeting our expectations the best property type in terms of coverage is the superordinate type (78 %), this being the only property type having the coverage over 50% in Wordnet. The relations that denote parts and that correspond to various types of Wordnet meronymy are relatively well-represented, occupying positions: 2, 3 and 7. The properties classified under External Surface Property property type occupy the fifth place. The high position in the table for the external surface properties can be explained by the fact that the definitions of many concepts denoting concrete objects list properties of their external surfaces (e.g. shape, color). For example, the definition of the concept **apple** contains the attributes *red*, *green* and *yellow*, all being external surface properties according to Wu and Barsalou taxonomy. The same explanation can be given for the rank of the Origin property types. The definitions of many concepts make reference to their origins (e.g. the **beehive** definition contains a reference to *bees*) Suprinsingly, the last property type, Evaluation, has no representation in Wordnet. The properties typed as evaluation reflect subjective assessment of objects or situations (for example the evaluation that a **bag** *is useful*, that a **blouse** *is pretty* or the **shark** *is dangerous*). I think that subjective assessments play important roles in the cognitive representation of some concepts. Arguably, we attach subjective evaluation to different things in the world (for example many people consider sharks being dangerous, bombs bringing destruction, etc.).

4.5.3 Per category comparison

In the previous sections we made a global comparison and a per property type comparison between each feature norm and Wordnet. In this section we want to find an appropriate classification for the concepts in the two feature norms and then to see which categories are better represented in Wordnet. The output of this comparison will be a list of categories (e.g. tool, animal, fruit, etc.) ordered by the percent of properties we find in Wordnet. To construct a classification of properties in the norms two methods are used. The first classification is derived with the help of Wordnet hyperonym relation. A second classification is derived from the information present in each feature norm. The final classification combines the Wordnet classification with the feature norm classification.

To derive Wordnet classification we generate the Wordnet tree along the hyperonym relation starting from the synsets in the projection set. We treat the synsets in the projection set as the objects to be classified and any category subsuming the synsets in the projection set as a potential classifier. To eliminate

property Type	CP pairs in feature norm	CP pairs in Wordnet	Percent in Wordnet
Superordinate	588	470	80%
External Component	926	442	48%
Internal Component	168	64	38%
Origin	59	16	27%
Contingency	91	24	26%
External Surface Property	1175	306	26%
Made Of	471	122	26%
Function	1089	281	25%
Participant	183	44	24%
Internal Surface Property	179	40	22%
Location	455	84	18%
Associated Entity	153	22	14%
Systemic Property	293	38	13%
Entity Behaviour	495	63	13%
Action	184	20	11%
Evaluation	105	0	0%

Table 4.3: Per property type comparison between McRae feature norm and Wordnet

some concepts in the classification tree we impose two requirements:

1. The chosen Wordnet concepts should form a maximum partition of the synsets in the projection set. The reason behind this demand is that we do not want a concept to vote for more than one category.

2. The chosen Wordnet concepts should be basic level categories. These are the categories which the subjects in property generation usually produce. For example, we accept the classifier *animal* for the concept **cat** but reject the classifier *placental*.

In figure 4.5 a part of the classification tree generated following the hyperonym relation in Wordnet is presented. The leaves of the tree are synsets in the projection set. For this simple example the problem of finding the best node to cut the tree at is stated in the following way. Should we cut the tree at the node *musical instrument* and classify the leaves of the tree as musical instruments? Or should we cut the tree at the nodes *free reed instrument* and *woodwind* and classify with these labels the nodes (**flute**) and (**harmonica** and **accordion**) respectively? To obtain the maximum partition we start by finding the smallest possible generalization of the concepts labeling the leaf nodes. A node gives the smallest possible generalization if it dominates at least two leaves. After collecting all the nodes satisfying the above condition we attempt a new generalization producing only those direct hyperonyms of the categories which maximize the partition. For the simple example in figure 4.5 we first generate the smallest generalization of the leafs: the node **free reed instrument**. Please notice we do generate the node **woodwind** because this node dominates only one synset from the projection set. The node **musical instrument** successfully generalizes **free reed instrument** because it maximizes the partition by adding to it the leaf **flute**. The criterion stating that the chosen Wordnet concepts should be basic level categories obliges us to cut the tree at the same node **musical instrument**. Unfortunately, we do not know a priori which concepts in the tree are basic level categories and therefore the chosen concepts are in some cases artificial categories.

The second method derives a classification from the feature norms themselves. The category field in Garrard feature norm and the Superordinate property type for the McRae give basic level classifiers for the concepts in the norms. Unfortunately, not all concepts in the norms have a classifier. Moreover, even in the case of concepts for which we find classifiers the resulting classification schema does not form a partition of the objects to be classified.

The first method gives a partition of the synsets in the projection set but not all categories of the partition are basic level concepts. The second method generates basic level classifiers that do not form a partition. Thus the final set of categories is derived starting from the categories produced by the subjects in each of the two experiments and inspecting the classification tree derived from Wordnet. The final basic level category partition came with the cost of not being able to cover the whole concept space:

1. For Garrard feature norm the following categories form a partition of 50

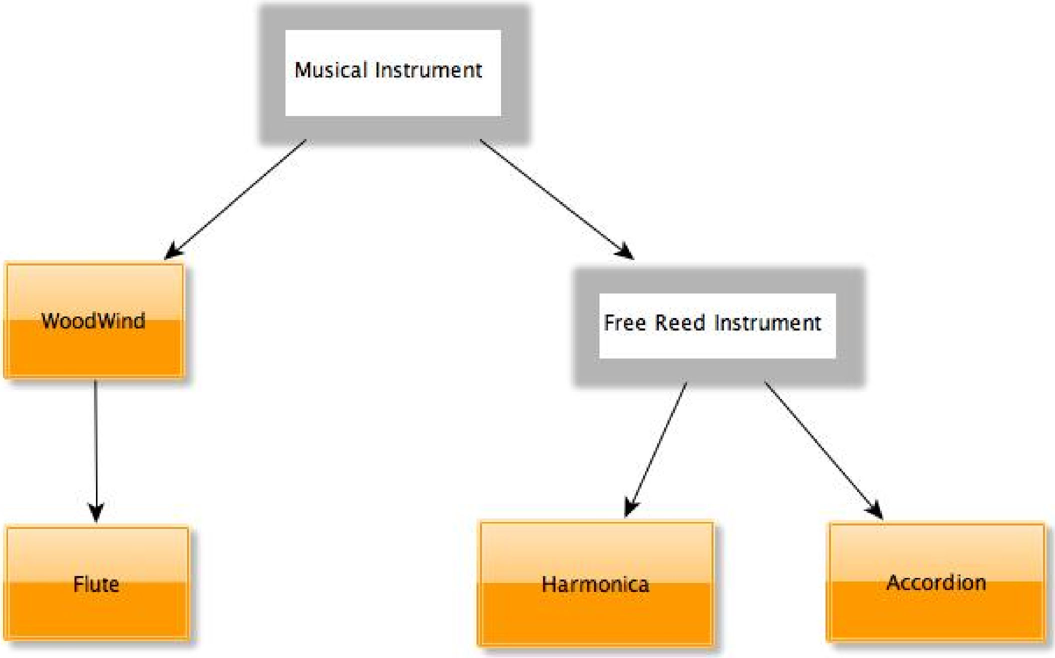

Figure 4.5: An example of a classification tree -

concepts: {**implement (tool)**, **bird**, **mammal**, **fruit**, **container**, **vehicle**, **reptile**}. In this partition the category of animals is split into reptiles, mammals and birds.

2. For the McRae feature norm the partition has 16 categories and covers 345 concepts: {**clothing**, **implement**, **fruit**, **furniture**, **mammal**, **plant**,**appliance**,**weapo** **container**, **musical instrument**, **building**, **vehicle**, **fish**, **reptile**, **insect**, **bird**}.

Category	Percent in Wordnet
Fruit	37%
Bird	34%
Implement	25%
Container	21%
Mammal	20%
Vehicle	20%
Reptile	17%

Table 4.4: Per category comparison between Garrard feature norm and Wordnet

Category	Percent in Wordnet
Fish	52%
Fruit	43%
Vehicle	43%
Bird	42%
Plant	37%
Musical Instrument	36%
Weapon	33%

Table 4.5: Per category comparison between McRae feature norm and Wordnet

The next two tables 4.4 and 4.5 give the percent of properties in Wordnet for each category in the partition of the two feature norms[1]. It can be noted that

[1]Only the 7 top most categories are presented

the order of the best represented categories in Wordnet slightly differs in the two feature norms. For example, the category **Bird** is the second well represented category for Garrard feature norm but only the 4th category in McRae feature norm. As expected, given the results of the global comparison the categories in McRae feature norm are better represented than the categories in Garrard feature norm. For example, the best represented category from McRae feature norm is **Fruit** having 52 % coverage in Wordnet. Instead the top represented category from Garrard feature norm **Fish** has 37 % coverage in Wordnet.

4.6 Discussion

The automatic comparison was performed under the hypothesis that the properties in the feature norms and Wordnet are registered using the same words. In this section we asses the extent to which this statement is true. For evaluating the accuracy of the automatic mapping procedure a manual comparison between 20 Wordnet concept descriptions and each of the two corresponding feature norms concept descriptions was performed. The concept set is a balanced one: 10 concepts have the highest properties overlap with the norms and 10 concepts that have the lowest property overlap with the norms. The mapping between the feature norm concept description and Wordnet concept description revealed that the number of concept-property pairs common to feature norms and Wordnet is bigger than the estimation given by our algorithm. There are three reasons for this fact.

1. The first reason is that some properties are expressed differently in Wordnet and in the feature norms. By way of example let's consider one of the properties of the concept **anchor** in McRae feature norm: *found on boats*. The definition of the concept **anchor** in Wordnet contains a semantically close word to **boat**: **vessel**[1]. Therefore from the fact that **boat** and **vessel** are semantically close it can be understood that the property *found on boats* is present in the gloss of the concept **anchor**. We could have exploited

[1]Actually, Wordnet says that **vessel** is a hyperonym of **boat**.

this information if the words in glosses would have been semantically disambiguated. However, even a perfect WSD of the words in the glosses does not completely solve this problem. It is a notorious fact that Wordnet makes very fine sense discriminations and therefore many near synonyms of the words in the glosses would not be found.

2. The second reason for the inaccuracy of our automatic procedure is related with a general problem of feature norms. It is assumed for methodological simplicity that properties listed in the property production task are independent, which, as we will show, is false. One of the most important relations linking the properties in the norms is entailment. For example, the concept **trolley** properties: *used for carrying things* and *used for moving things* are related by entailment. If someone carries some things with a trolley he will always move the things(carry implies moving). Moreover the entailment relation does not hold only between the properties in the feature norm but also holds between the properties in Wordnet and the properties in the norms. The property *has a container for transporting things* of the same concept **troley** is inherited from one of its Wordnet hyperonyms **wheeled vehicle**. This property entails both properties of the concept **trolley** in McRae feature norm: *used for carrying things* and *used for moving things* as it can be seen in figure 4.6. To give another example let's consider the functional property of the same concept **anchor**: *used for holding the boats still.* It is logically equivalent with the property *prevents a vessel for moving* found in the Wordnet gloss of the same concept.

3. The third reason why the automatic comparison fails to reveal the true overlap between feature norms and Wordnet is the incompleteness of Wordnet. A very salient property the subjects produce when they describe concrete objects are the parts of the respective objects. However many concepts from the projection set lack the meronyms in Princeton Wordnet. This is true especially for living things like fruits and animals (e.g. **apple**, **cheetah**). Tools, vehicles and musical instrument's parts are much better represented in Wordnet.

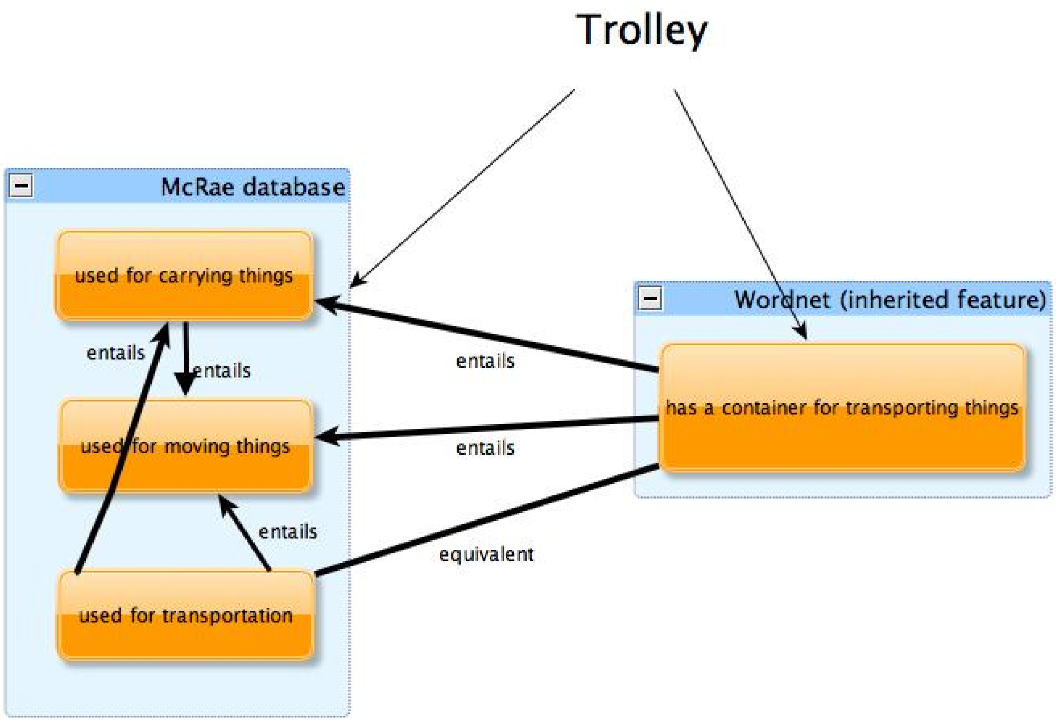

Figure 4.6: Entailment relation between properties -

The manual comparison between the 20 concepts in the two feature norms and Wordnet show an overlap of approximately 40 % with McRae feature norm and 30 % with Garrard feature norm. These figures raise by 10 percent the results of the automatic comparison in the table 4.1. The comparison between feature norms and Wordnet reveals some potential improvements for the future Wordnet versions. But before discussing some interesting properties Wordnet lacks we notice that except for the Superordinate property type and the property types related with parts Wordnet does not explicitly represent any other properties. Wordnet improvement means for us the explicit representation in Wordnet structure of a number of additional relations. The properties missing in Wordnet can be found inspecting the table 4.3. Next we briefly discuss three-property types present in feature norm but missing from or underrepresented in Wordnet: the evaluation, the associated entity and the function property types.

- **Evaluation**. As we demonstrate this property type is absent from Wordnet even if the evaluation properties are an important part of the semantic representation for some concepts. We do not think that every possible subjective evaluation should find a place in Wordnet, but only the most salient ones (what counts as a salient property should be decided by the Wordnet builders). In particular, we think that the evaluations: **sharks** *are dangerous* or **hyenas** *are ugly* should be part of the Wordnet entry for shark and hyena respectively. This property type is completely absent from Wordnet and, sadly, the glosses could not be exploited to make the property explicit.

- **Associated Entity**. Other interesting property type under-represented in Wordnet is the associated entity property type. The concepts in the property generation tasks usually denote concrete objects. It is no surprise that the mental representation for these concepts includes the entities associated with these objects. For example, we typically associate an anchor with the chains or ropes it is attached to or we associate an apple with the worms it may be infested by or even we associate bagpipes with Scotland. The major obstacle in explicitly adding this relation in Wordnet is the lack of a precise definition. Consider the three cases given as association entity examples. In the first two cases (**apple** associated with **worms** and **anchor** associated with **ropes**) the association takes place at the physical level (the worms are physically present in the apples and the ropes are tight together with the anchor). In the third case the nature of association is encyclopedic: the subjects in McRae experiment think that the bagpipes are representative for Scotland. This association is supported by the largest online encyclopedia. According to Wikipedia the bagpipes are widely used in Europe, Northern Africa, Persian Gulf and Caucasus, but the Scottish and Irish varieties have the widest spread visibility.

- **Function**. The function or role that an entity serves for an agent is an important part of the meaning of the respective entity. The keys are used for locking or opening the doors, the baskets are emptied and filled, the trolleys are used for transporting things and the garages are utilized for storing cars. The function relation is widely studied and in some cases it is even formalized. We think that at least for artifacts the function should be explicitly added as a Wordnet relation. In some cases the glosses can be exploited for a semi-automatic function relation generation (281 of the 1098 function properties in McRae feature norm are present in Wordnet glosses).

The main conclusion of the comparison between Wordnet and feature norms is that Wordnet should be redesigned if it is wanted to be a model of the human semantic memory. Based on the property-norm collection effort in cognitive psychology we suggested a set of new relations that Wordnet could implement. We also showed that some of these relations can be added exploiting the information in the glosses. The main obstacle in adding the relations to Wordnet is the lack of formalization of some of the relations.

4.7 Chapter Summary

In this chapter we devised a procedure for automatic property collection from Wordnet (4.4). Then we made three automatic comparisons between Wordnet and the feature norms (4.5). First we compared the resources at the global level, then we compared them at the property type level and finally we made a per category comparison. We argued that the Wu and Barsalou taxonomy is a better classification schema than the crude classification employed by Garrard and colleagues. The main results of the comparison are presented in table 4.3. Inspecting the table it can be seen which properties are absent or underrepresented in Wordnet and which properties can be automatically collected from Wordnet. We also performed a manual comparison (4.6) between the extracted Wordnet properties and feature norms. The manual comparison revealed that the mapping between Wordnet and norms is a more complex process and that the properties are interliked with the entailment relations. There are entailment relations which hold between properties in feature norms and there are also entailment relations holding between the properties in the norms and the properties in Wordnet. Moreover, we showed that the overlap between Wordnet properties and properties norm is higher than the automatically found overlap and the figures approach the overlap between the two feature norms (4.6).

5

Extracting feature norm like structures from corpora

5.1 Contributions

In this chapter we bring the following contributions:

1. We propose a method to acquire feature norms like structures from corpora using weakly supervised methods. To this end we map the property types in Wu and Barsalou taxonomy to a better set of relations for automatic learning.

2. In the process of learning the relations in the new schema we benchmark four association measures at the task of finding good patterns for 4 semantic relations. Two of the four relations are statistically based and are used for the first time for this task.

3. We also made some preliminary experiments for automatic acquisition of a set of relations using kernel based methods.

5.2 A new schema for Property Classification

We already discussed in chapter 3 a perceptually oriented classification of the properties in the norms: the Wu and Barsalou classification schema. A variant of this schema is used to annotate the properties in McRae norm. In this subsection based on McRae variant of Wu and Barsalou schema we build a new schema. We emphasize that the property types of the new schema can be learned much easier than the original property types.

First we try to persuade the reader that the original distinctions made in Wu and Barsalou schema, justified from a psychological perspective, make very difficult the learning process. Consider for example the distinctions that the schema operates between external and internal properties. According to the perceptual simulation theory there is a difference in how the mind accesses this two types of properties. The external properties are perceived on the surface of the objects while the internal properties are perceived beyond the object surface. According to Barsalou the subjects scan[1] the objects for finding their internal properties. Consequently the schema distinguishes between the external surface properties (e.g. color, texture, size) perceived on the surface of objects and internal surface properties (e.g. color, taste) perceived inside the object surface. In the case of apple for instance the color red is an external surface property (e.g. **the apple** *is red*) and the color white is an internal surface property (**the apple** *is white inside*). The same distinction between external and internal also operates in case of object components: the external components at least in some measure reside on the surface of objects in contrast with internal components which reside totally inside the object surface. We think that from a practical point a view the distinction between internal and external properties is misleading. Assume that someone asks the question: Is a part of an object in the interior of the object? Paradoxically we cannot answer this question using the distinctions in Wu and Barsalou taxonomy. In agreement with the above definition an external component could have or could have not internal parts. Moreover, there is very hard to automatically learn if a component or property is or is not external using shallow methods. The reason is that this information is coded in the world knowledge

[1]The mental scanning is akin to imagining, just that in Barsalou opinion the scanning can also be unconscious.

semantics[1].

In any case our task is to extract in a cheap way feature norms like descriptions but we are not tied in any way to a particular schema. Because of the above observations we decided to start with Wu and Barsalou schema and modify it so that it fits automatic property acquisition. The new schema has the following property types[2]:

1. **Superordinate**. The superordinate property types are those properties that classify a concept from a taxonomic point of view. For example, the **dog** (focal concept) *is an animal* (taxonomic property).

2. **Part**. The Property Type part includes components of an object. For example *blade* (part property) is a part of an **axe** (focal concept). Part properties are obtained by merging both internal and external components in Wu and Barsalou taxonomy .

3. **Stuff**. The property types labeled Stuff denote the stuff an object is made of. For example, **bottle** (focal concept) *is made of glass* (stuff property). This is a property type not present in the original Wu and Barsalou schema but added by McRae and collegues.

4. **Location**. The properties labeled with the type location denote typical places where instances of the focal concepts are found. For example, **airplanes** (focal concept) *are found in airports* (location property). It is equivalent with the Location property type in Wu and Barsalou category.

5. **Action**. This class of properties represents the characteristic actions defining the behavior of an entity (the **cat** (focal concept) *meow* (action property)) or the function, instances of the focal concepts typically fulfill (the **heart** (focal concept) *pumps blood* (function property)). This property type maps the Entity Behavior and Function in Wu and Barsalou taxonomy. It

[1] Usually it is assumed that people know that the engines are inside the cars, for example. However, it is possible that a richer semantic representation can be exploited to understand Wu and Barsalou distinctions.

[2] In this chapters the concepts in the norms will be called either focal concepts or target concepts.

is interesting that in the original taxonomy the Function is considered a Situation Property while Entity Behavior is labeled as an Entity Property. Functions are assigned to an entity by an agent and are goals for the respective agent. Behavior instead is the normal function of an entity and it does not serve any purpose for an agent. This distinction is hard to understand even for humans. Consequently, to formalize it and make the computers to grasp it is even harder.

6. **Quality**. This class of properties denotes the qualities (color, taste, etc.) of the objects instances of the focal concepts. For example, the **apple** (focal concept) *is red* (quality property) or *is sweet* (quality property). The category maps both internal and external surface properties in Wu and Barsalou Taxonomy.

The property types enumerated above do not cover all property types in WB taxonomy but only the most salient ones. A property type is salient if it classifies at least 100 properties in McRae feature norm.

In conclusion, the most relevant properties produced by the subjects in the property generation experiments are in the categories presented above. Thus, asked to list the defining properties of the concepts representing concrete objects subjects will typically: classify the objects (Superordinate), list their parts and the stuff they are made from (Parts and Stuff), specify the location the objects are typically found in (Location), their intended functions, and their typical behavior (Action), or name their perceptual qualities (Quality). At the semantic level the properties listed by subjects can be classified using the above schema. At the morphological level a property is expressed by a noun, verb or adjective. Table 5.1 gives the morphological and semantic classification (property types) of the properties in the McRae database. Obviously, part relation holds between nouns, action relations between nouns and verbs and qualities are generally expressed by adjectives.

Semantic Classification	Morphological Classification
Superordinate	Noun
Part	Noun
Stuff	Noun
Location	Noun
Action	Verb
Quality	Adjective

Table 5.1: The semantic and morphological classification of properties in McRae feature norm

5.3 Property Learning

For learning the property types in the norms we use the schema devised in the previous section together with the information in the table 5.1. In the learning process we employ two different strategies. Superordinate, Part, Stuff and Location properties are learnt using a pattern-based approach. Quality and Action properties are learnt using a novel method that quantifies the strength of association between the nouns representing the focal concepts and the adjective and verbs co-occurring with them in a corpus. Please note that the relations to be learnt are all binary.

The learning decision is motivated by a corpora based experiment. The purpose of the experiment is to see in what kind of contexts the seeds[1] appear in corpora. We took concept-property pairs from McRae feature norm and extracted sentences from a corpus where a pair concept - property appears in the same sentence. We noticed that, in general, the quality properties are expressed by the adjectives modifying the noun representing the focal concept. For example, for the concept property pair (**apple**, *red*) we find contexts like:

"She took the red apple."

The action properties are expressed by verbs. The pair (**dog**, *bark*) is conveyed by contexts like:

"The ugly dog is barking."

[1]The seeds are the words representing concepts linked by a binary relation.

where the verb expresses an action to which the dog (i.e. the noun representing the concept) is a participant.

The experiment suggests that to learn Quality and Action properties we should filter the adjectives and verbs co-occurring with the focal concepts. For the rest of the property types the extracted contexts recommend that the best learning strategy should be a pattern-based approach. Moreover, with the exception of the Location relation, that, to our knowledge, has not been studied yet, for the relations Superordinate, Part and Stuff some patterns are already known. The properties we try to find lexico-syntactic patterns for are classified at the morphological level as nouns (see Table 5.1). The rest of the properties are classified as either adjectives (Qualities) or verbs (Action). The following two subsections present the methodology for property type learning.

5.3.1 Learning Superordinate, Stuff, Location and Part relations

The idea of finding lexico-syntactic patterns expressing with high precision semantic relations was first proposed by Hearst (Hearst (1998)). Nowadays the patterns are ubiquitously used in a series of tasks like: relation extraction, ontology learning and question answering. Unfortunately, most of the patterns are formulated using intuitive judgments and very limited corpus analysis. I do not think that a fully automated procedure can be constructed for weakly supervised pattern learning. However, what we can hope is to find a procedure to limit the searching space for the patterns. More precisely, the procedure would propose to the linguist some n pattern candidates which ideally would include the most precise lexico-syntactic patterns. Upon inspection and testing the researcher can manually pick up the best candidates. In this section we try to automate to some extent the procedure for best pattern identification.

First we need to see what was the framework Hearst proposed in its seminal paper and which phase in the framework can be automatized. For identifying the most accurate lexico-syntactic patterns, that is the patterns which have high precision and possibly also high recall, she defined a bootstrapping procedure.

The procedure iterates between three phases called: pattern induction, pattern ranking and selection, and instance extraction:

1. **Pattern Induction** In the pattern-induction phase a relation of interest is choosen (for example hyperonymy) and a list of instances of the relation is produced. Subsequently, all contexts in corpus containing these instances are gathered and their commonalities identified. These commonalities form the list of potential patterns.

2. **Pattern Ranking and Selection** In this stage the most salient patterns expressing the semantic relation are identified. Usually, the best patterns are discovered inspecting the list of the potential patterns and computing the precision of the most salient ones.

3. **Instance Extraction** Using the best patterns one gathers new instances of the semantic relation. The algorithm continues from the first step and it finishes either when no more patterns can be found or the number of found instances is sufficient for a certain task.

There are some steps in Hearst framework that can be formalized. In the first place it can be asked what a context is. The answer seems self evident to the researchers and nowadays nobody asks this question. The context where the seeds should co-ocurr is the sentence. Moreover, due to the language syntactic structure the probability that a relation is expressed between two words in a sentence decreases with the distance. The second clarification has to do with the meaning of the term commonality. What does it mean the identification of commonalities between the contexts where the seeds occur? An answer to this question is provided by Ravichandran and Hovy (Ravichandran and Hovy (2002)) who defined the commonality as being maximum common substring that links the seeds in k distinct sentences. However, the hardest think to formalize is the procedure of pattern selection. The first formal answer to this question was given by the same Ravichandran and Hovy. They said that probably the best patterns are the most frequent substrings linking the seeds. Another answer is provided by Pantel and Pennachiotti (Pantel and Pennacchiotti (2006)) who used an information theoretic measure to find the best patterns[1] In the next section we will mathematically formulate the two measures proposed before (by Ravichandran and Hovy and Pantel and Pennachiotti[2]) and two new statistical association measures. Then we will benchmark all association measures with the aim of finding good patterns for schema relations holding between the nominals: Superordinate, Part, Stuff and Location.

5.3.1.1 Association measures for Lexico-Syntactic Pattern Learning

A binary semantic relation can be defined from an extensional point of view indicating the set of instances it denotes. In a corpus the strings linking the instances in the extension of the semantic relation may:

1. express the semantic relation with good precision. This is the case with patterns like : "NP such as NP" for the IS-A relation.

2. express many other semantic relations (the case of general patterns). The challenge in this case is to find a procedure for differentiating the expressed semantic relations. Usually, this involves deriving constraints for the relation arguments that can be used by supervised learning algorithms. Example of very general patterns are the possessive constructions like "NP of NP" which expresses parts (the handle of the door), possession (the leg of the table) and a bunch of other relations. Because the relation learning algorithm is weakly supervised we will get rid of many general patterns in the pattern induction phase.

3. express spurious constructions which appear with instances of the relation in the corpus. For example, the sentence: "John picked up the fruit and ate the apple" does not express a IS-A relation between the seeds fruit and apple but the seeds happened to occur in the sentence.

The hypothesis we pursue is that the best lexico-syntactic patterns are those highly associated with the instances representing the relation of interest. To

[1] In fact, their method is more complex. The patterns are voted by the seeds and the seeds vote the patterns in an iterative process.

