

Lexical Knowledge Structures

MTech Seminar Report

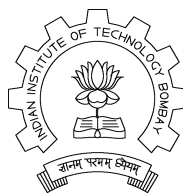
by

J. Ramanand, KReSIT

Roll No: 05329402

under the guidance of

Prof. Pushpak Bhattacharyya
Computer Science and Engineering
IIT-Bombay



Kanwal Rekhi School of Information Technology
Indian Institute of Technology, Bombay
Mumbai

Acknowledgments

*The best teacher is the one who suggests rather than dogmatizes,
and inspires his listener with the wish to teach himself*
- Edward Bulwer-Lytton (1803 - 1873)

Thanks, **Prof. Pushpak Bhattacharyya**, for all the suggestions and guidance.

J. Ramanand
MTech First Year
KReSIT, IITB

Abstract

This report outlines a survey of lexical knowledge structures (also known as lexical or semantic networks) carried out as part of a seminar study. Lexical knowledge networks help provide semantic information about words and concepts, and find use in several fields. Various approaches regarding their construction and representation have been proposed in research studies. This seminar surveyed some of these. In addition, mention is made of the allied field of upper ontologies.

Contents

1	Introduction To Lexical Knowledge Structures	1
1.1	Motivation	1
1.2	Definition	2
1.3	Purposes and Goals	2
1.4	Structure	2
1.5	Applications	3
1.6	Parameters of Lexical Networks	3
2	WordNet	4
2.1	Origin and Philosophy	4
2.2	Structure	5
2.3	Construction	5
2.4	Applications	6
2.5	Comments	6
2.6	Illustration	7
3	ConceptNet	8
3.1	Origin and Philosophy	8
3.2	OMCS	9
3.3	Structure	9
3.4	Construction	11
3.5	Applications	11
3.6	Comments	12
4	HowNet	13
4.1	Origin and Philosophy	13
4.2	Knowledge Base	14
4.3	Concept Representation	14
4.4	Construction	15
4.5	Applications	15
4.6	Comments	16

5	FrameNet	17
5.1	Origin and Philosophy	17
5.2	Frames	17
5.3	Structure and Construction	18
5.4	Comments	19
6	MindNet	20
7	Comparison of Lexical Knowledge Structures	22
7.1	Conceptual Basis	22
7.2	Principles of construction	22
7.3	Methods of construction	23
7.4	Representation	23
7.5	Quality of Database	23
7.6	Spread	23
8	SUMO	25
9	Conclusion	27

Chapter 1

Introduction To Lexical Knowledge Structures

1.1 Motivation

Fields such as natural language processing and information retrieval often seek lexicons that can provide information on words in tasks such as Word Sense Disambiguation, Question Answering, Context Generation etc. This information includes parts of speech for words, associations with other words (semantic, syntagmatic, paradigmatic etc.), meaning, glosses, example usages and involvement in larger concepts. Such lexical databases should be represented in machine-readable form so that they can be exploited by NLP tools. Conceptually, these concepts and words have a highly systematic structure underlying them which must be reflected in such databases. This structure consists of various kinds of relations among words and usually results in a directed acyclic graph (DAG), with the lexemes as nodes and edges as associations among them.

Such lexical knowledge structures are distinguished from ordinary dictionaries. The latter principally consist of definitions of various meanings associated with a word along with part-of-speech and examples whereas the former provide far richer semantic information about words. Similarly, thesauri contain only relations such as synonymy and antonymy, while lexical structures like WordNet provide far richer linkages such as hypernymy, meronymy etc. These lexical networks form rich ontological structures of concepts.

Several knowledge structures have been constructed in different research efforts. These have laid out many important design parameters which influence their quality, utility and maintenance. These structures also form part of recent trends in developing ontologies for various domains and are gradually becoming important members of the infrastructure of applications promising semantic intelligence.

1.2 Definition

A lexical knowledge structure can be defined as *a systematic collection of word or words along with labeled relations between them, usually in machine-readable form*. Such a structure may have different relations among them, depending on the underlying motivations behind the structure. The lexicon helps capture semantics of the words by observing their attributes and their relative position in the lexicon. The study of this systematic, meaning related structure is called *Lexical Semantics* ([3]). It is based on the perception that a word is an association between a lexicalised concept and an utterance that plays a syntactic role ([5]). Some Lexical Databases are also called “semantic networks” when they provide rich semantic information.

1.3 Purposes and Goals

Building knowledge structures initially involves a survey of words to be included and then discovering relations among them. Understanding the key principles behind construction helps the underlying structure to be discovered effectively. Initially, the primary goal of construction is to accurately discover basic concepts and the kind of relations to focus upon and then documenting actual relations among words. Completeness is usually a long-term and ongoing goal as most lexicons cannot claim or even aim to achieve full coverage, especially within a small period of time. Thus, coherence of principles is given primary importance in earlier stages.

Another goal is to define the process of growing the structure. A number of manual and automated methods have been devised in this regard. However, this decision usually involves a tradeoff between quality and speed of collection. Human knowledge collection tends to have a much higher degree of quality as compared to automated methods while being much slower in growing the structure.

A third goal relates to identifying ambiguities arising from issues such as polysemy, discriminating conceptual differences and handling these in the representation. An important goal in constructing lexical databases is defining the storage representation of concepts and relations. For these databases to be popular among NLP applications, they need to be machine-readable. This calls for decisions related to storage of words and concepts (whether textual, binary, symbolic), and relations (symbolic, pointer-based, file linkages).

