

Some experiments on Telugu

Siva Abhilash

Language Technologies Research Center
IIIT Hyderabad

August 27, 2008

Outline

- 1 WSD for telugu
 - Introduction
 - Resources
 - Approach
- 2 Building Wordnets
 - Introduction
 - Dutch Wordnet
- 3 Building Telugu Wordnet
 - Method
 - Collecting seeds
 - Using Search engine
 - Using Wikipedia

Outline

- 1 WSD for telugu
 - Introduction
 - Resources
 - Approach
- 2 Building Wordnets
 - Introduction
 - Dutch Wordnet
- 3 Building Telugu Wordnet
 - Method
 - Collecting seeds
 - Using Search engine
 - Using Wikipedia

Outline

- 1 WSD for telugu
 - Introduction
 - Resources
 - Approach
- 2 Building Wordnets
 - Introduction
 - Dutch Wordnet
- 3 Building Telugu Wordnet
 - Method
 - Collecting seeds
 - Using Search engine
 - Using Wikipedia

WSD for telugu

Definition

Given a word, its context and its possible senses, the aim of WSD is to determine the sense of the word in that context

Problem

Senses of a word can be known from wordnet. But Telugu does not have a wordnet.

WSD for telugu

Definition

Given a word, its context and its possible senses, the aim of WSD is to determine the sense of the word in that context

Problem

Senses of a word can be known from wordnet. But Telugu does not have a wordnet.

Resources Available

1 Resources

- Hindi Wordnet
- Hindi-Telugu Dictionary
- Telugu Corpus

2 Telugu-Hindi dictionary is built from Hindi-Telugu dict

Resources Available

- 1 Resources
 - Hindi Wordnet
 - Hindi-Telugu Dictionary
 - Telugu Corpus
- 2 Telugu-Hindi dictionary is built from Hindi-Telugu dict

Senses of a Telugu word w

- 1 Get the Hindi Translations of w using Telugu-Hindi dict.
- 2 For each hindi translation, get all its ontological categories from Hindi Wordnet.
- 3 Senses of w are subset of above senses.
- 4 Since we don't know which senses correspond to w , we use all the senses as senses of w

Senses of a Telugu word w

- 1 Get the Hindi Translations of w using Telugu-Hindi dict.
- 2 For each hindi translation, get all its ontological categories from Hindi Wordnet.
- 3 Senses of w are subset of above senses.
- 4 Since we don't know which senses correspond to w , we use all the senses as senses of w

Senses of a Telugu word w

- 1 Get the Hindi Translations of w using Telugu-Hindi dict.
- 2 For each hindi translation, get all its ontological categories from Hindi Wordnet.
- 3 Senses of w are subset of above senses.
- 4 Since we don't know which senses correspond to w , we use all the senses as senses of w

Approaches

Using the above senses of w , we can disambiguate w using the following approaches

Current existing algorithms

Yarowsky (1992)

Dekang Lin (1997)

Our approaches

Expectation Maximization

Path based approaches

Approaches

Using the above senses of w , we can disambiguate w using the following approaches

Current existing algorithms

Yarowsky (1992)

Dekang Lin (1997)

Our approaches

Expectation Maximization

Path based approaches

Approaches

Using the above senses of w , we can disambiguate w using the following approaches

Current existing algorithms

Yarowsky (1992)

Dekang Lin (1997)

Our approaches

Expectation Maximization

Path based approaches

Building Wordnets

- Hand Crafted Wordnets incur a lot of time and cost for building. This can be overcome by building automatic wordnets.
- We consider Hypernymy-Hyponymy relations only.
- The methods for building automatic wordnets employ pattern based approaches. They generally use fixed syntactic patterns (Hearst 1992)
- Some of the earlier attempts are Pasca 2004, Snow 2005 and Kim 2007.

Building Wordnets

- Hand Crafted Wordnets incur a lot of time and cost for building. This can be overcome by building automatic wordnets.
- We consider Hypernymy-Hyponymy relations only.
- The methods for building automatic wordnets employ pattern based approaches. They generally use fixed syntactic patterns (Hearst 1992)
- Some of the earlier attempts are Pasca 2004, Snow 2005 and Kim 2007.

Extending Dutch Wordnet, Kim 2007

For every noun pair N1 and N2, queries are generated and are given to a search engine.

Query formats

N1 infix N2

prefix N1 infix N2

prefix N1 infix N2 suffix

N1 infix N2 suffix

Definition

Pattern is Prefix * Infix * Suffix.

Extending Dutch Wordnet, Kim 2007

For every noun pair N1 and N2, queries are generated and are given to a search engine.

Query formats

N1 infix N2

prefix N1 infix N2

prefix N1 infix N2 suffix

N1 infix N2 suffix

Definition

Pattern is Prefix * Infix * Suffix.

Extending Dutch Wordnet, Kim 2007

For every noun pair N1 and N2, queries are generated and are given to a search engine.

Query formats

N1 infix N2

prefix N1 infix N2

prefix N1 infix N2 suffix

N1 infix N2 suffix

Definition

Pattern is Prefix * Infix * Suffix.