[2] Our measure is slightly different but is in the same class of measures.

calculate the association strength between the relation instances and patterns we test four association measures: frequency, point mutual information, Chi-squared and Log-Likelihood. We follow the framework introduced in previous section:

1. The pattern induction phase starts with a set of seeds instantiating one of the four semantic relations. We collect sentences where the seeds appear together at a distance of at most five words[1]. Every seed occurrence in the sentence is replaced with the label Noun. For example, if our seeds are apple and fruit and we extract a sentence like: "They prefer fruit like apples, but not bananas" after the replacement the sentence will be changed into: "They prefer **Noun** like **Noun** but not bananas".

2. The potential patterns are computed as suggested by Ravichandran and Hovy getting the longest common substring including the seeds between two extracted sentences. By way of example, consider the following two sentences after the seed replacement by nouns:

    (a) They prefer **Noun** like **Noun** but not bananas.

    (b) **Noun** like **Noun** are found at the zoo.

    The longest common substring is "Noun like Noun" which will be added to the potential pattern table together with a frequency equal to 2 because it appears in two sentences. Further we eliminate from the pattern table all the patterns which contain between the seeds only expressions in a stop list. This list contains conjunctions and very common expressions. Thus, patterns like : "Noun and Noun" or "Noun or Noun" will be eliminated.

The remaining patterns in the list are ranked using the above mentioned association measure. Next we present the association measures giving their mathematical formulas and briefly motivating their usage. Before presenting the mathematical formulas it is useful to introduce the following notation:

[1] The likelihood that a substring which connects two words in a sentence expresses a relation between the words decreases with the length of the string.

1. The set of relation instances in the training set is given by the equation 5.1. Because the relations under study are binary, each instance is a pair of seeds $i_j = \{s_{1j}, s_{2j}\}$, $j = 1 : n$

$$I = \{i_1, i_2 \dots i_n\} \tag{5.1}$$

2. The set of potential patterns is given by the equation 5.2. The potential patterns are the longest common substrings including the seeds and are the results of the pattern induction phase.

$$P = \{p_1, p_2 \dots p_k\} \tag{5.2}$$

3. The set of all pattern-instance pairs is denoted by S and illustrated by the equation 5.3.

$$S = \{\{i_1, p_1\}, \{i_1, p_2\} \dots \{i_m, p_k\}\} \tag{5.3}$$

4. If we consider an instance i from 5.1 and a pattern p from 5.2, then the following measures are defined and illustrated in the contingency table 5.2

   (a) O_{11} the number of occurrences the instance i has with the pattern p .

   (b) O_{12} the number of occurrences the instance i has with any other pattern except p .

   (c) O_{21} the number of occurrences any other instance except i has with the pattern p.

   (d) O_{22} the number of occurrences any instance except i has with any pattern except p.

   (e) R_1 and R_2 are the sum of contingency table 5.2 rows

   (f) C_1 and C_2 are the sum of contingency table 5.2 columns

   (g) N is the number of all instances with all patterns (the cardinality of S).

We are now able to mathematically define the four association measures:

	p	$\neg p$	Row Sum
i	O_{11}	O_{12}	$R_1 = O_{11} + O_{12}$
$\neg i$	O_{21}	O_{22}	$R_2 = O_{21} + O_{22}$
Column Sum	$C_1 = O_{11} + O_{21}$	$C_2 = O_{12} + O_{22}$	$N = R_1 + R_2$

Table 5.2: The contingency table

1. **Simple Frequency**. It is the cell 11 in the contingency table 5.2 and gives the number of occurrences of a pattern with an instance. The idea behind this association measure is that the most frequent lexico-syntactic patterns (the patterns occurring with many instances in I) express with good precision the semantic relation under study. With the increase of instances of a relation it is hoped that the spurious constructions will have low frequency and the good patterns will have high frequency.

$$O_{11} \tag{5.4}$$

2. **(Pointwise) mutual information** (Church and Hank (1990)). This measure should not be confused with the mutual information measure in information theory. The measure has the advantage that it quantifies the degree of association between the patterns and instances. A positive score means association and negative scores represent repulsion. It was found that this measure is biased toward infrequent events and that the bias increases for large corpora. In practice a correction is used to counterbalance the bias effect. We use the most popular correction in the family MI^k, MI^2 (equation 5.5). The advantage of this measure is that the pattern candidates which have low frequency but nevertheless are potentially good can be voted.

$$MI^2 = \log_2 \frac{O_{11}^2}{\frac{R_1 C_1}{N}} \tag{5.5}$$

3. **Chi-squared (with Yates continuity correction)** (DeGroot and Schervish (2002))

If the first two measures were used in the past for the automatic pattern acquisition, the next two measures are new to the task. Unlike the preceding measures the next two measures have a strong statistical background. The general statistical theory which has a solid mathematical background recommends that the measures of association of events should be based on the cross-classifications. A way of exploiting the idea of cross classification are the contingency tables (see the contingency table 5.2).

The first statistical association measure we use is the chi-Squared measure, a generalization of z-scores. Chi-Squared, like pointwise mutual information, exhibits a bias towards low-frequency events. Thus, in practice the Yates continutiy correction is used as in equation 5.6.

$$chi_{corr} = \frac{N(|O_{11}O_{22} - O_{12}O_{21}| - \frac{N}{2})^2}{R_1 R_2 C_1 C_2} \tag{5.6}$$

4. **Log-Likelihood** (Dunning (1993)) The Log-Likelihhod measure is one of the most used association measures in computational linguistics and its formula is given by the equation 5.7

$$log - likelihood = 2 \sum_{ij} \log \frac{O_{ij}}{\frac{R_i C_j}{N}} \tag{5.7}$$

Each of the above measures calculates the strength of association between each pattern in P and each instance in I.

The highest ranked patterns voted by the measures will be the patterns having the higher association score with the all instances in I. Therefore if we denote by $S_{p_j}^{assoc}$ the overall score of the patterns p_j with all instances in I for the association measure $assoc$[1] then we have (equation 5.8):

$$S_{p_j}^{assoc} = \sum_{jk} s_{p_j i_k} \tag{5.8}$$

where $s_{p_j i_k}$ represents the association strength between the pattern p_j and the instance i_k.

[1]Evidently $assoc$ is one of the four association measures presented.

5.3.1.2 Voted patterns for Superordinate, Stuff, Location and Part relations

For pattern learning we used BNC and for calculation of pattern precision we use uKWac (Ferraresi *et al.* (2008)). BNC is a balanced corpus having approximately 100 million words. uKWac is a very large corpus of British English, containing more than 2 billion words, constructed by crawling the web.

The instances for each binary relation (Superordinate, Stuff, Location and Part) are taken from the McRae feature norm and used as seeds for pattern learning. In fact, the concepts in McRae database are split into two sets: a test set comprising 44 concepts and a training set made by the rest of concepts. The instances for the learning phase are all concept-property pairs in the training set. All contexts where the seeds appear in BNC at a distance of at most 5 words are extracted . The context extraction and association measure computation were performed using the CWB(Oli (1994)) and UCS toolkits (http://www.collocations.de/software.html). The next table 5.3 shows the top voted patterns for each relation by each association measure. The column names : PMI stays for pointwise mutual information, Chi for Chi-Squared and LL for Log-Likelihood.

Because BNC is a clean corpus the punctuation is a good indicator of the topic. Therefore the patterns: "N such as N" and "N, such as N" are considered different because they differ by a comma.

The pattern precision is computed in the following way. A set of maximum 50 concept-property pairs is extracted from uKWac corpus using the voted patterns. Then we label a pair as a hit if the semantic relation holds between the concept and the property in the pair and a miss otherwise. The pattern precision is defined as the percent of hits. For example, to evaluate the precision of the pattern: "Noun made of Noun " for the Stuff relation we extract concept property pairs like **hammer** - *wood*, **bottle** - *glass*, **car** - *cheese*, etc. We have two hits: **hammer** - *wood* and **bottle** - *glass* and one miss: **car** - *cheese*. Thus, we have a pattern precision of 66 %.

For each relation only the most three significant patterns are reported. The first row of the table 5.3 contains the voted patterns for the superordinate relation.

Relation	Frequency	PMI	Chi	LL
	N such as N	N such as N	N be not the only N	N such as N
Superord	N like N	N, such as N	N other than the N	N, such as N
	N, such as N	N like a N	N or other N	N other than N
	N from the N	N make of N	N make of N	N make of N
Stuff	N be in N	N be in N	N be in N	N be in N
	N make of N	N from the N	N be N	N from the N
	N from the N	N from the N	N cold N	N from the N
Location	N have N	N out of the N	N black N	N cold N
	N to the N	N where at the N	N back from the N	N black N
	N have N	N round her N	N round her N	N have N
Part	N for N	N like a N	N make of N	N round her N
	N from the N	N from the N	N full of N	N make of N

Table 5.3: The top voted patterns for each relation

All voted patterns except "N like a N" and the variant "N like N" have a precision over 70 %. In this case no association measure wins the competition for patterns learning. A very small advantage has the **Chi-squared** and **Log Likelihood** measures mainly because of the presence of the high precision pattern "N be not the only N" and "N other than N", respectively. However, the difference with respect to the other measures is not significant. It is interesting that all measures seem to vote good patterns for this well studied relation.

The only potentially good pattern voted for the Stuff relation is "N make of N"; the other voted patterns are in the class of spurious constructions appearing with the relation instances. The association measures differ only by the weight they give to the voted patterns. All measures except the simple frequency vote the pattern in the first position.

None of the Location patterns expresses precisely enough the concept of location we are looking for. In many cases the patterns "N from (the) N" and "N out of (the) N" convey a very contextual concept of location (e.g. apple from the kitchen). Moreover, among the patterns extracted, there are some spurious constructions appearing with the instance of location relation in BNC corpus: "N black N", "N cold N".

It is known from previous studies on part-of relation that, in general, the lexico-syntactic constructions expressing this relation in corpus are general: possessives and statements with the verb "have". The pattern "N have N" is voted as the most significant one by the frequency and log-likelihood measures. Unlike the above three cases where the general rule is that the measures differently order the same patterns, in the case of part relation the voted patterns are more diverse. However, none of the voted patterns expresses in a precise way the part relation.

5.3.1.3 Extracting Superordinate, Stuff, Location and Part properties

In the last section we benchmark four association measures at the task of finding precise patterns expressing four relations. In this section, exploiting the findings in the last section, consulting the relevant literature and using our intuition we select the best lexico-syntactic patterns to represent the four relations above. The selected patterns are recorded in the table 5.4.

Relation	Pattern
Superordinate	Noun [JJ]-such [IN]-as Noun
	Noun [CC]-and [JJ]-other Noun
	Noun [CC]-or [JJ]-other Noun
Stuff	Noun [VVN]-make [IN]-of Noun
Location	Noun [IN]-from [DT]-the Noun
Part	Noun [VVP]-comprise Noun
	Noun [VVP]-consists [IN]-of Noun

Table 5.4: The selected patterns

The results of property extraction phase are reported in Table 5.5. The columns of the table represent in order: the property types to be extracted, the recall of our procedure and the pattern precision. The recall tells how many properties in the test set are found using the patterns in Table 5.4. The pattern precision states how precise the selected pattern is in finding the properties in a certain semantic class and it is computed as shown before. In case more than one pattern have been selected, the pattern precision is the average precision for all selected patterns.

Property Class	**Recall**	**Pattern Precission**
Superordinate	87%	85%
Stuff	21%	70%
Location	33%	40%
Part	0%	51%

Table 5.5: The results for each property class

The recall for the superordinate relation is very good and the precision of the patterns is not bad either (average precision 85%[1]). However, many of the extracted superordinate properties are roles and not types. For example, **banana**, one of the concepts in the test set, has the superordinate property: *is a fruit* (type). We find that **banana** *is a fruit* (a type) but also *is an ingredient* and *is a product* (roles). The lexico-syntactic patterns for the superordinate relation blur the type-role distinction. Other extracted pairs for the superordinates relation include (the left side of the pair contains a concept from the test set, while the right side lists its extracted superordinates): **cat**- (*pet*, *animal*), **potato-**(*vegetable*, *food*), **chicken**-(*bird*, *product*). In general, the extracted taxonomic knowledge is accurate if we neglect the type role distinction. However, for some applications the extracted roles can be of no import (e.g. **knife**-*equipment*). All patterns are proposed by all association measures except the second pattern in the table which, nevertheless, is a variant of the third pattern (”or” is substituted by ”and”).

The pattern used to extract Stuff properties has a bad recall (21 %) and an estimated precision of 70 %. To be fair, the pattern expresses better than the esti-

[1]To eliminate the contextual roles we impose that a property should be voted at least 3 times.

mated precision the substance an object is made of. The problem is that in many cases constructions of the type "Noun made of Noun" are used in a metaphoric way as in: "car made of cheese". In the actual context the car was not made of cheese but the construction is used to show that the respective car was not resistant to impact. Other examples of extracted relations are: **bottle**-(*glass*, *aluminum*), **ship** -(*oak*, *metal*), **cup**-(*stone*, *paper*). The extracted information should be carefully assessed because many times the properties extracted are highly contextual and do not qualify as common-sense knowledge. The difference between common-sense knowledge and non-common sense knowledge is notoriously difficult. Maybe everybody would agree that a cup is usually made of plastic, glass or porcelain. But in the last years with the eco-product trend a new material for building cups was introduced: the corn. Should we consider the fact that cups are made of corn common sense knowledge or not? As above the pattern "N make of N" is proposed by all association measures. Furthermore all measures except the simple frequency vote the pattern in the first position.

The pattern for Location relation has bad precision and bad recall. The properties of type Location listed in the norm represent typical places where objects can be found. For example, in the norm it is stated that **bananas** *are found in tropical climates* (the tropical climate being the typical place where banana-trees grow). However, what one can hope from a pattern-based approach is to find patterns representing with good precision the concept of Location in general. We found a more precise Location pattern than the selected one: "N is found in N". Unfortunately, this pattern has 0% recall for our test set. The extracted properties are in general imprecise: **duck**- (*exploit*), **hammer**-(*north*). There are two reasons why this pattern cannot be used for extraction of location relation. In the first place, as stated before, the pattern is in the class of general patterns. This means that it does not always express location (e.g. the dog from the police). Secondly the contextual facet is more pronounced here than for the Stuff properties. The objects can be found in many places at different moments of time. For example, a dog can be in the kennel, on the street, on the beech, on the meadow, etc. According to the definition of location only the kennel would be considered a good property for dogs. The pattern is voted by all association measures.

The patterns for Part relation have 0% recall for the concepts in the test set and their precision for the general domain is not very good either. As others have shown (Girju *et al.* (2006)) a pattern based approach is not enough to learn the part relation and one needs to use a supervised approach to achieve a relevant degree of success.

In conclusion, only two of the four relations can be learned using a pattern based approach: Superordinate and Stuff[1]. Part and Location relations cannot be learnt in this way. The contest of association measures for a good pattern selection marginally favors log-likelihood measure. Therefore, we cannot give a definitive answer to the question: which measure is better for weakly supervised pattern learning? An extensive testing with different kinds of corpora is needed. Nevertheless the experiment suggests that the best solution could be a majority voting procedure which considers the above association measures and perhaps others[2]. Further for some relation types like part, there is the case that the precise patterns are rarely found in general corpora and consequently they will not be found by any association measure.

5.3.2 Learning Quality and Action properties

5.3.2.1 Association measures

The hypothesis we pursue is that the properties of type quality and action found in feature norms should be among the strongest associates of the target concepts in a corpus. We previously noticed that, in general, the quality properties are expressed by the adjectives modifying the nouns representing the target concepts. The action properties instead are represented by verbs which encode actions into which the target concepts are participants. The association measures for quantifying the attraction between the nouns and apposite adjectives and verbs are the same used for pattern voting : simple frequency 5.4, (pointwise) mutual information (equation 5.5), Chi-squared (with Yates continuity correction) (equation 5.6)

[1] In the case of Stuff relation, as discussed, there is a difficulty in separating the common sense knowledge from the correct but not common sense knowledge.

[2] Obviously, each measure will vote for some candidate pattern.

and Log-Likelihood (equation 5.7). However, the meaning of the components of the formulas is changed accordingly, as the following enumeration shows.

1. $TC = \{tc_1, tc_2 \dots tc_n\}$ is the set of all occurrences of the target concepts in a corpus.

2. $M_{tc} = \{m_{tc}^1, m_{tc}^2 \dots m_{tc}^m\}$ is the set of all occurrences of the apposite adjective and verbs. Let tc_i be an occurrence of a target concept and m_{tc}^j an apposite adjective or verb modifying tc_i. Then to compute the attraction between tc_i and m_{tc}^j we define:

3. O_{11} the number of occurrences tc_i has with m_{tc}^j .

4. O_{12} the number of occurrences tc_i has with $\neg m_{tc}^j$.

5. O_{21} the number of occurrences $\neg tc_i$ has with m_{tc}^j.

6. O_{22} the number of occurrences any $\neg tc_i$ has with $\neg m_{tc}^j$.

Please observe that for each target concept and for each association measure we obtain a set of adjective and verb associates ordered by attraction strength. Only the most salient n candidates will be considered for evaluation.

5.3.2.2 Extracting Quality and Action properties

Unlike the procedure for learning Superordinate, Stuff, Location and Part relations which was weakly supervised the procedure for Quality and Action property learning is unsupervised. All occurrences of the concepts in the test set are identified in the uKWac corpus. The association strength between the nouns representing the concepts in the test set and the apposite adjective and verbs is computed as shown in the previous section. Only the 30 most powerful associates are evaluated. Because the best recall for the test set was obtained by Log Likelihood measure the results are reported for this measure. In the table 5.6 some associates for the concepts in the test set are presented. The extracted quality properties refer to color (*golden*, *green*), appearance (*bald*, *spotted*), perceived intelligence (*lame*), etc. and are the kind of qualities that the human subjects are

Concept	Quality	Action
Duck	*wild*, *tufted* *lame*, *ruddy*	*waddle*, *fly* *swim*, *quack*
Eagle	*golden*, *bald* *white-tailed*, *spotted*	*soar*, *fly* *perch*, *swoop*
Turtle	*marine*, *green* *giant*, *engendered*	*dive*, *nest* *hatch*, *crawl*

Table 5.6: Some quality and action properties for the concepts in the test set

likely to know. Even better are the actions extracted. For animals they represent the activities performed in a habitat: *dive*, *fly*, etc.

The results for Quality and Action properties are presented in table 5.7. The columns of the table represent in order: the property type, the Recall and the Property Precision. The Recall represents the percent of properties in the test set our procedure extracted from the corpus. The Property Precision computes the precision with which our procedure finds properties having a property type and is computed for the first 30 best associates. Because the number of potential properties is reasonable for hand checking, the validation for this procedure was performed manually.

Property Type	Recall	Property Precission
Quality	60%	60%
Action	70%	83%

Table 5.7: The results for Quality and Action property classes

The manual comparison between the corpus extracted properties and the norm properties confirms the hypothesis regarding the relation between the association strength of properties of type adjective and verbs and their degree of relevance as properties of concepts. This can be explained by the fact that all concepts in the test set denote concrete objects. Many of the adjectives modifying nouns denoting concrete objects express the objects qualities, whereas the verbs usually denote actions that different actors perform or to which various objects are subject.

Many of the properties found using this method encode pieces of common sense knowledge not present in the norms. For example, the semantic representation of the concept **turtle** has the following Quality properties listed in the norm: *green*, *hard*, *small*. The strongest adjectives associated in the uKWac corpus with the noun turtle ordered by the loglikelihood score are: *marine*, *green*, *giant*. The property *marine* carries a greater distinctiveness than any of the similar properties listed in the norms.

Likewise, the actions typically associated with the concept **turtle** in the McRae feature norm are: *lays eggs*, *swims*, *walks slowly*. The strongest verbs associated in the uKWac corpus with the noun turtle are: *dive*, *nest*, *hatch*. The *dive* action is more specific and therefore more distinct than the *swim* action registered in the feature norm. The *hatch* property is characteristic to reptiles and birds and thus is a good candidate for the representation of the concept turtle. In conclusion both Quality and Actions properties can be learnt using the simple framework introduced in this section. Moreover the framework can also be used for enriching the common sense knowledge gathered manually (e.g. property generation task).

5.4 Supervised experiments for relation learning

In this section we report some preliminary experiments for supervised relation extraction using some property types defined in Wu and Barsalou taxonomy. In the next subsection we give a brief introduction to Kernel methods and then discuss data collection and the experimental results. The results were obtained in collaboration with Caludio Giuliano and Lorenza Romano from FBK.

5.4.1 Short Introduction to Kernel Methods and Support Vector Machines

In the last years many fields like Natural Language Processing, Information Retrieval, computational biology and others showed an exponential expansion of kernel based methods. This section will introduce the main concepts behind the kernel based methods using a light mathematical apparatus. We do not give a tutorial nor a formal and rigorous presentation. Enough material will be surveyed to enable the reader understand the supervised relation extraction section. For a thorough examination of kernel based methods there are some excellent books (e.g. Shawe-Taylor and Cristianini (2004), Scholkopf and Smola (2001)) and Ph.D. thesis (Smola (1998)) or web resources(http://www.kernel-machines.org).

All systems based on machine learning techniques try to find regularities in the data. Based on these regularities the system makes predictions about unseen data. Kernel methods are robust and efficient approaches for finding patterns in data. A kernel method has two main components:

1. **Mapping module**. The mapping module performs the mapping of the data in the feature space. This function is needed because there is no guarantee that the input data exists in a feature space. It is required that the feature space is a dot product space (a vector space endowed with a dot product). Formally the mapping is given by the equation 5.9. X is the domain of the data and the feature space H usually is R^n

$$\phi : X \longrightarrow H \tag{5.9}$$

   The purpose of embedding data in H feature space is that in the new space a linear classifier can be used to discover the regularities.

2. **A learning Algorithm**. The task of the learning algorithm is to find the patterns in the feature space H. A nice characteristic of the algorithm is that it does not need the coordinates of the data in the feature space. It only needs the pairwise dot products[1]. The pairwise dot products are efficiently computed using a kernel function. The kernel function denoted

[1] The notation for the pairwise dot products is $<>$.

by k computes the similarity between each pair of input data x_1 and x_2 in the feature space X (equation 5.10). Nowadays the most used kernel algorithm is Support Vector Machines.

$$k(x_1, x_2) =< \phi(x_1), \phi(x_2) > \tag{5.10}$$

A major advantage of the kernel methods is that they give you the freedom to chose the mapping function. In many cases the selected mapping function is nonlinear. Another advantage is that they allow the use of linear algebra techniques with the learning algorithm. To facilitate the understanding of all introduced concepts we show a very simple example (the example is taken from the book Scholkopf and Smola (2001)).

Consider a binary classification problem. The data for this problem lies in R^2 and is represented in figure 5.1. A data point is either a circle or a cross. Please notice that the boundary separating the crosses from the circles is an ellipse and therefore is not linear. To make the circles and crosses linearly separable we map the data in a new feature space using the mapping given by the equation 5.11. The feature space H is a three dimensional space. In the second part of the figure 5.1 you can see how the ellipse boundary transformed in a hyper-plane. It can be easily shown that the kernel function can be computed in the original space without representing the coordinates in the feature space in conformity with equation 5.12.

$$\phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2) \tag{5.11}$$

$$k(x_1, x_2) =< \phi(x_1), \phi(x_2) >=< x_1, x_2 >^2 \tag{5.12}$$

The learning algorithm we will use in the next section is Support Vector Machines (SVM). SVM is based on a theory originating in statistics and on ideas by Vladimir Vapnik and Alexey Chervonenkis (Vapnik (1998)). The algorithm is best illustrated by reference to a binary classification task[1]. In figure 5.2 a set of training points is represented in a multi-dimenssional feature space. Each

[1]The algorithm can be applied for the separation of circles and crosses in the example above.

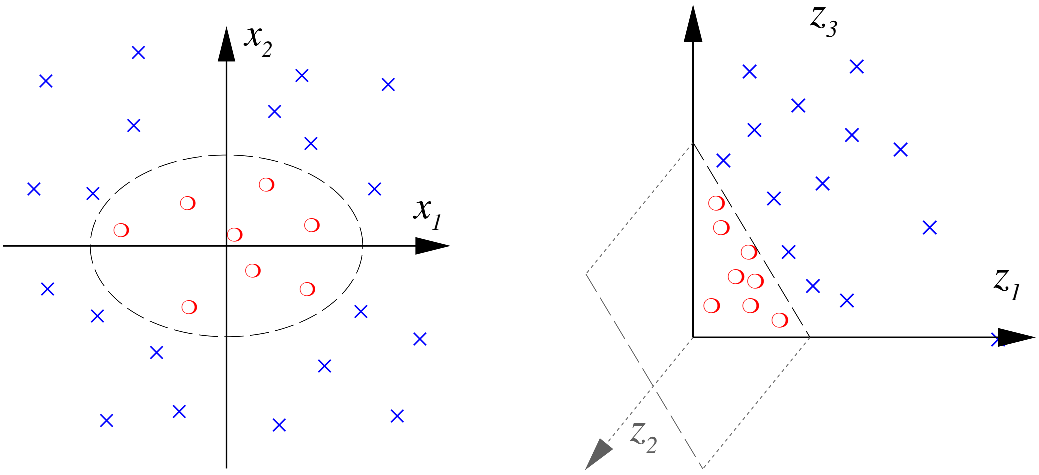

Figure 5.1: An example of mapping form a 2D space to a 3D space -

data point belongs to a one of two classes (Class 1 and Class 2 in figure 5.2) and it is assumed that the data is linearly separable. The closest examples to the separating hyper-plane are called Support Vectors (in figure 5.2 the Support Vectors lie on the planes H_1 and H_2). The learning idea implemented by SVM is to orient the hyperplanes as far as possible from the closest members of both classes.

5.4.2 Relations for supervised relation extraction

This section presents some experiments for supervised relation extraction performed for a subset of 6 relations in Wu and Barsalou taxonomy. The experiments should be regarded as preliminary. They were an initial verification by the authors of the usefulness of kernel based methods for general relations based on data extracted from a heterogenous corpus (uKWac). The 6 relations for which we tested the kernel methods are the following:

1. **Function**: Function relations denote the function an entity typically fulfills. For instance, "**airplane** *used for transportations*".

2. **Internal Component**: These properties denote internal components of an

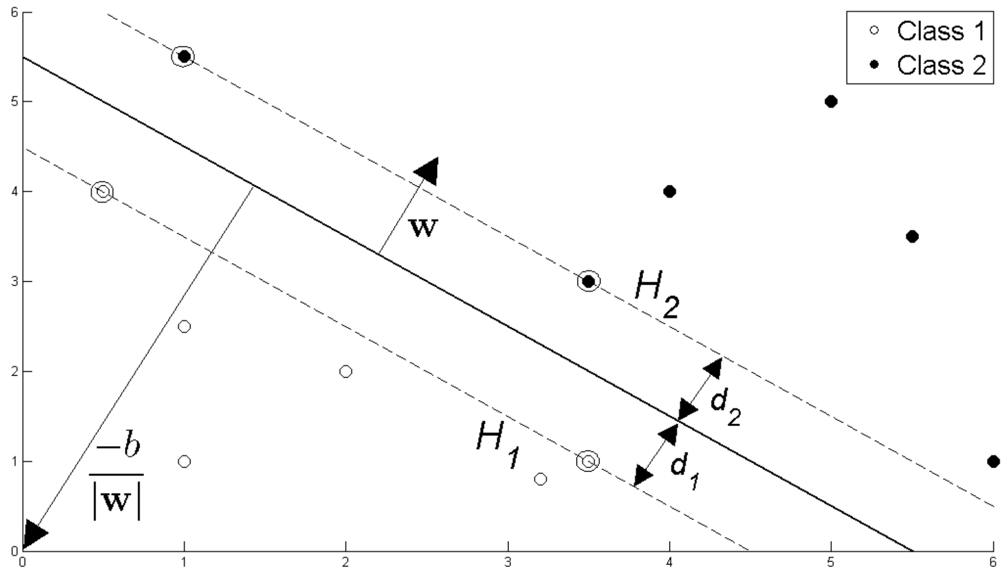

Figure 5.2: Support Vector Machines for a 2 class classification -

entity and are its hidden parts. Wu and Barsalou taxonomy distinguishes between the internal components of an object and its external components; from the point of view of relation learning we regard both of them as simply parts.

3. **Origin**: Origin properties are those properties denoting the origin of an entity, as in "**Cigar** *made in Cuba*".

4. **Participant**: Participant properties denote agents who typicality use an entity, performs an action on it or interacts with other participants. Example: "**desk** *used by students*".

5. **Superordinate**: Superordinate is the well known and well studied IS-A relation.

6. **External Surface**: According to Wu and Barsalou: "External surface properties are those properties of an entity that are perceived on or beyond the entity's surface, including shape, color, pattern, texture, size, touch,

smell, taste" (Wu and Barsalou (2006)). These properties usually denote the qualities of a concrete object, as in, e.g.: "**the car** is *red*". In this case *red* is a quality of the concept **car** and the color resides on the external surface of the car. We will not distinguish between internal and external properties; we regard both of them as qualities.

