Word N-Gram Based Approach for Word Sense Disambiguation in Telugu Natural Language Processing

Palanati Durga Prasad, K.V.N.Sunitha, B.Padmaja Rani

Abstract— Telugu (తెలుగు) is one of the Dravidian languages which is morphologically rich. As in the other languages it too contains polysemous words which have different meanings in different contexts. There are several language models exist to solve the word sense disambiguation problem with respect to each language like English, Chinese, Hindi and Kannada etc. The proposed method gives a solution for the word sense disambiguation problem with the help of n-gram technique which has given good results in many other languages. The methodology mentioned in this paper finds the co-occurrence words of target polysemous word and we call them as n-grams. A Telugu corpus sent as input for training phase to find n-gram joint probabilities. By considering these joint probabilities the target polysemous word will be assigned a correct sense in testing phase. We evaluate the proposed method on some polysemous Telugu nouns and verbs. The methodology proposed gives the F-measure 0.94 when tested on Telugu corpus collected from CIIL, various news papers and story books .The present methodology can give better results with increase in size of training corpus and in future we plan to evaluate it on all words not only nouns and verbs.

Index Terms— Joint probabilities, Machine translation, n-grams, Word Sense Disambiguation.

1. INTRODUCTION

Machine translation (MT) is playing an ample role in the present digitization era. Translation of text from one language to another is one problem of machine translation. The field of Natural Language Processing (NLP) being an impetus for machine translation models can have machine readable text or speech as input or output. In the case of text translation it can be done by fully understanding the text both semantically and syntactically. Major challenge in designing a translation model is understanding the semantics of the text. If a word considered one at a time in a sentence of source language text sometimes may have multiple mappings (meanings) in the target language. The number of mappings may be reduced and can have a unique mapping if some adjacent words are considered. A process of unique mapping is called resolving the ambiguity is the concept widely studied in NLP. Ambiguity may be at different levels like lexical (word) level and syntactical (parsing) level. Resolving the ambiguity at lexical level leads to the concept of Word Sense Disambiguation (WSD).Approaches of WSD assign a correct meaning to the target word by considering the context. Consider the following Telugu sentence in the given example:

పట్టు పురుగు జీవితం ఐదు స్థాయిల్లో జరుగుతుంది.

పట్టు	పురుగు	జీవితం	ఐదు	స్థాయిల్లో	జరుగుతుంది.
↓	↓	↓	↓	↓	↓
SILK or GRIP	INSECT	LIFE	FIVE	IN STAGES	SPANS.

Example 1.Mapping of word from Telugu to English

In the above example the word 'పట్టు' in Telugu language has several mappings in English language i.e., SILK, HIVE, or GRIP. In this case the decision of mapping to 'SILK' is more appropriate when considering the surrounding words INSECT, LIFE. By resolving these types of lexical level ambiguities WSD helps in producing accurate results by several language translation models. So the data driven approach like n-gram translation model which can produce probabilistic transformation rules is used in this paper for WSD. This paper details an approach that can improve accuracy of a translation model. The related work in this field is discussed in section II followed by methodology in section III, experimental work in section IV , results in section V, conclusion and future scope in section VI.

II. RELATED WORK

The process of word sense disambiguation as part of machine translation starts with dividing the text into chunks. Later various approaches are applied to get correct sense of the ambiguous word present in the given text. The word sense disambiguation problem can be solved using supervised, unsupervised and semi-supervised approaches. In [1] unsupervised knowledge-free WSD method proposed which solves WSD problem in five steps like extracting of context features, computing feature similarities, word sense induction, labeling induced senses and disambiguation in context. Sense of ambiguous word depends on the statistics of the words occurring before and after it. .Many language

Revised Version Manuscript Received on April 05, 2019.