1.4 Structure

Most databases can be abstracted as directed acyclic graphs. In these DAGs, lexemes or concepts or phrase fragments form the nodes. The edges repre-

sent associations between these nodes. The associations can be of various types. Some of them express syntagmatic relations such as contextual and domain relations, paradigmatic relations such as synonyms and other semantic relations such as cause-effect relations etc. The set of relations documented by the lexical database is to be defined by the designers based on the principles and motivations of the collection. Nodes can also be associated with attributes such as part of speech information, glosses etc.

1.5 Applications

Lexical databases find several uses in NLP tasks. A sample:

1. Word Sense Disambiguation: DBs like WordNet provide detailed sense distinctions for lexemes
2. Machine Translation: Lexicons can be constructed in various languages and linked to aid in MT
3. Gisting and summarising tasks
4. Context generation

1.6 Parameters of Lexical Networks

A study of lexical knowledge structures involves understanding the following aspects:

1. Domains addressed by the structure
2. Principles of construction
3. Methods of construction
4. Representation
5. Quality of database
6. Applications
7. Usability mechanisms for software applications and users: APIs, record structure, User interfaces
8. Size and coverage

The following chapters survey some lexical databases and attempt to understand some of the above aspects in them.

Chapter 2

WordNet

2.1 Origin and Philosophy

WordNet ([5]) is an extremely popular online lexical network which provides rich, semantic information about words. The design philosophy recognised the fact that the same word can have different senses to express different concepts. In WordNet, senses have primacy i.e. a single sense forms a node in the knowledge structure and associations (principally paradigmatic in nature) among them form the labeled edges. Thus the organisation is in terms of word meanings rather than word forms.

WordNet uses the idea of a *lexical matrix*, which can be viewed as a table with word senses as rows and actual words as columns. If an entry in a cell exists, it indicates that the word form can be used to express that word sense. For instance, the word “pen” would have entries for the writing instrument sense as well as the animal enclosure sense in this matrix. Thus if there are multiple entries in the same column, that word form is *polysemous* and if there are multiple entries in the same row, there exist *synonyms* that can express that sense.

The fundamental approach in WordNet is that of “differentiation” which influences how the lexicalised concepts are to be represented. If a constructive approach had been chosen, the representation would need to contain enough information to construct a concept. It was theorised that a differentiated representation was easier to implement. This leads to the idea of the *synset*.

A synset is short for “synonym set” and contains a set of synonyms that serve to identify a singular sense. For instance, “wire” has multiple word senses. By creating a synset containing both “wire” and “telegram”, we can clearly differentiate between this sense and other senses indicated by “wire”. Since languages, especially English, are usually rich in synonymy, synsets are easily constructed.

2.2 Structure

WordNet is organised as a lexical network of synsets and semantic relations between them. Some of the most common and important relations are that of synonymy, heteronymy/hyponymy and meronymy. Additionally, antonymy (a lexical relation) is also included. All words in a synset are considered synonyms of each other. Ideally, all synonyms in a synset should be perfect synonyms of each other, but these can be rare to find. Hence a weakened definition is also used.

A gloss is usually included with each synset that provides a definition for the concept. Semantic relations between senses are indicated by special symbols such as *vehicle* @ \rightarrow *jeep* and *jeep* $\sim\rightarrow$ *vehicle*

WordNet has separate sections for nouns, verbs, adjectives and adverbs.

2.3 Construction

The basic principles of construction of synsets revolve around applying *minimality*, *replaceability* and *coverage*. Minimality implies collection of the minimum number of words in a synset such that the sense can be unambiguously stated. The synset must contain all the important known words (ordered by frequency) for that sense - this is coverage. Finally, the most frequent words must be replaceable.

The following examples illustrate these principles:

Consider the word *back*. This polysemous word has about 9 word senses in the noun form. Suppose, the sense of interest is *the series of vertebrae forming the axis of the skeleton*. A synset is created for this sense solely with this word, giving '{back}'. Because of the polysemy of its only word, this synset does not represent a single unique sense yet and hence violates *minimality*.

Now, if the word *spine* (also polysemous) is added to this synset creating '{back, spine}' and if the two words do not have any other common senses, this minimal synset can sufficiently describe the required sense. The synset now satisfies *minimality*. In case a synset could not be minimally represented by two words, this process would continue until a minimal set of words is discovered to uniquely identify the sense. In the likely event that a polysemous word has a particular sense with no additional disambiguating synonyms, a gloss must be generated to provide distinction. An example of this in WordNet is for the principal sense of the word '{paper}' which is described by a gloss defining it as *a material made of cellulose pulp...*

The principle of *coverage*, on the other hand, pulls in the opposite direction. The aim here is to collect all words, or at least those that are in frequent use, which provide the sense governing the synset. In this example, words such *backbone* or *vertebral column* should appear in any acceptable

synset. Additionally, the words are ordered by frequency of appearance (corpus evidence is used to decide the frequency). For instance, the WordNet synset for this example is; ‘{spinal column, vertebral column, spine, backbone, back, rachis}’. The uncommon word *rachis* clearly is most infrequent among the six. The quest for coverage will be an ongoing one for many words.

Finally, *replaceability* suggests that words in the synset must be capable of replacing each other in (almost) all sentences indicating that sense. Again, corpus evidence helps decide this. A lower possibility of replaceability suggests the existence of more fine-grained senses that the current synset is unable to provide. For instance, consider creation of an incorrect synset ‘{ship, airship, naval ship, spaceship}’ to represent a *vessel carrying people*. Now clearly, *airship* cannot substitute for *naval ship* in most situations, which suggests further distinctions in senses are required.

Synset creation is usually a long and laborious process that is currently being done manually by researchers. In essence, it involves identifying senses from high quality dictionaries, obtaining synonyms for these senses satisfying the basic construction principles, and making entries in synset files as per the WordNet storage structure.

