Automatic Lexicon Generation through WordNet

Nitin Verma and Pushpak Bhattacharyya*

Department of Computer Science and Engineering
Indian Institute of Technology Bombay
{nitinv, pb}@cse.iitb.ac.in

Abstract. A lexicon is the heart of any language processing system. Accurate words with grammatical and semantic attributes are essential or highly desirable for any application- be it machine translation, information extraction, various forms of tagging or text mining. However, good quality lexicons are difficult to construct requiring enormous amount of time and manpower. In this paper, we present a method for automatically generating the dictionary from an input document- making use of the *WordNet*. The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with disambiguation information. The entries are associated with syntactic and semantic properties- most of which too are generated automatically. In addition to the WordNet, the system uses a *word sense disambiguator*, an *inferencer* and the *knowledge base (KB)* of the *Universal Networking Language* which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labour substantially.

Keywords: Lexicon, Lexical-Syntactic-Semantic Attributes, WordNet, Universal words, Universal Networking Language, Ontology

1 Introduction

Construction of good quality lexicons enriched with syntactic and semantic properties for the words is time consuming and manpower intensive. Also word sense disambiguation presents a challenge to any language processing application, which can be posed as the following question: *given a document D and a word W therein, which sense S of W should be picked up from the lexicon?*. It is, however, a redeeming observation that a particular W in a given D is mostly used in a single sense throughout the document. This motivates the following problem: *can the task of disambiguation be relegated to the background before the actual application starts? In particular, can one construct a* **Document Specific Dictionary** *wherein single senses of the words are stored*?

Such a problem is relevant, for example, in a machine translation context [2]. For the input document in the source language, if the *document specific dictionary* is available a-priori, the generation of the target language document reduces to essentially syntax planning and morphology processing for the pair of languages involved. The WSD problem has been solved before the MT process starts, by putting in place a lexicon with the document specific senses of the words.

In this paper we have addressed this problem by showing how the WordNet [5][3] can be used to construct a document specific dictionary. The entries in the dictionary are the *Universal Words (UWs)* which are language words (primarily English) concatenated with disambiguation information.

* Contacting Author

The entries are associated with syntactic and semantic properties- most of which too are generated automatically. In addition to the WordNet, the system uses a *word sense disambiguator*, an *inferencer* and the *knowledge base (KB)* of the *Universal Networking Language (UNL)* which is a recently proposed interlingua. The lexicon so constructed is sufficiently accurate and reduces the manual labour substantially.

Section 2 briefly describes the UNL system. Section 3 is on Universal Words [4]. Format of the UW Dictionary is described in section 4. Section 5 narrates the resources required for the dictionary generation and section 6 explains the methodology. Section 7 gives the results of experiments and charts the future directions.

2 Universal Networking Language (UNL)

UNL [4] is an interlingua for machine translation [2] and is an attractive proposition for the multilingual context. In this scheme, a source language sentence is converted to the UNL form using a tool called the *EnConverter* [4]. Subsequently, the UNL representation is converted to the target language sentence by a tool called the *DeConverter* [4]. The sentential information in UNL is represented as a hyper-graph with concepts as nodes and relations as arcs. The UNL graph is a hyper-graph because the node itself can be a graph, in which case the node is called a *compound word* (CW). Figure 1 represents the sentence *John eats rice with a spoon.*

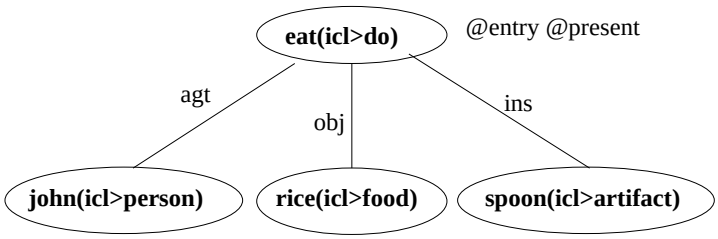

Fig. 1. UNL graph of *john eats rice with a spoon*

In the above graph the arcs denoting *agt* (agent), *obj* (object) and *ins* (instrument) are the relation labels as defined in the UNL specification. This graph is represented as a set of directed binary relations between two concepts present in the sentence. The relation *agt* stands for *agent*, *obj* for *object* and *ins* for *instrument*. The binary relations are the basic building blocks of the UNL system, which are represented as strings of 3 characters or less each. There are 41 relations in the UNL system.