Approach, Kim 2007

- They got around 32,83,492 unique patterns.
- These patterns are validated with already existing hypernymy-hyponymy pairs in DWN.
- Finally 1,18,306 patterns are selected which have recall greater than 0.
- The noun pairs which have a pattern among selected patterns are considered for hyper-hypo relation.
- Each noun pair is scored using hypernymy evidence score.

Hypernymy evidence score

$$\text{score}(h, w) = \frac{\text{freq}(h, w)}{\text{freq}(w)} + \sum_c \frac{\text{freq}(c, w)}{\sum_y \text{freq}(y, w)}$$

Building Telugu WN using Web as corpus

Resources

- Search engine (IIIT Webkhoj)
- English/Hindi Wordnet
- English/Hindi - Telugu dictionary

Definition

Seed is valid hypernymy-hyponymy pair.

Building Telugu WN using Web as corpus

Resources

- Search engine (IIIT Webkhoj)
- English/Hindi Wordnet
- English/Hindi - Telugu dictionary

Definition

Seed is valid hypernymy-hyponymy pair.

Approach

Steps

- 1 Collecting intial seeds
- 2 Validating and Extending seeds

Collecting Intial Seeds

- Using Wordnets and dictionary
- Using Linguistic Features

Approach

Steps

- 1 Collecting intial seeds
- 2 Validating and Extending seeds

Collecting Intial Seeds

- Using Wordnets and dictionary
- Using Linguistic Features

Using Telugu-Hindi Dict and Hindi WN

- For every noun pair N1 and N2 in telugu, get Hindi translation pairs.
- If one of the Hindi pairs has a hyper-hypo relation in Hindi WN, then N1 and N2 are more likely to have hyper-hypo relation.

Pros

We get a single seed pair each time
Easy to validate a pair.

Cons

Time complexity is high.

Using Telugu-Hindi Dict and Hindi WN

- For every noun pair N1 and N2 in telugu, get Hindi translation pairs.
- If one of the Hindi pairs has a hyper-hypo relation in Hindi WN, then N1 and N2 are more likely to have hyper-hypo relation.

Pros

We get a single seed pair each time
Easy to validate a pair.

Cons

Time complexity is high.

Using Telugu-Hindi Dict and Hindi WN

- For every noun pair N1 and N2 in telugu, get Hindi translation pairs.
- If one of the Hindi pairs has a hyper-hypo relation in Hindi WN, then N1 and N2 are more likely to have hyper-hypo relation.

Pros

We get a single seed pair each time
Easy to validate a pair.

Cons

Time complexity is high.

Using Hindi WN and Hindi-Telugu Dict

- For every hyper-hypo pair in Hindi WN, get the telugu translation pairs.
- Some of the telugu pairs are valid hyper-hypo pairs. Since we don't know which of them, we have to validate them.
- We got around 2,30,000 pairs in telugu. Using skip bigrams, we took only the pairs which are commonly used in the language

Pros

Less time complexity since the number of pairs generated are less

Cons

We get a candidate set in which solution exists.
So it is difficult to obtain a seed.

Some pairs

Example

swrl \Leftarrow BArya

sWalaM \Leftarrow BUmi

saMKya \Leftarrow Aru

xuMpa \Leftarrow ullipAya

Using Linguistic Features

Conjunctive construction can contain hyper-hypo relations.

- We extracted Conjunctive constructions with patterns **NN, NN,NN adj NN** and **NN adj NN, NN, NN** on 30 Mb corpus.
- Manually analyzed 5000 of these constructions. All the constructions does not have hyper-hypo relation.
- Most of the time we found the construction **NN1_NULL, NN2_NULL, NN3_NULL adj NN4** is having hyper-hypo relation. (NN1, NN2, NN3 are siblings and NN4 is the parent)
- The adj in the conjunctive construction is a strong cue for hyper-hypo relation.

Example

poVtla, xosa, xoVMda moVxalEna kUragAyalu
jalubu, xaggu, notipUwa vaMti vyAxulu
mAtallonU, ceRtallonU, veRaMlonU ituvaMti vyawyAsAlu

Connectors

moVxalEna
moVxalagu
vaMti
moVxalayina
oVka
IAMti

Validating and Extending Seeds and Patterns

Definition

Pattern is a regular expression of the syntactic structure of hyper-hypo construction. Eg: Conjunctive construction, NN vaMti NN

- From seeds, news patterns are generated. Patterns are generated using Kim 2007 queries
- From patterns, new seeds are generated.
- Both seeds and patterns are scored at each iteration.
- In this way best seeds and patterns are generated at each iteration.

Say about scoring Mechanisms

Using Wikipedia

Most hierachies of Wikipedia have hyper-hypo relation. To find such hierarchies, we can use the seeds and thus extend the seeds. Wikipedias structure.

Hierarchy of Vegetables

Potato
Tomato
Bringal
Ladies finger

Heirarchy of China Gold medals

In swimming 5
In athletics 10
In boxing 2