The set of the 6 relations contain 2 relations, superordinate and part, for which supervised frameworks were proposed before. The superordinate relation is by far the most studied relation in NLP. The part relation is less studied but nevertheless there were proposed various learning frameworks (e.g. Girju *et al.* (2006)). The function was studied as part of Qualia Structures (see chapter 2) but we are not aware of any supervised efforts attempting to learn it. For other two relations , Origin and Participant, no supervised framework was put forward. Finally, the External Surface property is interesting from a kernel perspective. The relations for which the kernel based methods are appropriate are binary relations linking two nouns. The External Surface Properties are generally expressed by adjectives.

5.4.3 Data Collection

Instances for each relation in the above set were collected from McRae feature norm. The words expressing the properties are normalized, lemmatized and part of speech tagged using TreeTagger. As an example of normalization consider the transformation of the instance (**dog** *is an animal*) in a pair of seeds (dog, animal). Then all sentences containing the seeds at a distance of at most 5 words including the punctuation are extracted from uKWac. For each relation we collected a set of 500 sentences. We made sure that each relation instance is present in approximately the same number of sentences.

Each sentence is annotated by two annotators as positive, negative or don't know. A sentence is a positive example if the binary relation is explicitly expressed in the sentence and negative otherwise. The don't know option is used when the annotator is unsure if the relation should be annotated as positive or negative. The annotation was performed with a platform independent GUI based annotation tool called "Context Relations" developed specially for the task. In figure 5.3 a view of the tool is presented.

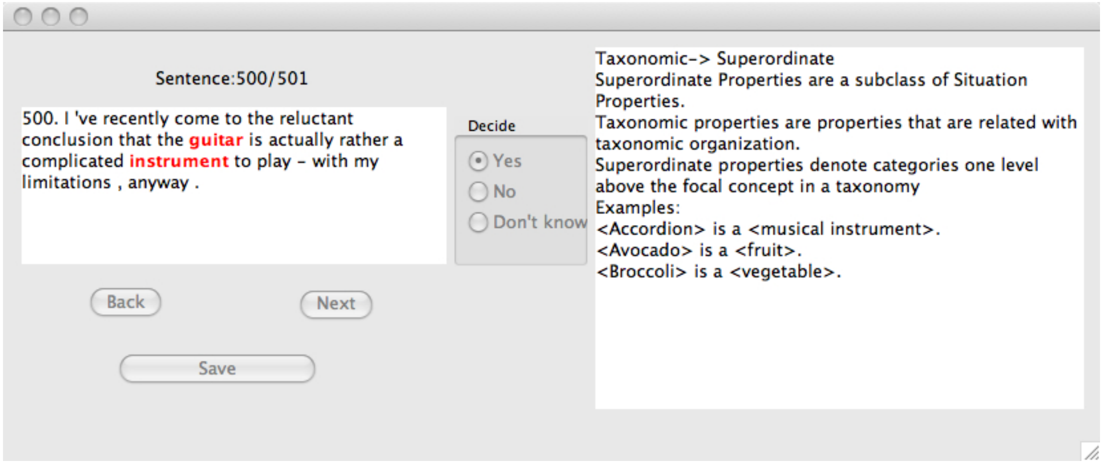

Figure 5.3: The view of Context Relation GUI tool -

In the left box the sentence containing the seeds of a binary relation is presented to the annotator. In figure 5.3 colored in red there are two seeds: **guitar** and **instrument** representing the superordinate relation. In the right box the annotator sees a short description of the relation to be annotated together with positive examples for the relation in case. A set of three radio buttons let the annotator select the appropriate label for the sentence. In the case above the answer should be yes because the sentence explicitly says that the guitar is an instrument: ...**the guitar** is rather a complicated **instrument** to play...

For a negative example of the same Superordinate relation consider the sentence extracted for the same instance (**guitar**, *instrument*):

1. He had a flute, a violin, a **guitar** and a Japanese stringed **instrument**.

Even if the seeds **guitar** and **instrument** appear in the same the sentence this does not say that a guitar is an instrument. The instrument in this case refers to another musical instrument.

The table 5.8 gives the inter-annotator agreement for each of the six semantic relations using the Kappa statistic (Siegel and Castellan (1988)).

The raters were not trained before the task. They only had a general explanation before starting the annotation and they also had the instructions in the

Relation	Kappa Score
External Surface	0.79
Function	0.62
Internal Component	0.78
Origin	0.63
Participant	0.5
Superordinate	0.73

Table 5.8: Inter-annotator agreeement on the six Wu-Barsalou relations

right panels of the tool in figure 5.3. Given this circumstance the results are good. The External Surface, Internal Component and Superordinate have scores in the range 70-80 % . Intuitively the judgment that a property is or not a component, a superordinate or a quality should be easier that the judgment that a property is a function or a participant. The reasons for this are: the imprecise status of function relation [1] and the fact that the participant relation heavily depends on the context [2].

5.4.4 Kernel methods for relation extraction

The experiments for relation learning starts with the exploitation of shallow linguistic information. Therefore the kernels used for the experiment make use of: tokenization, sentence splitting, part-of-speech tagging and lemmatization. Following the general trend the kernel algorithm we use is Support Vector Machines (Cristianini and Shawe-Taylor (2000)). Rather than defining individual kernels and using each kernel for relation extraction we employ a combination of kernels. The reason is that in NLP literature it is proven that a suitable combination of kernels improves the performance of the individual ones. Each kernel is calculated as in equation 5.13:

[1] There is a controversy in philosophy regarding its definition and the reader can see some shallow details in the next chapter.

[2] In many cases the context is not given only by the sentence but it is given by the larger discourse context.

$$K(x_1, x_2) = \frac{\langle \phi(x_1), \phi(x_2) \rangle}{\|\phi(x_1)\| \|\phi(x_2)\|}, \tag{5.13}$$

where ϕ is the mapping function discussed above and $\| \cdot \|$ is the 2-norm. Please notice that the kernel is normalized (divided) by the product of the norms of the vectors in the feature space.

We now turn to present the kernels our system uses:

1. **Global Context Kernels**. These kernels were used in the past in the extraction of relations between entities(Bunescu and Mooney (2005),Giuliano *et al.* (2006)). It has been shown (Giuliano *et al.* (2007)) that the same kernels can be successfully employed in relation extraction between nominals. The main idea behind using the global kernels is that the relation between two entities is expressed using words in one of the following positions [1]:

   (a) **Fore-Between** Tokens before and between the two entities, e.g. "the head of [ORG], Dr. [PER]". Please observe that the tokens are simultaneously used to express the affiliation of a person to an organization. In the context of chemical interaction an example is *"the reaction between* [S_1] *and* [S_2]*"* where [S_1] and [S_2] are two chemical substances. In our case a good example is found for location relation: *"the country of* [*emu*] *is* [*Australia*]*"*.

   (b) **Between** Only tokens between the two entities, e.g. "[ORG] spokesman [PER]" ,"[S_1] binds with [S_2]". Many of the relations holding between nominals are expressed with the words between the two concepts: "[*hare*] figured as a sacrificial *animal*".

   (c) **Between-After** Tokens between and after the two entities, e.g. "[PER], a [ORG] professor", "[S_1], and [S_2] interact". I could not find any example of relation expression with between-after tokens for the test relations.

[1] In this section Fore-Between, Between and Between-After will be called patterns.

The global context kernels can be formalized for a given relation R and a context C as a row vector like in equation 5.14:

$$\phi_C(R) = (tf(t_1, C), tf(t_2, C), \ldots, tf(t_l, C)) \in \mathbb{R}^l \tag{5.14}$$

In the equation above the function $tf(t_i, C)$ records how many times a particular token t_i is used in C. The kernel construction differs from a standard bag of words approach in that we do not count the nominal tokens in the context. The kernel performance can be further improved by extending ϕ_C to embed n-grams of (contiguous) tokens (up to $n = 3$). The n-gram kernel K_n counts uni-grams, bi-grams, ..., n-grams that two patterns have in common. In the literature, it is also called *n-spectrum* kernel. It is obtained substituting ϕ_C into equation 5.13. The global context kernel $K_{GC}(R_1, R_2)$ is then defined as

$$K_{FB}(R_1, R_2) + K_B(R_1, R_2) + K_{BA}(R_1, R_2), \tag{5.15}$$

where K_{FB}, K_B and K_{BA} are respectively the n-gram kernels that operate on the Fore-Between, Between and Between-After patterns and R_1 and R_2 are the examples compared by the kernel function.

2. **Local Context Kernel** Each local context is represented using the following basic properties: the token itself, the lemma, the PoS tag, the stem, and orthographic properties, within a text window. Formally, given a relation example R, a local context $L = t_{-w}, \ldots, t_{-1}, t_0, t_{+1}, \ldots, t_{+w}$ is represented as a row vector

$$\psi_L(R) = (f_1(L), f_2(L), \ldots, f_m(L)) \in \{0, 1\}^m, \tag{5.16}$$

where f_i is a property function that returns 1 if it is active in the specified position of L, 0 otherwise. In the reported experiments, we used a context window of ± 2 tokens around the candidate entity. For example, the orthographic function $CAP(t_0)$ is the jth component of the vector $\psi_L(R)$ and it is 1 if and only if the token t_0 in the local context L of R is capitalized. The local context kernel is defined as follows:

$$K_{LC}(R_1, R_2) = K_{left}(R_1, R_2) + K_{right}(R_1, R_2), \tag{5.17}$$

where K_{left} and K_{right} are defined by substituting the embedding of the left and right local context into Equation 5.13 respectively.

Notice that K_{LC} differs substantially from K_{GC} as it considers the ordering of the tokens and the feature space is enriched with PoS, lemma and orthographic properties.

3. **Shallow Linguistic Kernel** The shallow linguistic kernel $K_{SL}(R_1, R_2)$ is defined as
$$K_{GC}(R_1, R_2) + K_{LC}(R_1, R_2). \tag{5.18}$$

4. The bag-of-words kernel $K_{BoW}(R_1, R_2)$ is defined as the global context kernel but it operates on the whole sentence. We defined this kernel for comparison only; notice that it is not used in the shallow linguistic kernel.

5.4.5 Results

Sentences have been tokenized, lemmatized, and POS tagged with TextPro. We considered each relation as a different binary classification task, and each sentence in the data set is a positive or negative example for the relation. All the experiments were performed using jSRE. The results were obtained by 10-fold cross-validation. Table 5.9 shows the performance of the shallow linguistic kernel for 6 relations and the micro-average result. Majority and all true are the baselines that return the majority class and always true, respectively. BoW is the bag-of-words kernel, it provides a stronger baseline.

The most surprising result is that the Highest F_1 score was obtained for External Surface properties. The score is better than the score obtained for the traditional IS-A relation. This means that maybe the kernel combination used can be extended also to other relations than the relations holding between nominals. Our kernel combination clearly outperforms all three baselines. The results are very promising but we think that more experiments are needed before a large scale usage of the method. In particular we would like to see how the results change when we use different kinds of corpora. Moreover, we would like to catalogue the contexts for each relation type.

relation	Prec	Recall	F_1
External Surface	84.0	86.4	85.2
Function	87.2	81.5	84.2
Internal Component	70.4	75.0	72.6
Origin	72.2	73.2	72.7
Participant	85.1	84.0	84.5
Superordinate	79.0	86.8	82.7
Micro	82.1	82.8	82.5
Majority	65.9	73.9	69.6
All true	50.9	100.0	67.5
BoW	71.8	74.9	73.3

Table 5.9: Relation extraction performance.

5.5 Chapter Summary

In this chapter we propose a new classification schema for the properties stored in McRae feature norm. We showed that some of the distinctions operated by Wu and Barsalou schema hamper the learning process. Especially the distinction made between external and internal properties is only justified from a psychological point of view (5.2). For learning the property types in the new schema we used two methods. The binary relations holding between nominals are learnt with the aid of lexical patterns (5.3.1). To automatically identify the best lexico-syntactic patterns we use four association measures. We were able to find useful patterns automatically for two of the four relations. For the other two relations some good patterns exists, but, unfortunately, they have 0 recall for our test set. The main conclusion derived is that there is not a best measure for pattern identification. However for some relations the precise patterns could be very rare and never appear with the seeds even in a huge corpus. The properties expressed by adjectives and verbs are learnt computing the association strength between the nouns representing the target concepts and the apposite adjectives and verbs (5.3.2). The results are surprisingly good given that the method is unsupervised. Because the number of properties generated can be hand-checked the method can be used in property acquisition from large corpora[1]. In the last part of the chapter (5.4) a pilot experiment for supervised relation acquisition was presented. Sentences containing seeds for each of the 6 relations are extracted from a corpus and then they are annotated as positive or negative examples by two raters. We employ a kernel based method for learning the relations. The results are encouraging but some other experiments are needed before using the method for large scale relation acquisition.

[1]We hope that it will also be considered by psychologists.

6

Extracting Richer Knowledge Structures from Wikipedia

6.1 Contributions

If the first two chapters of this thesis explored wordnet and corpora to extract feature norms like description, this chapter is more ambitious. We introduce an unsupervised method for extracting knowledge from Wikipedia. The method seeks to extract lexico-syntactic structures from descriptions of similar concepts. The structures we extract are formalized as surface patterns linking the concepts with their properties. The hope is that the obtained lexico-syntactic structures can be mapped on semantic relations. Unlike other approaches in computational linguistics[1] we do not identify patterns by supplying seeds. Moreover, we do not make any assumption about the properties and relations which should be extracted. All the relevant relations should emerge from data.

6.2 What is Wikipedia?

Wikipedia is the largest multi-lingual encyclopedia ever built. The coverage of Wikipedia is impressive: at the time of writing this thesis Wikipedia surpassed for English language three millions articles covering a broad range of topics: from

[1]Hearst inspired approaches an example of which you saw in a preceding chapter.

history of science to football, from Renaissance art to fashion,etc. Unlike other resources in computational linguistics, like various corpora, Wikipedia is under constant change[1]. Another characteristic of Wikipedia is the style of writing of its articles. This style is more homogeneous than the mixed bag of styles one encounters in general corpora or in unrestricted text found on the web. In fact, there is a manual of style produced by the Wikipedia editors that article authors should conform to. The articles that do not conform with the manual directions are flagged. We give some guideline examples relevant for this chapter:

1. Each Wikipedia article should describe only one concept and article titles should conform to Wikipedia naming conventions. This fact is exploited by our mapping algorithm which seeks to disambiguate a set of words against the Wikipedia entries.

2. The equivalent terms are linked to an article using the system of redirects. Redirects are pages containing directives and pointing to an article related to a search term. Redirects are performed for synonyms, abbreviations, capitalize terms, etc.

3. When a term is ambiguous Wikipedia returns a disambiguation page containing links to all the pages corresponding to the searched term.

4. The very first sentences of an article should define its topic and the main concept the article explicates.

5. Each article is split in head sections. The headings should be informative and allow easy navigation.

6. Theoretically, the meaning of punctuation and of typographic style can be safely utilized. For example, the italic style is used when a new term is introduced or when the meaning of the term changes.

[1]The author of this thesis constantly follows the chess tournaments. We checked many times the Wikipedia articles about the main chess players after each major tournament. We were pleasantly surprised to see that the pages of the players winning the tournaments where suddenly changed to reflect the results.

In addition to rough text Wikipedia articles also contain infoboxes which present article information in a structured way. The figure 6.1 shows the infobox for the capital of Italy: Rome.

Given the extensive information coded in Wikipedia there is no surprise that Wikipedia is used in various NLP tasks. We do not attempt to survey the research which make use of Wikipedia, but we briefly review some research related to ours. For a good survey, though a little outdated, of Wikipedia as knowledge repository the reader can consult (Medelyan *et al.* (2009)).

Our method for extracting concept properties is based on the concept similarity. Some research uses Wikipedia to quantify concepts similarity and measures the agreement between machine derived similarity and the human notion of similarity (Strube and Ponzetto (2006)). Both the raw text of the articles and the infoboxes are used in Information Extraction. Among the methods which use the Wikipedia text for information we mention the work of Hebelot and Copestake who extract hyponym relations from sentences containing the verb to be. Ruiz-Casado uses Wikipedia for the relations structuring the noun hierarchy in Wordnet. Other research seeks to extract very particular relations like *has Artist* (which holds between an artist and the work it produces) or *has Director*(which holds between films and the film directors). In conclusion, the relation extraction from Wikipedia is performed under the assumption that a set of relations is given beforehand and the task of the research is to devise precise methods for relation extraction.

6.3 The System for Knowledge Extraction from Wikipedia

Unlike the previous work, we do not seek to extract a beforehand given set of relations, but we want that the relevant relations emerge from the data. The main idea we follow is that similar concepts (i.e. those classified under the same node in a taxonomy) are described in a comparable way in Wikipedia. More precisely, we suppose that the relevant knowledge of these similar concepts is expressed using equivalent surface patterns. The learning process starts with the generation of

Country	Italy
Region	Lazio
Province	Rome (RM)
Government	
- Mayor	Gianni Alemanno (PdL)
Area	
- Total	1,285.31 km^2 (496.3 sq mi)
Elevation	20 m (66 ft)
Population (30 April 2009)[1]	
- Total	2,726,927
- Density	2,121.6/km^2 (5,494.9/sq mi)
- Demonym	Romani
Time zone	CET (UTC+1)
- Summer (DST)	CEST (UTC+2)
Postal code	00121 to 00199
Dialing code	06
Patron saint	Saint Peter and Saint Paul
Saint day	June 29
Website	Official website

Figure 6.1: The infobox associated with the Wikipedia article Rome -

concept hierarchies. The concepts in each hierarchy are mapped onto Wikipedia pages and the knowledge appropriate to the concepts is automatically extracted. In the next subsections we describe the knowledge extraction process.

6.3.1 Taxonomy Extraction and Mapping

The higher nodes in an hierarchy define the similarity between all the sub-nodes they dominate. This is not surprising given the meaning of the IS-A relation. In a well-defined taxonomy all the properties of the concepts representing the upper nodes should be propagated to the concepts representing the lower nodes. Therefore the concepts representing the lower taxonomy nodes will be similar in so much as they share a common set of inherited properties. The main question we confront with is what kind of taxonomy to use: a manually built taxonomy or an automatically extracted one. There is much effort in NLP to automatically acquire taxonomies from text and many authors report a high degree of success. Despite the recent achievements the automatically extracted taxonomies are much less precise than the manually built ones. Moreover, we want to test the performance of the algorithm on multiple taxonomies belonging to different domains and acquiring them requires considerable effort.

The second problem we confront with is how to automatically map the concepts of the taxonomy onto Wikipedia pages. It is a notorious fact that the words used to represent concepts are in many cases ambiguous and that Wikipedia has many entries for the same word. Consequently, we expect the taxonomy we use to have a mechanism for specifying ambiguity. DBPEDIA repository (http://wiki.dbpedia.org/OnlineAccess) offers some mappings between various taxonomies and Wikipedia articles. We checked the reliability of the following classification schemas:

- **Collaborative Tagging System**. A classification schema derived from the collaborative tagging system employed by Wikipedia editors. The advantage of this classification schema is that it reflects the understanding Wikipedia editors have for the content of an article they authored. Unfortunately, the taxonomy contains many other relations different from IS-A.

Unless one finds a reliable procedure to tell which tags are taxonomic, the schema cannot be used for our purpose[1].

- **Wordnet Classes**. Wordnet Classes map synsets in Princeton Wordnet 1.6 (W3C version on wordnet) on Wikipedia. Besides linking Wikipedia with an outdated version of Wordnet, the mapping is not very accurate[2].

- **Yago Classes**. YAGO(Suchanek *et al.* (2007)) is a lightweight ontology that incorporates information extracted from the structured part of Wikipedia articles and Princeton Wordnet. It has more than 2 million entities and over 20 millions facts. YAGO does not attempt to map Wordnet synsets onto Wikipedia articles but it extended Wordnet coverage adding Wikipedia pages as instances of leaf synsets in Wordnet taxonomic tree. For example, Albert Einstein and Max Planck are added as instances of the synset scientist.

We decided to use taxonomies derived from Wordnet. The advantage of a Wordnet derived taxonomy is that there is a way to control the ambiguity of the words representing the concepts in the taxonomy. Moreover, choosing the first sense of a wordnet word results in 64% success in Wordnet to Wikipedia mapping (Pradhan *et al.* (2007)). Wordnet taxonomy is also integrated in YAGO and in case it is needed we can use YAGO to expand the Wordnet taxonomy with instances. The disadvantage of using Wordnet derived taxonomies is that the hyperonymy relation is not a true IS-A relation and the concepts on the same level of a taxonomy have different levels of generality (for a short discussion of this aspect see the chapter 4). We assume that an one to one mapping between concepts in taxonomy and Wikipedia pages exists. To generate the taxonomy of concepts and map the generated taxonomy onto Wikipedia articles we follow the next steps:

[1]After performing the experiments in this thesis we found out about the work of Ponzetto (Ponzetto and Strube (2007)) who derived a taxonomy from Wikipedia Categories. In the future we plan to use this taxonomy for a large scale testing of our algorithm.

[2]This is the conclusion after manually checking some random synsets.

1. First, we pick a concept of interest representing the higher level node of the taxonomy to be extracted and map it onto a WordNet synset. For example, if you have chosen the concept **dog** and you want to get the sense corresponding to the animal, you map the concept to the sense number 1 in WordNet.

2. Second, the hyponymy (sub)tree having as root the concept chosen in the previous step is produced and the concepts in the tree are mapped onto Wikipedia pages. The best mapping heuristic is to choose that member of a synset which has the sense number 1[1]. Even so, the ambiguity problem is not completely solved. For it is possible that concepts having low or no ambiguity in WordNet to be highly ambiguous in Wikipedia. Fortunately, in this case the Wikipedia server returns a page having a standard structure and allows us to reject the ambiguous concept or to guess the right mapping. The disambiguation is performed concatenating the ambiguous concept with each of its WordNet hyperonyms and searching again in Wikipedia until an unambiguous entry is found. For example, the concept **buckskin** appears in two synsets in WordNet and in 8 possible entries in Wikipedia. Because we are interested in the sense of buckskin having the hyperonym the concept **horse** we concatenate the two words (buckskin_(horse)) and send the new entry to Wikipedia server. Fortunately, in this case no ambiguity results and the correct mapping is automatically performed.

In the figure 6.2 you can see part of the taxonomy generated for the root category **horse**. The concepts **mustang** and **buckskin** are ambiguous and the disambiguation proceeds as explained above.

6.3.2 System architecture

Once we extracted the taxonomy and mapped the concepts onto the Wikipedia articles the generated taxonomy is given as input to a system for Knowledge Extraction. The system is constituted from a series of modules. Each of the

[1]There are other better mapping heuristics but they are much more complicated.

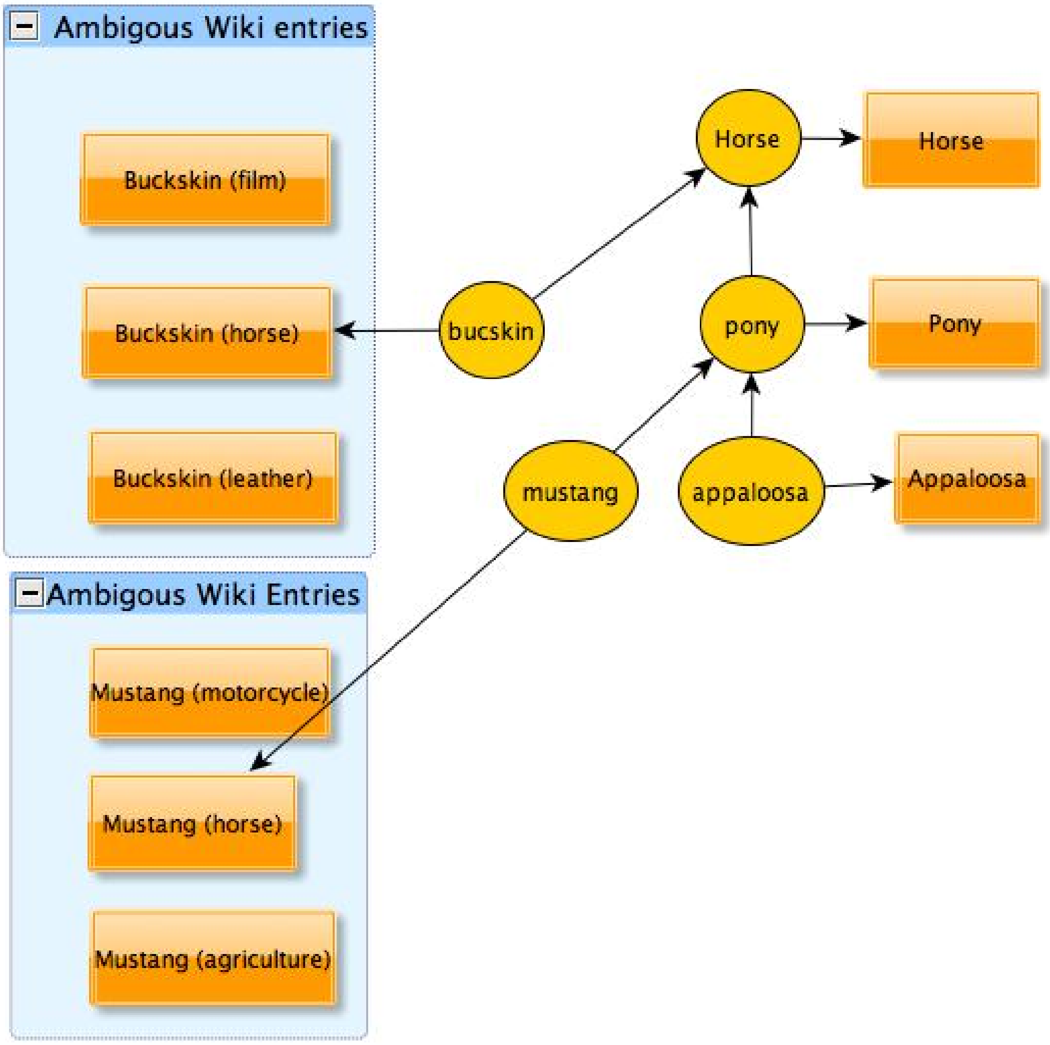

Figure 6.2: The mapping between the concepts in Wordnet Extracted Taxonomies and Wikipedia articles -

modules operates on the output produced by the preceding module in the pipeline. The high level system architecture is presented in figure 6.3.

The module **Article Downloader and Parser** downloads the Wikipedia articles corresponding to the concepts in the taxonomy. From the rough content of Wikipedia articles it eliminates the useless html tags and the head structure of the article is recovered (e.g. to each higher order head in the article its corresponding text is assigned). In addition, the module eliminates the content of some heads not used by the system, like: Links, Miscellaneous, See also, etc. The next example gives a part of the Wikipedia article for the concept **cat** parsed by the module:

Concept: **cat**

Head: *Abstract*

Content:The cat (Felis catus), also known as the Domestication Cat or 'house cat' to distinguish it from other felines and felids, is a small...

Head: *Physiology*

Content:The size of a male cat typically weighs between 2.5 and 7 kg...

Head: *Behavior*

Content:For cats, life in close proximity with humans (and other animals kept by humans) amounts to a "symbiotic social adaptation" which has developed over thousands of years...

