Hindi Semantic Category Labeling Using Semantic Relatedness Measures

Siva Reddy
LTRC
IIIT Hyderabad, India
gvsreddy@students.iiit.ac.in

Abhilash Inumella
LTRC
IIIT Hyderabad, India
abhilashi@students.iiit.ac.in

Navjyoti Singh
Center for Exact Humanities
IIIT Hyderabad, India
navjyoti@iiit.ac.in

Rajeev Sangal
LTRC
IIIT Hyderabad, India
sangal@iiit.ac.in

Abstract

In this paper, we evaluate and compare six semantic relatedness measures used for Hindi semantic category labeling. Our experiments show that the measure "adapted lesk" performed better than other measures. However, a simple baseline system achieved better accuracy than all the measures.

1 Introduction

The task of semantic category labeling has been introduced in (Reddy et al., 2009). Given a word, its admissible semantic categories and its context, the task of semantic category labeling is to assign the most appropriate semantic category to the word. Our language of interest is Hindi [1]. An example is shown in *Table 1*. We used Hindi WordNet Ontological categories (Narayan et al., 2002) as semantic category inventories rather than WordNet synsets which are conventionally used in word sense disambiguation.

1.1 Ontological Categories

Hindi WordNet Ontological Categories are coarse grained distinctions of word senses. These categories are organized in a hierarchical fashion based on 'is-a' relation. A separate ontological hierarchy exists for each syntactic category (noun, verb, adjective adverb). Total number of categories in noun, verb, adjective and adverb hierarchy are 101, 31, 25 and 11 respectively and the maximum depth of the hierarchy is 5. There are 28,663 synsets in Hindi WordNet. Every synset is mapped to a category in the ontological hierarchy. Figure 1 depicts ontological hierarchy of the word *billA*.

2 Related Work

Inspired by the original Lesk algorithm (Lesk, 1986), a number of WordNet based disambiguation algorithms were proposed. Lesk algorithm disambiguates a target word by assigning the sense whose gloss (definition) maximally overlaps with the neighbouring words gloss. Banerjee and Pedersen (2002) used hierarchical relationships in WordNet to include the glosses of words that are related to the target word and its neighbours. Patwardhan et al. (2003) takes the view that gloss overlaps are just another measure of semantic relatedness. They evaluated a number of semantic relatedness measures for English word sense disambiguation.

Our work builds on the earlier work of (Patwardhan et al., 2003). We evaluated a number of semantic relatedness measures in the light of Hindi Semantic Category Labeling.

3 Experimental Setting

In this section, we describe our method of labeling and various semantic relatedness mesures used.

3.1 Labeling Algorithm

We used a variation of simplified Lesk algorithm to label the semantic category for a given target word in a given context. Unlike simplified lesk algorithm which uses gloss overlap of target word's category and sentential context as a relatedness measure, our algorithm generalizes by using other semantic relatedness measures. It chooses the semantic category of the target word which is maximally related with its context. We used the immediate neighbours of the target word W_T, word to the left W_L and word to the right W_R, as context of the target word. The semantic category 'C' which maximizes the following equation is chosen to be the label of the target word.

$$\max_{C \in cat(W_T)} (leftRelatedness(C) + rightRelatedness(C))$$

where

cat(w) are the categories of the word 'w'

$$leftRelatedness(C) = \max_{L \in cat(W_L)} Rel(C, L)$$

[1] Hindi is the official language of India. Urdu is a close cousin to Hindi. Hindi and Urdu are spoken by approximately 500 million people in the world.

1. *kuwwe/Dog* (*Mammal*) ko *xeKawe/seeing* (*NaturalEvent*) hI **billA/cat** (*Mammal*) *pedZa/tree* (*NaturalObject*) para/on caDZa/climbed (*VerbOfAction*) gayA

2. *saBA/Meeting* (*Event*) meM/in *Aye/came* (*VerbOfAction*) saBI/all *svayaMsevaka/volunteers* (*Group*) **billA/badge** (*Artifact*) lagAye/wear (*VerbOfState*) hue We

Table 1: Examples showing the task of semantic category labeling. *wx-notation* is used here to write Hindi.

