

Towards Evaluating Lexico-Semantic Networks

M.Tech Project - Stage One Report

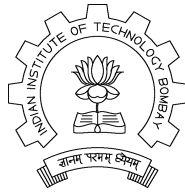
Submitted in partial fulfillment of the requirements
for the degree of

Master of Technology
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

Many thanks to **Prof. Pushpak Bhattacharyya** for all his kind suggestions and for steering me to the right corners of the content universe.

A word of gratitude to **Tim Berners-Lee** (it hardly matters that he does not know who I am) for inventing the *World Wide Web* without which very little of this would happen without difficulty.

J. Ramanand
M.Tech, First Year
KReSIT
I.I.T. Bombay
July 17, 2006

Abstract

This report builds approaches towards evaluating *lexico-semantic networks* by studying evaluation strategies applied to ontologies. It shows the lack of such methods for networks such as *WordNet*, and so builds a case for such evaluations. A brief introduction to *lexico-semantic networks*, a mention of the principles of evaluation and the successes in Machine Translation evaluation are also included in this report.

Contents

1	Introduction	1
1.1	Motivation	1
1.2	Outline	2
2	Introduction To Lexico-Semantic Networks	3
2.1	Motivation	3
2.2	Definition	4
2.3	Purposes and Goals	4
2.4	Structure	4
2.5	Applications	5
2.6	Lexico-Semantic Networks	5
2.6.1	WordNet	5
2.6.2	ConceptNet	6
2.6.3	HowNet	9
2.6.4	FrameNet	10
2.6.5	MindNet	11
2.7	Parameters of Lexical Networks	12
3	Evaluation in the NLP World	13
3.1	Motivation	13
3.2	Strategies and Metrics in Evaluation	14
4	Evaluating Ontologies	16
4.1	Introduction and Motivation	16
4.2	Strategies	17
4.3	A Proposed Formal Model For Evaluation	19
4.4	Metrics	20
4.4.1	Information Retrieval Metrics	20
4.4.2	Learning Accuracy	21
4.4.3	Lexical Comparison Level measure	22
4.4.4	Semiotic metrics	22
4.5	Examples of Evaluation	22
4.5.1	OntoMetric	23

4.5.2	OntoClean	23
4.5.3	An Example of Task Based Evaluation	24
4.5.4	OntoLearn	25
4.6	Issues	25
4.7	Conclusion	25
5	Evaluation Aspects of Semantic Networks	27
5.1	Introduction	27
5.2	WordNet: Statistics and Evaluations	27
5.3	ConceptNet: Statistics and Evaluations	30
6	Evaluations in the Machine Translation Field	32
6.1	Introduction and Motivation	32
6.2	Approaches and Challenges	32
6.3	String Matching Techniques	33
6.4	IR Techniques on N-grams	33
6.5	Criticisms	35
7	Evaluating Lexico-Semantic Networks	36
7.1	Conclusions from Literature Survey	36
7.2	Future Directions	37

Chapter 1

Introduction

1.1 Motivation

Lexico-semantic networks such as *WordNet* have burst into prominence because applications, especially those targeted for the Web, now aim to enhance the semantic dimensions of their performance. An example of such an application is in *Information Retrieval* where a lexical resource can help provide easy query keyword disambiguation and improve the quality of search results retrieved. This is especially due to the fact that the quantity of documents now available via the Web is extremely large, resulting in the need for further sophistication.

Alternatively, consider automatic generation of content for certain contexts such as tourist phrasebooks, or automatic sensing of emotion from text. Lexical resources that can potentially reveal, generate or help infer such content are being developed by various research groups. These are no longer simple dictionaries; rather they are rich in “semantic content” going far beyond the scope of mere lexicons.

Lexico-semantic networks can also be viewed as a reservoir of common-sense concepts arranged ontologically, hence describing the real-world through lexical knowledge. The bottomline is that such resources are being increasingly co-opted in applications involving language technology, and not just in English. Almost every major language now has a *WordNet* project, and efforts such as *ConceptNet* attempt to include those aspects not covered by WordNet.

The increasing production of such networks and their application in diverse areas call for evaluation methods to describe the quality of rival networks as well to set expectations about their likely performance in applications. This covers a gamut of criteria, which unfortunately have not been studied in detail. This report sets the stage for an investigation into evaluation strategies for lexico-semantic networks.

1.2 Outline

This first stage report contains a literature survey of lexical networks and evaluation approaches in order to build insights towards evaluating these kinds of networks. The report begins by briefly introducing the reader to lexico-semantic networks. The focus then moves to evaluations, by describing the general concepts involved in evaluating NLP entities. This is followed by a detailed look at the theories, metrics and practical methods used to evaluate ontologies at present. There have been no major evaluations of lexico-semantic networks, but there have been a couple of in-house evaluations by network constructors. As a sample, the report then discusses evaluation efforts aimed at some regional *WordNets* and the *ConceptNet* evaluation by its creators.

This report also summarises the key highlights in the domain of Machine Translation. Though not directly useful towards measuring the quality of lexical networks, the progress made in that field can have important lessons for evaluations in other NLP fields. The report looks at the likely directions to be pursued and the challenges likely to be faced while attempting to evaluate lexico-semantic networks in a more sophisticated manner. This aims at motivating the work for the next stage in this project.

Chapter 2

Introduction To Lexico-Semantic Networks

2.1 Motivation

Fields such as natural language processing and information retrieval often seek lexicons that can provide information on words and concepts 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), 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.

2.2 Definition

A lexico-semantic network can be defined as *a systematic collection of words along with labeled relations between them, usually in machine-readable form*. Such a structure may have different relations among the words, 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* ([13]). It is based on the perception that a word is an association between a lexicalised concept and an utterance that plays a syntactic role ([17]).

2.3 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 document 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 the 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 trade-off 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).

2.4 Structure

Most databases can be abstracted as directed acyclic graphs. In these DAGs, lexemes or concepts or phrase fragments form the nodes. The edges represent 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 *synonymy* and other semantic relations such *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.*

2.5 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*.

2.6 Lexico-Semantic Networks

The following sections briefly summarise some popular lexical networks.

2.6.1 WordNet

WordNet ([17]) is perhaps the most popular lexical network available. The design philosophy is based on the facts that a concept or a “sense” can be represented by more than one word (*synonymy*) and that the same word can have different senses to express different concepts (*polysemy*). Senses are given primacy, and each node in the network stands for a particular sense. The nodes are each represented by a unique *synset*.

A synset is short for “synonym set” and contains a set of synonyms that serves 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.

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.

The basic principles of construction of synsets are *minimality*, *replaceability* and *coverage*. Minimality implies the 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, there must be corpus evidence that the most frequent words are able to replace each other.

A simple illustration of *WordNet* is shown in figure 2.1.

2.6.2 ConceptNet

ConceptNet ([15]) 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, 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. The emphasis here is on everyday concepts rather than rigorous linguistic lexical differentiations.