The next module, **Sentence Extractor and Co-Reference Resolution**, extracts from the Wikipedia text of an article all sentences containing references to the title concept. The idea behind extracting all sentences containing the title concept is that these sentences express in a direct way relevant knowledge about the categories in the taxonomy. Moreover, the knowledge can be formalized as lexico-syntactic patterns linking the concepts with their relevant properties. Sadly, we cannot capture all occurrences of the main concept just extracting the sentences containing the word in the title of article. Consider for example the first two sentences in the entry of the concept **cat**:

"The cat (Felis catus), also known as the domestic cat or housecat to distinguish it from other felines and felids, is a small carnivorous mammal that is valued by humans for its companionship and its ability to hunt vermin and household

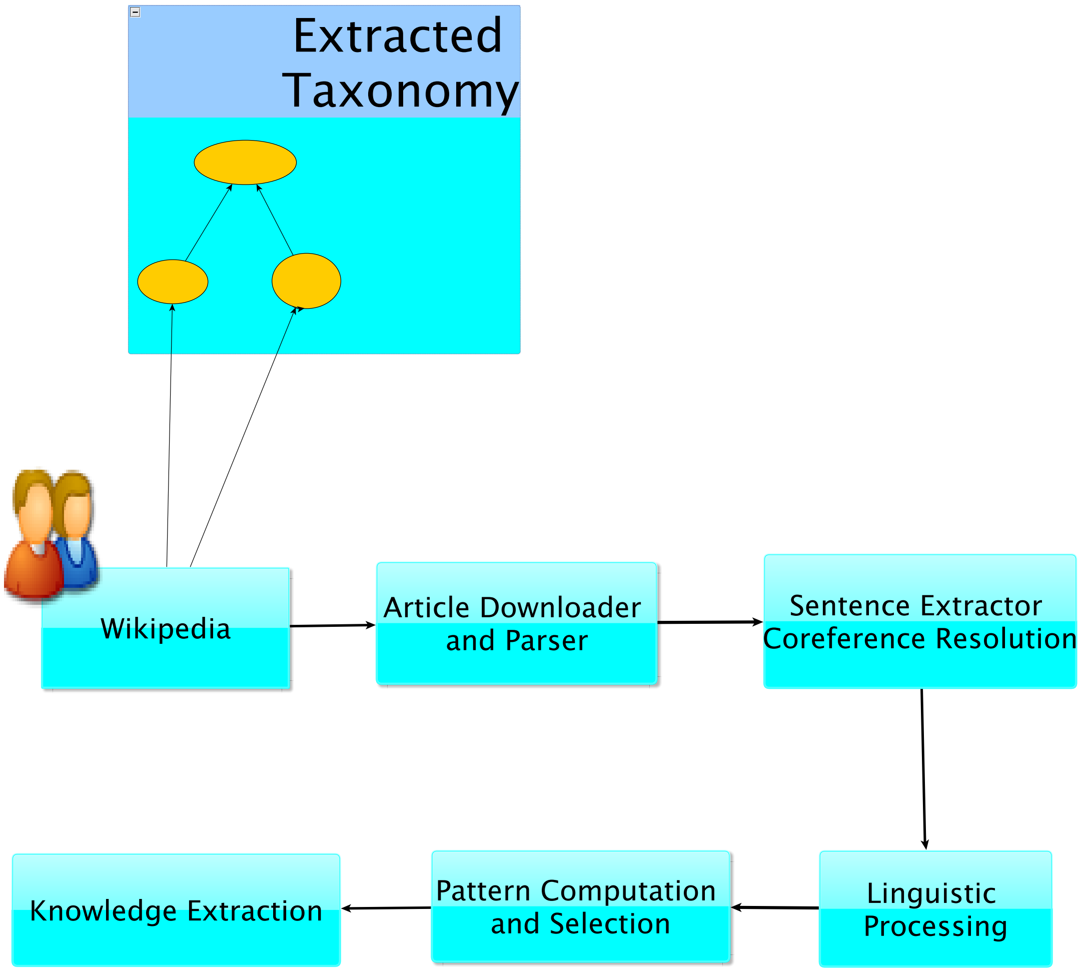

Figure 6.3: The pipeline of the system for knowledge extraction -

pests. It has been associated with humans for at least 9,500 years and is currently the most popular pet in the world."

In the second sentence first pronoun refers back to the concept **cat**. The features that can be extracted from the second sentence are: *associated with humans* and *the most popular pet in the world*. Therefore, to extend the range of the sentences extracted, the module identifies the other possible occurrences of the title concept using the following strategies:

1. **Synonym Expansion**. The title concept is expanded with its synonyms list derived from Princeton Wordnet. Moreover, the first noun occurence delimited by " is added to the synonyms list[1]. Any sentence containing an occurrence of the synonym of the title concept is extracted.

2. **Hyperonym Expansion**. We assume that any hyperonym up to and including the second level in Princeton Wordnet occurring in the first five words of the sentences of the title concept refers back to the title concept.

3. **Pronoun Coreference resolution**. All the sentences containing occurrences of the pronouns: their, it, he, they within the first three words are extracted.

Of course, the last strategies are accurate but not perfect. To illustrate the strategies in action, consider the following sentences from the article assigned to the concept **cougar**:

"The **cougar** (Puma concolor), also puma, mountain lion, or panther, depending on region, is a mammal of the Felidae family, native to the Americas. The large, solitary **cat** has the greatest range of any wild terrestrial mammal in the Western Hemisphere, extending from Yukon in Canada to the southern Andes of South America. **It** is the second heaviest cat in the American continents after the jaguar, and the fourth heaviest in the world, along with the leopard, after the

[1]The software API used by the module **Article Downloader and Parser** replaces the first hyperlink to the title concept in the text with the word delimited by ". Because of the system of Wikipedia redirects the title concept does not always coincide with Wordnet concept. This method is useful in identifying the actual title concept.

tiger, lion, and jaguar, although it is most closely related to smaller felines. The **animal** may be recolonizing parts of its former eastern territory."

The **Sentence Extractor and Co-Reference Resolution** modules extracts all the sentences in the example because each sentence passes one of the above-mentioned criteria. The first sentence contains an occurrence of the title concept **cougar** and therefore it is extracted. Two hyperonyms in the chain of the concept **cougar** are **cat** and **animal**, thus the second and the last sentences are extracted. Finally, the third sentence is also extracted because it contains one of the pronouns assumed to refer back to the title concept (**It**).

The module **Linguistic Processing** performs part-of-speech tagging, lemmatization and term identification for the extracted sentences. In order to harvest multi-word expressions and to achieve a better generalization across multiple similar sentences we use the term definition in equation 6.1:

$$\begin{aligned} (NPrep)?(((((Adv)?Adj)*(and)?)(Adv)?Adj)?|(Ger)*)(Noun)+) \\ where \\ Noun = NNS|NN|NPS|NP \\ Adv = RB \\ Adj = JJS|JJR|JJ \\ Ger = VBG \\ NPrep = DT \end{aligned} \tag{6.1}$$

This term definition is a slightly modified version of the term definition used by (Justeson and Katz (1995)). The modifications are made to incorporate the TreeTagger tagging set (in the equation all references after the "=" sign are to the tags in TreeTagger Tagset). In case a verb exists at the left of the term, the term is concatenated with this verb. The output of this module is a list of sentences where the terms containing the title concepts are replaced with the generic label "TitleConcept" and the rest of the terms are replaced with the label T. By way of example consider the transformation of a sentence from the Wikipedia article of the concept **snow leopard**:

1. The snow leopard or panthera uncia, sometimes known as ounce, is a moderately large cat native to the mountain ranges of Central Asia.

2. [the snow leopard] or [panthera uncia] sometimes know as [ounce] be [moderately large cat] native to [the mountain] of [Central Asia]

3. TitleConcept or T sometimes known as T be T native to T of T

The form in 1 is the original sentence in the Wikipedia article. The form in 2, which from now on will be called **Term Mapped Form**, groups together the terms in parentheses. Finally, the **Simplified Term Form** is the form in 3. In the **Simplified Term Form** the title concept **snow leopard** is replaced by the label TitleConcept and the rest of the terms by the label T. As said before, the main idea of the procedure for knowledge extraction is that the properties of similar concepts are stated using the same surface patterns. Inspecting the simplified term form it becomes clear how the patterns look like. The first two terms of the pattern TitleConcept or T sometimes known as T be T encode synonyms of the main concept, while the third term includes its taxonomic classification. Applying the pattern on the term mapped form in 2 the following knowledge for the title concept **snow leopard** is obtained: (*panthera uncia*, *uncia*, *moderately large cat*). The first two terms are synonymous of the title concept and the third term gives a more precise taxonomic classification of the concept. Note that we already knew from the taxonomic classification that the **snow leopard** is a cat but we did not have any information about its size: *moderately large*.

The next module in the pipeline is called **Pattern Computation and Selection**. Its task is to identify the patterns expressing relevant knowledge for the taxonomy concepts. This module is best understood as a collection of two sub-modules: **Pattern Generation** and **Pattern Ranking and Selection**.

The submodule **Pattern Generation** computes candidate patterns. The idea behind pattern generation is that the patterns originated should express knowledge characteristic to similar concepts. Now it is expected, given that the style of writing of dictionaries and encyclopedias is uniform and controlled, that similar concept properties should be expressed using the same lexico-syntactic patterns. Moreover, the language of Wikipedia articles being uniform, it is supposed that the properties characteristics to large classes of concepts (not necessarily similar) can be mapped on a set of patterns. An example of this kind of properties are the taxonomic classifications and parts. The taxonomic classifications can always be mapped on a pattern of type: TitleConcept be T and the part properties on a pattern of type: TitleConcept have T.

In this study two concepts are considered similar if they are classified under the same node in a taxonomy. Theoretically, in a well built taxonomy the most similar concepts should be sisters in the taxonomic tree. As we saw above, in a previous chapter, this is not true in the case of Wordnet derived taxonomies and certainly it is not true of any taxonomy built automatically. Of course, the quality of the results should increase with the quality of the taxonomy.

In this chapter we cut the taxonomic tree at a specific node and consider that all sub-concepts dominated by the respective node are similar. One expects that the patterns expressing relevant knowledge of these concepts appear in the extracted Wikipedia sentences for more than one concept. To produce candidate patterns the Cartesian Product between all sentences in simplified term form (as outputted by the **Linguistic Processing** module) belonging to each pair of similar concepts is performed. For each pair of sentences in the Cartesian product we consider as candidate patterns the longest common substring including the title concept between the sentences.

In figure 6.4 two sentences belonging to the similar concepts **caracal** and **panthera_pardus** are considered and both **Term Mapped Form** and **Simplified Term Form** are presented. Taking the longest common substring between the sentences in simplified term form belonging to the two concepts the candidate pattern "TitleConcept be T" is produced.

More formally, let's consider

$$C = \{C_1, C_2 \ldots C_n\} \tag{6.2}$$

a set of similar concepts obtained cutting a tree at a node. And let also

$$S_{Ci} = \{S_{1i}, S_{2i} \ldots S_{ki}\} \tag{6.3}$$

$$S_{Cj} = \{S_{1j}, S_{2j} \ldots S_{pj}\} \tag{6.4}$$

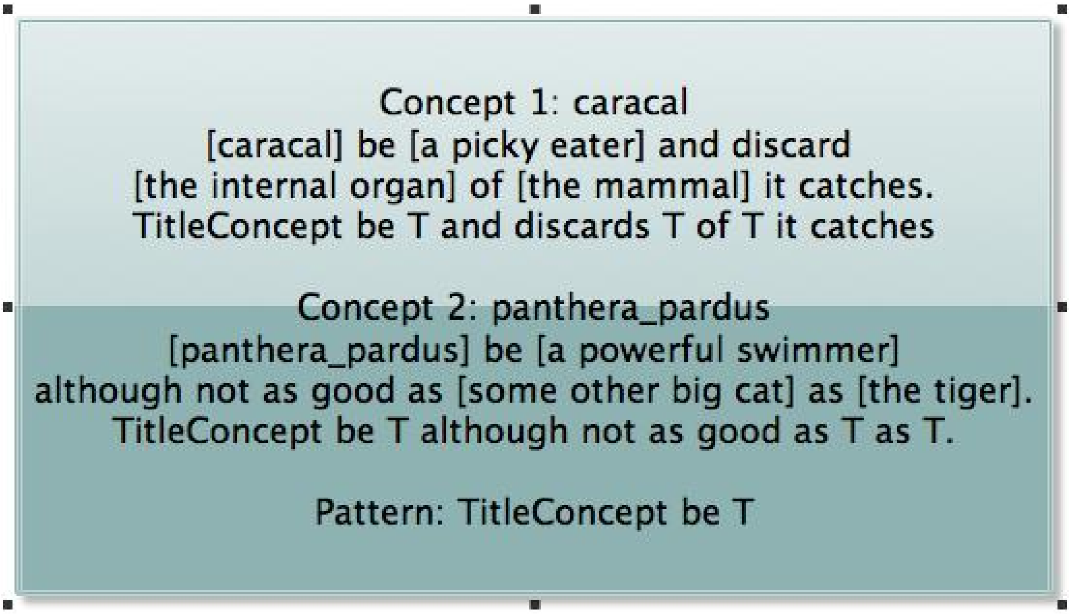

Figure 6.4: The generation of Candidate patterns by the sub-module Pattern Generation -

be the set of sentences in **Simplified Term Form** for the concepts C_i and C_j. To generate the candidate patterns for these two concepts we compute first the cartesian products for the sentences S_{Ci} and S_{Cj}.

$$S_{Ci} \times S_{Cj} = \{\{S_{1i}, S_{1j}\}, \{S_{1i}, S_{2j}\} \dots \{S_{ki}, S_{pj}\}\} \tag{6.5}$$

and for each pair in the cartesian product $S_{Ci} \times S_{Cj}$ we calculate the longest common substring (lcs) between the members of the pair to obtain the candidate patterns:

$$\begin{aligned} \{p_1, p_2 \dots p_t\} = \{lcs(\{S_{1i}, S_{1j}\}), lcs\{S_{1i}, S_{2j}\} \dots lcs(\{S_{ki}, S_{pj})\}\} \\ where \\ t = kp \end{aligned} \tag{6.6}$$

Please notice that the generated patterns in 6.6 need not be different. The same operation is performed for any pair of concepts in C. The generated patterns together with their frequency are recorded in the pattern table (6.1)

Pattern	**Frequency**
p_i	500
p_j	300
p_k	200
$\dots$	100
p_m	4

Table 6.1: The pattern table generated by the sub-module **Pattern Generation**

In the next phase the best patterns are selected. At this point two strategies for pattern selection are possible: a manual selection of patterns or an automatic one. The best strategy in terms of quality is the manual one. Ideally, the pattern precision is computed for each candidate pattern and the most precise patterns are then used in the knowledge extraction phase. Unfortunately, the manual selection of patterns is a labour intensive procedure and is not feasible in practice.

The automatic strategy attempts to find a formal procedure for pattern selections. This is the task of the sub-module **Pattern Ranking and Selection**. The sub-module ranks and filters the patterns produced by the sub-module **Pattern Generation** and stored in the pattern table 6.1. After testing a series of heuristics for pattern selection we deem as the best the following three:

1. **Pattern Shape**. All patterns not having the shape given by the regular expression equation 6.7 are eliminated. Thus, we accept the following patterns: "T of TitleConcept be T", "TitleConcept be T", "TitleConcept be design by T" and reject the next patterns: "in T , TitleConcept be", "of TitleConcept, T". While the former patterns have both topic (what is being talked about; it always contains the TitleConcept) and focus (what is being said about the topic), the latter are incomplete, missing either topic or focus, thus being useless for information extraction.

$$(TitleConcept|T)(.+)(T|TitleConcept). \tag{6.7}$$

2. **Pattern Grouping**. Interpreted as strings many patterns are substrings of other patterns. This heuristics groups the patterns in the pattern table in classes based on substring inclusion. In the table 6.2 we see examples of three classes of patterns extracted for the taxonomy concepts under the category **passerine**. The patterns grouped in Class 1 usually give taxonomic information ("TitleConcept be T") together with location information ("TitleConcept be T in T"). The second class of patterns extract information related to parts ("TitleConcept have T and T") and the third class of patterns obtain knowledge about what different kinds of birds mainly eat ("TitleConcept usually eat T").

3. **Frequency Heuristics**. The patterns appearing in less than 20 % of concepts in the test set are eliminated. The remaining patterns are used in the knowledge extraction phase.

While examining the automatic extracted patterns we noticed that some of the voted patterns are not accurate and produce noise in the final results. This observation reinforces our believe that the best pattern selection strategy is an

Pattern Class	Pattern Examples
Class 1	TitleConcept be T
	TitleConcept be T in T
	TitleConcept be T and T
	TitleConcept be T of T
Class 2	TitleConcept have T
	TitleConcept have T and T
	TitleConcept have T with T
	TitleConcept have T on T
Class 3	TitleConcept mainly eat T
	TitleConcept usually eat T
	TitleConcept eat T , T and T

Table 6.2: Examples of pattern classes as generated by the **Pattern Grouping** heuristics

automatic pattern selection (akin to that use above) followed by a manual phase in which the noisy patterns are eliminated.

The last module in the pipeline, **Knowledge extraction**, extracts knowledge for the concepts in the taxonomy using the patterns voted by the last three heuristics. As an example of extracted knowledge consider the knowledge resulting after applying the voted pattern "TitleConcept consists of T" to one of the sentences in the entry of the concept **knife**:

- **knife** *consist of a blade*

The extracted property *consist of a blade* will denote a part of the knife. Moreover, applying the pattern "TitleConcept be use in T" to the entries corresponding to the concepts **razor** and **sickle** we extract the following relations that can be interpreted as functions:

- **razor** *be_use_in carpentry*

- **sickle** *be_use_in druidic ritual*

6.4 Knowledge extraction and evaluation

In this section we select a set of concepts arranged in 6 taxonomies from Wordnet. Then we extract and evaluate the extracted knowledge for this set of concepts.

6.4.1 Wikipedia Initial Set

Wikipedia Initial Set consists in the concepts in six Wordnet taxonomies. The concepts are mapped onto Wikipedia articles as explained before. The root nodes of taxonomies are three animals (**Horse**, **Dog**, **Bird**), two vehicles (**Aircraft** and **Boat**) and one tool (**Cutlery**). The distribution of concepts for each taxonomy together with examples of concepts is given in Table 6.3. The number of concepts in the six taxonomies varies from a minimum of 34 concepts to a maximum 128 concepts with an average number of 64 concepts per category. The encyclopedia entries corresponding to the taxonomies categories are downloaded with the software module WWW::Wikipedia. The Wikipedia text is part-of-speech tagged and lemmatized with TreeTagger.

Taxonomic Root	Number of Concepts	Examples
Aircraft	34	**monoplane**, **seaplane** **airliner**, **stealth fighter**
Boat	30	**wherry**, **fireboat** **motorboat**, **steamboat**
Horse	34	**tarpan**, **shetland pony** **percheron**, **palomino**
Dog	128	**belgian sheepdog**, **collie** **rottweiler**, **dalmatian**
Bird	121	**crossbill**, **oscine** **nightingale**, **tailorbird**
Cutlery	34	**knife**, **chisel** **sickle**, **razor**

Table 6.3: The roots of the extracted taxonomies and concept examples

6.4.2 Pattern Voting

Table 6.4 shows examples of patterns voted for each of the six taxonomies.

Our study shows that the extracted patterns can be grouped into two distinct classes: patterns common to all categories and patterns specific to distinct taxonomies. Among the first class of patterns there are: patterns expressing taxonomic information, patterns expressing parts and patterns expressing synonym information.

For example, inspecting the table 6.4 we observe that a pattern voted in all taxonomies is "TitleConcept be T". This pattern is present in almost all articles in Wikipedia and it is usually found in the first three sentences of the abstract. Included in the term connected with the title concept by the verb "to be" there is a noun phrase giving the taxonomic classification of the title concept together with other important information. Interestingly, the taxonomic classification extracted with the help of this pattern is not always found among the superordinate terms in the taxonomy we started with. For example, the extracted superordinate for the concept **red_eyed_vireo** is **songbird**. In WordNet the relevant superordinates of the concept **red_eyed_vireo** are: **oscine**, **passerine** and **bird**, none of which is **songbird**.

Sometimes the patterns used to extract parts are taxonomy specific as is illustrated by the concepts under the category **Cutlery**. For these concepts the system selects the pattern "TitleConcept consist of T" which has 100 % precision for **Wikipedia Initial Set**. By far the most voted part-patterns are in the Class 2 in the table 6.2 above: "TitleConcept have T", "TitleConcept have T and T". On general corpora they are very imprecise, but in Wikipedia they have a satisfactory level of precision.

The third category of common patterns extracts synonyms and can be used in building or improving a thesaurus or wordnet: "TitleConcept known as T", "TitleConcept (T)", etc.

As expected, some of the voted patterns express knowledge specific to the concepts in certain taxonomies. For example, the pattern "T build TitleConcept" is related to concepts in the taxonomy **Aircraft** and the pattern "TitleConcept eat T" is specific to the concepts in the taxonomy **Bird**. Although we thought that the second pattern "TitleConcept eat T" will also appear in the concepts of the taxonomies **Dog** and **Horse** it turned out that it did not appear or it was not voted as relevant. In the first case, the knowledge extracted are constructors of aircraft models like *Pan Am One* or *Edison*. In the second case, the properties obtained are kinds of food (*insects*, *snail*) consumed by different types of birds.

Taxonomic Root	**Examples of voted Patterns**
Aircraft	TitleConcept be T T use TitleConcept T build TitleConcept
Boat	TitleConcept be T TitleConcept use T TitleConcept have T
Horse	TitleConcept be T TitleConcept be use in T TitleConcept require T
Dog	TitleConcept be T TitleConcept need T TitleConcept also know as T
Bird	TitleConcept be T TitleConcept forage on T TitleConcept eat T
Cutlery	TitleConcept be T TitleConcept consist of T TitleConcept be T with T

Table 6.4: Examples of extracted patterns for taxonomy classes

6.4.3 Knowledge Evaluation

In the table 6.5 we give examples of the generated knowledge for three concepts: **andean_condor**, **airship** and **knife** belonging to the taxonomies **Bird**,

Aircraft and **Cutlery**, respectively. Please observe that the extracted knowledge ranges from synonyms (*be_call the Argentinean_Condor* for the concept **andean_condor**) to Location (*be find in South_America* for the same **andean_condor**) or different kind of parts (*consists of a blade* and *make of copper* for the concept **knife**).

Concept	**Examples of Properties**
andean_condor	*be_find_in South_America* *be_call the Argentinean_Condor* *Vultur gryphus*
airship	*use dynamic helium volume* *have a natural buoyancy* *be_know_as dirigible*
knife	*consists_of a blade* *come_in many forms* *make_of copper*

Table 6.5: Examples of extracted properties for three concepts

The knowledge extracted for **Wikipedia Initial Set** is evaluated by two raters using a 3-point scale:

- Ideal Knowledge - (2 points). The extracted properties are necessary for the concepts in the taxonomy. They should be part of an ideal list of properties for the taxonomy concepts (e.g. *is omnivorous* for the concept **australian magpie** or *consists of a blade* for the concept **knife**).

- Partially Correct - (1 point) if the extracted properties correctly describe the taxonomy concepts but are not among their ideal list of properties (e.g. *is related to butcher birds* or *described by English Ornithologist John Latham* for the concept **australian magpie**).

- Incorrect Knowledge - (0 points) if the extracted properties do not apply in any way to the category (e.g. the property *number* for the concept **knife** or the property *be on average* for the concept **andean condor**).

We did not force the intuition of the raters and impose a formal definition for labeling the properties. Notwithstanding this fact the hardest thing for the raters was to distinguish between the ideal knowledge and partially correct knowledge. Consider for example the extracted property *has few natural predators* for the concept **bewick's swan**. Should it be labeled as Ideal Knowledge or Partially Correct Knowledge? On the one hand, it can be argued that the property defines an important characteristic of **bewick's swan**. The property differentiates this bird from other types of birds that have many predators. On the other hand, it can be maintained that the property *has few natural predators* is a contingent and not a necessary property. This is the main reason why we need a third rater, a judge, who solves the disagreements and adds the final label. A set of 60 concepts (approximately 10 concepts per taxonomy) is evaluated by the raters and the judge. In table 6.6 the inter-rater agreement is computed using the Kappa score (Siegel and Castellan (1988)) and the precision is computed for the judge scores.

Taxonomic Root	**Kappa Score**	**Precission**
Aircraft	0.62	0.55
Boat	0.65	0.57
Horse	0.62	0.63
Dog	0.65	0.66
Bird	0.68	0.60
Cutlery	0.79	0.61

Table 6.6: The inter-rater agreement and the precision for the extracted knowledge

The precision is computed using the equation 6.8.

$$Precission = \frac{2N_{IK} + 1N_{PC}}{2N_{Properties}} \tag{6.8}$$

where

- N_{IK} counts the number of ideal knowledge labels

- N_{PC} represents the number of partially correct labels
- $N_{Properties}$ counts all properties evaluated.

The best agreement is obtained for the concepts in the taxonomy Cutlery (Kappa Score 0.79). This is not surprising considering that the extracted properties refer in most cases to the physical characteristics of objects instances of concepts in this taxonomy (e.g. **scissors** *consist of a pair of metal blade*). The average Kappa score is 0.66, a good score considering that the raters were new to the task and had not been trained beforehand.

The average precision for **Wikipedia Initial Set** is 0.60. To compare this results with previous reported figures is not easy. The closest framework is a method proposed by Ruiz-Casado and all (Ruiz-Casado *et al.* (2007)). They acquire from Simple English Wikipedia (an Wikipedia variant intended for people whose first language is not English) patterns expressing the semantic relations linking nouns in Princeton WordNet 1.7 (hyperonymy, hyponymy, holonymy and meronymy). Then they gather new instances for these relations improving in this way the WordNet coverage. The reported precision for the newly extracted relationships is between 60 and 70 depending on the relation. A direct comparison between their system and our system is not possible because, in the first place, the framework they use is weakly supervised, while our framework is completely unsupervised. Secondly, their system is tuned to acquire certain kinds of relations (hyperonyms, parts), while our framework does not make any assumption about the relations that should be extracted. On the positive side, there is an important overlap between the patterns for hyperonyms and part relation generated by both methods[1].

6.4.4 Clustering the Wikipedia Initial Set

The concepts in each taxonomy in the table 6.3 (**Wikipedia Initial Set**) are clustered using CLUTO with the combination of clustering parameters introduced

[1]The patterns they found are more diverse because they used much more concepts covering wider topics.

in the next chapter[1].

Due to data sparseness we cannot use the whole properties extracted by the algorithm as features for clustering. To overcome this problem the clustering features will be any noun or adjective in the noun phrases representing properties. For example, in the case of the extracted property *adult female horse* for the concept **mare** we use as features for clustering the following nouns: {*adult*, *female*, *horse*}. We acknowledge that in some cases it does not make sense to consider any noun or adjective as a potential feature. For example, in the case of the extracted property *require routine hoof care* of the concept **cart-horse** the extracted feature *routine* in isolation is not a good feature[2].

In fact, there is a test that tells us when the adjectives can be used as clustering features. Let be $a_1, a_2 \ldots a_n$ a sequence of adjectives linked or not by conjunctions or disjunctions. Together with the noun they modify they form a property for the concept c. In the case we can infer that c is a_1 and c is a_2 and c is a_n, then the adjectives can be used as features. By way of example consider the extracted property *a calm and easygoing breed* for the concept **tennessee walker**. It can be inferred that the **tennessee walker** is *a breed*, *is calm* and *is easygoing* and therefore the adjectives can be used as features in the clustering process. Because we cannot see how this test can be performed automatically we just say that this is the price to pay for using a mechanical method. The results of clustering for the **Wikipedia Initial Set** are presented in table 6.7

Test Set	**Measure**	**Direct Method**		**Graph Method**	
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Wikipedia Initial Set	**Entropy**	0.322	0.329	0.620	0.390
	Purity	0.806	0.806	0.511	0.753

Table 6.7: Clustering results for **Wikipedia Initial Set**

The best clustering results ($Entropy = 0.322$ and $Purity = 0.806$) are obtained for the Direct Method with cosine similarity measure and this means that

[1]We strongly advise the reader to read first the introductory part of the next chapter where we present the combination of clustering parameters in CLUTO.

[2]There does not exist a relation to link the routine concept with the cart-horse concept.

the resulting clusters are globular. The tree built on the top of this clustering solution sometimes gives interesting concept groupings like **wild horse** and **tarpan**. In a taxonomy the concept **tarpan** is a subordinate of the concept **wild horse**. When there are clustering mistakes the groups under the generated tree are spurious. For example, the concepts **mustang** and **bomber** are grouped under the same node in the tree.

6.5 Chapter Summary

The main contribution of this chapter is the introduction of a new method for information extraction from Wikipedia. The chief idea behind the method is that in Wikipedia the knowledge of similar concepts is expressed using equivalent surface lexico-syntactic patterns. In this chapter we defined the concept similarity by reference to taxonomies of concepts. Two concepts are similar if they are dominated by the same concept in a taxonomy. First our method automatically maps the concepts onto Wikipedia pages (6.3.1). Then relevant knowledge is extracted (6.3.2). We evaluated the extracted properties for a set of concepts called **Wikipedia Initial Set**. The properties extracted for this set are manually evaluated by two raters and then they are evaluated in a clustering task. We showed that the properties extracted fall in two classes: a general class of properties which are present in all Wikipedia entries and specific properties to each taxonomy. The precision of our results is good considering that the method for knowledge extraction is unsupervised. In the next chapter we will more thoroughly evaluate the extracted properties in a clustering task and devise a schema for annotating the properties by their types.

7

The Knowledge Test Set

7.1 Contributions

In this chapter we will collect knowledge for a test set from all three sources: Wordnet using the method presented in chapter 4, web corpora using the combination of unsupervised and weakly supervised method introduced in chapter 5 and Wikipedia using the unsupervised procedure in chapter 6. The main contributions of this chapter are:

1. First we asses the extracted properties in a clustering task. We want to see how good the properties are in grouping the concepts for which they were extracted.

2. Second we annotate the knowledge extracted for the test set from Wikipedia. In this way we can better understand what kind of knowledge is generated for different categories.

7.2 The Knowledge Test Set

The test we collect knowledge for will be called from now on the **Knowledge Test Set**. It has 56 concepts and is assembled among the concepts in McRae feature norm. The concepts in the set belong to one of the 5 categories: {*fruit*, *mammal*, *tool*, *musical instrument*, *bird*}. To the set of concepts in McRae feature norm we add all the 5 categories above because we want to know how the properties generated for the rest of the set compared with the properties extracted for the categories themselves.