Palanati Durga Prasad , Academic consultant, CSE Department, Ucet, Mg University, NALGONDA,Telangana, India
Dr.K.V.N.Sunitha, Professor, Cse Department, BVRIT, JNTUH, Hyderabad, Telangana, India
Dr.B.Padmaja Rani, Professor,CSE Department, JNTUH, Hyderabad, Telangana, India

models are using those statistics of probable co-occurrence words derived by applying conditional probabilities. Thus use of n-grams helps in disambiguation of polysemous words. In natural language processing and information retrieval systems n-grams find use in many problems like automated text summarization, text categorization, spelling corrections, auto-fill of text and speech recognition. Performance of MT system is improved by integrating WSD system as proved in [2] and there is a statistically significant improvement in the translation performance of a state-of-the-art hierarchical phrase-based statistical MT system. Uni-gram based automatic evaluation of summaries[3] is somewhat correlated to human evaluation but not 100%.The system described in [4] for lexical disambiguation using n-gram web scale models states that the corpus used for training set must give more information about the words i.e., co-occurring words, and embedded words etc..Lexical attraction which shows the likelihood of two words occurring in a sentence helps in analyzing the meaning of a polysemous word. The concept of creating distributional semantic models means word vectors proposed in [5] for highly inflected languages increased the accuracy largely around 11%. The word vectors designed in [5] uses skip-grams for better results. The n-gram feature is used in [6] for text classification experiments. The unigram and bigrams together with lexicon were used as baseline in the system developed in [6] for twitter sentiment analysis gave better performance. Hybrid Machine Translation approach used in [7] determined the correct meaning of Hindi word with respect its context using N-gram approach. In [7] various categories of bigrams and trigrams were used to find the appropriate sense of the ambiguous word and accuracy of this system is found out to be 87.60%. The language model presented in [8] a joint probability model for statistical machine translation, which automatically learns word and phrase equivalents from bilingual corpora. A Tri-gram based model proposed in [9] has given excellent results for POS tagging in Marathi language. When given the previous two tags the proposed model in [9] chooses appropriate tag for the target word with an accuracy of 91.63%.The deleted interpolation method used in WSD algorithm [10] for disambiguating Hindi words combines different n-gram orders and this algorithm works with an accuracy between 60% to 70%.Two WSD methods proposed in [11] which are based on context expansion. One of the methods in [11] considers synonyms as contextual feature to train the Bayesian model. As this method produces some noise in the machine generated training data containing synonyms another method was proposed. In the second method the machine generated synonym set treated as pseudo training data combining with authentic training data which is used to train the Bayesian classifier. In [12] the Ordered Weighted Averaging (OWA) operators used for query expansion by assigning proper sense for the polysemous word present in the query. The method proposed in [12] leads to retrieval of more related documents by query expansion. This query expansion is done by replacing the polysemous word in the query by another word (with sense having highest similarity score generated by OWA operator).A skip-gram based model proposed in [13] develops a semantic network with linkages between words with nearby senses. This model applied to SWEDISH data set in which it finds the conditional probability of sense and context. In this approach instead of ranking the probabilities the senses are ranked and highest score sense is selected.

III. N-GRAM BASED TECHNIQUE FOR WSD

A. Methodology

The goal of n-gram based WSD is to find the sense of given ambiguous word that has highest joint probability with its co-occurring words collected using uni-gram, bi-gram and tri-gram approaches in the training corpus. Our system consists the following phases:

- **Training phase**: Training the corpus (training set) to collect n-gram probabilities by varying the size of n.
- **Test phase**: Test on the input data set.

Algorithm in training phase is as below:

Step 1: Tokenize the data set.

Step 2: Remove the stop words.

Step 3: Find the target polysemous word.

Step 4: Collect the uni grams, bi-grams and tri-grams (the system is trained using these n-grams).

Step 5: Find Joint probabilities of bi-grams and tri-grams in which the target polysemous word exists.(use the equation 1).
Bigram joint probability

$$P\left(\frac{W_n}{W_{n-1}}\right) = \frac{frequncy\,(W_{n-1}W_n)}{frequncy\,(W_{n-1})} \qquad (1)$$

P-joint probability
w_n - target polysemous word
w_{n-1} – immediate word left to target polysemous word

Step 6: Store the probabilities in a list. Name it as Joint Probability of Ambiguous Word (JPAW).

Algorithm in testing phase is as below:

Step 1: Tokenize the data set.

Step 2: Remove the stop words.