Structurally, *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

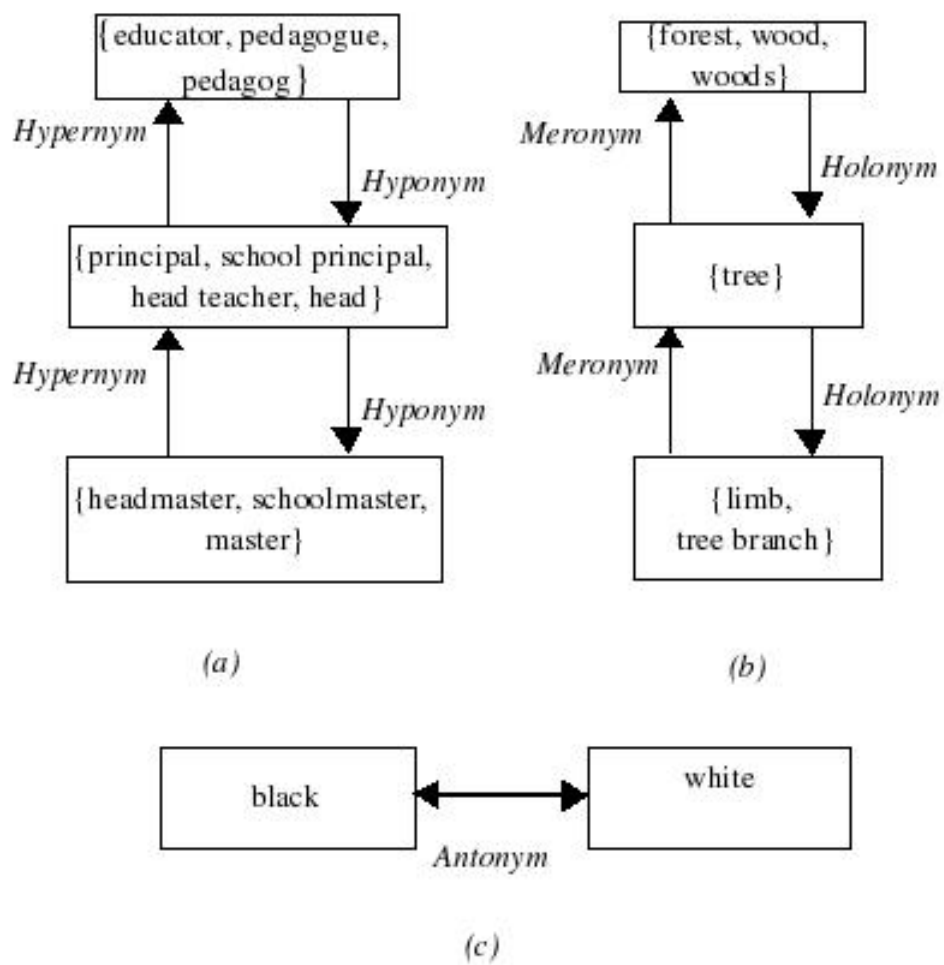


Figure 2.1: Examples of WordNet relations

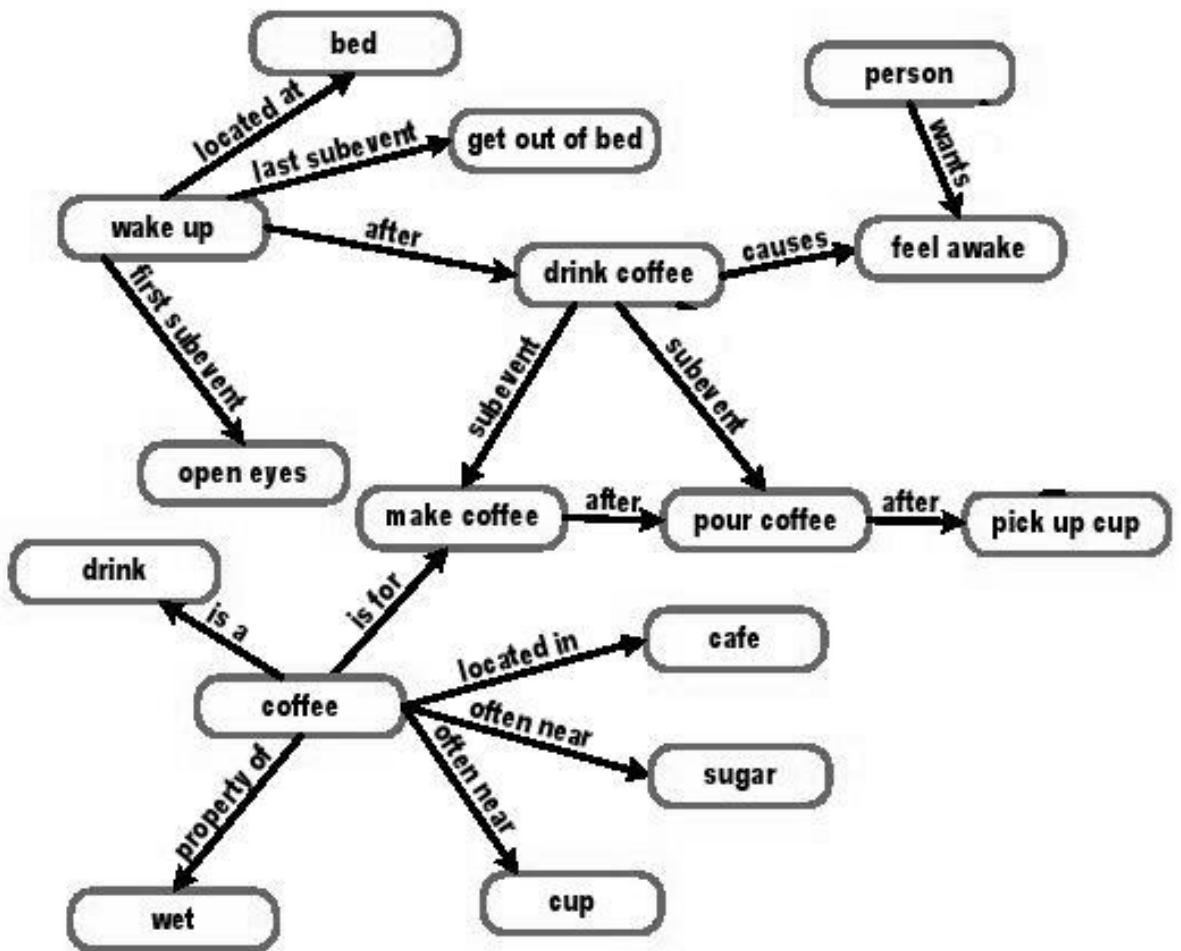


Figure 2.2: Example of ConceptNet nodes and relations

as *breathe air*. There are twenty relation types, some of which are *IsA*, *CapableOf*, *EffectOf*, and *LocationOf*. Due to its automatic construction method, the relations collected are largely syntagmatic in nature.

An example illustration of *ConceptNet* is shown in figure 2.2.

2.6.3 HowNet

HowNet([5]) 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. 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.

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 that identifying sememes can be a subjective process). *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. Around a few thousand sememes were identified by these means.

A larger concept is represented by combining these basic sememes by means of relations. An example is as follows:

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

- Hypernym: *human*
- Attribute(s): *education*
- A relation to *teach* is established; this is of the type *Agent*

Apart from its constituents, each concept is given a unique identifier, example gloss and syntactic class.

2.6.4 FrameNet

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 a word sense can be involved in.

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 on mainly representing 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* that 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.

2.6.5 MindNet

MindNet ([20]) 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 that 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 promi-

ment 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.