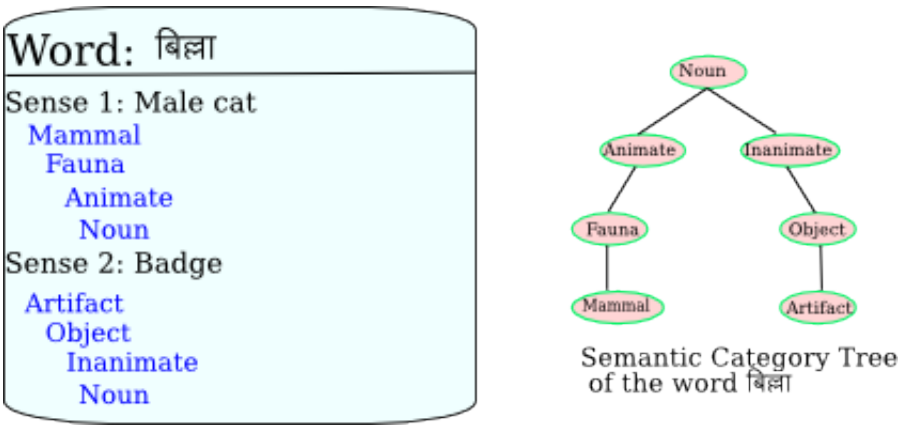

Figure 1: Hindi WordNet entry of the word *billA*. The word has two senses meaning male cat and badge. Ontological category mappings of the two senses are shown on the left side of the figure. On the right, the semantic category tree(SCT) of the word is shown.

$$rightRelatedness(C) = \max_{R \in cat(W_R)} Rel(C, R)$$

and *Rel() gives the semantic relatedness value between two categories measured using semantic relatedness measure*

The next section describes the semantic relatedness measures used by us.

3.2 Semantic Relatedness Measures

Semantic Relatedness Measure gives a metric to measure the relatedness of two concepts. A concept can either refer to a semantic category or a synset. We conducted our evaluation using the following Semantic relatedness metrics: Lesk (lesk), adapted Lesk (adpLesk), Leacock & Chodorow (lch), Wu & Palmer (wup), Lin (lin), and Jiang & Conrath (jcn). We provide below a short description for each of these six metrics.

We view gloss overlaps as just another measure of semantic relatedness. Simplified and adapted lesk relatedness measures are based on this assumption.

The Lesk Measure

Lesk relatedness between two concepts is the number of gloss overlaps of the two concepts. Hindi WordNet Ontological categories does not contain adequate gloss (and examples). To provide more gloss for an ontological category, we used the gloss of the synsets which correspond to this ontological category.

$$Rel_{lesk} = Overlap\ (\ gloss_{concept1},\ gloss_{concept2}\ )$$

The Adapted Lesk Measure

Adapted lesk relatedness between two concepts is defined as

$$Rel_{adpLesk} = Overlap\ (\ extendedGloss_{concept1},\ extendedGloss_{concept2}\ )$$

where extendedGloss of a concept is the total gloss of all the concepts in the hierarchy of the given concept (including itself).

The Leacock-Chodorow Measure

The (Leacock and Chodorow, 1998) relatedness between two concepts is determined as:

$$Rel_{lch} = -log\frac{length}{2D}$$

where length is the length of the shortest path between two concepts using node-counting, and D is the maximum depth of the hierarchy.

The Wu-Palmer Measure

The Wu and Palmer (Wu and Palmer, 1994) relatedness metric measures the depth of two given concepts in the WordNet taxonomy, and the depth of the least common subsumer (LCS), and combines these figures into a relatedness score:

$$Rel_{wup} = \frac{2 * depth(LCS)}{depth(concept_1) + depth(concept_2)}$$

The Lin Measure

The (Lin, 1998) relatedness between two concepts is defined as

$$Rel_{lin} = \frac{2 * IC(LCS)}{IC(concept_1) + IC(concept_2)}$$

where IC is defined as:

$$IC(c) = -logP(c)$$

and P (c) is the probability of encountering an instance of concept c in a large corpus. As we don't have a large sense tagged corpora, we calculated this probability by making the assumption done in (Patwardhan et al., 2003): each category of a word is equally likely. We used a Hindi web corpora of size 324 MB to collect these statistics.