The **Knowledge Test Set** is given in table 7.1.

Category	Concepts
Fruit	**fruit** , **grapefruit** , **cherry** **avocado** , **apple** , **cranberry** **blueberry** , **pineapple** , **pear** **prune** , **peach** , **banana** , **grape**
Mammal	**mammal** , **cheetah** , **lion** **elephant** , **zebra** , **coyote** **cow** , **dog** , **sheep**,**racoon**
Tool	**implement** , **drill** , **pen** **racquet** , **chisel** , **knife**, **axe** **wand** , **razor** , **skillet**,**scissors**
Musical Instrument	**musical instrument** , **drum** , **violin** **clarinet** , **accordion** , **saxophone**, **piano** **harp** , **guitar** , **trombone**,**trumpet**
Bird	**bird** , **swan** , **pelican** **ostrich** , **owl** , **duck**, **sparrow** **woodpecker** , **emu** , **penguin**,**eagle**

Table 7.1: The Knowledge Test Set

7.3 Clustering algorithms and clustering measures

The property based clustering of a set of concepts can be stated in the following way. Given a set of concepts belonging to some categories, how good are their properties for clustering back the concepts in their respective categories? The assumption behind this clustering test is that the concepts belonging to a certain category are more similar to each other than are to concepts belonging to different categories. For example, it is expected that based on the extracted properties **musical instruments** will be in a different cluster than **animals** or **tools**. We can represent the concepts to be clustered as rows of a matrix (see 7.1) and their properties as columns. A 0/1 in a matrix entry means that a property is absent/present. Of course instead of using binary values we can assign weights to the extracted properties. In this thesis we prefer to cluster using binary vectors. The reason for this decision is that the properties generated for Wordnet and Wikipedia are not weighted.

$$\begin{array}{c c} & \begin{array}{ccc} F0 & F1 & F2 \end{array} \\ \begin{array}{c} C0 \\ C1 \\ C2 \end{array} & \begin{pmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array} \tag{7.1}$$

All algorithms used for clustering are implemented in the CLUTO package (Zhao and Karypis (2002)). CLUTO has a multitude of clustering algorithms and each algorithm can be tuned across multiple dimensions. To decide which of these algorithms to use is not an easy task. In practice there arise two types of clusters: globular and transitive ones. For globular clusters there exists a subspace of the original dimension across which the objects to be clustered agree. To be successfully used with this type of clustering solution the dimensions of the objects to be clustered (in our case the dimensions are the concept properties) should be shared by a large fraction of the concepts. Unlike the globular clusters, the transitive clusters contain many sub-clusters that share small subspaces of the dimensions and there exists a path that connects these sub-clusters. This type of clustering will be profitably used in our case if and only if there are many subclasses of concepts sharing a high number of common properties and these subclasses will also be connected by a strong path. We do not want to make any assumption about the type of clusters that are better for concept clustering. Consequently, algorithms for both type of clusters are tested:

1. **The Direct Method**. The Direct Method in CLUTO produces globular clusters. The number of clusters (k) to be discovered is computed simultaneously. The quality of the obtained clusters for small values of k (k $<$ 11)

is generally better than the quality of clusters found via a more traditional method: repeated bisection. We compute the clustering solution using two similarity functions: the cosine function and the correlation coefficient.

2. **The Graph Method**. The Graph Methods produce transitive clusters. The concepts to be clustered are represented as vertexes of a graph and the most similar vertexes are connected by edges. The graph is split into k clusters based on a min cut partitioning algorithm. To compute the similarity between two concepts we used two functions: inversely proportional Euclidian distance and extended Jaccard coefficient.

We instruct CLUTO to compute an **Hierarchical Agglomerative Tree** on the top of the clustering solution produced by the Direct Method with cosine function. The nodes in the tree will be used to find some interesting concepts groupings.

The quality of each clustering solution is measured by two standard metrics: **Entropy** and **Purity**. The Entropy looks at the distribution of concepts in each of the categories and the Purity quantifies the extent to which a cluster contains concepts from only one category. More formally the two measures are defined as follows:

Entropy: The entropy of a clustering solution (7.2) is the sum of the entropies of each cluster belonging to the clustering solution(7.3).

$$Entropy = \sum_{r=1}^{k} \frac{n_r}{n} E(Sr) \tag{7.2}$$

$$E(S_r) = -\frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r} \tag{7.3}$$

where n_r is the size of the cluster S_r, q is the number of categories and n_r^i the number of concepts of the ith class that were assigned to the rth cluster.

Purity: The purity of a clustering solution (7.4) is the sum of the purities of each cluster belonging to the clustering solution(7.5).

$$Purity = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r) \tag{7.4}$$

$$P(S_r) = \frac{1}{n_r} \max_i (n_r^i) \tag{7.5}$$

A clustering solution is good if it has small entropy values and high purity values. Of course, a perfect clustering solution will have 0 entropy and 1 purity. Each cluster is also characterized by two sets of features called **descriptive** and **discriminating**. The descriptive features best summarize the cluster and the discriminating features best distinguish a cluster from other clusters.

7.4 The Knowledge Test Set and Wordnet

7.4.1 Property Collection from Wordnet

The collection of properties for **Knowledge Test Set** from Wordnet follows the procedure described in chapter 4 with a small difference. The original algorithms extracts all nouns together with all their modifying adjectives and the verbs from glosses and consider each of them a separate property. In chapter 4 we were interested in generating a high number of Wordnet properties, thus increasing the probability to find the properties in McRae or Garrard Feature Norms. In the property collection for **Knowledge Test Set** the informativeness is our priority and that is why here we consider gloss properties equivalent with terms. The term definition was introduced before but we give it here to ease reading (equation 7.6) .

$$\begin{aligned} &(NPrep)?(((((Adv)?Adj)*(and)?)(Adv)?Adj)?|(Ger)*)(Noun)+) \\ &\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad where \\ &\qquad\qquad\qquad\qquad\qquad Noun = NNS|NN|NPS|NP \end{aligned}$$

$$
\begin{aligned}
Adv &= RB \\
Adj &= JJS|JJR|JJ \\
Ger &= VBG \\
NPrep &= DT
\end{aligned} \tag{7.6}
$$

For example, from the gloss "any of various lithe-bodied roundheaded fissiped mammals, many with retractile claws" of the concept **feline** we extract the following terms: {*lithe-bodied roundheaded fissiped mammal*, *retractile claw*}. Because we are interested in how good the Wordnet properties are for clustering **Knowledge Test Set** we manually clean the automatically extracted properties. In case of Wordnet the precision of automatic property collection is quite high. To estimate the precision we manually compared the automatic extracted properties with the manually cleaned one for 10 concepts. The estimated precision is 85 %. Please notice that the only errors can result from the term extraction from glosses and some wrong property propagation. Despite this high precision the manual intervention is necessary to eliminate the few wrong properties generated or to slightly reformulate the extracted properties. The extracted set of properties is presented in annex 9.1.

7.4.2 Clustering Results

The 5-way clustering is performed for two set of concepts: **Knowledge Test Set** and the **Knowledge Test Set** with the categories removed (e.g without **musical intrument**, **fruit**, etc.).

Table 7.2 presents the clustering results of the **Knowledge Test Set** with the properties extracted from Wordnet. The best clustering results are obtained for the Direct Method with Correlation similarity function. However, the difference between the Correlation similarity function and the Cosine similarity function is not significant (0.062 vs 0.076 entropy). The results for the Graph Method are much worse and therefore we conclude that the resulting clusters are globular. This fact is not surprising considering the way we extract the Wordnet properties[1].

[1] We propagate the properties from higher hierarchy levels to lower hierarchy levels. Therefore there exists a core set of properties common to all concepts in a certain category.

Test Set	Measure	Direct Method		Graph Method	
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Knowledge Test Set	**Entropy**	0.076	0.062	0.491	0.191
	Purity	0.964	0.964	0.643	0.857

Table 7.2: The **Knowledge Test Set** clustered with Wordnet properties

Figure 7.1 represents the tree solution built on top of the Direct Method with Cosine Similarity Function. It can be observed that many concepts exhibit higher similarity and are rightly clustered under the same node in this tree (e.g. **blueberry** vs **cranberry** or **eagle** vs **owl**). On the other hand, there are only two obvious clustering mistakes (e.g. the concept **mammal** is grouped together with birds and not with other mammal animals and the concept **musical instrument** is clustered together with tools). The mistakes can be explained by the fact that both mammals and birds inherit the properties of the higher concept: **animal**. Likewise, musical instruments and tools are subsumed in the Wordnet hierarchy by the concept **instrumentation**. From the point of view of the property inheritance the properties produced for the categories themselves are at the right level of generality. This was to be expected given the procedure for property generation from Wordnet.

For comparison purposes the **Knowledge Test Set** without the categories is clustered using the properties extracted from Wordnet and the properties generated by subjects in McRae feature norms. Table 7.3 presents the clustering results. The first part of the table lists the clustering results for Wordnet properties. We obtained perfect clustering for Direct Method with Cosine Similarity function and almost perfect results for all other methods except Graph Method with Euclidian distance. The second part of the table lists the results for McRae properties. In this case perfect clustering is achieved for the Direct Method using any of the similarity measures.

The next list presents the descriptive features ordered by relevance for McRae and Wordnet for each of the produced clusters:

1. The cluster corresponding to the category **Bird**:

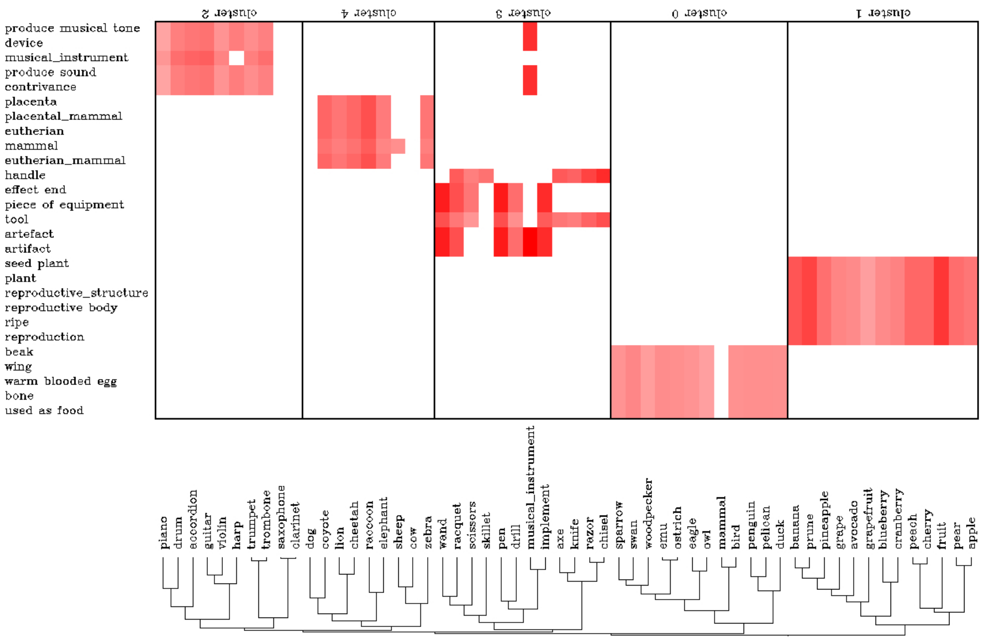

Figure 7.1: The Wordnet Tree Cluster -

Test Set	Measure	Direct Method		Graph Method	
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Wordnet	**Entropy**	0	0.041	0.504	0.041
	Purity	1	0.980	0.6580	0.980
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
McRae	**Entropy**	0	0	0.679	0.188
	Purity	1	1	0.521	0.863

Table 7.3: Clustering Comparison between Wordnet and McRae

- McRae: *has feathers , has a beak has wings, fly, lays eggs*
- Wordnet: *used as food , beak , wing , lay warm blooded egg , bone*

2. The cluster corresponding to the category **Musical Instrument**:
   - McRae: *produces music , used in bands , made of wood , has keys , used in orchestras*
   - Wordnet: *produce sound , device , contrivance , produce musical tone , musical instrument*
3. The cluster corresponding to the category **Fruit**:
   - McRae: *tastes sweet , grows on trees , tastes good , is juicy , is edible*
   - Wordnet: *reproduction, ripe , reproductive body , reproductive structure , seed plant*
4. The cluster corresponding to the category **Mammal**:
   - McRae: *has 4 legs , has a tail , hunted by people , lives in africa , is furry*
   - Wordnet: *mammal , placental mammal , eutherian mammal , eutherian , placenta*
5. The cluster corresponding to the category **Tool**:
   - McRae: *is sharp , made of metal , has a handle , used for cutting , is dangerous*

- Wordnet: *tool , artifact , handle , piece of equipment , effect end*

In general, the descriptive properties extracted from Wordnet correctly apply to the whole category whereas the descriptive properties in feature norms only apply to sub-categories of the main category. The best examples can be observed for the category **Fruit**. All the descriptive properties extracted from Wordnet are necessary properties for the category **fruit**. Instead the properties generated by subjects are not necessary (e.g. not all fruits taste sweet, for example).

In conclusion, the Wordnet derived properties are as good for clustering as are the properties generated in the norms. Both set of properties achieve a perfect clustering solution for the **Knowledge Test Set** intersected with McRae concepts.

7.5 The Knowledge Test Set and corpora

7.5.1 Property Collection from Corpora

The properties for the **Knowledge Test Set** were acquired as detailed in chapter 5. Only the most precise types of properties are extracted: Superordinates and Stuff using the patterns in the table 5.4 and Quality and Action using the association strength method presented in section 5.3.2.2. All the automatically extracted properties were manually cleaned and the resulting set is presented in the annex 9.2.

7.5.2 Clustering Results

Because in a general corpus the probability that the higher order categories like **musical instrument** refer to the whole category is very low[1] we eliminate from **Knowledge Test Set** the higher order categories. We cluster the resulting set with all the properties (table 7.4) and with superordinate properties removed (table 7.5).

[1]In particular we checked all extracted properties of the category **musical instrument** and almost all refer back to a particular musical instrument

Measure	Direct Method		Graph Method	
	Cosine	**Correlation**	**Euclidian**	**Jaccard**
Entropy	0.183	0.101	0.863	0.251
Purity	0.900	0.940	0.386	0.840

Table 7.4: The **Knowledge Test Set** clustered with all properties

According to table 7.4 for all clustering methods with the exception of the Graph Method with Euclidian similarity measure the results are very good. The best results are obtained with the Direct Method and Correlation Similarity Measure (Entropy=0.101 and Purity=0.940). In the following list we give the descriptive features ordered by significance for each cluster:

1. Cluster 0 corresponds to **Tool**: *sharp*, *handle*, *tool*, *cut*, *steel*

2. Cluster 1 corresponds to **Musical Instrument**: *play*, *sound*, *instrument*, *solo*, *classical*.

3. Cluster 2 corresponds to **Fruit**: *fruit*, *grow*, *ripe*, *red*, *juicy*

4. Cluster 3 corresponds to **Bird**: *fly*, *nest*, *bird*, *feed*, *breed*

5. Cluster 4 corresponds to **Mammal**: *animal*, *run*, *eat* , *farm*, *white*

Among the descriptive features for each cluster one can notice the superordinate properties. Some of the descriptive properties correctly apply to all concepts in a certain category (e.g. *play* and *sound* for **Musical Instrument**). Other properties characterize only subcategories of the higher-order category (e.g. the property *farm* for the category **mammal**).

Also good results (but not as good as clustering with all properties) were obtained when we removed the superordinates from the feature set meaning that the weight of superordinate properties is significative (see table 7.5). This time the best results are given by the same Direct Method but with Cosine similarity method. The descriptive features are largely the same as above, as the next list shows. The new descriptive features are underlined.

1. Cluster 0 corresponds to **Tool**: *sharp*, *handle*, *cut*, *steel*, *plastic*
2. Cluster 1 corresponds to **Musical Instrument**: *play*, *sound*, *solo*, *classical*, *accompany*
3. Cluster 2 corresponds to **Fruit**: *grow*, *ripe*, *red*, *juicy*, *slice*
4. Cluster 3 corresponds to **Bird**: *fly*, *nest*, *feed*, *breed*, *wild*
5. Cluster 4 corresponds to **Mammal**: *run*, *eat*, *farm*, *white*, *black*

Measure	**Direct Method**		**Graph Method**	
	Cosine	**Correlation**	**Euclidian**	**Jaccard**
Entropy	0.175	0.330	0.823	0.251
Purity	0.880	0.800	0.386	0.840

Table 7.5: The **Knowledge Test Set** clustered without Superordinate properties

The tree built on top of the clustering solution using all the properties is depicted in the figure 7.2. In most cases the nodes in the tree group together highly similar concepts (e.g. **owl** and **eagle**, **razor** and **knife**). Other nodes show some interesting groupings like **pelican** and **penguin**. The two concepts are similar maybe because they are both water-birds.

7.6 The Knowledge Test Set and Wikipedia

7.6.1 Knowledge Collection

The algorithm for knowledge extractions follows the same steps as the general algorithm presented before (chapter 6). However there are two particularities that the **Knowledge Test Set** has with respect to the **Wikipedia Initial Set**:

1. We give as input to the system in figure 6.3 trees with two levels. The top level is represented by each of the five higher ordered categories and the second level by the rest of concepts in the **Knowledge Test Set**.

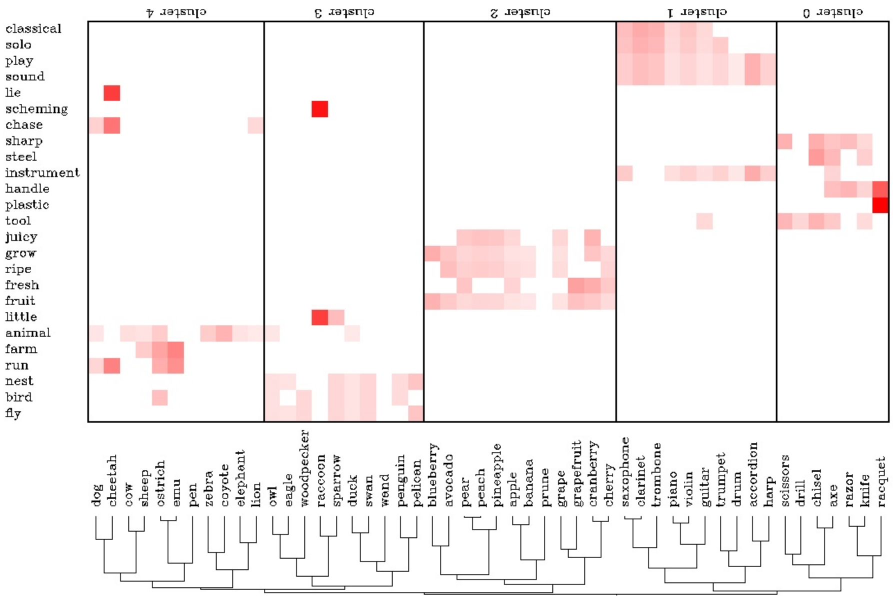

Figure 7.2: Tree clustering of the concepts based on all properties -

2. The concepts in **Knowledge Test Set** are mapped onto their respective Wordnet synsets. Therefore, the synonym expansion procedure and the disambiguation procedure exploit the Wordnet information.

Because we extracted knowledge for Wikipedia for yet another test set[1] we want to see how the voted patterns for similar or identical categories compare. The following categories in the two sets are similar or identical: the **Bird** category is identical in both taxonomies, the **Cutlery** category in **Wikipedia Initial Set** is subsumed by **Tool** category in **Knowledge Test Set**, the **Horse** and **Dog** categories in **Wikipedia Initial Set** are subsumed by the **Mammal** category in **Knowledge Test Set**. We are mainly interested in the conservation of patterns related to a specific category (the patterns in class 3 in table 6.2) and not in general patterns (patterns in classes 1 and 2). The comparison results for the most salient 3 patterns are presented in the table 7.6. The short meaning of the lines of the table is: **B. vs B.** stays for **Bird vs. Bird**, **C. vs T.** stays for **Cutlery vs. Tool** and **HD. vs M.** means **Horse & Dog vs. Mammal**

Category	Specific Patterns Voted for Wikipedia Initial Set	Specific Patterns Voted for Knowledge Test Set
B. vs B.	*TargetConcept Adv eat T*	*TargetConcept feed on T*
	TargetConcept Adv build T	*TargetConcept make T*
	TargetConcept breed in T	*TargetConcept be consider T*
C. vs T.	*TargetConcept consist of T*	*TargetConcept be use in T*
	TargetConcept can be make from T	*TargetConcept be make of T*
	TargetConcept require T	*T use TargetConcept*
HD. vs M.	*TargetConcept name T*	*TargetConcept live in T*
	TargetConcept require T	*TargetConcept be one of T*
	TargetConcept need T	*TargetConcept prefer T*

Table 7.6: Pattern Comparison between **Wikipedia Initial Set** and **Knowledge Test Set**

Some of the differences in pattern conservation can be explained by the different number of concepts in the two sets. If the average number of concepts in the

[1] The **Wikipedia Initial Set**

taxonomies belonging to **Wikipedia Initial Set** set is 64 for the **Knowledge Test Set** the same figure is 6 times lower. Our hope was that the patterns voted should have shown a good degree of conservation for the category **Bird** which is the only category present in both sets. For other categories we thought that the degree of pattern conservation would not be so good. The expectations are confirmed by the results in table 7.6:

- For the category **Bird** the first two patterns for the two sets are equivalent: *TargetConcept Adv eat T* vs *TargetConcept feed on T* and *TargetConcept Adv build T* vs *TargetConcept make T*. Both pairs of patterns extract the same type of information. The first pair of patterns extracts types of food the birds consume and the second pair of equivalent patterns extracts different artifacts produced by birds (e.g. nests).

- For the categories **Cutlery** and **Tool** we found only one equivalent pair of patterns (*TargetConcept can be make from T* vs *TargetConcept be make of T*). This pair of patterns extract properties representing the substance in the composition of different kinds of tools.

- The patterns for the categories **Horse** and **Dog** do not match the patterns voted for the general category mammal.

7.6.2 Clustering Results

The results of clustering are given in the table 7.7. We used three sets of features for clustering, each set corresponding to one of the rows of table 7.7:

1. **Terms**. The clustering features are the whole terms extracted by the algorithm. For example, some of the features of the concept **axe** are: *pick-shaped pointed poll*, *steel head*, *deep angle*. Due to data spareness, a problem we had already anticipated when we clustered **Wikipedia Initial Set**, the results are in general bad (the entropy goes towards 1 and the purity towards 0). Apparently, the best results are obtained for the Graph Method with Jaccard similarity measure. Unfortunately, this method is able to cluster only 31 of the 56 concepts.

2. **Words**. The clustering features are identical with those used in clustering the **Wikipedia Initial Set**. As in that case, the resulting clusters are globular and the best results are obtained for the Direct Method. Due to the manual processing, the results are slightly better in comparison with the results obtained for **Wikipedia Initial Set**: the entropy is decreasing from 0.322 to 0.224 and the purity is increasing from 0.806 to 0.893.

3. **Propagation**. The propagation features are term features to which the features of higher order category are added. For example, to the features extracted for the concept **swan**: *pure white plumage*, *duck family Anatidae*, etc. we add all features belonging to the upper category **bird**: *special flexible lens*, *poor sense*, etc. As you can see all methods with all the similarity measures except the euclidian distance produce perfect clustering.

Clustering Features	Measure	Direct Method		Graph Method	
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Terms	**Entropy**	0670	0.778	0.783	0.382
	Purity	0.500	0.411	0.463	0.718
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Words	**Entropy**	0.343	0.224	0.818	0.454
	Purity	0.786	0.893	0.362	0.655
		Cosine	**Correlation**	**Euclidian**	**Jaccard**
Propagation	**Entropy**	0	0	0.423	0
	Purity	1	1	0.612	1

Table 7.7: Clustering results for **Knowledge Test Set** for Wikipedia

The assumption we made for **Wikipedia Initial Set** that term features are not good for small concept sets is proven correct by the first row of the table 7.7. Moreover, as in case of feature norms and Wordnet extracted properties the perfect clustering solution is due to the presence of higher level properties among the properties generated for lower taxonomy levels. The ordered descriptive features for the form clusters when the clustering features are words are the following:

1. **Tool**: *blade*, *head*, *steel*, *cooper*

2. **Musical Instrument**: *instrument*, *music*, *musical*, *sound*

3. **Fruit**: *fruit*, *cultivar*, *ingredient*, *vitamin*

4. **Mammal**: *animal*, *predator*, *small*, *species*

5. **Bird**: *bird*, *large*, *small*, *symbol*

The agglomerative tree built on top of the clustering solution gives some interesting groupings (e.g. **pear** and **apple**). Mainly due to clustering mistakes some nodes in the tree do not group similar concepts (**penguin** and **elephant**).

7.6.3 The Property Types generated for the Knowledge Test Set

We have a debt to the reader because in the preceding chapter we haven't said what kind of properties are extracted from Wikipedia for each category. In this section we assign a type to the properties produced for the **Knowledge Test Set**. The whole **Knowledge Test Set** annotated with the next property type can be consulted in annex 9.3

1. **Classification**. We encounter two kinds of classifications: either a type concept subsumes another type concept or a role concept subsumes a type concept:

   (a) *Types subsuming Types.* This is the right taxonomic relation. As an example consider the following two examples: **bird** such as *hummingbird* and **the blueberry** *is a false berry.* There is not possible for an instance of the concept *hummingbird* not to be a **bird**. Likewise, it is not possible that blueberry not to be a false fruit. As expected from a resource like Wikipedia, some of the subsuming types are scientific names: *class Aves* for the concept **bird** or *large procyonid* for the concept **raccoon**.

(b) *Roles subsuming Types.* By way of example, consider the next extracted IS-A relations where roles subsume types: **avocado** is **an ingredient** and **elephant** is *a protected species.* Intuitively, it is possible for a certain instance of the concept **avocado** not to be an ingredient. In the case of elephants we can well imagine a situation in the future when in a certain part of the planet the elephants are not declared protected species. Strictly speaking, role concepts can never properly subsume type concepts[1].

The adjectives modifying the head nouns help inferring other properties of the concepts in the **Knowledge Test Set**. In the example: **the coyote** is one of *the few medium-to-large-sized animals* above we can infer the size of coyotes relative to that of the class of animals: medium-to-large.

2. **Behaviour**. Behaviour labeled properties describe the behaviour of animals. It is obvious that these properties apply only for **Bird** and **Mammal** categories. The animal behavior is described along various dimensions like: the food consumed (e.g. the diet of **a pelican** consist of *fish*), hunt related behavior (e.g. **lion** is *a major killer*) , social behaviour (e.g. **sheep** has *a strong lead-follow tendency*, **the domestic dog** has *social intelligence*).

3. **Part**. In general, parts represent the components of the instances of the concepts in the **Knowledge Test Set**. The best examples of parts can be illustrated for the categories **Tool** and **Musical Instrument**. All the extracted Parts have a role in the general functioning of tools or musical instruments: **axe** *has a pick-shaped pointed poll*, **accordion** *a left hand button-board*. Observe again that also in this case the extracted property brings to the table more information than simply enumerating the components. For example, the axe has not simply a poll but the poll is pick-shaped pointed and the accordion has not only a button-board but a button-board for left hand. As expected, the parts extracted for mammals and birds are anatomical components: **swan** has *pure white plumage*, **elephant** has

[1]Nowadays the major-part of the taxonomies built in NLP have not well defined IS-A relations: roles subsume types.

small ear and the parts obtained for fruits are biological components: **the banana** *contain rather large seed*.

4. **Substance**. This properties mainly denote the substance the objects instances of the concepts are made of. Further for the concepts in the category **Fruit** the properties also represent some not physical components: *vitamin B6* which is contained by bananas or *manganese* which enters the composition of **pineapples**[1]. For the categories **Musical Instrument** and **Tool** we extract examples like: **the saxophones** *are made of brass* and **racquets** *are made of composite materials*.

5. **Function**. The concept of function is one of the most debated concepts in Philosophy, Artificial Intelligence or Biology. In philosophy, for example, there are three schools of thought that try to define the concept of function. The functions are either: dispositions (Wright (1973)), explained by reference to etiological theories (Vermaas and Houkes (2003)) or they are considered rather subjective assignments than objective features of reality (Searle (1995)). It is not our purpose to step up in this debate. Suffices to say that in our opinion the function is better defined for tools. In this case functions are interpreted in terms of the designer or the user intention. For example, it was the designer's or user's intention that the **scissors** *are used to cut hair* or are used to cut *various thin material*. Interestingly, the functions for the **Musical Instrument** category have a much stronger social component: **trombone** *is used in outdoor events*, **the accordion** *is used in folk music*. There are a few functions assigned to animals and they show the intended usage of animals by humans: **cow** *give milk*.