2.7 Parameters of Lexical Networks

This section looked at some lexico-semantic networks that contained words along with various kinds of semantic information. Currently, there are no measures of quality to evaluate or differentiate among these. A study of lexical networks could involve understanding the following aspects:

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

Chapter 3

Evaluation in the NLP World

3.1 Motivation

Evaluations of NLP techniques and resources are studied so as to help answer questions such as the following:

- How to select one method or resource over another?
- Is this technique or resource sound, useful, and accurate?
- Is this technique or resource usable, scalable, and deployable?
- Is this NLP entity suitable for a particular domain or application?
- How does this entity compare with another?
- Does the entity meet quality attributes set by different stakeholders, and does it meet given specifications?

Evaluations have been applied to different NLP tasks such as parsing, machine translation, ontologies, spell checking, entity extraction *etc.* Evaluations usually consist of defining metrics with a theoretical basis, defining measurement strategies and then finding values for the metrics from the candidate entities. Evaluation can be subjective in some cases, but the ideal goal is to come up with objective criteria to measure these. However, the interpretations occasionally remain subjective.

Some of the key issues in NLP evaluations are:

- Scale of operations (such as various different possibilities of translation), size of domain
- Lack of complete consensus on quality attributes
- Lack of agreement on intended functionality

- Instability of domain
- Complexity and variety of natural language
- Lack of standardisation

3.2 Strategies and Metrics in Evaluation

Strategies of evaluation have usually been one of (or variants of) the following:

- Human evaluation
- Comparison to a “gold standard”
- Evaluation in the context of an application
- Use of a corpus in evaluation
- Semi-automated evaluation

Given the size, scale and scope of NLP entities usually involved, the ideal is to move to a completely automated method of evaluation, but this is a difficult goal to achieve in its entirety.

[14] mentions the relevant ISO standards in the world of software evaluation, in which the metrics are coarsely classified into ‘internal’, ‘external’ and ‘quality of use’. Equivalents can be found for these specifically in the NLP space. Internal metrics apply to static properties of software, *i.e.* software considered independently of its execution. These can be likened to intrinsic properties of an entity, independent of use, such as the quality of content held in a resource. External metrics apply to software when it is being executed, to the behaviour of the system as seen from outside. Thus they may measure the accuracy of the results, the response time of the software, the learnability of the user interface and a host of other attributes that go to make up the quality of the software as a piece of software. Quality-in-use metrics are appropriate when the software is being used to accomplish a particular task in a particular environment. They are more concerned with the effects of using the software than with the software itself. Quality-in-use metrics are therefore very dependent on a particular environment and a particular task.

[14] also discusses the desirable properties of a good metric. These are:

1. A metric should reach its highest value for perfect quality (with respect to the attribute being measured), and, reciprocally, only reach its highest level when quality is perfect.
2. A metric should reach its lowest level only for the worst possible quality (again, with respect to the attribute being tested).

3. A metric should be *monotonic*: that is, if the quality of software A is higher than that of software B, then the score of A should be higher than the score of B.
4. A metric must be clear and intuitive.
5. It must correlate well with human judgements under all conditions.
6. It must measure what it is supposed to measure.
7. It must be reliable, exhibiting as little variance as possible across evaluators or for equivalent inputs.
8. It must be cheap to prepare and to apply.
9. It should be automated if possible.

There are competitions that provide a standard data set for testing evaluation methods on; there are standards emerging to provide a fair basis for comparing evaluation methods; in some fields like Machine Translation, certain metrics such as *BLEU* have emerged as the leading measure which to emulate or improve.

Chapter 4

Evaluating Ontologies

4.1 Introduction and Motivation

An Ontology is *an explicit formal conceptualization of some domain of interest* ([2]). In simpler terms, it is a structure that describes entities, classes of entities, attributes and associations among them for a particular conceptual domain. Creating a machine-readable representation of a particular domain provides very useful information in tasks such as meaning-rich information retrieval, information extraction (especially in focused domains), knowledge management and in particular, the idea of the Semantic Web, where information is annotated with appropriate tags. Ontologies are important resources in the world of human language tasks used by higher level applications. Concept taxonomies tend to form a significant component of ontologies.

As an example, consider a simple ontology for a typical computer science lab. This would have terms such as *student, desktop PC, laptop, server, printer, network switch, chairs etc.* There could be relations between *desktop PC* and *student* as *assigned-to*, between *lab-administrator* and the *lab* as *serviced-by etc.*

Several efforts to create ontologies that express domains such as manufacturing, processes, wildlife, tourism, politics *etc.* are currently underway. Allied ontology technologies such as decision engines, annotation tools, ontology-based search engines and mining tools are also being built on top of ontologies. The increasing number of ontologies, especially in the context of the Semantic Web, has raised the question of “quality”, especially during selection of comparable ontologies for functional tasks. This has provoked discussion on the key aspects and requirements of ontologies.

Ontologies belong to the class of data models and are unlike software code or processes. Hence, the quality requirements tend to be tenuously defined and specific to the needs of the stakeholders ([11]). Probably the most important requirement of ontologies in the current scenario is that of

interoperability i.e. ontologies should be exported and imported with ease among applications and should be fairly compatible with each other. Being data models, ontologies can be evaluated for intrinsic worth (*is the content accurate and relevant?*) and on a comparative basis (*is this ontology better than that one?*). These issues are discussed in succeeding sections.

The principal motivation of studying evaluation methods in the ontology space is the fact that ontologies can be considered a form of semantic networks, as they are organised representations of concepts and relations. Lexical knowledge structures are similar to ontologies in that they describe a certain conceptual or lexical space, and are also used in similar types of applications.

Secondly, owing to the great enthusiasm for ontologies in industrial applications, several efforts to evaluate them have been proposed and studied. These two reasons provide adequate interest for understanding how evaluations have been conducted in this knowledge space.

4.2 Strategies

At the highest level, ontologies can be evaluated for content and methodology. Content evaluation helps avoid use of incorrect or unsuitable information, while methodology of construction, usability and maintenance has a bearing on deployment and operability.

[2] provide a classification of ontology content evaluation strategies as follows:

1. Compare the ontology to a “gold standard” (usually itself an ontology).
2. Task based evaluation *i.e.* use the ontology in an application and evaluate the results of performance of the application.
3. Compare the contents of the ontology to a source of data about the domain covered.
4. Human evaluation to ratify the ontology on a variety of standards pertaining to requirement fulfillment, domain coverage and usability.

[2] also provides an additional perspective to evaluation based on the “level” of appraisal. The belief behind this is that since the typical ontology is complex, it is often more practical to focus on the evaluation at different levels of the ontology separately rather than trying to directly evaluate the ontology as a whole. The following levels are proposed:

1. *Lexical, vocabulary, or data layer*: Evaluation at this level judges whether the necessary concepts and data items have been included in the ontology. String matching measures are traditionally used in such tasks, comparing the elements to a known set of strings expected.

2. *Hierarchy or taxonomy*: Elements involved in “is-a” relationships form a significant part of ontologies and specific evaluation of this relation may be in order. *OntoClean* is an example of a method that focuses primarily on this relation.
3. *Other semantic relations*: precision/recall measures, philosophical notions like *identity*, *rigidity*, and *unity* have been proposed.
4. *Context*: If an ontology is part of a larger system, or is referenced by others, this context may be important to consider in the assessment. “Popularity” or “Importance” of references to the ontology may be one measure of gauging the worth of an ontology at this level. Task-based evaluation schemes measure the utility of the ontology via the functionality of an application that provides context for the ontology.
5. *Syntactic level*: A check whether the representation of the ontology is valid in the chosen language or notation.
6. *Structure, architecture, design*: Sometimes ontologies are constructed (typically by hand) to satisfy certain pre-defined standards and design needs. The evaluation at this level verifies adherence to these standards and is usually a manual effort.