In the above figure the nodes such as *eat(icl>do), John(iof>person), rice(icl>food)* and *spoon(icl >artifact)* are the *Universal Words (UW)*. These are language words with *restrictions* in parentheses. *icl* stands for *inclusion* and *iof* stands for *instance of*. UWs can be annotated with attributes which provide further information about how the concept is being used in the specific sentence. Any of the three restriction labels, *viz., icl, iof* and *equ*, is attached to an UW for restricting its sense.

3 Universal Words

Universal Words constitute the *vocabulary* of the UNL [8]. A UW represents a *unique concept* by combining an *English word* along with a *restriction*. For example, the UW *spring(icl>tool)* describes a *tool*, and the UW *spring(icl>season)* stands for a *season*.

A UW is created using the *specifications* of the *UNL Knowledge Base (KB)*. UNL KB organizes the UWs in a *hierarchy*. A *part* of the UW hierarchy for *nouns* in the UNL KB is shown in figure 2 which is self-explanatory.

```
Depth   UW
-----    --------------
 0    thing
      |
 1    |__abstract thing{(icl>thing)}
 2    |  |___activity(icl>abstract thing)
 3    |  |    |___broadcasting(icl>activity{>abstract thing})
 3    |  |    |___defense(icl>activity{>abstract thing})
 3    |  |    |___development(icl>activity{>abstract thing})
 2    |  |___art(icl>abstract thing)
 3    |  |    |___craft(icl>art{>abstract thing})
 3    |  |    |___fine arts(icl>art{>abstract thing})
 4    |  |    |   |___picture(icl>fine arts{>art})
      ............
      ............
```

Fig. 2. Hierarchy of *noun* UWs in the UNL KB (a snapshot)

For verbs, the hierarchy is not so deep. All the verbs are organized under three categories, *viz.*, *do, occur* and *be*. The first two are *aktionstat verbs* and the last one is the set of *stative verbs*. The adjective, adverb and preposition hierarchies too are quite shallow. The adjectives that are both *attributive* and *predicative* are given the restriction ($aoj > thing$), where *aoj* is a semantic relation denoting *attribute of the object* and *thing* denotes a nominal concept. The adjectives which are only *predicative* are given the restriction ($mod > thing$) where *mod* is the *modifier* relation. The adverbs are uniformly expressed through ($icl > how$).

4 L-UW Dictionary

The dictionary maps the *words* of a natural language to the *universal words* of the UNL system [6]. For example,

```
[kuttaa] "dog(icl>mammal)" (...attributes...)
[bhOka] "bark(icl>do)" (...attributes...)
```

are the entries in a Hindi-UW dictionary [7]. Similarly

```
[dog] "dog(icl>mammal)" (...attributes...)
[bark] "bark(icl>do)" (...attributes...)
```

are the entries is an English-UW dictionary. When the sentence *The dog barks* is given to an UNL-based English-Hindi MT system, the UWs *dog(icl>mammal)* and *bark(icl>do)* are picked up. *If the*

L-UW dictionary contains only document specific UWs, the analyser and the generator systems do not commit error on account of WSD.