6. **Culture**. The concepts do not only allow us to identify and classify the objects in the world but they play important roles in our culture and society at large. The Culture type should be thought as a specialization of the Function type. The cultural function of birds is constituted by the symbolism attached to them: some birds are depicted on various flags, other

[1]Until we find a better classification for this kinds of properties we consider them also Substance properties.

birds are represented on the coins of different countries (e.g. **a pelican** *is depicted on the 1 lek albanian coin*). Many tools have ritual functions (e.g. **wands** *are used in ceremonies*).

7. **Location**. For animals these properties indicate the place where they live, whereas for fruit the places where they grow. The locations have either highly imprecise boundaries (*very hot climate* is indicated as the place where **wild elephants** live), more precise but not clear-cut limits (southern of Brazil) or are precisely delimited geographical regions like Hawaii or United States.

8. **Synonyms**. As the name says, this kind of knowledge spells out synonymous terms for the concepts. The synonyms are either scientific names like *Canis lupus familiaris* for the concept **dog**, synonyms which can be found in a synonym dictionary: (**young swan** *is known as cygnet*) or regional variants (*murawung* or *birabayin*).

9. **Logic**. Strictly speaking, this is not a property type, but it is a characteristic belonging to the language expressing the knowledge. I wanted to stress it specifically because it is the major quality distinguishing the language constructs in the norms from the language constructs extracted from Wikipedia. There are three types of logical constructs we encountered:

   (a) *Quantifier*. This is the most common type of logical construct met. It says how many instances of a concept have a certain property. As illustrations of extracted knowledge consider the following examples: **most birds** *have a poor sense*, **some trombones** *have valve*, **many vegetables** *are botanical fruits*. The employed quantifiers include: all, many, most or cardinal quantifiers (one, two, three).

   (b) *Negation*. Most properties predicate affirmative facts about the individuals. In opposition, the Negation properties establish negative facts: (e.g. **penguins** *have no land predator*).

   (c) **Defined Classes**. This is the same concept of defined classes as in logic (concepts for which the necessary and sufficient conditions are specified). For example: **mare** *is an adult female horse*.

10. **Quality**. For animals the extracted qualities refer to auditory, tactile or visual performances: **eagle** *have extremely keen eyesight.* For fruit the qualities express the colors (greenish flesh), taste(sweet, sour), texture (wrinkly) or shape (oval).

11. **Hyponym**. The algorithm also extracts knowledge which applies to subclasses of the main concept (e.g. **the electric guitar** *is used in jazz* where the information about the concept **the electric guitar** is extracted from the entry of the concept **guitar**).

12. **Similar**. To understand a concept, one usually tries to give comprehensive definitions in the style of **Defined Classes** above. In case the definitions do not work, a similar concept assumed to be understood can be compared with our concept. In general, the similar concept is a co-hyponym of the concept to be explained. Properties labeled Similar differentiate the similar concept from our concept along some dimensions (e.g. **northern coyote** *is typically larger than southern subspecies*).

13. **Physiology**. These properties are specific to animals and denote various physiologic characteristics like:

    (a) Diseases: **dog** *is prone to certain genetic ailment.*

    (b) Peculiar characteristics of body functioning: **lion cub** *is born with brown rosette.*

14. **Functioning**. The set of properties labeled functioning are defined for tools and musical instruments. These properties specify distinct characteristics of tools - functioning: **small head size racquet** *offers more control* or the way the musical instruments are played: **drum** *is played by the hand.*

15. **Cultivation**. Cultivation properties are specific to the plants producing fruit. The properties refer to: the diseases the plants are prone to (**young apple tree**— *is also prone to mammal pest*), the specific mandatory nutrients requiered in tree cultivation (**peach tree** *require a constant supply of water*), etc.

16. **Larger Whole**. Designate a larger organizational structure the instances of concepts participate in: **the clarinet** *is a standard fixture in the orchestra.*

The property types listed above are the most salient property types found. In addition to them the reader will find in the corresponding annex (annex 9.3) the following less important property types:

1. **Time**. The properties denote different temporal aspects (e.g. the time when the fruits are ripe)

2. **Producer**. In case of musical instruments there are listed some musical instruments producers.

3. **DN**. We did not know how to classify some properties and all these properties received the label DN a shortcut for Don't kNow.

Table 7.8 shows the number of property types in the schema for each category in the **Knowledge Test Set**. The last column in the table called "Total" gives the total number of properties annotated. If we consider the property types as parameters and the number of property types as weights assigned to the parameters, then each category in the set can be described in terms of parameter weights. The parameters ordered according to the weight constitute a **signature** for a category. A signature will be denoted as a set of attributes value pairs $ScatName = \{p1 : w1, p2 : w2 \dots pn : wn\}$ where $p1, p2 \dots pn$ are the parameters and $w1, w2 \dots wn$ the parameter weights.

The category **Bird** has the signature given by the equation 7.7[1]. For this category the most prominent property type is Behavior followed by Part and Classification. It is interesting to compare the categories and see which property types are category specific and which are conserved across categories. We expect that birds and mammals belonging to the same superclass animal to be more similar than are birds to fruit or mammals to fruit. Looking to the first four most representative property types for mammals we observe that they are Behavior, Physiology, Classification and Part. Three of them are shared with the four most

[1]In fact, to read the signature for any of the category the corresponding column in table 7.8 should be ordered according to the parameter weight.

Property Type	Bird	Fruit	Mammal	Musical Instrument	Tool	Total
Part	19	8	16	39	8	100
Culture	14	12	12	26	11	75
Classification	14	18	17	11	4	64
Logic	18	10	9	20	1	58
Hyponym	8	3	9	24	12	56
Behavior	23	0	24	0	0	47
Substance	0	17	0	9	11	37
Function	1	9	4	5	14	33
Location	9	16	1	0	0	26
Functioning	0	0	0	16	8	24
Quality	7	8	8	0	1	24
Physiology	3	0	19	0	0	22
Synonymous	2	7	6	0	5	20
Similar	4	3	8	1	0	16
Larger Whole	0	1	3	6	0	10
Cultivation	0	8	0	0	0	8

Table 7.8: Quantitative Evaluation of Property Types for the **Knowledge Test Set**

prominent property types for category Bird but only one is shared with the first four most salient Fruit property types.

As we said before in this chapter, parts and classifications are present in many Wikipedia entries and our **Knowledge Test Set** is no exception. Surprisingly overall the Cultural aspect is very important, the parameter culture existing in all categories, on a par with Classification and Part. Hyponyms and Function are also present in all categories but the most distinguishing feature of the knowledge extracted from Wikipedia is the importance of logic statements also present in all categories. You can notice that the number of Logic statements is quite high. Some properties are specific to particular categories: Functioning to musical instruments and tools, Physiology to Mammals and Birds and Cultivation to Fruits.

$$\begin{aligned} S_{bird} = \{&Behavior : 23, Part : 19, Culture : 14 \\ &Classification : 14, Logic : 18, Location : 9 \\ &Hyponym : 8, Quality : 7, Similar : 4 \\ &Physiology : 3, Synonymous : 2, Function : 1\} \end{aligned} \tag{7.7}$$

The produced property types are far from being perfect and in the future we are going to build an annotation schema with several levels. However, the types are useful for helping us to better evaluate the extracted knowledge.

7.7 Chapter Summary

In this chapter we define a set of concepts called the **Knowledge Test Set** for which we collected knowledge from Wordnet, web corpora and Wikipedia using the methods presented in previous chapters. The extracted properties were used in a clustering task. Based on the extracted properties we should be able to find clusters of concepts that are close to the five higher order categories. We showed that a perfect clustering solution can be obtained if we propagate the properties from higher order categories to concepts in the set. If we do not propagate the properties, a good but not a perfect clustering solution is achieved.

The properties generated in an unsupervised way from Wikipedia are annotated with a list of property types and the distribution of property types in the higher-order categories is quantified. Each higher order category has a signature formed by weighted property types. One of the most relevant types in all Wikipedia entries is the Cultural aspect. Unlike the language used to express knowledge, in feature norms the language used to express knowledge in Wikipedia has many logical constructs.

8

Summary, Conclusions and Further Work

The main subject of this thesis was the acquisition of conceptual structures. Although the term conceptual structure is general enough to accommodate anything from property list to neural networks, from simple conceptual relations to heavily formalized ontologies, following the work on concepts in psychology we described in chapter 2 three conceptual structures. Misleadingly, these conceptual structures are called concept theories and they are named: the classical theory of concepts, the prototype theory and the theory-theory. The conceptual structure in the classical theory is represented by a set of necessary and sufficient conditions. The offspring of prototype theory of concepts are feature norms: collections of concepts and their most salient properties produced in the feature generation task. No formal definition of the theory-theory has ever been proposed but we think that the language used to represent the theory-theory should include logical constructs. In this thesis we acquire feature-norm like structures from Wordnet and web based corpora. Furthermore we go beyond the feature norm structures and acquire more powerful constructs from Wikipedia.

In the chapter 3 we discussed two feature norms (Garrard norm and McRae norm) and a schema which classify the properties in the norms: the Wu and Barsalou taxonomy. Wu and Barsalou taxonomy is based on Barsalou theory of concept representation in the brain. According to this theory the concepts are represented in the brain by perceptual symbols. Further this theory states that the subjects produce properties in feature generation task using simulations. Based on the perceptual simulation view Wu and Barsalou produced a taxonomy of property types. A slightly modified version of this taxonomy is used to annotate the properties generated in McRae feature norm. In the same chapter we answer a question about the stability of semantic memory. If the feature norms are models of semantic memory then we expect the content of semantic memory to be preserved across different feature norms. We showed that this is not the case comparing Garrad and McRae feature norms. Despite this negative result we offered some reasons for the statement that feature norms might be the best model of semantic memory.

Chapter 4 develops a method of extracting conceptual structures from Wordnet. We evaluate Wordnet as a model of semantic memory comparing Wordnet extracted properties with the properties in the two norms:

1. **Global Comparison**. Here we see how the two sets of properties (the properties in each norm and Wordnet Extracted properties) overlap. The found overlap is low: 22 % with Garrard norm and 30 % with McRae norm.

2. **Per Property Type Comparison**. This comparison answers the question : "Which type of properties lack from Wordnet and which ones can be extracted from glosses?". A short answer is that except for the superordinates and partially various kinds of meronymy all other relations are scarcely represented and that glosses can be exploited for the extraction of functional or quality properties. A detailed answer is provided by the table 4.3.

3. **Per category comparison**. This comparison orders a set of categories according to their property representation in Wordnet. The best represented category is **Fruit** followed by **Bird**. Nevertheless, the ordering of the categories is feature norm dependent.

A manual mapping between each norm and Wordnet revealed that the property overlap is larger than estimated automatically. Even if the norms are presented as independent lists of properties the truth is that the properties are interconnected by entailments. Therefore an accurate comparison between the norms and Wordnet is made considering that the properties are interlinked by equivalence and entailment relations. The new manual comparison for a small set of concepts shows that the overlap is at least 10 % larger than estimated automatically.

The chapter 5 exploits very large corpora to extract feature norm like structures. To this end we devise a schema for relation learning starting from Wu and Barsalou taxonomy. According to this schema when asked to list the defining properties of the concepts representing concrete objects the subjects in feature generation task will typically: classify the objects (Superordinate), list their parts and the stuff they are made from (Parts and Stuff), specify the location the objects are typically found in (Location), their intended functions, and their typical behavior (Action), or name their perceptual qualities (Quality). To learn the property types in the schema we employ two different strategies. Superordinate, Part, Stuff and Location properties are learnt using a pattern-based approach. We tried to semi-automatize the pattern-learning framework testing four association measures for the pattern learning task. Quality and Action properties are learnt using a novel method that quantifies the strength of association between the nouns representing the focal concepts and the adjective and verbs co-occurring with them in a corpus. The main conclusion is that only Superordinate, Stuff , Quality and Action properties can be learnt with a reasonable degree of precision. For the rest of properties probable supervised approaches should be tried. The competition for automatic pattern learning does not favor any tested association measure. In the second part of the chapter we used a supervised kernel algorithm implemented at FBK by Claudio Giuliano to learn some relations in the original Wu and Barsalou taxonomy. The results are encouraging but more research is needed for large scale experiments.

If the last two chapters of this thesis explored wordnet and corpora to extract feature norm like descriptions the next chapter exploits Wikipedia to extract richer conceptual structures. Further, in chapter 6 we do not assume that the properties to be extracted are given beforehand. Instead we expect that they emerge from Wikipedia data in an unsupervised way. The method we developed seeks to extract lexico-syntactic surface patterns linking the concepts with their properties. Unlike other approaches in computational linguistics we do not identify patterns using Hearst framework. The main idea we follow is that similar concepts (i.e. those classified under the same node in a taxonomy) are described in a comparable way in Wikipedia. More precisely, we suppose that the relevant knowledge of these similar concepts is expressed using equivalent surface patterns. The learning process starts with the generation of concept hierarchies. The concepts in each hierarchy are mapped onto Wikipedia pages and the knowledge appropriate to the concepts is automatically extracted at a precision ranging from 0.55 to 0.66 depending on the taxonomy. Moreover, we annotate the extracted properties using a simple property type list in chapter 7. The properties extracted range from Superordinate to various kind of Parts, from Behavior to Function. Surprisingly, we found that the cultural aspect is very important, the parameter culture existing in all explored categories, on a par with Classification and Part. The most distinguishing feature of the knowledge extracted from Wikipedia is the presence of logic statements. Unlike the properties in the norms which do contain quantifiers, negation, etc. the logical statements are ubiquitous in the conceptual structure of all tested categories.

In chapter 7 we define a unified concept set called the **Knowledge Test Set** and collect knowledge from all three sources: Wordnet using the method presented in chapter 4, web corpora using the combination of unsupervised and weakly supervised method introduced in chapter 5, and Wikipedia using the unsupervised procedure in chapter 6. The extracted properties are evaluated in a clustering experiment. The main idea behind the clustering experiment is that the extracted properties should be similar for members of a specific category. The clustering results obtained with the properties extracted from Wordnet are as good as the clustering results for properties generated in feature production task. We think that this is due to the fact that the concepts to be clustered share a core set of inherited properties. The clustering results for the properties extracted from corpora and Wikipedia are good but not perfect. The nodes in the tree built above the clustering solution give in many cases appealing concept groupings.

I think the intelligent reader noticed that every resource we used for concept structure extraction gives interesting information[1]. In general, the information extracted from Wordnet and Wikipedia is general and has a scientific flavor. The information extracted from corpora is much more contextual. The very next thing we are going to do is to asses the generality of extracted information. In the future we would like to extend the method for knowledge collection from Wikipedia by using better formalized taxonomies. We would also like to improve the flat property list used to annotate the properties derived from Wikipedia. On the theoretical side we also want to make more formal the notion of theory because it is our belief that the concepts cannot exist in isolation: we do not know about cases when it can be said about a person that she/he possesses exactly one concept.

[1]The curios reader can glance over the thesis annexes.

9

Annexes

9.1 Properties extracted From Wordnet for Knowledge Test Set

Category:**bird**

swan: *craniate*, *neck*, *salt water*, *bird*, *diaphragm*, *foot*, *fowl*, *tail*, *vocal organ*, *swan's down*, *external covering*, *finger*, *feather*, *swim*, *pennon*, *dive*, *notochord*, *used as food*, *uropygium*, *hindquarters*, *aquatic bird*, *heavy body*, *air sac*, *thorax*, *forelimb*, *digit*, *furcula*, *large brain*, *rib*, *bone*, *warm blooded egg*, *wing*, *wade*, *chordate*, *beak*, *tail feather*, *caudal appendage*, *flesh*, *bird's foot*, *skull*, *spinal column*, *usually white plumage*, *trunk*, *lung*, *cartilaginous skeleton*, *belly*, *phylum Chordata*, *vertebrate*, *very long neck*, *animal*, *cranium*, *toe*, *uropygial gland*, *syrinx*,

pelican: *salt water*, *bird*, *distensible pouch*, *foot*, *fowl*, *tail*, *vocal organ*, *seafowl*, *four toed webbed foot*, *external covering*, *feather*, *swim*, *pennon*, *dive*, *used as food*, *large fish*, *frequentopen ocean*, *uropygium*, *hindquarters*, *sea bird*, *aquatic bird*, *air sac*, *furcula*, *forelimb*, *bone*, *wing*, *warm blooded egg*, *wade*, *beak*, *tail feather*, *flesh*, *bird's foot*, *lung*, *pelecaniform seabird*, *frequent coastal water*, *vertebrate*, *large bill*, *warm water seabird*, *syrinx*, *uropygial gland*, *eat fish*,

ostrich: *toed foot*, *craniate*, *most bird*, *ratite*, *neck*, *Africflightless bird*, *bird*, *diaphragm*, *flat breastbone*, *foot*, *fowl*, *tail*, *vocal organ*, *flightless bird*, *external covering*, *finger*, *feather*, *pennon*, *notochord*, *used as food*, *uropygium*, *ratite bird*, *hindquarters*, *thorax*, *air sac*, *forelimb*, *digit*, *furcula*, *large brain*, *rib*, *bone*, *warm* *blooded egg*, *wing*, *chordate*, *beak*, *tail feather*, *caudal appendage*, *flesh*, *bird's foot*, *struthio camelus*, *skull*, *spinal column*, *vertebrate foot*, *lackkeel attachment of flight muscle*, *trunk*, *large living bird*, *lung*, *cartilaginous skeleton*, *belly*, *phylum Chordata*, *vertebrate*, *animal*, *cranium*, *toe*, *uropygial gland*, *syrinx*,

owl: *craniate*, *hunt animal*, *most bird*, *bird of night*, *neck*, *bird*, *claw*, *diaphragm*, *hawk*, *foot*, *fowl*, *tail*, *vocal organ*, *nocturnal bird*, *external covering*, *finger*, *feather*, *pennon*, *notochord*, *used as food*, *uropygium*, *hindquarters*, *air sac*, *large head*, *thorax*, *forelimb*, *large brain*, *digit*, *furcula*, *kill animal*, *rib*, *bone*, *warm blooded egg*, *wing*, *bird of prey*, *chordate*, *beak*, *tail feather*, *caudal appendage*, *flesh*, *bird's foot*, *skull*, *spinal column*, *vertebrate foot*, *trunk*, *carnivorous bird*, *bird of minerva*, *lung*, *cartilaginous skeleton*, *belly*, *phylum Chordata*, *raptor*, *raptorial bird*, *vertebrate*, *hooter*, *eye*, *prey*, *animal*, *cranium*, *toe*, *syrinx*, *uropygial gland*,

duck: *most bird*, *swimming bird*, *broad bill*, *salt water*, *bird*, *foot*, *fowl*, *tail*, *vocal organ*, *external covering*, *feather*, *swim*, *pennon*, *dive*, *depressed body*, *used as food*, *uropygium*, *hindquarters*, *waterfowl*, *aquatic bird*, *short leg*, *freshwater aquatic bird*, *air sac*, *furcula*, *forelimb*, *bone*, *wing*, *warm blooded egg*, *wade*, *beak*, *tail feather*, *flesh*, *waterbird*, *bird's foot*, *anseriform bird*, *lung*, *web foot*, *vertebrate*, *duck down*, *uropygial gland*, *syrinx*,

woodpecker: *craniate*, *strong bill*, *climbing bird*, *neck*, *bird*, *diaphragm*, *any animal*, *foot*, *fowl*, *tail*, *vocal organ*, *external covering*, *finger*, *feather*, *boring wood*, *pennon*, *notochord*, *used as food*, *nonpasserine*, *uropygium*, *hindquarters*, *thorax*, *air sac*, *strong claw*, *digit*, *furcula*, *large brain*, *forelimb*, *rib*, *bone*, *warm blooded egg*, *wing*, *chordate*, *beak*, *tail feather*, *caudal appendage*, *flesh*, *bird's foot*, *skull*, *spinal column*, *vertebrate foot*, *trunk*, *peckerwood*, *piciform bird*, *stiff tail*, *lung*, *pecker*, *cartilaginous skeleton*, *belly*, *phylum Chordata*, *vertebrate*, *chisel like bill*, *animal*, *cranium*, *bore for insect*, *toe*, *uropygial gland*, *syrinx*, *bore into wood*,

emu: *craniate*, *ratite*, *neck*, *bird*, *diaphragm*, *flat breastbone*, *foot*, *fowl*, *tail*, *vocal organ*, *dromaius novaehollandiae*, *flightless bird*, *external covering*, *finger*, *feather*, *pennon*, *notochord*, *used as food*, *uropygium*, *ratite bird*, *hindquarters*, *thorax*, *air sac*, *forelimb*, *digit*, *furcula*, *large brain*, *rib*, *bone*, *warm blooded egg*, *wing*, *ostrich*, *chordate*, *beak*, *tail feather*, *caudal appendage*, *flesh*, *bird's foot*, *emu novaehollandiae*, *skull*, *spinal column*, *vertebrate foot*, *lackkeel attachment* *of flight muscle, trunk, lung, cartilaginous skeleton, belly, phylum Chordata, large Australiflightless bird, vertebrate, animal, cranium, toe, syrinx, uropygial gland,*

bird: *craniate, most bird, animate being, neck, diaphragm, animal tissue, foot, fowl, tail, vocal organ, beast, creature, head, external covering, finger, feather, pennon, brain, notochord, tissue, used as food, uropygium, hindquarters, voluntary movement, thorax, air sac, large brain, digit, furcula, forelimb, rib, wing, bone, warm blooded egg, beak, chordate, tail feather, caudal appendage, being, flesh, body, face, bird's foot, skull, spinal column, cell, vertebrate foot, trunk, fauna, organ, lung, cartilaginous skeleton, belly, phylum Chordata, brute, vertebrate, animal, cranium, toe, syrinx, uropygial gland,*

penguin: *salt water, bird, flightless cold water seabird penguin, foot, fowl, tail, vocal organ, seafowl, sphenisciform seabird, external covering, feather, swim, pennon, dive, used as food, frequentopen ocean, uropygium, hindquarters, sea bird, aquatic bird, air sac, furcula, forelimb, wing, bone, warm blooded egg, wade, beak, tail feather, flesh, seabird, wing asflipper, bird's foot, short legged flightless bird, lung, frequent coastal water, cold southern eAntarctic region, web foot, vertebrate, syrinx, uropygial gland,*

sparrow: *craniate, feed on seed, neck, songbird, bird, diaphragm, foot, fowl, tail, vocal organ, live nearground, feed on insect, external covering, finger, feather, pennon, true sparrow, notochord, used as food, passerine, uropygium, hindquarters, 4 toe, air sac, thorax, digit, furcula, large brain, forelimb, gripperch, rib, bone, warm blooded egg, wing, passeriform bird, chordate, beak, tail feather, caudal appendage, flesh, bird's foot, skull, spinal column, vertebrate foot, trunk, lung, cartilaginous skeleton, belly, phylum Chordata, vertebrate, animal, cranium, toe, uropygial gland, syrinx,*

eagle: *craniate, hunt animal, broad wing, most bird, neck, bird, diaphragm, any animal, foot, fowl, tail, bird of jove, vocal organ, external covering, finger, feather, various large keen sighted diurnal bird, pennon, notochord, used as food, uropygium, hindquarters, soar flight, thorax, air sac, forelimb, digit, furcula, large brain, kill animal, rib, wing, bone, warm blooded egg, chordate, beak, tail feather, caudal appendage, flesh, bird's foot, bird of prey, skull, spinal column, vertebrate foot, trunk, carnivorous bird, lung, cartilaginous skeleton, phylum Chordata, belly, raptor, raptorial bird, vertebrate, animal, cranium, toe, uropygial gland, syrinx,* Category:**fruit**

grapefruit: *grapefruit peel*, *reproduction*, *produce*, *ripe*, *large yellow fruit*, *reproductive body*, *strip*, *peel*, *grow in warm region*, *cook sugar syrup*, *reproductive structure*, *citrus fruit*, *edible reproductive body*, *genus Citrus*, *fresh fruit*, *citrus*, *grapefruit peel*, *green goods*, *vegetable*, *garden truck*, *thick rind*, *section*, *acid juicy pulp*, *edible fruit*, *sweet flesh*, *plant*, *sugar*, *juicy pulp*, *grow formarket*, *seed plant*, *fruit*, *rind*, *green groceries*,

cherry: *reproduction*, *ripe*, *plant organ*, *reproductive body*, *edible fruit*, *stone fruit*, *sweet flesh*, *plant*, *peel*, *reproductive structure*, *seed plant*, *single seed*, *fruit*, *red fruit*, *fleshy indehiscent fruit*, *edible reproductive body*, *drupe*, *rind*, *single hard stone*, *vegetable*,

avocado: *reproduction*, *produce*, *garden truck*, *ripe*, *blackish skin*, *plant organ*, *reproductive body*, *edible fruit*, *single large seed*, *sweet flesh*, *aguacate*, *plant*, *peel*, *avocado pear*, *reproductive structure*, *grow formarket*, *seed plant*, *tropical fruit*, *fruit*, *alligator pear*, *edible reproductive body*, *fresh fruit*, *rind*, *rich yellowish pulp*, *vegetable*, *green goods*, *green groceries*, *pear shape*,

apple: *reproduction*, *pome*, *fleshy fruit apple*, *ripe*, *seed chamber*, *haveouter fleshy part*, *plant organ*, *reproductive body*, *edible fruit*, *sweet flesh*, *plant*, *peel*, *reproductive structure*, *seed plant*, *fruit*, *edible reproductive body*, *pear*, *false fruit*, *rind*, *green skin*, *tart crisp whitish flesh*, *vegetable*,

cranberry: *reproduction*, *ripe*, *berry*, *plant organ*, *reproductive body*, *pulpy edible fruit*, *use for sauce*, *various structure*, *plant*, *make jam*, *use for juice*, *make dessert*, *small fruit*, *reproductive structure*, *cbe preserve*, *seed plant*, *fruit*, *make jelly*, *very tart red berry*,

fruit: *reproduction*, *seed plant*, *plant structure*, *natural object*, *ripe*, *plant part*, *plant organ*, *reproductive body*, *plant*, *reproductive structure*,

blueberry: *reproduction*, *produce*, *garden truck*, *ripe*, *berry*, *reproductive body*, *edible fruit*, *pulpy edible fruit*, *low growing*, *sweet flesh*, *plant*, *make jam*, *peel*, *make dessert*, *sweet edible dark blue berry*, *reproductive structure*, *cbe preserve*, *grow formarket*, *seed plant*, *fruit*, *grow on blueberry plant*, *edible reproductive body*, *fresh fruit*, *rind*, *high growing*, *make jelly*, *vegetable*, *green goods*, *green groceries*,

pineapple: *large sweet fleshy tropical fruit*, *haveterminal tuft*, *reproduction*, *produce*, *garden truck*, *ripe*, *plant organ*, *reproductive body*, *edible fruit*, *sweet flesh*, *plant*, *peel*, *reproductive structure*, *grow formarket*, *stiff leave*, *seed plant*, *fruit*, *edible reproductive body*, *fresh fruit*, *rind*, *vegetable*, *green goods*, *green groceries*, *ananas*,

pear: *come many variety*, *reproduction*, *pome*, *fleshy fruit apple*, *ripe*, *seed chamber*, *haveouter fleshy part*, *plant organ*, *reproductive body*, *edible fruit*, *sweet flesh*, *plant*, *peel*, *sweet juicy gritty textured fruit*, *reproductive structure*, *seed plant*, *fruit*, *edible reproductive body*, *false fruit*, *rind*, *vegetable*,

prune: *reproduction*, *produce*, *garden truck*, *ripe*, *reproductive body*, *edible fruit*, *sweet flesh*, *plant*, *peel*, *reproductive structure*, *grow formarket*, *seed plant*, *fruit*, *dried fruit*, *edible reproductive body*, *fresh fruit*, *rind*, *green goods*, *vegetable*, *green groceries*,

peach: *reproduction*, *ripe*, *plant organ*, *reproductive body*, *edible fruit*, *stone fruit*, *sweet flesh*, *plant*, *peel*, *downy juicy fruit*, *reproductive structure*, *seed plant*, *single seed*, *fruit*, *fleshy indehiscent fruit*, *whitish flesh*, *edible reproductive body*, *drupe*, *rind*, *vegetable*,

banana: *reproduction*, *produce*, *garden truck*, *ripe*, *elongate crescent shaped yellow fruit*, *plant organ*, *reproductive body*, *edible fruit*, *plant*, *peel*, *soft sweet flesh*, *reproductive structure*, *grow formarket*, *seed plant*, *fruit*, *edible reproductive body*, *fresh fruit*, *rind*, *green goods*, *vegetable*, *green groceries*,

grape: *grow in cluster*, *reproduction*, *produce*, *garden truck*, *ripe*, *plant organ*, *reproductive body*, *edible fruit*, *sweet flesh*, *purple skin*, *plant*, *peel*, *juicy fruit*, *fermented juice*, *reproductive structure*, *grow formarket*, *seed plant*, *genus Vitis*, *make of grape*, *fruit*, *edible reproductive body*, *fresh fruit*, *rind*, *wine*, *vegetable*, *green goods*, *green groceries*,