The following table from [2] summarises the strategies chosen for these levels:

Level	Golden Standard	Application-based	Data-driven	Assessment by humans
Lexical, vocabulary, concept, data	x	x	x	x
Hierarchy, taxonomy	x	x	x	x
Other semantic relations	x	x	x	x
Context, application		x		x
Syntactic	x			x
Structure, architecture, design				x

Some additional aspects to ontology evaluation are:

1. Evaluation of ontologies during the life-cycle of creation and maintenance.
2. Ontology tools such as RDF Schema, DAML+OIL, and OWL checkers, validators, ontology importers and parsers are also to be evaluated.
3. Scalability, navigability, usability evaluations.

4. A weighted and hybrid approach of many methods listed above.

4.3 A Proposed Formal Model For Evaluation

[11] describe an attempt at creating a formal model of an ontology with respect to specifying a given vocabulary's intended meaning. The most important observation here is that ontologies are only approximate specifications of conceptualizations. Therefore, it is fair to evaluate them on the basis of the degree of such approximation. This does not translate easily to a practical implementation because the relationship between an ontology and a conceptualization is rather delicate and requires some technical clarification. The model is shown in figure 4.1.

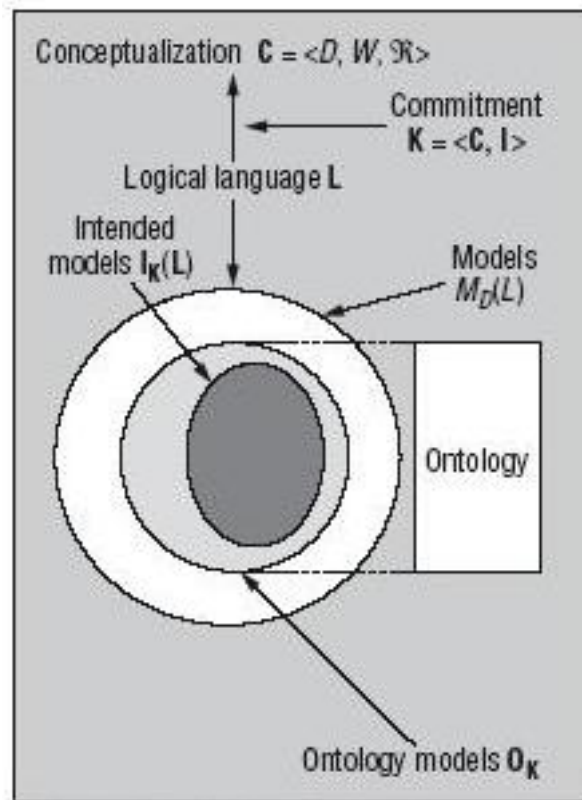


Figure 4.1: Relationship between an ontology and conceptualisation. Here, given a logical language L that commits to a conceptualisation C , an ontology's purpose is to capture those models of L that are compatible with C ([11])

Here, a desired conceptualisation C (made up of a set of entities, relations and arrangements of these items) is translated using a representation L to

a set of possible models O_k . The intended set of models that make the best sense are denoted by I_k , and may intersect or overlap with O_k in a variety of ways.

If this model is likened to an IR scenario where the ontology is seen as a method to retrieve a conceptualisation, Coverage (or recall) can be defined as:

$$C = | I_K \cap O_K | / | I_K |$$

while Precision can be defined as:

$$P = | I_K \cap O_K | / | O_K |$$

In case of infinite domains, it is impossible to obtain coverage numbers, so instead, methods using a list of examples and counter-examples are proposed whereby the evaluation task can be practically carried out by teams containing domain experts.

4.4 Metrics

Metrics defined for ontology evaluation have been applied to gold standard and corpora-driven comparisons. Here, concepts embedded in an ontology have been measured for coverage and accuracy in methods based on IE/IR precepts. This is especially so because competitions like TREC have driven evaluation efforts in these academic areas. Many of these metrics are therefore derived from the traditional metrics in these fields. Some of these from ([16]) are:

4.4.1 Information Retrieval Metrics

1. The well-known *Precision* and *Recall* measures.
2. *F-Measure* is often used in conjunction with *Precision* and *Recall*, as a weighted average (harmonic mean) of the two. If the weight is set to 0.5 (which is usually the case), *Precision* and *Recall* are deemed equally important. *F-measure* is formally defined as:

$$F - measure = ((\beta^2 + 1)P * R) / ((\beta^2 R) + P)$$

3. *Error Rate*: the number of wrongly identified answers divided by some fixed criteria such as document length. This is sometimes preferred in the IE field because, unlike precision, it is not dependent on relative document richness. Relative document richness is unimportant if evaluation is done using a single document, because it is equal for all systems. When comparing a single system's performance on different

documents, however, it is much more crucial, because if a particular document type has a significantly different number of any type of entity, the results for that entity type can become skewed ([7]).

4. *Cost Based Metrics*: For a cost-based error model, a cost would typically be associated with a miss and a false alarm (spurious answer), and with each category of result (e.g. recognising *Person* might be more important than recognising *Date* correctly). Expected costs of error would typically be based on probability (calculated on a test corpus). This makes the assumption that a suitable test corpus is available, which has the same rate of entity occurrence (or is similar in content) to the evaluation corpus. If necessary, the final score can be normalised to produce a figure between 0 and 1, where 1 is a perfect score.

4.4.2 Learning Accuracy

([16]) also describe a method called *Learning Accuracy* (LA) which augments Precision/Recall metrics with a cost tied to the degree of error involved based on semantic distance weights. The *Learning Accuracy* method measures how well a concept has been added in the right level of the ontology. It can be thought of being applied to a system that predicts the key concept that subsumes a new concept that is being added in the ontology. By checking the validity of this prediction, a measure of how correctly concepts have been placed is computed.

Learning Accuracy uses the following measurements:

- *Shortest Path (SP)*: the shortest length from root to the key concept.
- *FP*: the shortest length from root to the predicted concept. If the predicted concept is the ancestor of the new concept, then $FP = 0$, *i.e.* FP is only considered in the case that the answer given by the system is wrong.
- *Common Path (CP)* = shortest length from root to the MSCA (*Most Specific Common Abstraction*, *i.e.* the lowest concept common to SP and FP paths).
- *DP* = the shortest length from the MSCA to predicted concept.

Learning Accuracy is defined as follows:

- If the predicted concept is correct, *i.e.* if $FP = 0$, $LA = (CP/SP) = 1$
- If the predicted concept is incorrect, $LA = CP/(FP + DP)$.

If $LA = 0$, the concept is missing; if $LA = 1$, it has been perfectly placed.

4.4.3 Lexical Comparison Level measure

This measure is used to compare the contents of two ontologies without considering their conceptual structure, in a direct attempt to combat the problem of precision and recall and their restrictive binary nature. The measure is based on the well-known method of *edit distance* which measures the minimum number of insertions, deletions and substitutions needed to transform one string into the other. These scores are summarised for all the instances of a concept in the hierarchy, and averaged over the whole ontology.

