International Conference on Machine Learning and Computational Intelligence-2017
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2017 IJSRCSEIT | Volume 2 | Issue 7 | ISSN : 2456-3307

Sequence Labeling for Two Word Disambiguation in Telugu Language Sentences

Jinka Sreedhar*[1], A. Jagan[1], SK Althaf Hussain Basha[2] , Baijnath Kaushik[3] , D. Praveen Kumar[4]

[1]Department of Computer Science and Engineering B.V.Raju Institute of Technology, Narsapur, Telangana, India

[2]Department of Computer Science and Engineering Gokaraju Rangaraju Institute of Engineering and Technoloy,Hyderabad, India

[3]Department of Computer Science and Engineering Shri Mata Vaishno Devi University, Katra, Jammu and Kashmir, India

[4]Department of Computer Science and Engineering Institute of Technology, Dhanbad, Jharkhand, India

ABSTRACT

This paper is intended to apply sequence labelings which are introduced to find out the ambiguity in two-words. These words appear to give rise to ambiguity. They seem to be sequence words and this method can be applied only to these types of words. There is another theory of automata which is a mathematical model. By implementing this model to disambiguate the words of sequence it is found there is a kind of mathematical accuracy equal to that of sequence labeling method. The main aim of finding out these methods, is to find out solution to the problem of ambiguity in two-words sequences. Here designing of automata theory for two-words is dealt with the Two-Words disambiguation rules are explained with examples.

Keywords : Natural Language Processing(NLP),Information Retrieval Systems(IRS), MachineTranslation(MT), Finite Automata(FA), Two-Words disambiguation rules.

I. INTRODUCTION

To explain this theory clearly, four states have been identified. In this process state one may have more than one tag, state two may have more than one tag. Now one tag has been retained in the word one deleting the remaining tags. In the similar manner, the same procedure is continued in the second word order[1,2,3,4]. By doing so, the problem of ambiguity has been resolved. When this process comes to state three, it is treated as completed since it gives a complete sense. In the fourth state regarded as a dead state, all the unwanted tags will be appear[5,6,7,8].

With the help of transitional diagrams and transitional tables, the rules are explained. Transitional diagrams contain states, Parts-of-Speech(POS) tags, start state and final state[5]. These diagrams can be represented with the symbols like Q, ∑, S, F. Here Q stands for states one, two, three and dead states, ∑ contains POS tags, S contains the starting state, that is, state one, and F contains the final state, that is, state three[13,14,15,16,17,18].

Transitional Table is also framed to show how these tags appear in different states and give a picture representation.

W1 :: W2 => W1 :: W2

Where

W1 and W2 are sequence of words in that order.

II. DESIGNING AUTOMATA THEORY FOR TWO WORDS RULES

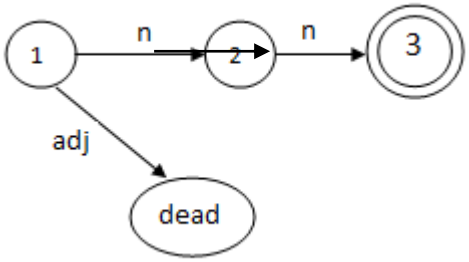

Figure 2.1. Two-word disambiguation for n, adj :: n

Where
1, 2, 3 and dead state belongs to Q and n, adj belongs to ∑.
Here n denotes noun and adj denotes adjective.
Q: {1, 2, 3, dead}
∑: {n,adj}
S: {1}
F: {3}

Table 2.1. Two-word disambiguation for n, adj :: n

∂	n	adj
1	2	dead
2	3	-
3	-	-
dead	-	-

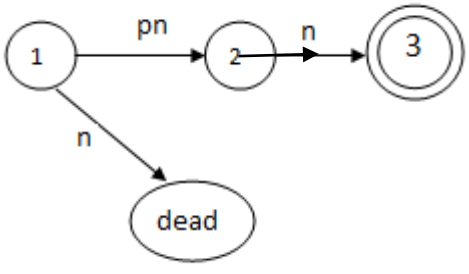

Figure 2.1. Two-word disambiguation for n, pn :: n
Where
1, 2, 3 and dead state belong to Q and n, pn belongs to ∑.
Here n denotes noun and pn denotes pronoun.
Q: {1,2,3,dead}
∑: {pn,n}
S:{1}
F:{3}

Table 2.2. Two-word disambiguation for n, pn :: n

∂	pn	N
1	2	Dead
2	-	3
3	-	-
dead	-	-

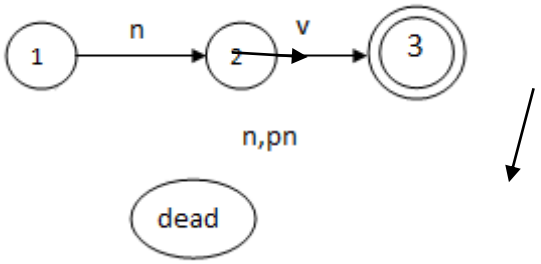

Figure 2.3. Two-word disambiguation for n :: v, n, pn

Where
1, 2, 3 and dead state belong to Q and n, v, pn belongs to ∑.
Here n denotes noun, v denotes verb and pn denotes pronoun.
Q: {1,2,3,dead}
∑: {n,v,pn}
S: {1}
F: {3}