The *attributes* attached to each entry in the L-UW dictionary are the *lexical, grammatical,* and *semantic* properties of the language specific words (*NOT of the UWs*). The syntactic attributes include the word category- *noun, verb, adjectives, adverb etc.* and attributes like *person* and *number* for nouns and *tense* for verbs. The *Semantic Attributes* are derived from an *ontology.* Figure 3 shows a part of the *ontology* used for obtaining semantic attributes [6].

```
Part of ontology and Semantic attributes for nouns
==================================================

Animate (ANIMT)
  o Flora (FLORA)
     =>Shrubs (ANIMT, FLORA, SHRB)
  o Fauna (FAUNA)
     =>Mammals (MML)
         1. Person(ANIMT, FAUNA, MML, PRSN)
         2. Ape (ANIMT, FAUNA, MML, APE)
     =>Birds (ANIMT, FAUNA, BIRD)
      .....
      .....

Part of ontology and Semantic attributes for verbs
==================================================

Verbs of Action (VOA)
  o Change (VOA,CHNG)
  o Communication (VOA,COMM)
  o Motion (VOA,MOTN)
  o Completion (VOA,CMPLT)
Verbs of State (VOS)
  o Physical State (VOS,PHY,ST)
  o Mental State (VOS,MNTL,ST)
    .....
    .....

Part of ontology and Semantic attributes for adjectives
=======================================================

Descriptive (DES)
  o Weight (DES,WT)
  o Shape (DES,SHP)
  o Quality (DES,QUAL)
  o Temperature (DES,TEMP)
Relational (REL)
.....
.....

Part of ontology and Semantic attributes for adverbs
====================================================

Time (TIME)
Frequency (FREQ)
Quantity (QUAN)
Manner (MAN)
Direction (DRCTN)
.....
.....
```

Fig. 3. Ontology and Semantic attributes

5 Resources for dictionary generation

For generating the document specific dictionary we use the *WordNet, a WSD System*, the *UNL KB* and an *inferencer*. The approach is *Knowledge Based*. The UNL KB as shown in figure 2 is stored as a *mysql* database. The table *UNL-KB-table* in figure 4 shows a part of this storage structure for nouns.

The word sense disambiguator [1] works with an accuracy of about 70% for nouns. The essential idea is to use the *noun-verb* association- as given in a co=occurence dictionary- to obtain a set of semantic clusters for the noun in question. The densest cluster denotes the most likely sense of the word. Taking the example of *the crane flies* we get two semantic clusters involving the hypernyms and the hyponyms of the *bird* sense and the *machine sense*. Since the former has much larger association with *fly*, it becomes the winner.

For other parts of speech, the first sense as given in the WordNet is chosen, which as per the WordNet is the most frequently used sense.

The semantic attributes are generated from a rule-base linking the lexico-semantic relations of the WN with the semantic properties of the word senses. To take an example, if the hypermyny is *organism*, then the attribute *ANIMT* signifying *animate* is generated. We have more than 1000 such rules in the rule base.

6 Methodology for dictionary generation

As discussed so far, there are two parts to the dictionary entry generation, *viz.*, creating UWs and assigning the syntactic and semantic attributes. The following subsections discuss this.

6.1 POS tagging and sense disambiguation

The document is passed to the word sense disambiguator [1]. This picks the correct sense of the word with about 70% accuracy. As a side effect the words are POS tagged too. The output of this step is a list of entries in the format **Word:POS:WSN**, where POS stands for part of speech and WSN indicates the WordNet sense number. The *syntactic* attributes are obtained at this stage.

6.2 Generation of UWs

The WN and UNL KB are used to generate the restriction for the word. If the word is a noun, the WN is queried for the hypernymy for the marked sense. All the Hypernymy ancestors $H_1, H_2, \ldots, H_n$ of W up-to the *unique beginner* are collected. If $W(icl > H_i)$ exists in the UNL KB, it is picked up and entered in the dictionary. If not, $W(icl > H_1)$ is asserted as the dictionary entry.

for example, for *crane* the *bird*-sense gives the hypernyms as *bird, fauna, animal, organism* and finally *living_thing*. $crane(icl > bird)$ becomes the dictionary entry in this case. Figure 4 illustrates this process.