Category:**tool**

drill: *bit*, *make object*, *artifact*, *workpiece*, *repeated blow*, *center*, *cutting edge*, *artefact*, *tool*, *use in practice of vocation*, *instrumentality*, *bitstock*, *lathe*, *piece of equipment*, *chuck*, *part*, *implement*, *sharp point*, *brace*, *holding device*, *drill press*, *adjustable jaw*, *instrumentation*, *effect end*, *make hole in hard material*,

implement: *object*, *piece of equipment*, *make object*, *assemblage*, *part*, *artifact*, *section*, *physical object*, *single entity*, *whole*, *unit*, *artefact*, *instrumentation*, *effect end*, *tool*, *instrumentality*,

pen: *nib*, *ink flow*, *piece of equipment*, *writing implement*, *make object*, *artifact*, *with point*, *artefact*, *instrumentation*, *tool*, *effect end*, *use to write*, *instrumentality*,

racquet: *surface*, *interlaced network of string*, *used in various game*, *artifact*, *oval frame*, *face*, *artefact*, *tool*, *use to strike ashuttlecock*, *instrumentality*, *mmake object*, *piece of equipment*, *sports implement*, *racket*, *use to strikeball*, *handle*, *effect end*, *instrumentation*, *sport*,

chisel: *edge tool*, *cutter*, *cutting tool*, *cutting implement*, *tool*, *use in practice of vocation*, *use for slice*, *cutting tool*, *edge*, *tool for cut*, *cutlery*, *use for cut*, *flat steel blade*, *handle*, *edge tool*, *knife edge*, *cutting implement*,

knife: *edge tool*, *cutter*, *cutting tool*, *cutting implement*, *pointed blade*, *tool*, *use in practice of vocation*, *use for slice*, *knife blade*, *cutting tool*, *edge*, *sharp end*, *tool for cut*, *instrument*, *sharp edge*, *cutlery*, *weapon*, *use for cut*, *edge tool*, *handle*, *point*, *knife edge*, *cutting implement*, *blade*, *haft*,

wand: *make of wood*, *mmake object*, *piece of equipment*, *implement*, *artifact*, *use by water diviner*, *use bymagician*, *make of metal*, *artefact*, *instrumentation*, *tool*, *effect end*, *instrumentality*, *rod*,

razor: *edge tool*, *cutter*, *cutting tool*, *cutting implement*, *razorblade*, *very sharp edge*, *tool*, *use in practice of vocation*, *use for slice*, *cutting tool*, *tool for cut*, *cutlery*, *use for cut*, *handle*, *edge tool*, *knife edge*, *cutting implement*, *blade*, *use in shaving*,

skillet: *kitchen utensil*, *pan*, *use in prepare food*, *consist ofwide metal vessel*, *use inhousehold*, *use for fry food*, *frypan*, *kitchen utensil*, *utensil*, *cookware*, *use for cooking*, *handle*, *frying pan*, *cooking utensil*, *practical use*, *material*, *cooking pan*,

axe: *edge tool*, *cutter*, *cutting tool*, *ax*, *cutting implement*, *heavy bladed head*, *ax handle*, *tool*, *use in practice of vocation*, *use for slice*, *ax head*, *edge*, *head*, *tool for cut*, *cutlery*, *use for cut*, *weapon*, *edge tool*, *handle*, *knife edge*, *cutting implement*, *blade*, *haft*,

scissors: *edge tool*, *pivot*, *cutting tool*, *compound lever*, *used as fastening*, *bar*, *lever*, *tool*, *pair of scissors*, *two cross pivot blade*, *edge*, *make of wood*, *rigid piece*, *pair of lever*, *piece of equipment*, *rigid bar*, *implement*, *used as weapon*, *make* *of metal, pivot aboutfulcrum, edge tool, handle, knife edge, effect end, used as obstruction, hinge atfulcrum, blade, fulcrum,*

Category:**mammal**

cheetah: *chetah, roar, quadruped, Africa, eutherian mammal, swift mammal, mammal, cat, paw, train for game, foot, long legged spotted cat, placental, eat mammal, carnivore, retractile claw, eutherian, roundheaded fissiped mammal, felid, acinonyx jubatus, aquatic flesh, placental mammal, feline, southwestern Asia, big cat, placenta, large cat, animal,*

lion: *roar, quadruped, Africa, long coarse hair, eutherian mammal, mammal, cat, paw, foot, animal's neck, placental, panthera leo, king of beasts, eat mammal, carnivore, crest, mane, retractile claw, eutherian, roundheaded fissiped mammal, aquatic flesh, felid, placental mammal, feline, big cat, tawny coat, India, placenta, large cat, animal, shaggy mane,*

elephant: *craniate, nourish with milk, pachyderm, neck, eutherian mammal, diaphragm, mammal, long trunk, tail, finger, any warm blooded vertebrate, tusk, massive herbivorous mammal, placenta, five toed pachyderm, thorax, digit, large brain, rib, caudal appendage, various nonruminant, coat, skull, hair, spinal column, vertebrate foot, skin, placental, trunk, mammalian, eutherian, thick skin, long pointed tooth, placental mammal, belly, cartilaginous skeleton, long flexible snout, vertebrate, proboscidian, proboscidean, cranium, animal, toe, proboscis, very thick skin,*

zebra: *leg, odd number of toe, several fleet, slender leg, odd-toed ungulate, perissodactyl, hoof, neck, eutherian mammal, hoofed mammal, equid, mammal, cannon, fetlock, hock, white striped Africequines, foot, placental mammal, hoofed mammal, placental, perissodactyl mammal, equine, narrow mane, eutherian, ungulate mammal, placental mammal, ungulate, placenta, flat coat,*

coyote: *canid, quadruped, prairie wolf, Eurasia, eutherian mammal, mammal, paw, foot, hunt in pack, canis latrans, western North America, placental, carnivore, eat mammal, brush wolf, eutherian, nonretractile claw, canine, aquatic flesh, small wolf native, placental mammal, placenta, various fissiped mammal, animal, various predatory carnivorous canine mammal, wolf,*

mammal: *craniate, nourish with milk, neck, diaphragm, animal tissue, tail, beast, creature, finger, head, brain, notochord, tissue, living organism, any warm* *blooded vertebrate, voluntary movement, thorax, large brain, digit, rib, chordate, caudal appendage, body, face, coat, skull, hair, cell, spinal column, vertebrate foot, skin, trunk, fauna, mammalian, organ, phylum Chordata, belly, cartilaginous skeleton, brute, vertebrate, cranium, animal, toe,*

cow: *calf, rumen, cud, beef, bovid, poll, psalterium, hoofed mammal, head, stomach divide into four compartment, bovine animal, bovine, reticulum, bos taurus, abomasum, stomach, mammary gland, cattle, digestion, mouth, kine, various cud, cows, udder, genus Bos, food, hollow horned ruminant, oxen, ear, domestic cattle, ruminant, moo-cow, meat,*

dog: *flag, animal tissue, domestic dog, paw, foot, tail, prehistoric time, beast, creature, head, brain, tissue, living organism, voluntary movement, various fissiped mammal, canid, quadruped, body part, many breed, body, face, canis familiaris, genus Canis, domesticate by man, domestic animal, cell, organism, fauna, common wolf, fit forhumenvironment, nonretractile claw, organ, canine, brute, animal,*

sheep: *rumen, cud, trotter, neck, bovid, draft animal, artiodactyl, mammal, fetlock, foot, placental mammal, goat, psalterium, hoofed mammal, stomach divide into four compartment, ungulate mammal, ungulate, abomasum, reticulum, leg, hoof, stomach, hoofed mammal, cannon, ruminant mammal, hock, digestion, mouth, various cud, withers, even-toed ungulate, hollow horned ruminant, food, artiodactyl mammal, ruminant, even number functional toe on each foot,*

raccoon: *nourish with milk, racoon, eutherian mammal, mammal, coat, omnivorous nocturnal mammal, native to North America, hair, plantigrade carnivorous mammal, placental, native to Central America, skin, carnivore, procyonid, eat mammal, mammalian, eutherian, aquatic flesh, placental mammal, any warm blooded vertebrate, placenta,*

Category:**musical instrument**

drum: *strike percussion instrument, musical instrument, strike kettledrum, membrane stretch across each end, contrivance, particular purpose, mallet, membranophone, percussion instrument, produce sound, strike marimba, strike glockenspiel, musical instrument, instrumentality, drumhead, light drumstick, hollow cylinder, tympan, instrument, device, rounded head, musical percussion instru-*

ment, *percussive instrument*, *produce musical tone*, *instrumentation*, *various device*,

violin: *wooden support*, *hollow body*, *strike*, *unfretted fingerboard*, *string*, *neck*, *narrow strip*, *violin family*, *thin board*, *bowed stringed instrument*, *particular purpose*, *produce sound*, *pitch*, *stretch cord*, *instrumentality*, *sound*, *fiddle*, *bow*, *produce musical tone*, *music resonator*, *wood*, *contrivance*, *fiddlestick*, *pluck*, *musical instrument*, *violinist*, *chin rest*, *taut string*, *play withbow*, *sounding board*, *instrument*, *device*, *bridge*, *fingerboard*, *stringed instrument*, *peg*, *four string*, *stringed instrument*,

clarinet: *single-reed woodwind*, *sound produced by enclose column air*, *beating-reed instrument*, *wood*, *thumbhole*, *straight tube*, *woodwind instrument*, *musical instrument*, *wind instrument*, *wind*, *woodwind*, *reed*, *player blow*, *different frombrass instrument*, *finger*, *mouthpiece*, *finger hole*, *bell*, *single-reed instrument*, *vibrate reed*, *tubular device*, *beating reed instrument*, *reed instrument*,

accordion: *instrumentality*, *piano keyboard*, *wind instrument*, *portable box*, *keyboard instrument*, *musical instrument*, *vibrate by air*, *bank of key*, *free-reed instrument*, *contrivance*, *produce sound*, *musical instrument*, *keyboard*, *bellow*, *instrument*, *device*, *aperture*, *control byplayer*, *squeeze box*, *produce musical tone*, *instrumentation*, *piano accordion*,

saxophone: *single-reed woodwind*, *sound produced by enclose column air*, *beating-reed instrument*, *wood*, *thumbhole*, *single reed woodwind*, *woodwind instrument*, *musical instrument*, *finger*, *wind*, *conical bore*, *woodwind*, *reed*, *player blow*, *different frombrass instrument*, *mouthpiece*, *finger hole*, *sax*, *bell*, *single-reed instrument*, *vibrate reed*, *tubular device*, *beating reed instrument*, *reed instrument*,

harp: *wooden support*, *strike*, *musical instrument*, *string*, *neck*, *thin board*, *contrivance*, *group*, *particular purpose*, *stringed instrument*, *soundbox*, *produce sound*, *pillar*, *pitch*, *pluck*, *pluck with finger*, *stretch cord*, *instrumentality*, *sound*, *triangular frame*, *taut string*, *sounding board*, *instrument*, *board*, *device*, *bridge*, *peg*, *produce musical tone*, *bow*, *chordophone*, *music resonator*,

guitar: *wood*, *wooden support*, *strike*, *six string*, *play by strum*, *string*, *neck*, *narrow strip*, *contrivance*, *thin board*, *produce sound*, *pitch*, *stretch cord*, *musical instrument*, *instrumentality*, *sound*, *taut string*, *sounding board*, *instrument*, *device*, *bridge*, *fingerboard*, *peg*, *stringed instrument*, *bow*, *produce musical tone*, *instrumentation*, *play by pluck*, *music resonator*,

musical instrument: *assemblage*, *make object*, *part*, *artifact*, *instrument*, *section*, *device*, *single entity*, *contrivance*, *whole*, *particular purpose*, *produce musical tone*, *artefact*, *instrumentation*, *produce sound*, *instrumentality*,

trombone: *sound produced by enclose column air*, *wind instrument*, *wind instrument*, *contrivance*, *wind*, *variable length*, *brass instrument*, *particular purpose*, *pitch*, *produce sound*, *valve*, *musical instrument*, *instrumentality*, *brass tube*, *player blow*, *long tube*, *brass wind instrument*, *mouthpiece*, *tone*, *instrument*, *device*, *bell*, *U shaped slide*, *brass*, *tubular device*, *brass instrument*, *produce musical tone*, *cup shaped mouthpiece*, *funnel shaped mouthpiece*, *air column*,

trumpet: *flare bell*, *brass instrument*, *particular purpose*, *pitch*, *produce sound*, *valve*, *instrumentality*, *brass tube*, *brass wind instrument*, *mouthpiece*, *tone*, *bell*, *produce musical tone*, *trump*, *air column*, *horn*, *sound produced by enclose column air*, *narrow tube*, *wind instrument*, *contrivance*, *wind instrument*, *wind*, *variable length*, *play by mean of valve*, *brilliant tone*, *player blow*, *musical instrument*, *cornet*, *instrument*, *device*, *brass*, *tubular device*, *cup shaped mouthpiece*, *funnel shaped mouthpiece*,

piano: , *piano action*, *strike*, *wooden support*, *piano keyboard*, *keyboard instrument*, *string*, *play by depressing key*, *thin board*, *particular purpose*, *mallet*, *lever*, *percussion instrument*, *produce sound*, *pitch*, *soft pedal*, *strike glockenspiel*, *strike marimba*, *stretch cord*, *instrumentality*, *damper*, *percussive instrument*, *hammer*, *pianoforte*, *bow*, *produce musical tone*, *instrumentation*, *protective covering*, *keyboard instrument*, *music resonator*, *strike percussion instrument*, *strike kettledrum*, *contrivance*, *pluck*, *fallboard*, *musical instrument*, *light drumstick*, *keyboard*, *set of key*, *taut string*, *sounding board*, *play*, *device*, *rounded head*, *bridge*, *peg*, *stringed instrument*, *forte-piano*, *sustaining pedal*,

9.2 Properties extracted From Corpora for Knowledge Test Set

Category:**bird**

swan: *breed*, *fly*, *neck*, *bird*, *sink*, *white*, *float*, *fowl*, *wildfowl*, *nest*, *sing*, *majestic*, *feed*, *swim*, *black-necked*, *beautiful*, *wild*, *glide*, *snow-white*, *injured*, *walk*, *graceful*,

pelican: *cross*, *feed*, *dive*, *fly*, *pink-backed*, *skim*, *seabird*, *waterbird*, *white*, *nest*, *brown*, *spot-billed*,

emu: *pure*, *white*, *good*, *run*, *farm*,

penguin: *breed*, *eat*, *loveable*, *trained*, *bird*, *seabird*, *fat*, *yellow-eyed*, *bear*, *fall*, *march*, *nest*, *not fly*, *seal*, *friendly*, *swim*, *cuddly*, *dive*, *royal*, *waddle*, *Antarctic*, *comical*, *walk*, *huddle*,

eagle: *two-headed*, *stirreth*, *scream*, *amr*, *fly*, *heraldic*, *nest*, *legal*, *swoop*, *perch*, *majestic*, *circle*, *spotted*, *bald-headed*, *martial*, *wedge-tailed*, *short-toed*, *soar*, *wheel*, *eat*, *booted*, *drop*, *double-headed*, *imperial*, *peck*, *predator*, *allude*, *bald*, *hunt*, *golden*, *royal*, *glide*, *raptor*, *eye*, *white-tailed*, *scoop*, *hover*,

owl: *breed*, *screech*, *hoot*, *fly*, *hear*, *bird*, *white*, *wise*, *nest*, *brown*, *swoop*, *perch*, *eared*, *feather*, *frighten*, *spotted*, *roost*, *tawny*, *grey*, *eat*, *stare*, *call*, *seal*, *hunt*, *whistle*, *wild*, *glide*, *short-eared*, *pet*, *raptor*, *patrol*, *animal*, *prey*, *long-eared*, *sleep*,

duck: *wader*, *breed*, *long-tailed*, *fly*, *dabble*, *bird*, *migrate*, *cram*, *nest*, *diving*, *yellow*, *tufted*, *swim*, *dive*, *rillettes*, *waddle*, *flock*, *ruddy*, *walk*, *white-headed*, *stir*, *lame*, *muscovy*, *float*, *winter*, *quack*, *spot-billed*, *feed*, *golden*, *wander*, *domestic*, *wild*, *ring-necked*, *white-winged*, *flight*, *paddle*, *animal*, *rear*, *bill*,

woodpecker: *blood-coloured*, *great-spotted*, *lineated*, *red-bellied*, *drum*, *three-toed*, *feed*, *ivory-billed*, *spotted*, *fly*, *ladder-backed*, *peck*, *bird*, *green*, *red-cockaded*, *black*, *buff-spotted*, *fine-spotted*, *white-backed*,

sparrow: *small*, *breed*, *eat*, *grey-headed*, *black-throated*, *black-striped*, *white-crowned*, *twitter*, *fly*, *bird*, *chirp*, *nest*, *sing*, *perch*, *hop*, *feed*, *orange-billed*, *stripe-headed*, *little*, *flock*, *clay-colored*,

ostrich: *stick*, *feed*, *farm*, *bird*, *large*, *run*, *white*, *big*, *black*, *bury*, *red*, *animal*,

Category:**fruit**

grapefruit: *pink*, *white*, *juice*, *fruit*, *red*, *fresh*, *size*,

banana: *small*, *eat*, *produce*, *taste*, *ripe*, *ripen*, *grow*, *slice*, *fried*, *sell*, *green*, *straight*, *chop*, *fair-trade*, *peel*, *export*, *cook*, *yellow*, *rotten*, *split*, *crop*, *fruit*, *Caribbean*, *bent*, *baked*, *top*, *flavour*, *product*, *leave*,

grape: *Red*, *ripe*, *purple*, *ripen*, *crushed*, *grow*, *green*, *white*, *noble*, *blend*, *red*, *hand-picked*, *hang*, *mature*, *fruit*, *pile*, *pick*, *wild*, *brandy*, *cultivate*, *harvest*, *seedless*, *black*, *juicy*, *sour*,

avocado: *small*, *serve*, *fruit*, *ripe*, *halve*, *slice*, *grow*, *baked*, *tangy*,

pineapple: *shake*, *crop*, *fruit*, *grilled*, *ripe*, *warm*, *ring*, *slice*, *grow*, *delicious*, *wedge*, *juicy*, *top*, *sweet*, *candied*,

cherry: *ornamental*, *ripe*, *ripen*, *grow*, *delicious*, *flowering*, *sweet*, *candied*, *red*, *flavor*, *intertwine*, *tart*, *fruit*, *pick*, *fresh*, *wild*, *black*, *flavour*, *sour*, *dark*,

blueberry: *food*, *fruit*, *berry*, *grow*,

peach: *colour*, *fruit*, *ripe*, *warm*, *soft*, *slice*, *grow*, *pale*, *white*, *sweet*, *juicy*, *delicate*, *infuse*, *yellow*,

apple: *taste*, *ripe*, *ripen*, *red-skinned*, *grow*, *stewed*, *rosy*, *delicious*, *green*, *sweet*, *fall*, *peel*, *red*, *crisp*, *rotten*, *dainty*, *golden*, *fruit*, *fresh*, *cooking*, *bitter*, *baked*, *juicy*, *dabinett*, *cox*, *sour*, *sweeten*,

cranberry: *small*, *tart*, *fruit*, *fresh*, *grow*, *wild*, *frozen*, *white*, *juicy*, *red*, *sun-dried*,

pear: *ripe*, *ripen*, *drop*, *grow*, *slice*, *stewed*, *delicious*, *cut*, *sweet*, *red*, *brown*, *crumble*, *yellow*, *spiced*, *fruit*, *fresh*, *wild*, *delicate*, *juicy*, *shape*,

prune: *French*, *stewed*,

Category:**tool**

wand: *magic*, *enable*, *hand-held*, *stretch*, *beautiful*, *hazel*, *craft*, *wave*, *magical*,

drill: *stick*, *electric*, *rotary*, *flint*, *flexible-drive*, *design*, *tool*, *practise*, *machine*, *pneumatic*, *perform*, *percussive*, *mount*, *radial*, *exercise*, *corded*, *equipment*, *shed*, *cordless*,

pen: *scribble*, *felt-tipped*, *scratch*, *mark*, *trace*, *fill*, *red*, *tape*, *refillable*, *clutch*, *attach*, *dip*, *write*, *draw*, *light*, *supply*, *push*, *spin*, *bronze*, *point*, *highlighter*, *coloured*, *refill*,

racquet: *plastic*, *handle*,

axe: *head*, *throw*, *flint*, *greenstone*, *instrument*, *stone*, *Neolithic*, *steel*, *handle*, *tool*, *sharp*, *brake*, *wield*, *orthogonal*,

chisel: *use*, *tool*, *splitting*, *hold*, *sharp*, *skew*, *steel*,

knife: *serrated*, *stick*, *insert*, *fight*, *artifact*, *ceremonial*, *slash*, *bladed*, *steel*, *slice*, *cut*, *use*, *tool*, *sharp*, *wield*, *sharpen*, *sacrificial*, *edge*, *pierce*, *throw*, *strap*, *utensil*, *stab*, *weapon*, *handle*, *protrude*, *thin-bladed*,

razor: *electric*, *slash*, *cut-throat*, *glide*, *cut*, *handle*, *straight*, *traditional*, *sharp*, *edge*, *pleat*,

scissors: *surgical*, *cut*, *tool*, *pointed*, *implement*, *sharp*, *sterile*,

Category:**mammal**

lion: *crouch*, *lead*, *growl*, *fight*, *African*, *chase*, *kill*, *mammal*, *guard*, *paw*, *beast*, *overpower*, *stand*, *carnivore*, *head*, *prowl*, *golden-headed*, *black*, *derive*, *eat*, *roar*, *man-eating*, *attack*, *purple*, *roam*, *black-maned*, *cat*, *devour*, *predator*, *tear*, *stalk*, *tame*, *hunt*, *heart*, *golden*, *wild*, *recall*, *animal*, *hungry*, *sleep*,

coyote: *wail*, *animal*, *call*, *zoomorphs*, *wild*,

raccoon: *little*, *scheming*,

elephant: *eat*, *pink*, *African*, *roam*, *lumber*, *bath*, *mammal*, *white*, *ride*, *lumbering*, *endangered*, *trumpet*, *creature*, *trample*, *head*, *weigh*, *wild*, *dance*, *game*, *straight-tusked*, *tusked*, *walk*, *animal*, *munch*, *giant*, *stomp*,

sheep: *breed*, *eat*, *shear*, *roam*, *graze*, *mammal*, *rustle*, *steal*, *pasture*, *stray*, *infected*, *domestic*, *wander*, *scrapie-infected*, *farm*, *slaughter*, *shag*, *ungulate*, *flock*, *black*, *merino*, *bleat*, *animal*, *ruminant*, *rear*,

cheetah: *run*, *use*, *lie*, *chase*,

cow: *eat*, *breed*, *lame*, *graze*, *mad*, *milk*, *sacred*, *stupid*, *inject*, *brown*, *home-bred*, *chew*, *holy*, *feed*, *moo*, *wander*, *jump*, *animal*, *ruminant*, *mate*,

zebra: *spotty*, *trot*, *cross*, *strip*, *game*, *striped*, *animal*, *yellow*,

dog: *eat*, *breed*, *lick*, *howl*, *faithful*, *kennel*, *trained*, *bark*, *chase*, *dangerous*, *shaggy*, *run*, *stray*, *beagle*, *chew*, *carnivore*, *groom*, *foul*, *domestic*, *jump*, *canine*, *bite*, *pet*, *black*, *ear*, *rabid*, *sledge*, *walk*, *animal*, *train*,

Category:**musical instrument**

drum: *interweave*, *stick*, *clatter*, *heavy*, *roll*, *tribal*, *acoustic*, *bash*, *pound*, *beat*, *sound*, *rotate*, *spin*, *thunder*, *trigger*, *cylindrical*, *resonate*, *thundering*, *thunderous*, *reverberate*, *electronic*, *revolve*, *spur*, *loud*, *bang*, *punchy*, *plod*, *play*, *instrument*, *pulsate*, *resound*, *metronomic*,

harp: *swirling*, *instrument*, *play*, *heavenly*, *triple*, *stringed*, *Gaelic*, *chromatic*, *hang*, *sound*,

trumpet: *loud*, *mournful*, *member*, *ring*, *instrument*, *play*, *blow*, *blare*, *orchestral*, *brasswinds*, *proclaim*, *solo*, *sound*,

violin: *mournful*, *electric*, *classical*, *scratchy*, *instrument*, *play*, *scrape*, *acoustic*, *bow*, *practise*, *accompany*, *sweeping*, *tune*, *solo*, *sound*,

guitar: *lead*, *electric*, *classical*, *acoustic*, *weave*, *tool*, *toy*, *tune*, *pluck*, *sing*, *catchy*, *solo*, *sound*, *loud*, *track*, *rhythmic*, *instrument*, *pick*, *play*, *hook*, *strum*, *distorted*, *sling*, *melodic*, *alder*, *equipment*,

trombone: *four-part*, *growl*, *classical*, *play*, *symphonic*, *brass*, *sound*, *solo*,

clarinet: *lead*, *classical*, *wooden*, *play*, *squeaky*, *accompany*, *sound*, *solo*,

accordion: *diatonic*, *instrument*, *play*, *virtuoso*, *sound*,

saxophone: *soulful*, *wail*, *mellow*, *classical*, *instrument*, *play*, *jazzy*, *American*, *accompany*, *sound*, *solo*,

piano: *lead*, *electric*, *classical*, *sparse*, *gloss*, *acoustic*, *hinge*, *tune*, *sing*, *smash*, *tinkly*, *solo*, *sound*, *romantic*, *perform*, *instrument*, *play*, *chime*, *tinkle*, *hammer*, *ting*, *tinkling*, *accompany*, *upright*,

9.3 Properties extracted From Wikipedia for Knowledge Test Set

Category:**bird**

Behaviour:

the white **pelican** fish in group,**duck** make a wide range of call,**ostrich** live_in nomadic group,**woodpecker** be cavity nester,the diet of a **pelican** consist_of fish,**penguin** form monogamous pair,**owl** make different sound,**duck** feed_on the surface of water,the diet of the **ostrich** consist_of plant matter,**sparrow** be seed-eaters,**sparrow** scavenge_for food,**emu** eat a variety of plant species,**swan** form monogamous pair bond,

Culture:

swan feature_in mythology,the **eagle** as a symbol of the Holy_Roman_Empire,**owl** be messenger,**owl** be_consider a companion,**eagle** feather be_use_in various ceremony,**swan** be a symbol,the **pelican** become a symbol of the Passion of Jesus,**owl**

have_be a feature of falconry,the two-headed **eagle** be the emblem of the Byzantine Empire,**ostrich** have a conservation status,a **pelican** be_depict_on the reverse of_Albanian_1 lek coin,**sparrow** be the bringer of the living death,the **eagle** be a sacred bird,

Classification:

the **ostrich** be the large living specie,**penguin** (order Sphenisciformes,**swan** be_consider a distinct subfamily,**penguin** belong_to a clade of Neoaves,a **pelican** be a large water bird,the **swan** be the large member of the **duck** family Anatidae,the **ostrich** belong_to the Struthioniformes order of ratite,**eagle** be large bird,**duck** be mostly aquatic bird,the **swan** be the large member of the duck family Anatidae,the **woodpecker** be arboreal bird,**pelican** be large bird,

Part:

eagle have very large powerful hooked beak,**penguin** have a thick layer of insulating feather,**swan** have a patch of unfeathered skin ,**emu** have small vestigial wing,**emu** have a soft bill,**swan** have pure white plumage,**ostrich** have unique pubic bone,**penguin** be_characterize_by hairy yellow ornamental head feather,**owl** have large forward-facing eye,

Location:

emu live_in most habitat across Australia,Southwest_Africa **ostrich** inhabit the semidesert,true **sparrow** be_indigenous_to Europe,**pelican** occur_in warm region,**emu** be_introduce_to Maria_Island,the **woodpecker** have a mostly cosmopolitan distribution,the **duck** have a cosmopolitan distribution,

Hyponym:

the small **woodpecker** be the Bar-breasted_Piculet,the small **owl** be the Elf_Owl,a **duck**ling be a young duck,the small **penguin** species be the Little_Blue_Penguin,the large **woodpecker** be the Imperial_Woodpecker,the Emperor **penguin** (the large penguin,

Quality:

sparrow tend_to_be small,**swan** be a dark blackish grey colour,the **emu** be pale blue,**ostrich** be glossy cream-coloured,adult **duck** be fast flier,

Physiology:

owl have binocular vision,**eagle** have extremely keen eyesight,

Behaviour_(Quantifier):

most **penguin** feed_on krill,**duck** have many predator,

Similar:

sparrow be_physically_similar_to other seed-eating bird,**penguin** egg be_small_than any other bird specie,