Step 3: Find the target polysemous word.

Step 4: Choose any of the uni- grams, bi-grams and tri-grams model (using which the system is trained).

Step 5: Search for the bi-grams or tri-grams in which the target polysemous word exists.

Step 6: Search for the highest joint probability in the list JPAW (which is obtained in training phase).

Step 7: Fetch the sense assigned to that probability.

Step 8: Assign the sense obtained in Step 7 to the target polysemous word.

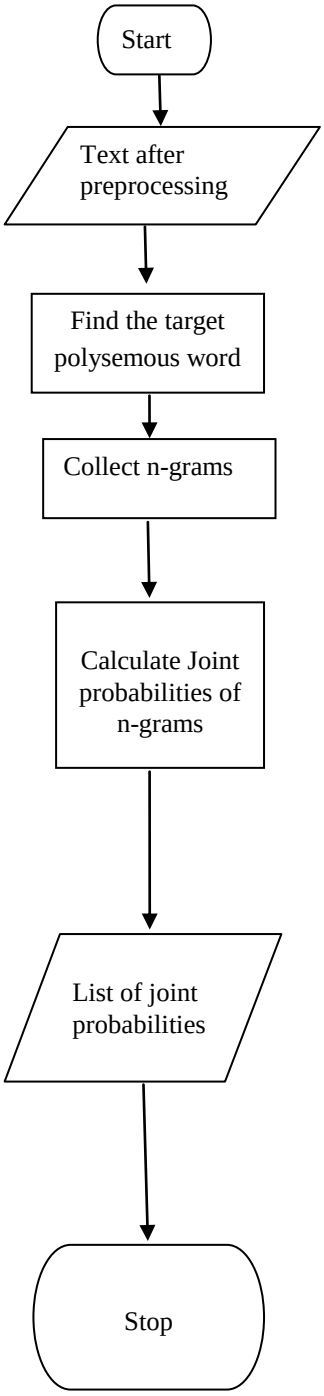
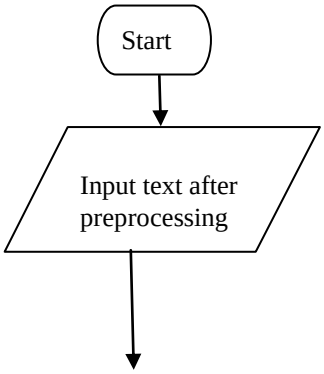


Figure 1.Flow chart of Training phase process



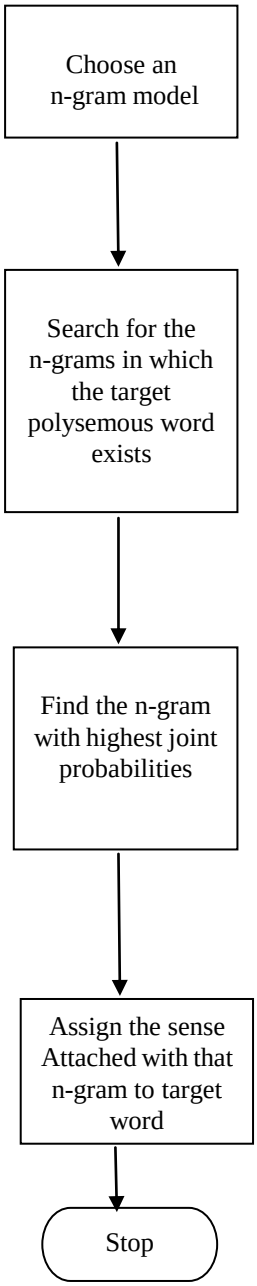


Figure 2.Flow chart of Testing phase process

IV.EXPERIMENTAL WORK

The below tables I, II, III show the results during training phase where the target polysemous Telugu word is 'కొట్టు' with the sense meaning 'shop'. The frequency (f) is used to find the joint probabilities of target polysemous word with the context.