Synsets are grouped into source files based on syntactic and semantic criteria and indices are maintained on these to help enable access using APIs or the User Interface.

2.4 Applications

The motivating application for WordNet was “Word Sense Disambiguation”. Additionally, it can also be applied to tasks such as Lexicon Generation, Question Answering and Text Summarisation.

2.5 Comments

WordNet is considered the leading online lexical resource for use in NLP applications. It has a simple design methodology for construction. Since creation and maintenance is currently carried out by researchers, the process is a gradual one but has yielded high quality in its coverage and accuracy. WordNet provides an easy-to-use set of APIs. The effort in English has been replicated in other languages as well, and efforts such as EuroWordNet have been initiated to link individual WordNets. These are testament to the growing spread of WordNet.

2.6 Illustration

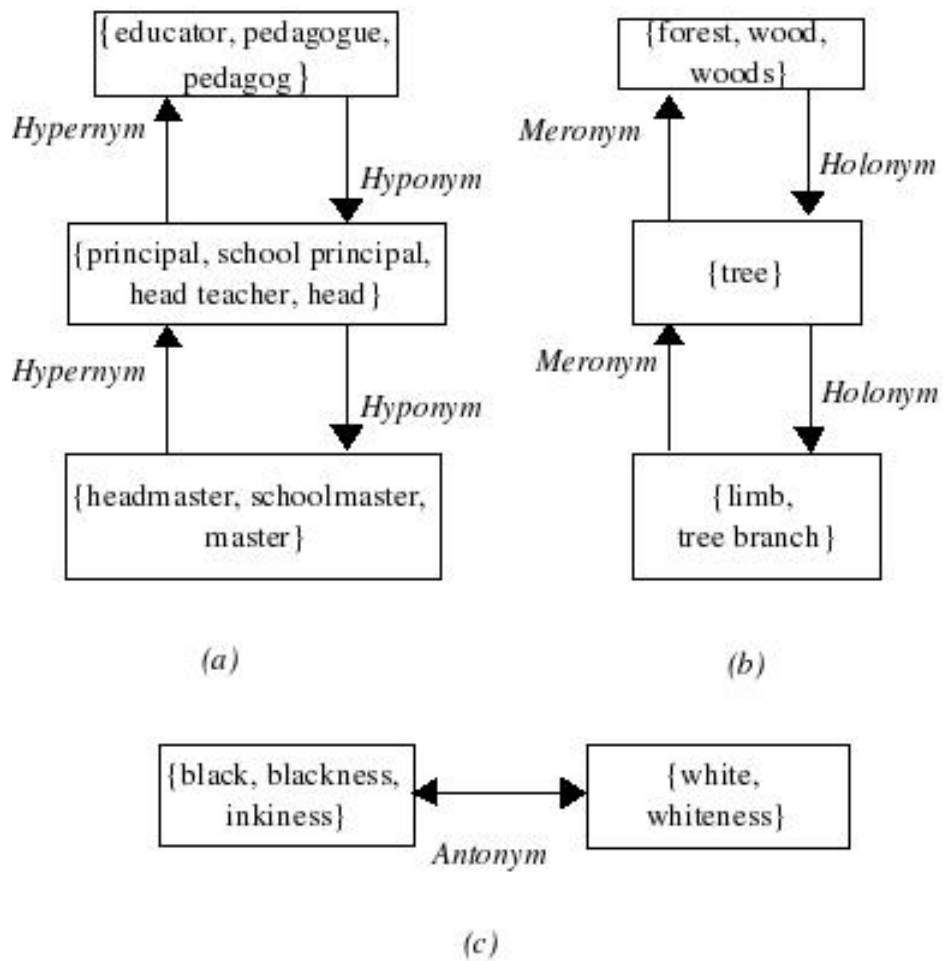


Figure 2.1: Examples of WordNet relations

Chapter 3

ConceptNet

3.1 Origin and Philosophy

ConceptNet ([4]) is a lexical knowledge base from MIT's NLP group, that aims to capture “commonsense” information and relationships among such information. Here “commonsense knowledge” indicates semantic information that enables humans to understand everyday commonplace events. An example: To understand a sentence such as “I borrowed ‘Treasure Island’ for a couple of weeks” requires the following commonsense information:

1. ‘Treasure Island’ is the name of a book
2. People borrow books to read
3. The book was most likely borrowed from a library
4. The book has to be returned to the lender in 14 days time

The people behind ConceptNet take the view that though keyword-based and statistical approaches have achieved some success in assisting tasks such as information retrieval, data mining and NLP systems, these approaches can be shallow in understanding. To further make progress, further and greater amounts of knowledge are required to give software the capacity for more meaningful understanding of textual data.

The commonsense knowledge to be collected has several flavours to it. These could be emotive (“I feel awful” is a negative emotion), functional (“Cups hold liquids”), cause-effect (“Extracting a tooth causes pain”), spatial (“Horses are usually found in stables”) and several more. ConceptNet aims to build a ontological store of commonsense data in which “concepts” rather than an individual lexical entity (as in WordNet) form the nodes in the semantic network. The emphasis here is on everyday concepts rather than rigorous linguistic lexical differentiations.

3.2 OMCS

Most lexical knowledge bases are built painstakingly by hand by hiring researchers and knowledge engineers from the fields of linguistics, psychology and computer science to collect linguistic data for construction. This approach, though usually a good guarantee of quality, has been seen to be poorly scalable in practice. In contrast, ConceptNet has been inspired by the success of distributed and collaborative projects on the Web. A website under a project called **Open Mind Common Sense (OMCS)** ([9]) was launched, where volunteers can contribute simple commonsense assertions. OMCS consists of eliciting data using 30 different activities which help collect simple one-sentence assertions of everyday life, descriptions of ordinary activities, stories about typical situations etc. The data is collected and stored in natural language (English) without translation into a more structured format. As a result, these OMCS sentences are not directly computable by themselves.