4.4.4 Semiotic metrics

[6] have developed a set of metrics, which attempt to identify the internal attributes of ontologies that give rise to external quality attributes. These are based on *Semiotics*, which studies the properties of signs. It assesses, for example, whether the sign used for *Chair* is good or bad, clear (unambiguous) or unclear (ambiguous). Ontologies use symbols, or signs, to describe terms. The metrics are:

1. *Lawfulness*: Correctness of syntax
2. *Richness*: Breadth of syntax used
3. *Interpretability*: Meaningfulness of terms
4. *Consistency*: Consistency of meaning of terms
5. *Clarity*: Average number of word senses
6. *Comprehensiveness*: Number of classes and properties
7. *Accuracy*: Accuracy of information
8. *Relevance*: Relevance of information for a task
9. *Authority*: Extent to which other ontologies rely on it
10. *History*: Number of times ontology has been used

These can also be combined into a hybrid measure by weighting the metrics. These metrics can be used to make useful relative comparisons between ontologies.

4.5 Examples of Evaluation

Quite a few approaches to evaluation have been implemented as a system. A sample of some of them follows in this section:

4.5.1 OntoMetric

OntoMetric ([7]) is a method that helps users pick an ontology for a new system. It presents a set of processes that the user should carry out to obtain the measures of suitability of existing ontologies, regarding the requirements of a particular system. A set of basic decision criteria to be considered by the user before choosing an ontology are decided based on the content, language, the methodology followed, the software environments used for building the ontology, the costs of using the ontology in the system. Each of these dimensions contains a set of factors, which are used to determine the suitability of the ontology regarding the needs of the project. *OntoMetric* computes a measure of suitability for every candidate ontology using a framework of 160 characteristics that describe the ontology domain.

4.5.2 OntoClean

The *OntoClean* ([12]) methodology is based on philosophical notions for a formal evaluation of taxonomical structures. It focuses on the cleaning of taxonomies. Central to the methodology are ontological notions of *rigidity*, *unity* and *identity*. These notions are meta-properties, and are attached to properties and classes used in an ontology. These help represent the core characteristics of a concept, mainly from a taxonomy perspective. They also help make logical inferences that affect the quality of the system and can be used to detect inconsistencies in the use of the *subsumption* relationship.

Philosophical Notions

Rigidity: A property has the meta-property *rigid* if it is *essential* to every instance of that property. *Essence*, in turn, is when a property is permanently present for an instance as well is fundamentally necessary for the instance. An example is the property of *being a flower*. All flowers will have this property forever. In contrast, a property such as *being wet* is not rigid because if a cloth is wet, it could dry up later and lose the property, and so isn't essential. Properties can be classified as *rigid*, *semi-rigid* (essential to some instances and not to others), and *anti-rigid*.

Unity: This is related to whole-part relationships; detecting parts and wholes, and conditions where an instance is a whole. An individual is a whole if and only if it is made by a set of parts unified by a relation R. A property P is said to carry *unity* if there is a common unifying relation such that all the instances of P are wholes under R. A property carries *anti-unity* if all its instances can possibly be non-wholes. A *Constellation* is an example of an entity that is a whole, as it is a collection of a set of stars.

Identity: A property has *identity* associated with it if there are identity criteria that help uniquely identify each of its instances. Therefore given two instances of the class or property, if they can be separated in case they are

different or shown to be one and the same if they are identical, the property has *identity*. Instances of *Television Set* are always unique.

Constraints

These notions are used to define constraints on the taxonomy that a “clean” ontology should satisfy. A sample of these constraints are listed below. These are applicable when, given two properties, p and q , there is a subsumption relation such that q subsumes p , the following constraints hold:

1. If q is *anti-rigid*, then p must be *anti-rigid*.
2. If q carries an *identity* criterion, then p must carry the same criterion.
3. If q carries a *unity* criterion, then p must carry the same criterion.
4. If q has *anti-unity*, then p must also have *anti-unity*.

Backbone Taxonomy

All the rigid properties of the ontology are organised according to the identified subsumption relationships. These can be considered to be the most important properties covering the domain being modeled. This constitutes the *backbone taxonomy*.

Evaluation Process

The process of evaluation in *OntoClean* consists of recognising the meta-properties associated with the classes and properties in the ontology. After that, the set of axioms is applied on the backbone taxonomy to identify violations of the axioms. These violations indicate errors where the subsumption relationships are used incorrectly in place of instantiation, part-whole relations, polysemy *etc.* The quantity of errors can help indicate taxonomic cleanliness of the ontology.

4.5.3 An Example of Task Based Evaluation

[19] describe a task-based evaluation scheme to examine ontologies with respect to three basic levels: *vocabulary*, *taxonomy* and *non-taxonomic semantic relations*. A score based on error rates was designed for each level of evaluation.

In a task-based evaluation, the results are to show the following shortcomings:

- insertion errors indicating superfluous concepts, is-a and semantic relations.
- deletion errors indicating missing concepts, is-a and semantic relations.

- substitution errors indicating off-target or ambiguous concepts, is-a and semantic relations.

The error rates corresponding to specific ontological shortcomings are calculated. A task is chosen to evaluate the performance. As an example, the tagging of entities in a test set with role information provided by an ontology was chosen as a task. A gold standard was used to compare the results of the test of using different competing ontologies in the task.

4.5.4 OntoLearn

[8] describe an ontology evaluation scheme that makes it easier for domain experts to evaluate the contents of an ontology. This scheme, called *OntoLearn*, consists of algorithms to extract terms from documents, build relationships between them and thus construct an ontology for a domain. To evaluate the concepts and relations by domain experts (who are not computer scientists), different glosses are generated from them and presented to the experts for evaluation.

4.6 Issues

1. Clearly, scale and complexity of ontologies pose several challenges in evaluating ontologies. Evaluations have to consider to what extent the size of the contents, the (mildly) dynamic nature of the contents, different relations *etc.* can be verified or measured for properties. Instead, sampling may be required. This aspect also demands automated or semi-automated evaluation methods to cut down evaluation time.
2. Ontologies may be specified in different representation schemes. Evaluation methods that make use of language specific features may not be generally applicable. However, the increasing use of languages like *OWL* may obviate this in the future due to increasing standardisation.
3. As ontologies are ultimately resources to be absorbed in applications, different tasks may have specific needs and may prefer one set of quality criteria over others. Evaluation results need to highlight innate aspects of the ontologies over more application-subjective results.

4.7 Conclusion

This chapter looked at different theoretical dimensions and some of the practical approaches that have been applied in the field of ontology evaluations so far. Some of these have been applied in passing to *WordNet* among other ontologies, treating it as a concept ontology. The main lesson to learn from

these efforts is that complexity is prevalent and evaluation may have to be broken down into sub-tasks. Additionally, as ontology is a resource, the application context is sometimes more important. There is, as yet, no unified approach that combines structural, contextual, content and methodology measures.

Chapter 5

Evaluation Aspects of Semantic Networks

5.1 Introduction

There are, currently, no common and standardised evaluations that can be applied to all or even a subset of lexico-semantic networks. This chapter looks at some of the individual evaluation results from *WordNet* and *ConceptNet*. It also records some statistics that attempt to capture characteristics about these networks, which may provide a starting point for some commonly realisable quality characteristics to tie evaluation efforts across these semantic networks.