Table 2.3. Two-word disambiguation for n :: v, n, pn

∂	n	v	pn
1	2	-	-
2	dead	3	dead
3	-	-	-
dead	-	-	-

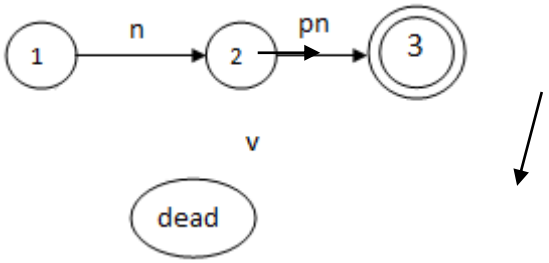

Figure 2.4. Two-word disambiguation for n :: pn, v

Where
1, 2, 3 and dead state belong to Q and n, v, pn belongs to ∑.
Here n denotes noun, v denotes verb and pn denotes pronoun.

Q: {1,2,3,,dead}
∑: {n,v,pn}
S: {1}
F: {3}

Table 2.4. Two-word disambiguation for n :: pn, v

∂	n	pn	v
1	2	-	-
2	-	3	dead
3	-	-	-
dead	-	-	-

Table 2.5. WSD Two-Word Rules with Sentence id's in the Telugu Corpus

S.NO	SENTENCE ID	BEFORE DISAMBIGUATION RULE	AFTER DISAMBIGUATION RULE (RESULT)
1	14784	n,adj :: n => n :: n	n :: n
2	274	n,pn :: n =>pn :: n	pn :: n
3	153	n :: n,pn,v => n :: v	n :: v
4	2291	n :: v,pn => n :: pn	n :: pn
5	10349	avy :: v,pn => avy :: v	avy :: v
6	21560	v ,pn :: avy => v :: avy	v :: avy
7	16646	v,n :: n => n :: n	n :: n
8	24355	n :: n,v => n :: v	n :: v
9	13677	v,pn :: avy =>pn :: avy	pn :: avy
10	442	n :: v,n,pn =>n :: pn	n :: pn
11	531	n :: v,pn => n :: v	n :: v
12	4552	n :: v,pn => n :: pn	n :: pn
13	25974	n :: v,n => n :: n	n :: n
14	12455	pn :: v,pn => pn :: pn	pn :: pn
15	656	avy :: v,pn => avy :: v	avy :: v
16	1893	pn,v :: v => pn :: v	pn :: v
17	590	pn :: adj,n => pn :: n	pn :: n
18	560	n :: v,pn => n :: v	n :: v
19	18714	n,adj :: n => adj :: n	adj :: n

Here n is noun, v is verb, pn is pronoun, adj is adjective and adv is adverb.

From rule 2 when a word carries tags (n,pn) and is followed by another word carrying the tag n, then the tag pn is retained eliminating the n from (n,pn).

From rule 9 a word carrying the tag, such as(n,pn) followed by avy, then most of the times pn will be retained and v will be eliminated. Depending on the context, the linguist will decide which tag will be retained and which one has to be eliminated. These are mostly contextually based syntactic rules. If two-word sequences are unable to resolve unique tags, then three-word, four-word sequence rules may be used for disambiguation.

III. CASE STUDY FOR TWO WORD AMBIGUITY

A Telugu sentence may have ambiguous words from Telugu corpus, like

Sentence: Adaxi aNacivewaku alavAtu padipoyiMxi.

MORPH OUTPUT:

Adaxi	Ada /adj,n
aNacivewaku	aNacivewa/n
alavAtu	alavAtu /n
padipoyiMxi	padu/v,adv,pn,n

Before Applying Disambiguation Rule:

W1 = Ada
W2 = aNacivewa

w1	:: w2	=>	w1 :: w2
n,adj	:: n	=>	n :: n

In the given Telugu sentence the word carries tags (n,adj) and is followed by another word carrying the tag n. Then the tag adj is retained eliminating the n from (n,adj), so from the above sentence adj is eliminated and n is retained.

After Applying Disambiguation Rule:

Adaxi	aNacivewaku	alavAtu	padipoyiMxi	.
n	n	n	v	
			punc	

Where punc is punctuation.

IV. ANALYSIS OF TWO WORD DISAMBIGUATION

The following figure 4.1 gives an analysis of the Accuracy. While X-axis indicates the number of test sessions, Y-axis indicates the Accuracy. As a result, the proposed method can disambiguate nearly 98% of ambiguity [59].