Table I. Unigram Database

Unigram	Frequency (f)
కిరాణా	15
బట్టల	4

పచారీ	4
కొనడానికి	3
మూస్తారు	2

Table II. Bigram Database

Bigram	Frequency (f)
కిరాణా కొట్టు	14
బట్టల కొట్టు	4
పచారీ కొట్టు	3
కొనడానికి కొట్టు	2

Table III. Trigram Database

Trigram	Frequency(f)
పంచదార కొనడానికి కొట్టు	2
చిన్న పచారీ కొట్టు	2

Conditional probability using uni-gram and bi-grams.

P(కొట్టు|కిరాణా)=f(కిరాణా కొట్టు)|f కిరాణా)=14|15=0.93

Conditional probability using bi-gram and tri-grams.

P (కొట్టు | పంచదార కొనడానికి) =f (పంచదార కొనడానికి కొట్టు) | f (పంచదార కొనడానికి) =2|2=1.0000

Example:

పండగల వేళ బట్టల కొట్టు రద్దీగా ఉంటుంది.

Festivals during Cloth Shop crowded is

After rearranging the sentence: **During festivals cloth shop is crowded**.

Implementation of WSD algorithm for above example using bi-gram:

Step 1: Input the sentence

పండగల వేళ బట్టల కొట్టు రద్దీగా ఉంటుంది.

Step 2: Tokenize the sentence (after removing special characters like periods, commas, etc.)

Tokens= {పండగల, వేళ, బట్టల, కొట్టు, రద్దీగా, ఉంటుంది }

Step 3: Remove stop words (Stop words for example ఈ, అనే, అంటే,ఉన్న,,ఓ)

Tokens={ పండగల, వేళ , బట్టల, కొట్టు, రద్దీగా, ఉంటుంది}

Step 4: Find ambiguous word

కొట్టు

Step 5: Find the frequency of unigrams, bigrams and trigrams (bigrams and trigrams that contain ambiguous word).

Uni-grams: **పండగల, వేళ , బట్టల, కొట్టు , రద్దీగా, ఉంటుంది**

Bi-grams: **(పండగల, వేళ),(వేళ , బట్టల),(బట్టల, కొట్టు), (కొట్టు , రద్దీగా),(రద్దీగా, ఉంటుంది)**

Tri-grams: **(పండగల, వేళ , బట్టల),(వేళ , బట్టల, కొట్టు), (బట్టల, కొట్టు , రద్దీగా),(కొట్టు , రద్దీగా, ఉంటుంది)**

Examples are shown in Table 1, 2, 3.

Step 6: Find the joint probability (using equation 1) of ambiguous word with the help of bigram and unigram

Bigram ⟶ (బట్టల, కొట్టు)

P (కొట్టు | బట్టల) =f(బట్టల, కొట్టు) |f(బట్టల) =4|4=1.000

The above frequencies are from the unigram, bigram and trigram databases created during training phase.

Step 7: Fetch the sense related to the n-gram and assign it to the ambiguous word (shown in Table IV).

Table IV. Output

Bigram	Ambiguous word	Sense
బట్టల, కొట్టు	**కొట్టు**	**Shop**

V. RESULTS

To evaluate our system a Telugu corpus collected from CIIL, newspapers and some story books. This corpus is the input for training phase. The output of training phase consists the joint probabilities of ambiguous words and context. These probabilities are stored in a list called JPAW in descending order. A sentence is sent for testing phase algorithm as input.

We have considered 150 Telugu polysemous words. These belong to different parts-of-speech like nouns, verbs, adjective etc. Even though a word belongs to one parts-of-speech it may have different senses (meanings) which are the WSD task.

Table V. Some example Telugu polysemy words.

Polysemous word	Meanings
అంబరము	ఆకాశము/వ్యసనము/వస్త్రము/ పరిమళద్రవ్యము
అనువు	అనుకూలము/తీర్పు/అవకాశము, ఉపాయము/విధము
తీర్థము	రేవు/పుణ్యక్షేత్రము/పవిత్రజలము/అగ్ని
వర్షము	సంవత్సరము /వాన/ మబ్బు/జంభూదీపము

The training phase output statistics of our system are shown in Table VI and VII.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

Table VI: Training phase statistics

Number of Telugu Ploysemy words	Sense per word (on an average)	Training corpus size (excluding stop words)
150	4	30000 words(Approximately)

Table VII: n-grams statistics

Uni-grams	30000
Bi-grams	29999
Tri-grams	29998

The testing phase output statistics of our system are shown in Table VIII.