However, the sentences collected in OMCS are usually semi-structured because of the manner of collection. For instance, most usable sentences entered through the OMCS interface follow identifiable patterns because volunteers tend to follow examples placed in the "help" for the interface or use templates. As a result, entries like "Treasure Island is a kind of book" (fitting a (*specific item*) isA (*generic item*) pattern) and "Books are found in libraries" (fitting a (*item*) isLocated (*place item*) rule) are collected. If a high percentage of entries do indeed follow such rules, then they can be mined to generate and augment the ConceptNet network.

3.3 Structure

ConceptNet is a directed acyclic graph formed by linking together over 1.5 million assertions into a semantic network of about 300,000 nodes. Each node is a fragment of text corresponding to a "concept". These nodes could thus be noun phrases such as "watermelon" or verb phrases such as "breathe air". There are twenty relation types, which are grouped into various "thematics". These are:

1. K-Lines: ConceptuallyRelatedTo, ThematicKLine, SuperThematicKLine
2. Things: IsA, PropertyOf, PartOf, MadeOf, DefinedAs
3. Agents: CapableOf
4. Events: PrerequisiteEvent, FirstSubEventOf, LastSubEventOf, SubEventOf
5. Spatial: LocationOf

- 6. Causal: EffectOf, DesirousEffectOf
- 7. Functional: UsedFor, CapableOfReceivingAction
- 8. Affective: MotivationOf, DesireOf

The following figure ([4]) provides an illustration of the ConceptNet graph:

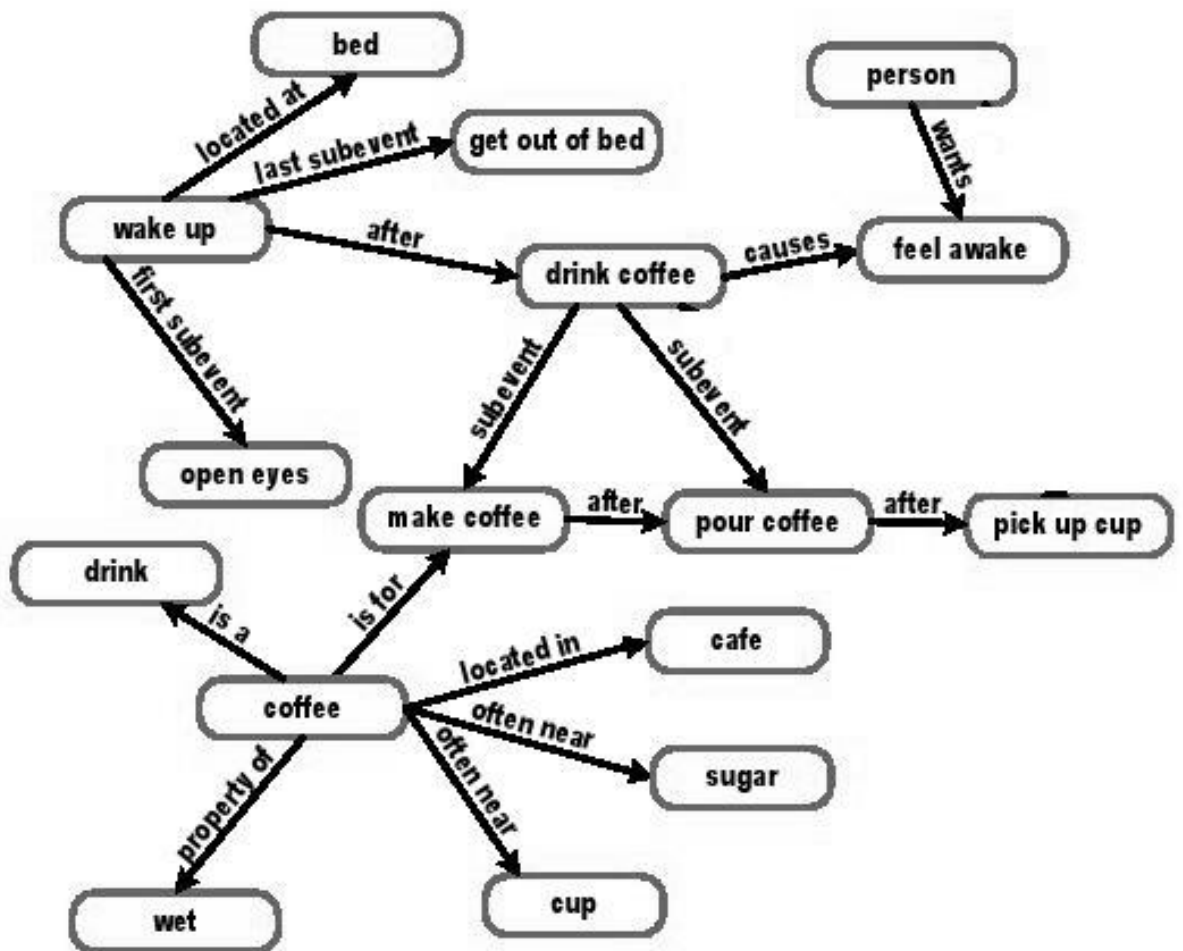


Figure 3.1: Example of ConceptNet nodes and relations

3.4 Construction

It is instructive to study the steps involved in using a sentence to generate ConceptNet entries. These are:

1. Extraction: About fifty regular expression rules are run on the sentences collected in OMCS. This helps extract nodes and relations. The actual storage is in text files. Examples of storage:

(a) (IsA “spider” “bug” “f=3; i=0;”)

(b) (LocationOf “waitress” “in restaurant” “f=2; i=0;”)

The frequency of seeing the assertion is recorded as is the number of times this relation was found through an inferencing procedure.