5.2 WordNet: Statistics and Evaluations

This section looks at some topological measures of WordNet outlined by [4]. These studies were carried out on an earlier *WordNet* version (1.1.7), which though comparatively old, provide some useful statistics about the topology of this lexical resource. These may be useful in developing evaluations in the future.

The key numbers are as follows:

1. *WordNet 1.1.7* had 74488 noun synsets, of which 78.8% were leaf nodes, giving 15902 internal nodes. The current version (2.1) has 81426 noun synsets ([10]).
2. The dimensional distribution of the lexical network is shown in the bar chart in figure 5.1. *Entity* is by far the most dominant class of nouns in *WordNet*. Examples of *Entity* are *person*, *ocean* and *ball*.
3. *Branching Factor* is defined as:

$$\text{BranchingFactor} = \text{NoOfDescendants} + 1(\text{for the node itself})$$

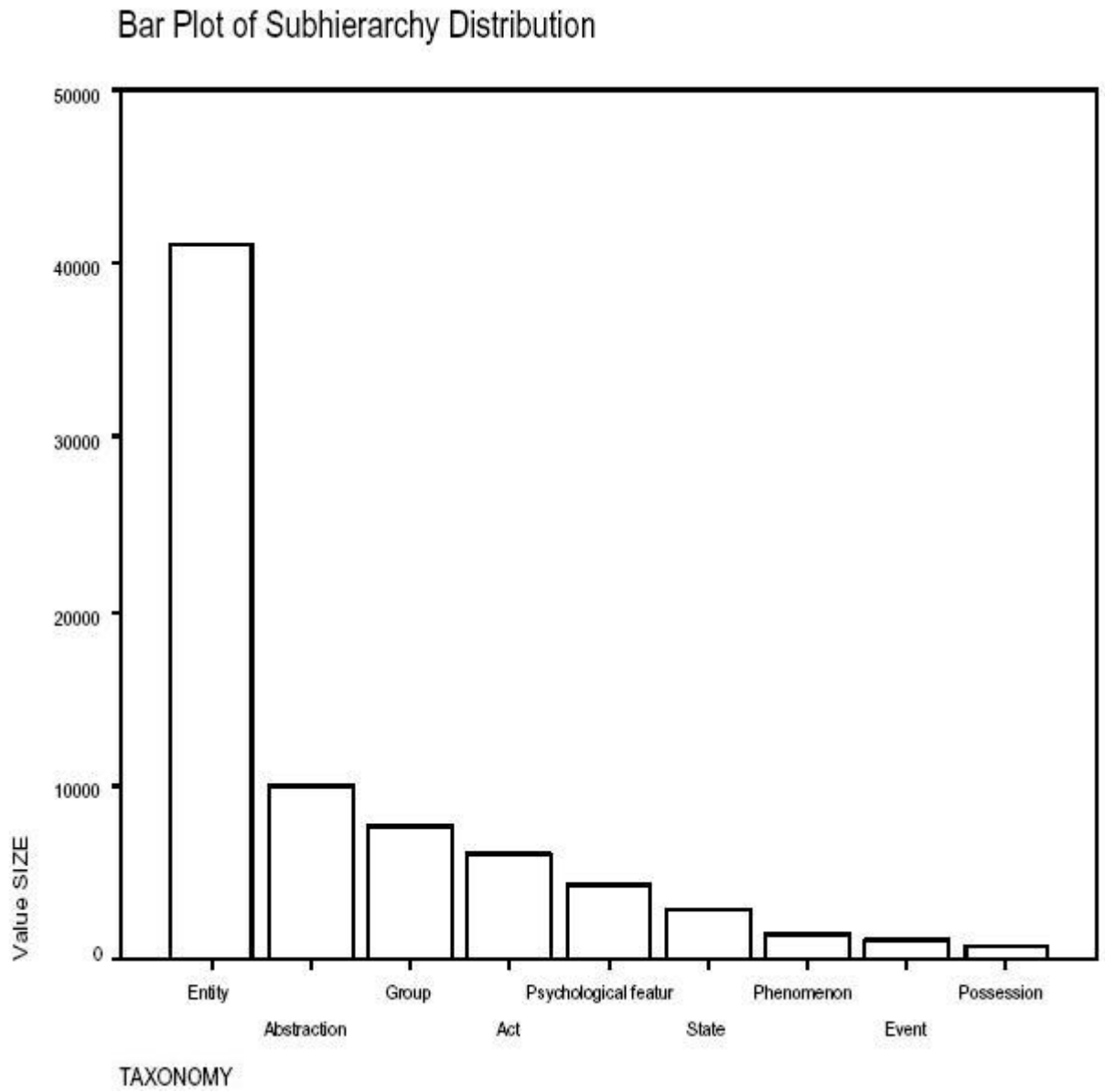


Figure 5.1: Bar Chart of synset distribution in top hierarchies ([4])

The Branching Factor has been found to range from 1 to 573 with an average value of 2.023. If leaf nodes are excluded, the average branching factor increases to 5.793, though 75% of the internal nodes have a branching factor of less than 5. This indicates that the taxonomy is not shallow and there is a degree of content specification embedded as a result.

4. Measures based on *Depth* and *Height* are defined as follows:

- *Maximum depth*: The longest path from a top taxonomy node to a given node
- *Minimum depth*: The shortest path from a top taxonomy node to a given node
- *Maximum height*: The longest path from given node to a leaf node
- *Minimum height*: The shortest path from a given node to a leaf node

The above measures have been used by [4] to investigate measures of content specificity.

5. *WordNet Evaluations*: Evaluations of different *WordNet* creation efforts have not been very prevalent. We describe one minor effort in the context of the **BalkaNet** project, which created WordNets for six Eastern European languages ([22]). The main aspects of the quality control methods applied were:

- A check on the quality of documentation of synsets, relations, data and resources provided with the *WordNet* package was carried out.
- A check whether *Base Concepts* from EuroWordNet were covered in these WordNets.
- These WordNets were represented in XML. So a number of simple XML validation steps to check *data consistency* by verifying existence of empty tags and attribute values, redundant repetition of literals, uniqueness of IDs, dangling links *etc.* were scripted.
- Semi-automated checks using additional language resources such as dictionaries were used to spell-check literals and definitions. Coverage and inconsistencies of highly frequent words and relations were also checked.
- The thrust of the evaluation was to provide feedback to the involved lexicographers and computer scientists during construction.

As can be seen, these efforts were driven by errors in actual physical representation and minimising mistakes in construction. A driving theory behind this was not derived and used.

5.3 ConceptNet: Statistics and Evaluations

As described earlier, *ConceptNet* is a commonsense knowledge base organised as a lexico-semantic network. [15] describe the methods employed by the creators to evaluate it. These measured the quality of the knowledge structure and its characteristics, using a mixture of automated methods and human evaluation.

1. The first result is that of complexity of nodes in *ConceptNet*. The statistic chosen to measure this is *word-lengths of nodes*. The results are shown in the figure 5.2. The results show that 70% of the nodes have a word-length less than 4, which can be inferred to indicate that a majority of concepts recorded are atomic rather than complex compound words.
2. The second result looks at the amount of assertions that are obtained directly from source statements (“direct assertions”) culled from the OMCS [21] knowledge corpus and those “inferred” from the direct assertions. The results are as in figure 5.3. These results show that most assertions and utterances are obtained very few times during the collection phase.
3. The third result looks at node degrees in the semantic network. The figure 5.4 illustrates the findings. 65% of nodes have at least 2 links.
4. A human evaluation was carried out to judge whether concepts are comprehensive or incorrect. This was essentially a coverage and accuracy finding exercise with judges looking at 100 concepts of their choice and ranking the results on a scale of 1 to 5. The results gave an overall score of 3.4/5 for comprehensiveness, 1.24/5 for noisiness and that 90% of concepts sought were found in the knowledge network.