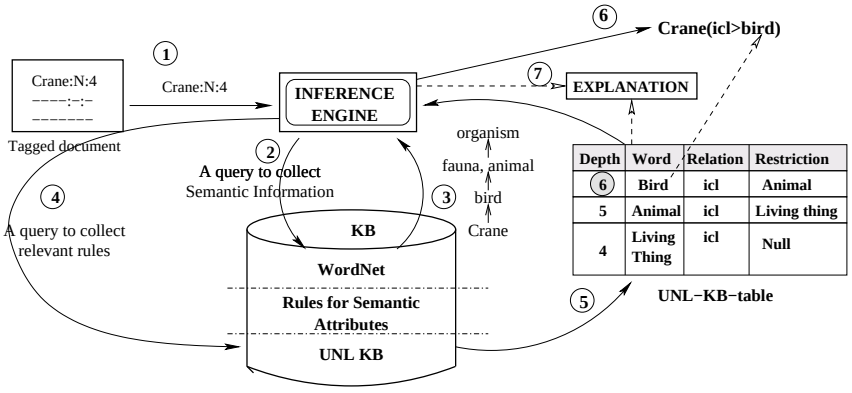

Fig. 4. Universal Word Creation: an example

For verbs, the hypernymy ancestors are collected from the WN. If these include concepts like *be, hold, continue etc.*, then we generate the restriction $(icl > be)$ (case of *be* verb). If not, the

corresponding *nominal word* (for example, the nominal word for the verb *rain* is *rain* itself) of the verb is referred to in the WN. If the hypernyms of the nominal word include concepts like *phenomenon, natural_event etc.*, then we generate the restriction ($icl > occur$) signifying an *occur* verb. If both these conditions are not satisfied, then the restriction ($icl > do$) is generated.

For adjectives, use is made of the *is_a_value_of* semantic relation in the WN. For example, for the adjective *heavy* the above relation links it to *weight*. If this relation is present then the restriction ($aoj > thing$) is generated. Else we generate ($mod > thing$) (please refer back to section 3).

For adverbs, ($icl > how$) is by default generated, as per the specifications of the UNL system.

6.3 Creation of semantic attributes

As explained in section 5, WN hypernymy information and the rule base is used to generate the semantic attributes of nouns. The tables in the figure 5 shows sample of such rules for all the POS words. The first entry in the table 1 corresponds to the rule: IF hypernym = *organism* THEN generate *ANIMT* attribute.

Table 1: Rules for noun

HYPERNYM	ATTRIBUTE
organism	ANIMT
flora	FLORA
fauna	FAUNA
beast	FAUNA
bird	BIRD

Table 2: Rules for verb

HYPERNYM	ATTRIBUTE
change	VOA,CHNG
communicate	VOA,COMM
move	VOA,MOTN
complete	VOA,CMPLT
finish	VOA,CMPLT

Table 3.1: Rules for adjectives

IS_VALUE_OF	ATTRIBUTE
weight	DES,WT
strength	DES,STRNGTH
qual	DES,QUAL

Table 3.2: Rules for adjectives

SYNONYMY	ATTRIBUTE
bright	DES,APPR
deep	DES,DPTH
shallow	DES,DPTH

Table 4: Rules for adverbs

SYNONYMY	ATTRIBUTE
backward	DRCTN
always	FREQ
frequent	FREQ
beautifully	MAN

Fig. 5. Rules for generating Semantic attributes

For example for the *bird* sense of *crane* (**crane:N:4**),

```
[crane]"crane(icl>bird)"(N,ANIMT,FAUNA,BIRD);
```

will be generated.