2. Normalisation: The mined nodes are normalised by correcting spelling, removing function words if unnecessary, lemmatisation etc. These are automated processes.
3. Relaxation: This involves the following sub-tasks:
 - (a) Merge duplicate assertions - update frequency values for these
 - (b) Apply properties from child to parent in a IsA relation for certain relations
 - (c) Do inferencing - i.e. construct newer nodes wherever possible. For instance, if apple has the property of being edible, perhaps a sibling node such as orange can be given the same property

3.5 Applications

ConceptNet comes with an API (variants in different programming languages) and a NLP toolkit called MontyLingua. There are two kinds of API methods: one that helps query the ConceptNet database and one that operates at the semantic level.

A sampling of the DB methods:

- *getAnalogousConcepts*: for a given node, returns a set of nodes such that each of these nodes have a large intersection of incoming links with it
- *getAllProjections*: Given a node, a graph traversal is made on all transitive relations and all resulting nodes (the extent of traversal can be specified)

A sampling of the natural language methods:

- *guessConcept*: given some natural language text, this returns a guess of the key concept(s) involved
- *guessMood*: given text, this guesses an emotion out of {happy, sad, angry, fearful, disgusted, surprised}

Some of the following applications can use ConceptNet:

- Topic generation
- Topic summary
- Analogy making
- Get context for a situation
- Disambiguation and classification

3.6 Comments

The quality of ConceptNet essentially depends on the data collected as part of OMCS, as nodes and relations are mined using automated processes with some clean-up operations. The data is therefore susceptible if OMCS is spammed. Currently however, the authors report that 75% to 80% of the database was judged by a set of human judges as useful.

Inferencing procedures seem to be useful in growing the database apart from the OMCS contributions. The Conceptnet approach is quite a novel way of growing a lexical knowledge structure, but will always be more vulnerable in terms of quality as compared to other hand-crafted lexicons.

The main challenges in ConceptNet would be to add new rules that further capture user input, in adding filtration techniques to improve quality and to grow the database while retaining its flexibility. Currently, using the tools on the DB seem to be a bulky affair in terms of resource usage owing to the plaintext file representation. Additionally, so far, applications using ConceptNet have been generated within the MIT community that is behind this effort. More objective evidence of its use will emerge when researchers outside that community make use of it.

Chapter 4

HowNet

4.1 Origin and Philosophy

HowNet ([2]) is a Chinese research project. It aims to create a semantic network that focuses on both lexical and conceptual entries. As a result, the relationships expressed in HowNet are similar in nature to both WordNet and ConceptNet. HowNet is based on a concept knowledge base and is built manually. The storage is in a well-defined record structure, which is therefore machine usable, an important goal of the project.

The underlying principles in HowNet are:

1. *Composition*: Concepts can be composed to form other concepts, i.e. whole-part relationships abound in the conceptual space. Inter-conceptual relationships express this.
2. *Evolution*: Entities have properties that may not be shared by specialisations of entities. These can be expressed using host-attribute relations.

HowNet takes an ontological view of the objective world and uses the following conceptual domains while building its structure:

1. Thing (sub-divided into physical and mental)
2. Part
3. Attribute
4. Time
5. Space
6. Attribute-value
7. Event

HowNet takes a “constructive” approach to building its semantic network. As will be explained in the following sections, a set of the most fundamental concepts are identified first. Other higher-order entities are then composed by combining these basic concepts and other previously available higher-order entities, and relations appropriately added between concepts. This bottom-up approach can be contrasted with the WordNet model, where words are differentiated in a top-down manner until their different senses have been categorised in keeping with its principles.

4.2 Knowledge Base

The HowNet Knowledge database is a carefully handcrafted database of “sememes”. A sememe is defined as a basic unit of meaning which cannot be further decomposed. (Note however that identifying sememes can be a subjective process.) These sememes fall under the conceptual categories defined in the previous section. Examples of a sememe is “human” or “food”. Concepts such as “doctor”, “teacher” etc. can thought to be a combination of sememes such as “human”. HowNet hypothesises that a closed set of sememes is sufficient to describe an open set of concepts.

HowNet has used the Chinese language character set as a starting point in sememe identification. Most Chinese characters represent a basic concept rather than a meaningless alphabet letter as in other languages. Each such Chinese syllable was examined and a few thousand sememes were identified. After further analysis, currently about 2000 sememes are available.

4.3 Concept Representation

Let the Concept to be represented be “Teacher”. In HowNet, the concept is expressed as a combination of the sememes for “human” (entity), “teach” (event) and “education” (entity). Thus, the HowNet record for “teacher” will have:

- Hypernym: “human”
- Attribute(s): “education”
- “Agent” relation to “teach”

HowNet has a Knowledge Dictionary as the knowledge base of the system. In this Dictionary, every concept of a word or phrase and its description form one entry. Each Concept is represented by the following record structure:

- NO: Concept number - e.g. “023249”

- W-X: Word or phrase - e.g. “teacher”
- G-X: Syntactic class (noun, verb etc.) - e.g. “N”
- E-X: Usage example gloss - e.g. “Teacher teaches in a educational institute”
- DEF: Concept Definition - e.g. “{human:HostOf={Occupation},domain={education}, {Agent = {teach}}, {teacher:agent={~}}”

The Concept Definition field contains those concepts that compose the defined concept (in this case “teacher”), and semantic relations linking these concepts to it. In this example, these entities are human, teach and education. Similarly attributes can also be associated (for instance “ice” and “cold”). HowNet uses symbols such as ‘~’, ‘%’, ‘*’ etc to represent relations. Further examples of relations are “synonym”, “part-whole”, “time-event” etc.