	Mean Score	Std. Dev	Std. Err
Comprehensiveness	3.40/5.00	1.24	1.58
Noisiness	1.24/5.00	0.99	1.05
%Concepts attempted but not in ConceptNet	11.3%	6.07%	0.37%

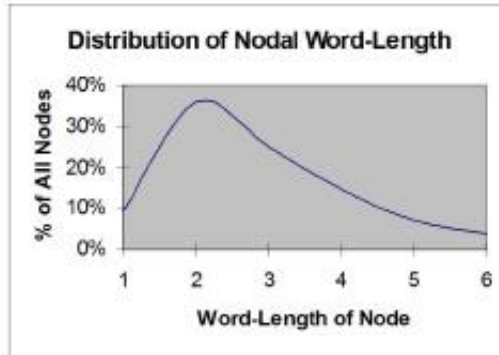


Figure 5.2: Word Length of Nodes as Indicator of Complexity ([15])

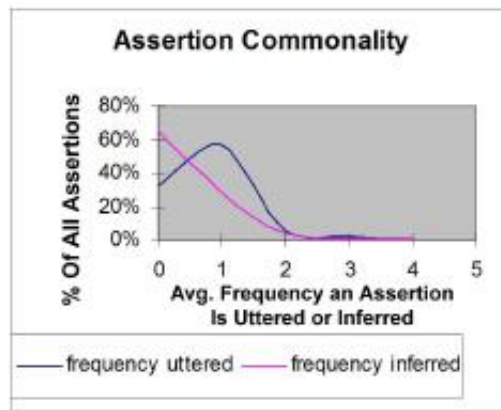


Figure 5.3: Origin of Assertions in ConceptNet ([15])

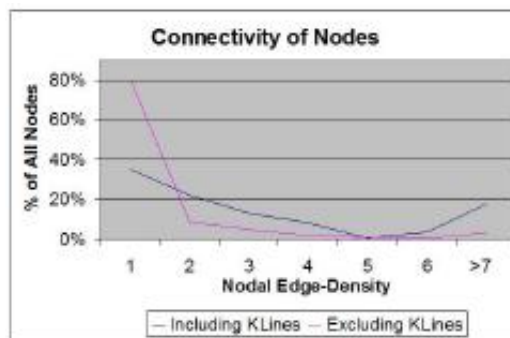


Figure 5.4: Connectivity of nodes in ConceptNet ([15])

Chapter 6

Evaluations in the Machine Translation Field

6.1 Introduction and Motivation

Evaluations in the domain of Machine Translation (MT) have seen more progress than other sub-fields of Natural Language Processing. This chapter describes some of the key approaches applied to evaluating the results of a Machine Translation activity. Our aim in studying these techniques is motivated not towards directly adapting any of these methods to lexico-semantic networks. Instead, there may be lessons to be learnt from the origins of these methods, the successes enjoyed and criticisms levelled at these methods. This is especially because there are certain strategies like **BLEU** that have established themselves as the leading methods against which further research in the field is compared with.

The following sections use a survey paper by [9] to study the field.

6.2 Approaches and Challenges

The two main dimensions to evaluating the translation of a source text to a target language are the concepts of *adequacy* and *fluency*. These are defined as follows:

1. **Adequacy** is the question of whether the target translation conveys the same meaning as the source. *e.g.* If the source text in a language has the English meaning *Sachin waved to his fans* and is translated to English as *Sachin acknowledged his supporters with a wave*, the translation would be deemed “adequate”. However, if the translated sentence turned out to be *Sachin fans his wave*, this would be highly inadequate.

2. **Fluency** is whether the translation is grammatically correct. *e.g.* Using the example above, if the translation was made as *waves Sachin to fans his*, this would be clearly invalid. A less extreme case is of *Sachin waves to their fans*.

Given the variations in acceptable translations due to reasons of synonymy, acceptable phrasal substitutes, ambiguity in words used in the source text, and even lack of equivalent terms from one language to another, machine translation is a very difficult process. Designing useful evaluation measures is also difficult as a result. These measures are expected to help make comparisons between MT methods as well as provide absolute scores of quality of evaluation.

The two popular families of MT evaluation have both been based on making comparisons to a gold standard. A set of source texts and reference translations are assembled forming a benchmark compilation. This basic idea may limit the quality of evaluation to only a given domain and make it hard to be certain of a method's performance on a very different domain. This is also a significant challenge.

The two families mentioned above are based on string matching techniques and *n-gram* based IR measures respectively.

6.3 String Matching Techniques

The idea in these techniques is to consider the produced translation and the reference translation as strings, and to compute the *minimum edit distance* or *Levenshtein distance* to convert the target to the reference. The editing actions are insertions, deletions and substitutions.

Some of the techniques are summarised as follows:

1. *Word Error Rate* (WER): WER is the sum of edit actions divided by the length of the reference sentence. A WER of zero indicates no changes were made and hence perfect translation.
2. *Position-Independent Error Rate*: this does not consider ordering of words during matches. Instead, the target and reference text are taken as bags-of-words and compared.

[9] mentions that studies show string matching techniques have broadly not shown good correlation with human judgement, especially when compared to the next family of techniques.

6.4 IR Techniques on N-grams

IR-based techniques involve a variation on the precision and recall metrics. A target translation is compared against one or more reference translations

for a sentence, and a metric is computed. The definitions of these metrics form the basis of the various approaches. Furthermore, instead of a simple bag-of-words approach, *n-grams* are used. Two of the most important metrics are described below.

BLEU

BLEU ([18]) is a precision-based metric. The intuition behind this measure is that a good translation is likely to share many words and phrases with the reference translation(s). The *n-grams* of the target translations are compared against *n-grams* of the references.

For instance, consider a source sentence that has the following possible reference translations:

1. Amit was wearing a red cap *or*
2. Amit wore a cap that was red in colour *or*
3. A red cap was worn by Amit

Any valid candidate translation by a MT method is likely to have words such as *wear* (or valid variants thereof), *Amit*, *cap* and *red*. Further, *n-grams* such as *red cap* are likely to be part of such valid translations.

The creation of the precision metric evolved as follows:

- In its simplest form, precision can be computed as the number of words in the target translation sentence that appear in the set of reference translations for that sentence divided by the total number of words in the target translation. This is the *unigram precision* measure.
- The above definition suffers from the problem of translations in which the valid words are repeated more often than they should (*e.g.* repetition of *cap* several times in the previous example) leading to high-precision but ultimately invalid translations. To obviate this, the number of times a word appears in the reference translations is counted. The total count of the same word in the target translation is “clipped” or normalised by this count. The precision is now calculated by summing all the clipped counts over the total word count of the target translation and is called the *modified unigram precision measure*.
- Finally, this measure can be extended from a unigram measure to *n-grams*.
- The score is calculated for each sentence in the source text using the modified precision scores for each word. Without discussing the details, the scores are combined by taking a geometric mean and multiplying the result by a factor to penalise target translations which are shorter than their reference translations. The metric ranges from 0 to 1, where a score closer to 1 indicates good quality of translation.