7 Experiments and Results

We have tested our system on documents from various domains like agriculture, science, arts, sports *etc.* each containing about 800 words. We have *measured* the *performance* of this system by calcu-

lating its *precision* in every POS category. The precision is defined as

$$Precision = \frac{Number\ of\ entries\ correctly\ generated}{Total\ entries\ generated}$$

figure 6 shows the results. The average precision for nouns is **93.9%**, for *verbs* **84.4%**, for *adjectives*

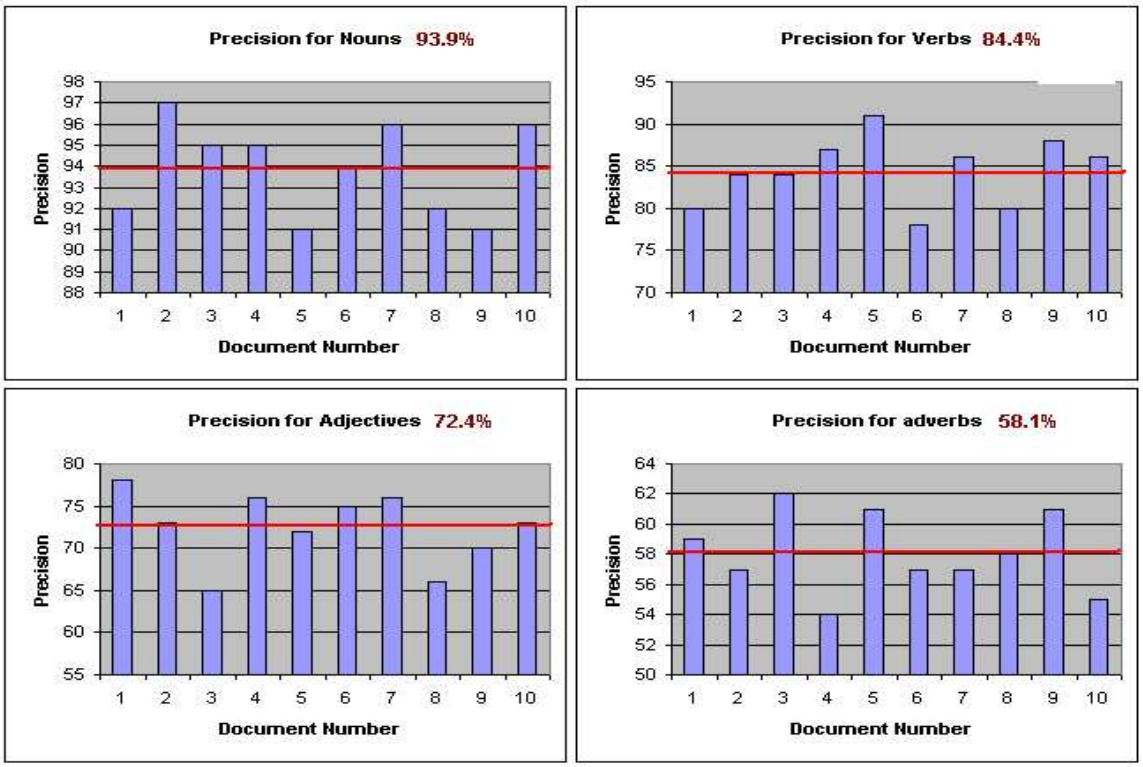

Fig. 6. Experiments and Results

72.4% and for *adverbs* **58.1%**.

The dictionary generated by the above methodology performs well in case of nouns and verbs. The reason for low accuracy for adjectives and adverbs is the shallowness in the hierarchy and lack of many semantic relations for these parts of speech. The system is being routinely used in our work on machine translation in a tri-language setting (*English, Hindi* and *Marathi*), it has reduced the burden of lexicography considerably. The incorrect entries- which are not many- are corrected manually by the lexicon makers. Figure 7 shows the dictionary generated (the wrong entries are marked by *) from a document containing the following paragraph.