Table VIII: Metrics

Number of Telugu Ploysemous words	Precision (P)*	Recall (R) **	F-Measure $2xPxR \div (P+R)$
150	1.0	0.88	0.94

*Precision = (number of correctly disambiguated words/number of disambiguated Words)
**Recall = (number of correctly disambiguated words/ number of tested set words)

VI. CONCLUSION AND FUTURE SCOPE

Word sense disambiguation (WSD) task is place vital role in machine translation. While translating one word of one language to another language it may have various meanings. For example, Telugu word 'చరణము' translated into English as 'leg' or 'line of a poem'. This problem can be best solved by considering the surrounding context which is done in the present work using n-grams. The polysemous word is assigned a correct sense based on the highest joint probability using bi-gram and tri grams in our work. Due to limited machine readable corpus in Telugu language uni-grams alone are not able to solve WSD problem using our algorithm which is one limitation of our work. Only few highly polysemeous Telugu words are considered presently. We plan to improve the number of polysemy words as the collection of Telugu corpus increases. The algorithm can be applied for best results after deciding the optimal size of n-gram window which is our future task. The performance can be improved by handling morphological inflections correctly and by considering more glosses for the ambiguous words.

REFERENCE

1 Alexander Panchenko et al.," Unsupervised Does Not Mean Uninterpretable: The Case for Word Sense Induction and Disambiguation", Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 86–98, Valencia, Spain, April 3-7, 2017
2 Yee Seng Chan, Hwee Tou Ng and David Chiang "Word Sense Disambiguation Improves Statistical Machine Translation" Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics pp. 33-40, 2007. (Book style with paper title and editor)
3 Chin-Yew Lin, "Automatic Evaluation of Summaries Using N-gram Co-Occurrence Statistics". pp. 71-78, Proceedings of HLT-NAACL 2003, Edmonton, May-June 2003.
4 Shane Bergsma, "Web-Scale N-gram Models for Lexical Disambiguation", pp 1507-1512, Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence, IJCAI-09.
5 Pranjal Singh, Amitabha Mukerjee," Words are not Equal: GradedWeighting Model for building Composite Document Vectors", Proc. of the 12th Intl. Conference on Natural Language Processing, pages 11–19, Trivandrum, India. December 2015.
6 Efthymios Kouloumpis et al.,"Twitter Sentiment Analysis: The Good and Bad and the OMG!"
7 Vishal Goyal, Gurpreet Singh Lehal,"Hindi to Punjabi Machine Translation System", Proceedings of the Association for Computational Linguistics System Demonstrations, pages 1 –6, Portland, USA, 21 June 2011.
8 Daniel Marcu and William Wong, "A Phrase-base Joint Probability Model for Statistical Machin Translation", Proceedings of EMNLP, 2002.
9 Jyothi et al.," Parts of speech tagging of marathi text using trigram method", International Journal of Advanced Information Technology (IJAIT) Vol. 3, No.2, April 2013.
10 UmrinderPal Singh et al.," Disambiguating Hindi Words Using N-Gram Smoothing Models", An International Journal of Engineering Sciences, Issue June 2014, Vol. 10 ISSN: 2229-6913
11 Zhizhuo Yang, Heyan Huang, "Chinese Word Sense Disambiguation based on Context Expansion" Proceedings of COLING 2012: Posters, pages 1401–1408, COLING 2012, Mumbai, December 2012
12 Kanika Mittal, Amita Jain, "Word sense disambiguation method using semantic similarity measures and owa operator", ICTACT journal on soft computing, ISSN: 2229-6956 (ONLINE) Volume: 05, Issue: 02, January 2015
13 Richard Johansson, Luis Nieto Pi ˜na," Combining Relational and Distributional Knowledge For Word Sense Disambiguation", pages 69-78,Proceedings of the 20th Nordic Conference of Computational Linguistics,NODALIDA,2015