Each HowNet entry has both English and Chinese values represented (above, only the English values have been shown).

4.4 Construction

A basic set of sememes were extracted from characters from the Chinese alphabet. This takes advantage of the fact that these characters represent a concept by themselves. Combining these characters further generates newer concepts. So HowNet records can be used to construct these newer concepts. The HowNet knowledge base is extended by identification of concepts followed by manual creation of corresponding records in the structure outlined above.

4.5 Applications

- Given a record, a gloss can be generated. e.g. using the “doctor” record, a sentence “a human can have an occupation of doctor in the medical domain who performs doctoral activities”
- Word Sense Disambiguation
- Machine Translation (especially between English and Chinese)
- Analogical Reasoning

4.6 Comments

The principle issues in HowNet emanate from the hand-crafted nature of the database. These involve the following:

- How to identify newer concepts efficiently?
- How are relations and entities maintained and extended?
- Can it be extended to other languages since it seems to be based on the Chinese alphabet system?

Quality is comparatively protected due to the fact that well-qualified researchers are currently in charge of HowNet’s maintenance.

In terms of usability, an important goal of HowNet is make the lexical knowledge base available for use in NLP applications, which explains the highly structured form of the database. However for popular use, a Mini HowNet is available for download which only provides lexicon from A-D. A Dictionary, Taxonomy and Concept Similarity Tool is included as part of Mini HowNet.

HowNet’s principal attraction is its “constructive” approach to building newer concepts, which could translate into a fairly scalable and evolutionary approach.

Chapter 5

FrameNet

5.1 Origin and Philosophy

FrameNet ([1]) is a lexical language resource developed at Berkeley. It has as its basis the idea of “frame semantics”. The process of constructing this lexical knowledge structure involves the production of frames for different English word senses. The structure provides information about semantic and syntactic generalisations along with corpus evidence for these frames. Frame semantics help record the different sentence variations the a word sense can be involved in.

The project produces these frame-semantic descriptions for several thousand lexical items along with semantically annotated endorsements from English corpora. These descriptions are generated from hand-tagged semantic annotations of example sentences extracted from text corpora by lexicographers and linguists. The primary aim of FrameNet is to encode semantic knowledge in machine-readable form and is essentially done by using a manual process aided by some NLP tools

5.2 Frames

Frames are well documented in the NLP literature. [7] define each frame as a “collection of attributes (“slots”), associated values and constraints”. Theoretically, such a definition implies that any concept is represented by all the attributes and constraints that uniquely define a class of entities. An example of this would be the concept “teddy bear” which could potentially be defined as a combination of entities representative of “toy”, “animal-like” etc and by attributes such as “cuddly” etc. This would depend on whether the requisite entities and attributes are available.

In FrameNet, a frame is a “conceptual structure that describes a particular type of situation, object or event, and participants in it”. However, the architects of FrameNet have chosen to focus to mainly represent situations

instead of entities, which implies that verb-sense oriented phrases receive primary attention. As an example, a frame can represent a concept such as “The cook baked a cake” or “The person made a telephone call”.

A Frame is made of “frame elements” which describe the sub-parts of that frame. A frame merely represents the base concept independent of specialised variations. For e.g. the APPLYHEAT frame is the concept behind “bake” with frame elements:

- COOK: ‘cook’
- FOOD: “cake”

In FrameNet, Lexical Units are words such as ‘cook’, ‘bake’, ‘cake’ that help evoke a concept, and are represented by frames such as ApplyHeat, CookingCreation, Food etc. Thus, the evocations of a frame represent these more familiar actions and nouns.

Theoretically, the ultimate goal in FN should be a frame per word sense, but current work continues to be about situational frames i.e. frames about verbs.

5.3 Structure and Construction

FrameNet’s database consists of three components:

1. *Lexicon*: This consists of an entry for each lexical unit (not frame) in FrameNet. This entry is made up of:
 - (a) A dictionary-style definition
 - (b) Formulae for syntactic realisations. For instance, this would help generate variations on “a person bakes a cake” such as “a person bakes a cake at 100 degrees” or “a person bakes a cake in an oven” or even “a person bakes a cake in an oven at 100 degrees” and so on. The entry for “bake” would point to formulae that involve it.
 - (c) Links to applicable frames (in the Frame Database) involving this lexical unit.
 - (d) Links to example sentences in the corpus for the different realisations.
2. *Frame Database*: This contains the different frames. Each frame consists of the name and description of its frame elements (core and non-core) and lexical units evoking frame and attributes.
3. *Annotated Example Sentences* : Annotated example sentences: marked with syntactic and semantic information. e.g.: Bake [(food)the potatoes] in a [(container) medium-sized pan]

In addition, various relations among the frames are also recorded. Example of relations are inheritance and uses.

Four processing steps are required produce the FrameNet database of frame semantic representations:

1. *Preparation*: generating initial descriptions of semantic and syntactic patterns for use in corpus queries and annotation
2. *Corpus Extraction*: extracting good example sentences
3. *Annotation*: marking (by hand) the constituents of interest
4. *Entry Writing*: building a database of lexical semantic representations based on the annotations and other data

As is quite evident, this is likely to be a very slow and laborious process.