NIST

This is a variant of the BLEU metric where the sentence scores are combined by an arithmetic mean. Additionally, less frequent *n-grams* are considered more informative, and given a higher weight.

The principal advantage of such metrics has been their simplicity in computation.

6.5 Criticisms

1. Despite the fact that metrics like **BLEU** and **NIST** have had some success, there continues to be lack of consensus on their acceptance. One major reason is that these are precision-based measures and do not measure recall over the set of reference translations well. To see why, consider the case where a candidate translation includes all words from a set of alternative reference translations by including all variations of a word or phrase. Recall is high in this case, but the resulting translation is usually unsuitable. [9] quote empirical evidence to suggest that recall correlates best with human judgement.
2. [3] provide a set of examples in which the words in the candidate translations are permuted and modified such as to actually yield a higher BLEU score without an increase in translation quality.
3. [3] also point out that these metrics handle synonyms and paraphrases only if they appear in one of the reference translations.

Chapter 7

Evaluating Lexico-Semantic Networks

7.1 Conclusions from Literature Survey

Having made a study of evaluations in ontologies and MT, here are the key learnings from them:

1. No major evaluation strategies have been designed for lexico-semantic networks in general, or even for specific knowledge structures. Evaluations in *WordNet* have been very localised, checking for errors in construction rather than at a higher level for coverage and accuracy. *ConceptNet's* method involved human appraisal on sample nodes, which is clearly not scalable.
2. The structural properties of networks can also be a key factor in rating these networks.
3. The progress in ontologies has shown that evaluation has several dimensions and the current state of art involves breaking this complexity into different evaluation criteria. Evaluations must aim to answer questions of relative comparison, intrinsic value, accuracy, consistency and coverage of a domain and suitability to applications.
4. Unlike ontologies, MT evaluations have settled upon IR-based metrics as a standard method of rating MT algorithms. Though there are criticisms of specific methods, there is a broad consensus on this direction of measurement.
5. The leading method of MT evaluation **BLEU** began with an intuitive statement of what good evaluation for MT must do. This idea is that *The closer a machine translation is to a professional human*

translation, the better it is ([18]). The metrics were designed from this principle.

6. Suitably designed and clearly defined metrics are necessary to succinctly represent the results of evaluation. Simplicity of a metric must be kept in mind though it may not be the most important virtue of the metric.
7. Evaluations have to be as automatic as possible. In many ontology ratings and in MT evaluations, this is achieved by developing algorithms to compare the unrated entities to a gold standard. In case of lexico-semantic networks with a large and dynamic content, this could be a key challenge.
8. The structural properties of *WordNet* seen so far may not be sufficient or useful to evaluation efforts. New properties may need to be derived based on ontological properties such as *Rigidity* and *Unity* proposed by [12].
9. There have been no studies on computational efficiency and usability (such as on querying for information), completeness and consistency aspects of these knowledge bases.

7.2 Future Directions

The increasing use of lexico-semantic networks in various applications as a source of not merely words and their properties, but also of semantic content, is the main driving force behind studying how to evaluate them. The directions of scientific investigation in this area could be based on the following questions, directions and challenges:

1. Establish criteria to measure intrinsic quality of the content held in these lexical networks and common-sense ontologies.
2. Establish criteria to make useful comparisons between different lexico-semantic networks.
3. Investigate methods to check if a network's quality has improved or declined before and after content updates.
4. The domain of common-sense is infinite; so aims such as coverage or recall will have to be tailored accordingly.
5. Content in the nodes, relationships between nodes and the collection as a whole should be targeted.
6. The size and scope of these lexico-semantic networks will be a key challenge.

7. Methods used in ontology appraisal could be used or adapted to lexico-semantic networks. Whether taking an ontological view of such networks will be sufficient needs to be verified.
8. Most appraisals are done *w.r.t.* a practically assembled gold standard. If applied to lexico-semantic networks, what form should such a gold standard take?
9. Like in *BLEU*, is there a central intuitive idea that can be stated and used to drive discovery of metrics?
10. Mere checking of structure and organisation without reference to content may be a hollow strategy. However, structure may provide key insights to be used, so it should not be abandoned.
11. Questions of usability, efficiency and scalability should also be explored for reasons of practical worth.
12. Some of the lexico-semantic networks (such as *MindNet* and to an extent, *HowNet*) are not readily available for use outside their parent research bodies. Investigation may be restricted to examples such as *WordNet* and *ConceptNet*.
13. A more focussed appraisal may be conducted by restricting the context of the network's use to a particular task.

Bibliography

- [1] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In *Proceedings of the COLING-ACL, 1998*.
- [2] Janez Brank, Marko Grobelnik, and Dunja Mladenic. A survey of ontology evaluation techniques. In *Proceedings of the Conference on Data Mining and Data Warehouses (SiKDD 2005)*, 2005.
- [3] C. Callison-Burch, M. Osborne, and P. Koehn. Re-evaluating the Role of BLEU in Machine Translation Research. In *In Proceedings of EACL*, 2006.
- [4] Ann Devitt and Carl Vogel. The Topology of WordNet: Some Metrics. In *Proceedings of GWC-04, 2nd Global WordNet Conference*, 2004.
- [5] Zhendong Dong and Qiang Dong. An Introduction to HowNet. Available from <http://www.keenage.com>.
- [6] Andrew Burton Jones et al. Metrics for Evaluation of Ontology based Information Extraction. In *Data and Knowledge Engineering*, 2004.
- [7] Jens Hartman et al. D1.2.3 Methods for ontology evaluation. Deliverable for Knowledge Web Consortium in 2005.
- [8] Paolo Velardi et al. Automatic Ontology Learning: Supporting a Per-Concept Evaluation by Domain Experts. In *Workshop on Ontology Learning and Population (OLP), in the 16th European Conference on Artificial Intelligence*, 2004.
- [9] Cyril Goutte. Automatic Evaluation of Machine Translation Quality. A Xerox Research Centre Europe Publication, 2006.
- [10] WordNet Group. WordNet 2.1 database statistics. <http://wordnet.princeton.edu/man/wnstats.7WN>.
- [11] Nicola Guarino. Toward a Formal Evaluation of Ontology Quality - (Why Evaluate Ontology Technologies? Because It Works!). *IEEE Intelligent Systems*, Vol. 19, No. 4:74–81, July/August 2004.

- [12] Nicola Guarino and Christopher Welty. Evaluating ontological decisions with OntoClean. *Communications of the ACM*, Volume 45, Number 2:61–65, February 2002.
- [13] D. Jurafsky and J. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. Prentice Hall, 2000.
- [14] Margaret King. Living up to standards. In *Proceedings of the EACL 2003 Workshop on Evaluation*, 2003.
- [15] 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.
- [16] Diana Maynard, Wim Peters, and Yaoyong Li. Metrics for Evaluation of Ontology based Information Extraction. In *EON2006 at WWW*, 2006.
- [17] 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.
- [18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002.
- [19] Robert Porzel and Rainer Malaka. A Task-based Approach for Ontology Evaluation. In *ECAI Workshop on Ontology Learning and Population*, 2004.
- [20] 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.
- [21] 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.
- [22] Pavel Smrz. Quality Control for Wordnet Development. In *Proceedings of GWC-04, 2nd Global WordNet Conference*, 2004.