Modern agriculture depends heavily on engineering and technology and on the biological and physical sciences. Irrigation, drainage, conservation, and sanitary engineering- each of which is im-

portant in successful farming- are some of the fields requiring the specialized knowledge of agricultural engineers.

```
[Modern]{}"modern(aoj>thing)"(ADJ,DES,APPR)<E,0,0>
[agriculture]{}"agriculture(icl>business)"(N,INANI,EVENT,ABS)<E,0,0>
[depend]{}"depend(icl>be(aoj>thing))"(VRB,CONT,VOS-PHY-ST)<E,0,0>
[heavily]{}"heavily"(ADV,QUAN)<E,0,0>
[engineering]{}"engineering(icl>subject)"(N,INANI,PSYFTR,ABS)<E,0,0>
[technology]{}"technology(icl>subject)"(N,INANI,PSYFTR,ABS)<E,0,0>
[biological]{}"biological(mod<thing)"(ADJ,REL)<E,0,0>
[physical]{}"physical(mod<thing)"(ADJ,DES,SHAPE)<E,0,0>
[scienc]{}"science(icl>skill)"(N,INANI,PSYFTR,ABS)<E,0,0>
[Irrigation]{}"irrigation(icl>act)"(N,INANI,EVENT,ABS)<E,0,0>
* [drainage]{}"drainage(icl>change)"(N,INANI,EVENT,ABS)<E,0,0>
[conservation]{}"conservation(icl>improvement)"(N,INANI,EVENT,NAT,ABS)<E,0,0>
* [sanitary]{}"sanitary(aoj>thing)"(ADJ)<E,0,0>
[important]{}"important(aoj>thing)"(ADJ,DES,NUM)<E,0,0>
[successful]{}"successful(aoj>thing)"(ADJ,DES,SND)<E,0,0>
* [field]{}"fields(icl>person)"(N,ANIMT,FAUNA,MML,PRSN,PHSCL)<E,0,0>
[requir]{}"require(icl>necessitate(agt>thing,gol>place,src>place))"(VRB,VOA-COMM,VOA-POSS)<E,0,0>
  [specialized]{}"specialized(mod<thing)"(ADJ)<E,0,0>
[knowledge]{}"knowledge(icl>cognition)"(N,INANI,PSYFTR,ABS)<E,0,0>
[agricultural]{}"agricultural(aoj>thing)"(ADJ,REL)<E,0,0>
[engineer]{}"engineer(icl>person)"(N,ANIMT,FAUNA,MML,PRSN,PHSCL)<E,0,0>
```

Fig. 7. UW Dictionary generated after running the system on a sample document

The future work consists in generating restrictions involving *iof (instance-of)*, *equ (equivalent to)*, *pof (part of)* and such other constructs. Efforts are also on to migrate the system to WordNet 2.0 which has the very useful relations of *derived_from* and *domt* doing cross POS linkage in the WN. It is hoped that this will mitigate the problems arising from the low accuracy of the WSD system and the shallowness of the non-noun hierarchies.

References

1. Dipak K. Narayan and Pushpak Bhattacharyya.: *Using Verb-Noun association for Word Sense Disambiguation.* International Conference on Natural language processing, November 2002.
2. W. John Hutchins and Harold L. Somers.: *An Introduction to Machine Translation.* Academic Press, 1992.
3. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K.: *Five papers on WordNet.* Available at URL: *http://clarity.princeton.edu:80/ wn/*, 1993.
4. The Universal Networking Language (UNL) Specifications, United Nations University. Available at URL: *http://www.unl.ias.unu.edu/unlsys/*, July 2003.
5. Christiane Fellbaum.: *WordNet: An Electronic Lexical Database.* The MIT Press, 1998.
6. P. Bhattacharyya.: *Multilingual information processing using UNL.* in Indo UK workshop on Language Engineering for South Asian Languages LESAI, 2001.
7. Shachi Dave, Jignashu Parikh and Pushpak Bhattacharyya,: *Interlingua Based English Hindi Machine Translation and Language Divergence*, Journal of Machine Translation, Volume 17, September, 2002. (to appear).
8. Hiroshi Uchida and Meiying Zhu. *The Universal Networking Language beyond Machine Translation.* UNDL Foundation, September 2001.