5.4 Comments

FrameNet is currently more verb-sense-oriented with nouns usually being dependent objects in “situations”. This makes the lexical structure more suitable for describing events and occurrences. Use of frames and formulae of realisations help in generative tasks such as sentence generation.

Addition of new frames to FrameNet depends significantly on linguists involved and is inherently an intense manual process. As is common with such a style of maintenance, quality is relatively assured but the pace of construction is likely to be quite gradual.

The FrameNet Database comes with a rich set of annotations along with illustrations of combinatorial possibilities of lexical units, marking it from other knowledge structures studied in this report. However, FrameNet does not seem to be widely in use outside the parent university.

Chapter 6

MindNet

MindNet ([8]) is an initiative in the field of lexical structures from the Microsoft Research NLP group. A MindNet is a collection of semantic relations automatically extracted from text data using a broad coverage parser. This broad coverage parser is the same as present in Microsoft Office applications. The parser is applied on data present in Machine Readable Dictionaries which principally consist of words and definitions. The extraction is by a fully-automated process, though the group behind MindNet has not ruled out the need for inspection of information to ensure accuracy and quality.

MindNet has 24 semantic relations, examples of which are Hypernym, Location, Size, Part, Time etc. Construction of the network involves the collection of “semrels” (short for ‘semantic relations’) from sentences. The semrels are highly structured. The automatic extraction process extracts these from a definition or example sentence and produces a hierarchical structure of these relations, representing the entire definition or sentence from which they came. Such structures are stored in their entirety in MindNet and provide crucial context for some of the procedures described in later sections of this paper.

As an example, consider the definition of “Car” as “a vehicle with 3 or usually 4 wheels and driven by a motor and used for carrying people”. From this, the following semrel structure can be extracted:

Car:
 Hypernym: vehicle
 Part: wheel
 TObj:
 drive
 Means: motor
 Purpose:
 Carry
 TObj: people

In addition to this extraction, other inferencing methods are also run on this structure to extract potentially useful assertions. The most prominent is “inversion”. An example of this using the “car” semrel structure would be extracting a structure where the entry for “drive” is linked to “car”. In this way, the network is enhanced by synthetically adding relevant associations between every relevant word that appears in these semrels. MindNet construction also contains a method to weight the paths between semrel entries.

The MindNet project has also done work on similarity methods over semrels that can be used in disambiguation tasks that also help improve MindNet itself. This includes looking for similarity in words on both paradigmatic as well as syntagmatic levels.

In summary, MindNet is a lexical resource whose key distinctions are a completely automated process of collecting semantic relations mainly from machine readable dictionaries and some novel ideas in inferencing over the collected data to augment the lexical knowledge structure.

Chapter 7

Comparison of Lexical Knowledge Structures

7.1 Conceptual Basis

WordNet principally addresses paradigmatic relations of lexemes rather than syntagmatic or semantic relations like “IsLocatedNear”. In contrast, ConceptNet relations tend to be more syntagmatic apart from the common relations such as hypernymy. Another difference among the two is that ConceptNet has concepts for nodes represented by text fragments, while WordNet strictly has words in nodes.

HowNet attempts to capture the best of both approaches, by being a semantic network of concepts with both kinds of relations and by representing its concepts as being built out of previous concepts. MindNet shares its relation space largely with HowNet. FrameNet uses the more classical concept of frames as nodes in its network, which are harder to represent and identify. These are about conceptual situations rather than noun entities.

7.2 Principles of construction

These principles characterise these networks. WordNet is about differentiating word senses while HowNet takes an opposite, bottom-up view by combining concepts and sememes to generate new concepts. FrameNet depends on identification of frames, while MindNet generates and merges words based on parsing text, mainly dictionary like definitions. ConceptNet works on semi-structured descriptions of everyday life without any strictly particular top-down or bottom-up process.

7.3 Methods of construction

Construction of these networks is an expensive process with this decision strongly affecting the tradeoff between quality of the network and the speed at which the collection is constructed and new entries added.

WordNet, HowNet and FrameNet principally use manual methods to identify and maintain their lexical networks. Some tools are used, but these are related to tidying operations. On the other hand, ConceptNet and MindNet are almost completely automated processes that use regular expressions or parsers to identify nodes and the nature of associations among them.

7.4 Representation

A principal goal of each network is to produce an effective machine representation for later use by NLP tools. In WordNet, synsets are grouped together semantically and put into different source files. The system designers also have built indices on syntactic values and on frequency of use of words. ConceptNet lumps all values into files broadly classified on the basis of thematics. HowNet has a well-defined record structure representing concepts as illustrated in an earlier section.

Relations are either denoted by using special symbols or by actually writing down the name of the relation or an abbreviation of it. The former scheme is obviously more economical in storage and may be the ideal way given that a more human-readable form is not necessary.

7.5 Quality of Database

By virtue of being handcrafted by a team of lexicographers and linguists, the WordNet and FrameNet databases are thought to have high quality. Most of the other networks have had quality studies done internally. In case of the networks with automated collection, the studies have revolved around the coherence and accuracy of the collected data. These studies show that though the collection quality is fair, this is mainly because the environment is still very sanitised. The prospects for spurious entries in ConceptNet could be a major concern.

7.6 Spread

WordNet is clearly the most well-known and widely used of these lexical knowledge structures, forming part of many NLP and IR tasks, particularly in sense disambiguation tasks that it is ideally suited for. In some sense, it is the herald of the idea and power of these networks. HowNet has found use in certain research efforts. The rest still remain localised to their parent

research groups and are yet to make an impact outside. Evidence of quality and ease of usage will only emerge when there are reports of more objective and intense applications of these lexical networks.