The Jiang-Conrath Measure

The Jiang and Conrath (1997) Measure used by us is similar to the one used in (Patwardhan et al., 2003)

$$Rel_{jcn} = \frac{1}{IC(concept_1) + IC(concept_2) - 2 * IC(LCS)}$$

4 Evaluation

In our experiments, we labeled only nouns. We evaluated our experiments on manually annotated sense (semantic category and synset) tagged data developed by Indian language machine translation consortium(ILMT). It comprises of articles from news and tourism domain. In all, there are 7200 manual annotated sentences covering 133 semantic categories. The average semantic category ambiguity of a word is 2.18 i.e. on an average each word can have 2.18 semantic categories whereas synset ambiguity of a word is 2.57.

4.1 Results

The results of the experiment are shown in tables 2 and 3. In table 2, the accuracies for the task of semantic category labeling are shown and in table 3, the results for word sense disambiguation using synsets are shown. As we can see from the results the semantic categories are coarse grained and hence it turns out to be an easier task compared to synset assignment. This is expected because in a number of cases multiple synsets correspond to same semantic category in Hindi WordNet.

Also, the accuracies of baseline system, which assigns the first sense is considerably higher than others. WordNet senses are listed in the order of its frequency from which the sense inventory is created. Our testing corpus might have fallen along these lines of WordNet creation. This might be the reason of having higher accuracies for baseline. On a different corpus(domain), the first sense in WordNet might not be the frequent sense in the domain of interest. This effects the accuracy of the baseline system. This is not the case with the algorithms based on relatedness measures.

4.2 Observations

It is interesting to observe that adapted Lesk performs well on semantic category labeling than on synset labeling whereas Lesk performs well on synset labeling. This gives an insight that gloss information of Hindi WordNet ontological category is not sufficient for semantic category labeling and has to depend on the ontological heirarchy. In the case of synsets, the gloss information is adequate and the addition of hierarchial information may create noise.

Results show that lesk and adapted lesk are performing well (recall) compared to other semantic relatedness measures. This might be due to the reason that lesk and adapted lesk can relate two words across the syntactic categories (part-of-speech tags) which is not the case with other relatedness measures used in this paper.

Model	Precision	Recall
Baseline	84.76	84.76
lesk	74.75	72.73
adpLesk	76.05	74.09
lch	74.87	61.30
wup	75.33	61.68
lin	74.11	60.17
jcn	71.93	51.43

Table 2: Semantic Category Labeling Evaluation of Nouns

Model	Precision	Recall
Baseline	78.23	78.23
lesk	65.27	63.51
adpLesk	63.36	61.73
lch	67.14	54.98
wup	67.45	55.23
lin	65.05	52.81
jcn	62.52	44.70

Table 3: Synset Assignment Evaluation of Nouns

5 Conclusion

In this paper, we evaluated a number of semantic relatedness measures in the light of semantic category labeling. We hope these statistics will be helpful in providing insights for any work which aim to use semantic relatedness measures. For the task of semantic category labeling, the measure **adapted Lesk** performs better than all other measures.

Acknowledgements

We would like to thank Prof. Ted Pedersen for his valuable suggestions on semantic relatedness measures.

References

S. Banerjee and T. Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. *Computational Linguistics and Intelligent Text Processing*, pp. 117-171.

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *International Conference Research on Computational Linguistics (ROCLING X) (September 1997)*

C. Leacock and M. Chodorow. 1998. Combining local context and WordNet similarity for word sense identification. *WordNet: An electronic lexical database*, 49(2):265–283.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM New York, NY, USA.

D. Lin. 1998. An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pages 296–304.

D. Narayan, D. Chakrabarty, P. Pande, and P. Bhattacharyya. 2002. An experience in building the Indo WordNet-a WordNet for Hindi. In *International Conference on Global WordNet*.

S. Patwardhan, S. Banerjee, and T. Pedersen. 2003. Using measures of semantic relatedness for word sense disambiguation. *In Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics (CICLING-03) (2003)*

S. Reddy, A. Inumella, R. Sangal, and S. Paul. 2009. All Words Unsupervised Semantic Category Labeling for Hindi. In *Proceedings of Recent Advances in Natural Language Processing*.

Z. Wu and M. Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics Morristown, NJ, USA.