Behaviour_(Negation):

the **ostrich** have no crop,**penguin** have no land predator,

Synonymous:

emu be_know_as murawung,young **swan** be_know_as cygnet ,

Similar_(Quantifier):

most **eagle** be_large_than any other raptor,

Culture_(Quantifier):

penguin have_be the subject of many book,

Location_(Quantifier):

all **penguin** specie be_native_to the southern hemisphere,

Part_(Quantifier):

ostrich have some small feather,

Category:**fruit**

Part_Substance:

avocado have the high fiber content,the **avocado** have a markedly high fat content,**cranberry** have moderate level of Vitamin C,**prune** contain dietary fiber,**pear** be_rich_in Vitamin_A,**avocado** leave contain a toxic fatty acid derivative,**cherry** contain anthocyanin,**grapefruit** contain naringin,**grapefruit** be a good source of vitamin C,**avocado** be_rich_in B vitamin,**grapefruit** be a good source of vitamin C,**grapefruit** contain naringin,pine**apple** be a good source of manganese,**banana** be a valuable source of vitamin B6,**blueberry** contain anthocyanin,**cranberry** juice contain a chemical component,pine**apple** contain a proteolytic enzyme bromelain,**prune** have a high antioxidant content,

Function:

the **banana** plant's trunk be_use_in Telugu,**grapefruit** peel oil be_use_in aromatherapy,**apple** be an important ingredient in many desert,**grapefruit** peel oil be_use_in aromatherapy,**avocado** be an ingredient,pine**apple** be_use_as an anti-helminthic agent,**avocado** be_use_for milk-shake,**prune** be_use_in cooking,other small- **pear** specie be_use_as rootstock,**prune** be a frequent ingredient,**cherry** tree provide food for the caterpillars,pine**apple** be_use_in dessert,**banana** heart be_use_as a vegetable,world **grape** production be_use_for wine,the **avocado** be_eat_as a fruit,**cranberry** juice be a major use of cranberry,

Location:

banana be_native_to the tropical region,the first pine**apple** grow_in England,native and non-native **cherry** grow_in Canada,the pine**apple** be_introduce_to Hawaii,sweet **cherry** be_grow_in Washington,the cultivation of the **pear** in cool temperate climate,the **peach** be_bring_to America,pine**apple** be_native_to the southern part of Brazil,these sterile **cherry** be the cultivar Kanzan,sweet pine**apple** grow_in Brazil,**pear** grow_in English medieval garden,organic **apple** be_produce_in the United_States,**pear** grow_in the sublime orchard,**blueberry** be_native_to North America,**banana** be_grow_in Hawaii,**pear** be_native_to coastal and mildly temperate region,

Classification:

the **grapefruit** be a subtropical citrus tree,the **banana** plant be a pseudostem,the **cranberry** be a small pale pink berry,**cranberry** be a group of evergreen dwarf shrub,**cranberry** be a major commercial crop,fresh **prune** be freestone cultivar,the **banana** be a highly sustainable crop,the **cranberry** be an epigynous berry,the **pear** be an edible pomaceous fruit,the **grapefruit** be a subtropical citrus tree,a **grape** be the non-climacteric fruit,the pine**apple** be a herbaceous perennial plant,pine**apple** be the only bromeliad fruit,the **avocado** be a climacteric fruit,the **blueberry** be a false berry,**peach** be a deciduous tree,

Culture:

the large and best-known **prune** producer be Sunsweet_Growers,the **banana** have an extensive trade history,**apple** in the book of Genesis,the antiquity of **banana** cultivation in Africa,native Americans use **cranberry**,**apple** ap**pear**_in many religious tradition,**avocado** be_popular_in chicken dish,**apple** appear_in many religious tradition,the **avocado** be_very_popular_in vegetarian cuisine,**peach** flower be the signal of spring,the wild **blueberry** be the official fruit of Maine,**avocado** be a commercially valuable crop,

Quality:

the **peach** have_yellow_or whitish flesh,the **grapefruit** be yellow-orange,**banana** come_in a variety of size and color,a medium **peach** (75g,White **grape** be_actually_green_in color,the **grapefruit** be yellow-orange,**prune** be_wrinkly_in texture,ripe **blueberry** have_white_or greenish flesh,**blueberry** have a sweet taste,

Synonymous:

the **prune** be_know_as Munacca,pine**apple** (Ananas comosus,the **grapefruit** be_know_as the shaddock,the **peach** (Prunus persica,the **avocado** (Persea Americana,the **grapefruit** be_know_as toronja,the **grapefruit** be_know_as the shaddock,**pear** juice be_call perry,the **grapefruit** be_know_as toronja,

Part:

grapefruit produce white four-petaled flower,the **grape** skin provide benefit,**cherry** have attractive flower,the original **banana** contain rather large seed,all culinary **banana** have seedless fruit,the **banana** have numerous string,**pear** have grit,**grapefruit** produce white four-petaled flower,

Cultivation:

peach tree require a constant supply of water,the **peach** be_very_susceptible_to brown rot,**cranberry** be_susceptible_to false blossom,**peach** have a high nutrient requirement,young **apple** tree be_also_prone_to mammal pest,**grape** grow_in cluster,

Similar:

the pear be_very_similar_to the **apple**,**grapefruit** be_close_to the orange,cooking **banana** be_very_similar_to potato,the **pear** be_very_similar_to the **apple**,**grapefruit** be_close_to the orange,

Hyponym:

sour **cherry** include Nanking,seedling **apple** be an example of Extreme_heterozygotes,important sweet **cherry** cultivar include Bing,

DN:

grapefruit come_in many variety,the pine**apple** be an example of a multiple fruit,**grapefruit** come_in many variety,

Part_Substance_(Quantifier):

grapefruit be an excellent source of many nutrient,**grapefruit** be an excellent source of many nutrient,

Cultivation_(Quantifier):

most **grape** come_from cultivar of Vinis vinifera,the **avocado** fruit be_poisonous_to some bird,

Time:

blueberry grow_in April & May,

Production:

the large exporter of **apple** be China,

Defined Class:

the **apple** be the pomaceous fruit of the apple tree,

DN_(Quantifier):

banana have some antacid effect,

Form:

prune have an oval shape,

Larger Whole:

cranberry be_find_in acidic bog,

——————————————————————-

Category:**mammal**

Behaviour:

cow be ruminant,the **cheetah** have an average hunting success rate,most **sheep** be seasonal breeder,the domestic **dog** have social intelligence,**elephant** have a very long childhood,**sheep** be prey animal,**coyote** be persistent hunter,**elephant** prefer forested area,an **elephant** be_consider an allomother,**lion** have the loud roar,**dog** eat plant,**dog** have ear mobility,**cheetah** be a sprinter,the **raccoon** be_sometimes_active_in daylight,the heavy known **lion** be a man-eater,**sheep** be exclusively herbivorous mammal,**sheep** have a strong lead-follow tendency,**zebra** be generally social animal,**dog** be predator,**lion** be major killer,**coyote** eat small mammal,**cheetah** be a hard chase,

Classification:

sheep be grazing herbivore,**dog** be pyometra,**zebra** be African equids,**cow** be the rumen,**cow** be large animal,**elephant** be the large land animal,the **cheetah** be a vulnerable species,domestic **sheep** be relatively small ruminant,the **raccoon** be the large procyonid,**lion** be_consider an ambassador species,**sheep**

be_one_of the few livestock animal,**zebra** have_two subspecies,**coyote** be an important furbearer,**elephant** be large land mammal,the **coyote** be_one_of the few medium-to-large-sized animal,the **cheetah** be a carnivore,

Physiology:

elephant be_born_with few survival instinct,**zebra** have night vision,**dog** be_prone_to certain genetic ailment,**zebra** have great hearing,**lion** be albino,the **lion** be_consider a vulnerable species,**sheep** have poor depth perception,**sheep** have an excellent sense of smell,**lion** cub be_born_with brown rosette,**sheep** have a gestation period,the **cheetah** have unusually low genetic variability,**dog** be dichromats,**sheep** have good hearing,

Culture:

the **elephant** be a protected specie,the **raccoon** be a protected species,**raccoon** be a festive meal,**sheep** be the heraldic animal,the **lion** be the biblical emblem,**lion** be_in Babylon,domestic **sheep** be_use_in medical research,the Nemean **lion** be_symbolic_in Ancient_Greece,the family have **dog**,**zebra** stripe be a popular style,**lion** be_use_in art,the **raccoon** appear_in Native_American art,

Part:

elephant have big ear,**sheep** have a complex digestive system,**lion** have minimal mane,**elephant** have relatively large ear,**elephant** have tusk,**elephant** have small ear,**cheetah** have dark tawny spot,female African **elephant** have large tusk,**raccoon** have short leg,**sheep** have horizontal slit-shaped pupil,the **cheetah** have a small head,**cheetah** have pale red spot,

Hyponym:

Grevy's **zebra** have a donkey-like skull,gray **coyote** be a significant predator,the large known **dog** be an English_Mastiff,the small known adult **dog** be a Yorkshire_Terrier,the white **lion** be a distinct subspecies,the tall **dog** be a Great_Dane,feral **dog** be scavenger,**sheep** milk include the Feta,

Similar:

coyote scat tend_to_be_small_than wolf scat,**coyote** pack be_generally_small_than wolf pack,northern **coyote** be_typically_large_than southern subspecies,domestic **dog** be_good_than chimpanzee,**zebra** be_generally_slow_than horse,the African **elephant** be_typically_large_than the Asian elephant,**dog** differ_from wolf,

Quality:

cheetah cub have a high mortality rate,**dog** be_highly_variable_in height,**elephant** be_typically_greyish_in colour,**elephant** be very light grey,**cheetah** be the fast land animal,**cheetah** can_reach high speed,**cow** be red-green color,

Synonymous:

the **lion** (Panthera leo,the **dog** (Canis lupus familiaris,the **cheetah** (Acinonyx jubatus,the **raccoon** (Procyon lotor,**elephant** be_call pachyderm,adult female **sheep** be_refer_to_as ewe,

Function:

cow give milk,the domestic **sheep** be a multi-purpose animal,**dog** be pack animal,**cow** be_use_as draft animal,

Larger Whole:

sheep be_call a flock,**elephant** live_in a structured social order,**zebra** be herd animal,

Behaviour_(Negation):

adult **lion** have no natural predator,healthy adult **elephant** have no natural predator,

DN:

zebra specie may_have overlapping range,**dog** have different shape,

Time:

male **lion** can_reach an age of 15 or 16 years,

Physiology_(Quantifier):

some **cheetah** have a rare fur pattern mutation,

Quality_(Quantifier):

some **zebra** have brown shadow_stripes,

Part_(Quantifier):

some **zebra** have pure white belly,

Location:

wild **elephant** live_in very hot climate,

Part_(Negation):

domestic **sheep** may_have no horn,

————————————————————————-

Category:**musical_instrument**

Culture:

Jerry Martini play **clarinet**,Steel use **accordion**,dizzy Gillespie be a **trumpet** virtuoso,the **clarinet** be_prominent_in Bulgarian wedding music,Beethoven use **trombone**,Billy Joel use a **clarinet**,the other great modern jazz **trumpet** player be Clifford_Brown,Radiohead use a **clarinet**,Tale Ognenovski play the **clarinet**,the **accordion** appear_in popular music,**trombone** be_also_common_in swing,Patti Smith use **clarinet**,notable natural **trumpet** player include Valentine_Snow,Aerosmith use the **clarinet**,**harp** be_prominent_in Welsh,**guitar** be_use_in Calabria,Wolfgang_Amadeus_Mozart use the **trombone**,the **clarinet** be a central instrument in early jazz,Branford Marsalis play the **clarinet**,the **piano** be a crucial instrument in Western classical music,the **clarinet** be_equally_famous_in Turkey,a **harp** be_common_in popular music,

Part:

a grand **piano** action have a repetition lever,**trumpet** do_have valve,**guitar** have a wide fingerboard,**harp** have a wide and deep soundbox,**accordion** have a left hand buttonboard,**piano** key be_make_of spruce,**harp** do_have forepillars,diatonic button **accordion** use a buttonboard,the **saxophone** use a single-reed mouthpiece,**harp** have a single-action mechanism,high-end **guitar** have inlay,**piano** have a lever,the **trumpet** be_construct_of brass tubing bent,the **saxophone** consist_of an approximately conical tube,piano **accordion** use a musical keyboard,**violin** consist_of a spruce top,Single-headed **drum** consist_of a skin,tenor **trombone** have a bore,the **clarinet** have an approximately cylindrical bore,the very small **accordion** have treble switch,**piano accordion** use a musical keyboard,the **trombone** consist_of a cylindrical tube bent,

Hyponym:

the pocket **trumpet** be a compact B trumpet,a silent **piano** be an acoustic piano,digital **piano** use digital sampling technology,the square **piano** have horizontal string,Double-strung **harp** have lever,high-end classical **guitar** have fretboard inlay,true **harp** such_as Mauritania's ardin,the early **saxophone** have_two separate octave vent,musical_instrument such_as a **piano**,the electric **guitar** be_use_in jazz,the bass **saxophone** be_use_in classic jazz recording,the B **clarinet** have nearly identical tonal quality,German **trombone** include long water key and snake decoration,fret Mighty_Wing **guitar** feature an altogether 6-octave range,diatonic **accordion** be the instrument,musical_instrument such_as the **violin**,the twelve string **guitar** have steel string,an acoustic **guitar** have a sound hole,the modern slide **trumpet** be a B trumpet,

Part_(Quantifier):

all **accordion** have reed rank,all **saxophone** use the same key arrangement,most **violin** have_four string,Some **trombone** have valve,most **saxophone** have key_touches,some **trumpet** have a slide mechanism,all piccolo **trumpet** have_four valve,some **accordion** use a chromatic buttonboard,some **clarinet** have a single joint,**drum** with_two head,Modern **trumpet** have_three piston valve,some **guitar** have a filler strip,all **harp** have a neck,**saxophone** use a single reed,grand **piano** have_three pedal,modern upright **piano** have_three pedal,electric **guitar** have_two neck,

Play:

clarinet use additional tone hole,the **violin** produce loud note,**harp** use the mechanical action,**guitar** enable simple fifth,classical **harp** technique use a fingering,**violin** be_make_in so-called fractional size,the **trumpet** can_be_play_in several different valve combination,**violin** have a somewhat harsh sound,percussionist play a **drum**,the **accordion** in_their distinctive sound,**drum** be_play_by the hand,the **accordion** be an aerophone,the bass **trumpet** be_play_by a **trombone** player,old **clarinet** be_tune_to meantone,the bass **trumpet** be_play_by a trombone player,the **accordion** enable the air flow,

Classification:

the **clarinet** be a musical instrument in the woodwind family,the **clarinet** family be the large such woodwind family,the **guitar** be a musical instrument,the **trombone** be_consider a cylindrical bore instrument,the **drum** be a member of the percussion group,the **trumpet** be a musical instrument,the **trombone** be a musical instrument in the brass family,the **piano** be a musical instrument,a **harp** be a stringed instrument,the **accordion** be a portable box-shaped musical instrument of the hand-held bellows,**saxophone** be a straight conical tube,

Part_Substance:

Modern **guitar** string be_construct_of metal,early **piano** be_make_with thin string,**saxophone** be_make_of brass,**trombone** bell be_make_from solid sterling silver,modern inexpensive **clarinet** be_make_of plastic resin,**trombone** be_make_of gold brass,a **piano** be_make_of a steel core,**piano** be_make_of hardwood, Larger Whole:

clarinet be part of standard orchestral instrumentation,Italian and German opera use **harp**,the **clarinet** be a standard fixture in the orchestra,**violin** appear_in ensemble,the **trombone** gain a regular footing in the orchestra,the **violin** have_be a part of British folk-rock music,

Function:

drum be_use_in music therapy,the **accordion** be_use_in folk music,the **accordion** be_use_in solo,**trombone** be_use_in outdoor event,the **harp** be_use_in ballad,

Time:

violin have a limited lifetime,the **accordion** be_one_of several European invention of the early 19th century,electric **violin** date_to the late 1930s,

DN:

the **violin** section use the col legno technique,repair **violin** be_call a luthier,

Producer:

the **violin** make_by the Stradivari,stringed electric **violin** be_available_from several manufacturer,

Play_(Quantifier):

most **drum** be_consider untuned_instruments,

Quantifier:

the **clarinet** be the only wind instrument among string instrument,

Similar:

violin be the small and highest-pitched member of the violin family,

Part_(Negation):

trombone have no seventh slide position,

Part_substance:

a **violin** be_make_from different type of wood,

Culture_(Quantifier):

many traditional culture **drum** have a symbolic function,

Category:**tool**

Part:

axe have a shallow wedge angle,a **drill** press consist_of a base,a **knife** may_be a fixed-blade or a folding version,Japanese **chisel** have hollow,kitchen **scissors** have the fulcrum,splitting **axe** have a deep angle,modern **axe** have steel head,**razor** have a shaving head,a **razor** be a bladed tool,**axe** have a pick-shaped pointed poll,**scissors** have overlapping blade,**scissors** consist_of a pair of metal blade,straight **razor** consist_of a blade,modern **knife** consist_of a blade,**axe** have head,**scissors** have blade,

Function:

the villainess use silver **wand**,**razor** be_use_in car**pen**try,dip **pen** be_use_in illustration,sculptor use a spoon **chisel**,**knife** use_to_cut electrical wire,**axe** use_for chopping,**razor** be_use_in bread production,**skillet** be_necessary_for cook,**scissors** use_to_cut hair,**chisel** be_use_for_cut groove,**drill** be_use_in medicine,razor be_use_in car**pen**try,**drill** be_use_in woodworking,**scissors** be_use_for_cut various thin material,

Part_Substance:

axe be_make_of a resilient hardwood,a **pen** make_from a flight feather,**axe** make_from ground stone,the **drill** be_make_from polythene,long **racquet** be_make_of wood,**axe** be_make_of copper,**skillet** be_make_of cast iron,cold **chisel** be_make_of steel,**skillet** make_from copper,**razor** be_make_of ivory,most **racquet** be_make_of composite material,

Culture:

Quill **pen** be_use_in medieval time,the quill **pen** be_use_in Qumran,**wand** be ceremonial,a **wand** to_fight the goblin,**knife** be_include_in some Anglo-Saxon burial rite,**skillet** be_use_in ancient Mesopotamia,**wand** be_also_common_in the fictional fantasy world,**razor** be_also_popular_in travel wash kit,the **wand** of Moses be a hazel wand,**wand** be_associate_with the element of fire,

Hyponym:

chisel be the slick,cordless **drill** be_make_from polymorph,the 1970s **skillet** be_know_as a multicooker,high end **pen** include archaic type,composite **racquet** be the contemporary standard,mid-plus **racquet** be the general standard,the early **drill** be bow drill,tin snip be **scissors**,Specialized **scissors** include sewing scissors,

Functioning:

a **drill** press will_have an throat distance,a cordless **drill** with a high torque,big head size **racquet** offer more power,**drill** be_power_by various power source,**drill** with a percussive action,the **razor** be_power_by a small DC motor,small head size **racquet** offer more control,a table tennis **racquet** be_use_by player in the game table tennis,

Synonymous:

this **skillet** be_call a sauteuse,a **pen** (Latin pinna,the double **axe** (labrys,the electric **razor** (_know_as the electric dry shaver,a **drill** press (belt,

DN:

the **axe** have many form,**knife** come_in many form,cold **chisel** come_in a variety of size,

Classification:

a **chisel** be a tool,the **wand** be a pre-Norman unit,**scissors** be a first-class double-lever,

Quality:

the extra long **racquet** be_light_in weight,

Part_(Negation):

a dip **pen** have no ink reservoir,

Bibliography

Agirre, E. and Martinez, D. (2000). Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the Semantic Annotation And Intelligent Annotation workshop organized by COLING*. 31

Agirre, E. and Rigau, G. (1996). Word sense disambiguation using conceptual density. In *Proceedings of the 16th International Conference on Computational Linguistics*. 31

Almuhareb, A. and Poesio, M. (2005). Concept learning and categorization from the web. In *Proc. of Annual Meeting of Cognitive Science Society*, Stresa (Italy). 14

Amaro, R., Chaves, R. P., Marrafa, P., and Mendes, S. (2006). Enriching wordnets with new relations and with event and argument structures. In *Seventh International Conference on Intelligent Text Processing and Computational Linguistics*. 33

Armstrong, S., Gleitman, L., and Gleitman, H. (1983). What some concepts might not be. *Cognition*. 10

Ashcraft, M. H. (1978). Property norms for typical and atypical items from 17 categories: A description and discussion. *Memory and Cognition. (6) 227-232*. 18

Barbu, E. (2008). Combining methods to learn feature-norm-like concept descriptions. In *In the proceedings of Worksop Lexical semantics: bridging the gap between semantic theory and computational simulations*. 1

Barbu, E. (2009). Acquisition of common sense knowledge for basic level concepts. In *In proceedings of Recent Advances in Natural Language Processing*. 1

Barbu, E. and Poesio, M. (2008). A comparison of wordnet and feature norms. In *In proceedings of Global Wordnet Conference*. 1

Barbu, E. and Poesio, M. (2009). Unsupervised knowledge extraction for taxonomies of concepts from wikipedia. In *In proceedings of Recent Advances in Natural Language Processing*. 1

Barsalou, L. (1985). Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition. 11, 629-654*. 18

Barsalou, L. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **358**(1435), 1177–1187. 19, 20

Bunescu, R. and Mooney, R. J. (2005). Subsequence kernels for relation extraction. In *Proceedings of the 19th Conference on Neural Information Processing Systems*, Vancouver, British Columbia. 85

Caramazza, A. and Berndt, R. (1978). Semantic and syntactic processes in aphasia:a review of the literature. In *Psychological Bulletin 85*. 32

Charles, W. G. and Miller, G. A. (1989). Contexts of antonymous adjectives. *Applied Psycholinguistics 10*. 32

Church, K. W. and Hank, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics(16)*. 67

Cimiano, P. and Wenderoth, J. (2005). Automatically learning qualia structures from the web. In *Proc. ACL workshop on Deep Lexical Acquisition*, Ann Arbor, MI. 13

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines*. Cambridge University Press, New York, NY, USA. 84

DeGroot, M. H. and Schervish, M. J. (2002). *Probability and Statistics*. Addison-Wesley Publishing Company. 67

Dorr, B. J. and Katsova, M. (1998). Lexical selection for cross-language applications: Combining lcs with wordnet. In *In Proceedings of the Third Conference of the Association for Machine Translation in the Americas*. 31

Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics (19)*. 68

Ferraresi, A., Zanchetta, E., Baroni, M., and Bernardini, S. (2008). Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proc. of the 4th Workshop on the Web as a Corpus*, Marrakesh. 69

Fodor, J. (1981). The present status of the innateness controversy. In *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press. 11

Gangemi, A., Guarino, N., Masolo, C., and Oltramari, A. (2003). Sweetening wordnet with dolce. *AI Mag.*, **24**(3), 13–24. 33

Garrard, P., Lambon, R., Hodges, M. A., and Patterson, J. R. (2001). Prototypicality, distinctiveness, and intercorrelation analyses of the semantic attributes of living and nonliving concepts. *Cognitive Neuropsychology*. 20

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic discovery of part-whole relations. *Computational Linguistics*, **32**(1), 83–135. 74, 82

Giuliano, C., Lavelli, A., and Romano, L. (2006). Exploiting shallow linguistic information for relation extraction from biomedical literature. In *Proceedings of the Eleventh Conference of the European Chapter of the Association for Computational Linguistics (EACL-2006)*, Trento, Italy. 85

Giuliano, C., Lavelli, A., Pighin, D., and Romano, L. (2007). Fbk-irst: Kernel methods for semantic relation extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 141–144, Prague, Czech Republic. Association for Computational Linguistics. 85

Guarino, N. (1992). Concepts, attributes and arbitrary relations. *Data and Knowledge Engineering*, **8**, 249–261. 14

Hearst, M. A. (1998). Automated discovery of wordnet relations. In C. Fellbaum, editor, *WordNet: An Electronic Lexical Database*. MIT Press. 13, 62

H.Liu and Singh, P. (2004). Conceptnet: A practical commonsense reasoning toolkit. *BT Technology Journal*. 19

Justeson, J. and Katz, S. (1995). Technical terminology:some linguistic properties and algorithms for identification in text. In *Proceedings of ICCL-95*. 102

Kamp, H. and Partee, B. (1995). Prototype theory and compositionality. *Cognition (57)*. 11

Kintsch, W. (1974). *The Representation of Meaning in Memory*. Lawrence Erlbaum Associates. 9

K.Knight (1993). Building a large ontology for machine translation. In *Proceedings of the ARPA Human Language Technology Workshop*. 31

Kronlid, F. (2003). *Modes of explanation - aristotelian philosophy and pustejovskyan linguistics*. Master's thesis, University of Goteborg. 13

Langer, S. and Hickey, M. (1997). Using semantic lexicons for intelligent message retrieval in a communication aid. In *Proceedings of the ACL/EACL Workshop on Natural Language Processing for Communication Aids*. 31

Lenat, D. B. (1995). Cyc: a large-scale investment in knowledge infrastructure. *Commun. ACM*, **38**(11), 33–38. 12

Magnini, B. and Strapparava, C. (2001). Using wordnet to improve user modelling in a web document recommender system. In *Proceedings of the NAACL 2001 Workshop on WordNet and Other Lexical Resources*. 31

Magnini, B., Negri, M., Prevete, R., and Tanev, H. (2002). A wordnet-based approach to named-entites recognition. In *Proceedings of the Coling 2002 Workshop "SemaNet'02: Building and Using Semantic Networks*. 31

Mandala, R., Tokunaga, T., and Hozumi, T. (1998). The use of wordnet in information retrieval. In *Proceedings of the COLING/ACL Workshop on Usage of WordNet in Natural Language Processing Systems*. 31

Margolis, E. and Laurence, S. (1999). *Concepts: Core Readings*. The MIT Press. 11, 12

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, **37**, 547–559. 20, 25

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from wikipedia. *International Journal of Human-Computer Studies*. 93

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: an on-line lexical database. *Int J Lexicography*, **3**(4), 235–244. 32

Moss, H., Tyler, L., and Devlin, T. (2002). The emergence of category-specific deficits in a distributed semantic system. In *Category-specificity in brain and mind*. Psychology Press. 18

Murphy, G. L. (2002). *The Big Book of Concepts*. MIT Press, Cambridge, MA. 12

Oli, C. (1994). A modular and flexible architecture for an integrated corpus query system. In *Proceedings of COMPLEX'94,*. 69

Pantel, P. and Pennacchiotti, M. (2006). Espresso: Leveraging generic patterns for automatically harvesting semantic relations. In *Proc. of ACL / COLING*, pages 113–120, Sydney. 63

Poesio, M. and Almuhareb, A. (2005). Identifying concept attributes using a classifier. In T. Baldwin and A. Villavicencio, editors, *Proc. of the ACL Workshop on Deep Lexical Semantics*, pages 18–27, Ann Arbor, Michigan. 14

Poesio, M., Barbu, E., Giuliano, C., and Romano, L. (2008). Supervised relation extraction for ontology learning from text based on a cognitively plausible model of relations. In *In the proceedings of Worksop Ontology Learning and Population*. 1

Ponzetto, S. and Strube, M. (2007). Deriving a large scale taxonomy from wikipedia. In *Proceedings of AAAI*. 96

Pradhan, S., Loper, E., Dligach, D., and Palmer, M. (2007). Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the 4th Workshop on Semantic Evaluations*. 96

Priss, U. (1998). The formalization of wordnet by methods of relational concept analysis. In *WordNet: An Electronic Lexical Database*. MIT Press. 33

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press. 14

Ravichandran, D. and Hovy, E. (2002). Learning learning surface text patterns for a question answering system. In *Proceedings of ACL*. 63

Rosch, E. (1999). Reclaiming concepts. *Journal of Consciousness Studies 6(11)*. 10

Rosch, E. and Mervis, C. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*. 9, 17

Ruiz-Casado, M., Alfonseca, E., and Castells, P. (2007). Automatising the learning of lexical patterns: An application to the enrichment of wordnet by extracting semantic relationships from wikipedia. *Data and Knowledge Engineering*, **61**(3), 484–499. Advances on Natural Language Processing - NLDB 05. 114

Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA. 78, 79

Searle, J. (1995). *The Construction of Social Reality*. Free Press. N.Y. 135

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, New York, NY, USA. 78

Siegel, S. and Castellan, N. J. (1988). *Nonparametric statistics for the Behavioral Sciences*. McGraw-Hill, 2nd edition. 83, 113

Smola, A. J. (1998). *Learning with Kernels*. Ph.D. thesis, Technische Universität Berlin. 78

Strube, M. and Ponzetto, S. (2006). Wikirelate! computing semantic relatedness using wikipedia. In *Proceedings of AAAI*. 93

Suchanek, F. M., Kasneci, G., and Weikum, G. (2007). Yago: a core of semantic knowledge. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 697–706, New York, NY, USA. ACM. 96

Vapnik, V. (1998). *Statistical Learning Theory*. John Wiley and Sons, New York, NY. 79

Vermaas, P. and Houkes, W. (2003). Ascribing functions to technical artefacts: A challenge to etiological accounts of functions. *British Journal for the Philosophy of Science*. 135

Wright, L. (1973). Functions. *Philosophical Review*. 135

Wu, L.-L. and Barsalou, L. (2006). Grounding concepts in perceptual simulation: Evidence from property generation. In press. 24, 82

Zhao, Y. and Karypis, G. (2002). Criterion functions for document clustering: Experiments and analysis. Technical report. 119