Chapter 8

SUMO

An Ontology is a structure that describes the entities, attributes and associations among them for a particular conceptual domain. Efforts to create ontologies that express domains such as manufacturing, processes, wildlife etc. are currently in progress. Lexical knowledge structures are similar to ontologies as they describe a conceptual or lexical space. An “Upper Ontology” is an ontology that describes general concepts that are domain-independent. For instance, an upper ontology would have high level entities such as “Animal” or “Country” or “GovernmentalOrganisation” while a domain-specific (lower) ontology would have “Leopard” or “India” or “Planning Commission”.

Several upper ontologies have been proposed, but in the interests of standardisation, a IEEE working group to evolve a standard “upper ontology” has been set up. Its aim to come up with a “Suggested Upper Ontology”. As a starting point, SUMO ([6]), which stands for “Suggested Upper Merged Ontology”, has been proposed. This has merged some of the important aspects of existing upper ontologies using both syntactic and semantic themes.

The reason for designing a standard version of an upper ontology are:

1. It would help in design of new knowledge bases and databases
2. Interoperability with other compliant systems
3. Reuse/integration with legacy DBs provided a one-time mapping can be done to a standard ontology
4. Interoperability of domain ontologies

SUMO has “Entity” as its root node, followed by a Hypernymy/Hyponymy tree structure. Axioms and Assertions express constraints (e.g. “An instance of the entity “Collections” must be non-empty”). Figure 8.1 shows some of the highest entities.

As an illustration of mapping SUMO with existing ontologies, a SUMO-WordNet association has been created. The idea was to tag each WordNet

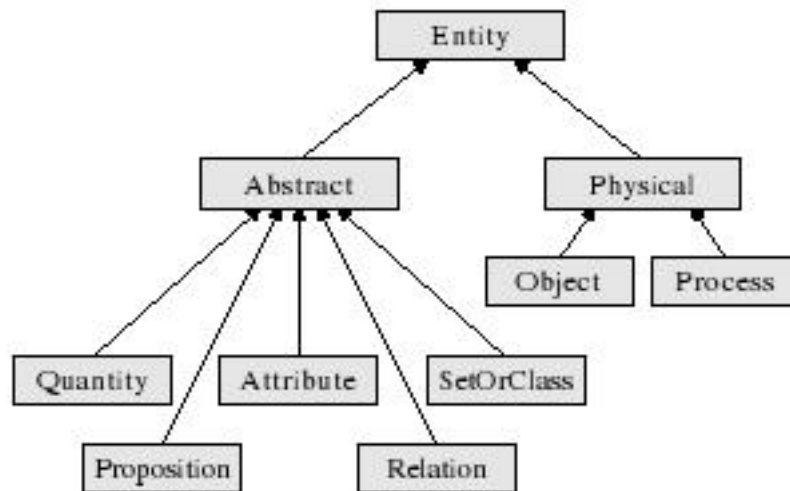


Figure 8.1: High level SUMO entities

synset with the corresponding SUMO concept/concepts. This method is currently available only for noun synsets. It consists of the following:

1. Look at each WordNet synset
2. Identify a SUMO concept that subsumes/is equivalent/is closely related to the synset
3. Add the concept and a symbol to the synset

Example of equivalent concepts: the synset for '{plant, flora, plant life}' tagged with '&%Plant=' Example of the SUMO concept being broader: synset for '{Panther}' tagged with '&%Animal+' Example of the synset being an instance of a class of SUMO concept: synset for '{United Nations}' tagged with '&%Organisation@'

The principle issues in SUMO are that it is hand-crafted and hence the pace of creation is slow. There have been conflicts on which entities are domain-independent and to be included, and on expression. SUMO is represented in a structured form called KIF (Knowledge Interchange Format - a set of prefix predicate rules) to make it useful for applications.

Chapter 9

Conclusion

This report introduced some lexical knowledge structures, which contained various interesting aspects in terms of content, construction and utility. A comparison among these lexical networks was also presented. Finally, a recent initiative in the allied field of upper ontologies was also presented.

These lexical structures are clearly important resources for various applications, but the field is still maturing. Issues involving scalable methods of collection versus quality are still being debated. Various approaches have been proposed with competing advantages. In the coming days, with trends like the Semantic Web and with service providers of search and web content aiming to provide a more pleasurable user experience with regards to divining information need, a study of these lexical knowledge structures is warranted.

Bibliography

- [1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL, 1998*.
- [2] Zhendong Dong and Qiang Dong. An Introduction to HowNet. Available from <http://www.keenage.com>.
- [3] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.
- [4] H. Liu and P. Singh. Commonsense Reasoning in and over Natural Language. In *Proceedings of the 8th International Conference on Knowledge-Based Intelligent Information and Engineering Systems, 2004*.
- [5] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography* 3 (4), pages 235–244, 1990.
- [6] A. Pease, I. Niles, and J. Li. The Suggested Upper Merged Ontology: A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web, 2002*.
- [7] Elaine Rich and Kevin Knight. *Artificial Intelligence (2nd Edition)*. Prentice Hall, 1999.
- [8] S.D. Richardson, W.B. Dolan, and L. Vanderwende. MindNet: acquiring and structuring semantic information from text. *ACL'98: 36th Annual meeting of the Association for Computational Linguistics*, 2:1098–1102, 1998.
- [9] P. Singh, T. Lin, E.T. Mueller, G. Lim, T. Perkins, and W.L. Zhu. Open Mind Common Sense: Knowledge Acquisition from the general public. In *Proceedings of the First International Conference on Ontologies, Databases, and Applications of Semantics for Large Scale Information Systems, 2004*.