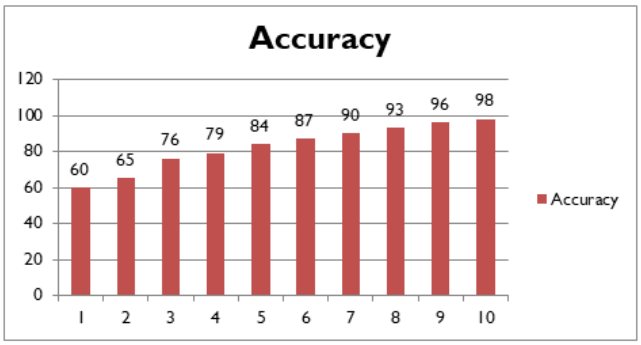

Figure 4.1. Two word disambiguation rules accuracy

V. CONCLUSION

Here dealing with the designing of two-word rules for Telugu Language Sentence word order. To make things easy to understand some rules have been made which can be applied for the word order of Telugu Language Sentences and it clarifies the ambiguity. All these things are so vividly explained with the help of case studies and theoretical explanations. When these rules are applied whenever needed, they help the user easily to eliminate the ambiguity. These theories help understand the study of disambiguation. By applying disambiguation rules it is found that the proposed method can disambiguate nearly 98% of the ambiguity. The theoretical explanation and disambiguation rules have resulted in the accuracy of evidences.

VI. REFERENCES

[1]. Noam Chomsky, "Logical syntax and semantics: their linguistic relevance", vol.31, No.1, pp: 36-45, 1955.

[2]. Noam Chomsky, "On Certain Formal Properties of Grammars", Information and Control, Vol. 9, pp: 137-167, 1959.

[3]. Nou, Chenda and WataruKameyama,Khmer, "POS Tagger: A Transformation-based Approach with Hybrid Unknown Word Handling", Proceedings of the First IEEE International Conference on Semantic Computing (ISCS), Irvine, CA. pp: 482-492, 2007.

[4]. PawanGoyal,LaxmidharBehera,Thomas Martin McGinnity, "Query Representation through Lexical Association for Information Retrieval", IEEE Transactions on Knowledge and Data Engineering, pp: 2260-2273, 2012.

[5]. PengJin,XingyuanChen,"A Word Sense Probabilistic Topic Model", 9th International Conference on Computational Intelligence and Security (CIS), pp: 401-404, 2013.

[6]. PengYuan Liu, "Another View of the Features in Supervised Chinese Word Sense Disambiguation", 9thInternational Conference on Computational Intelligence and Security, ISBN: 978-0-7695-4584-4, pp: 1290-1293, 2011.

[7]. Pengyuan Liu, YongzengXue, Shiqi Li, Shui Liu, "Minimum Normalized Google Distance for Unsupervised Multilingual Chinese-English Word Sense Disambiguation", International Conference on Genetic and Evolutionary Computing, ISBN: 978-0-7695-4281-2, 2010.

[8]. Ping Chen,BowesC,Wei Ding, et..al, "Word Sense Disambiguation with Automatically Acquired Knowledge",IEEE INTELLIGENT SYSTEMS, 2012.

[9]. PrashanthMannem, "Bidirectional Dependency Parser for Hindi, Telugu and Bangla", Proceedings of ICON09 NLP Tools Contest: Indian Language Dependency Parsing, India, 2009.

[10]. Quinlan, J. R, "Induction of decision trees", Mach. Learn. 1, 1, pp: 81–106, 1986.

[11]. Quinlan, J. R, "Programs for Machine Learning", Morgan Kaufmann, San Francisco, CA, 1993.

[12]. R.M.K.Sinha and K. Sivaraman, "Ambiguity Resolution in Anglabharati",TRCS-93-174, Department of Computer Science and Engineering, IIT,Kanpur, India, 1993

[13]. R. Mahesh K. Sinha,"Learning Disambiguation of Hindi Morpheme 'vaalaa' with a Sparse Corpus" 4th International Conference on Machine Learning and Applications, ISBN: 978-0-7695-3926-3, pp: 653-657, 2009.

[14]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar "Word Sense Disambiguation : An Empirical Survey" International Journal of Soft Computing and Engineering(IJSCE), Volume-2,Issue-2,May-2012,ISSN:2231-2307.

[15]. J.Sreedhar, S. Viswanadha Raju, A. Vinaya Babu, Amzan Shaik, P.Pavan Kumar"A critical Approaches to Identification of Disambiguation Words in NLP : A Current State of the Art" International Journal of Engineering Trends and Technologies (IJETT), Volume-3,Issue-3,May-2012,ISSN:2231-5381.

[16]. P.Pavan Kumar, J.Sreedhar "Innovative Techniques and Technologies in Translation in a Multilingual Context" 3rd International Conference on Translation Technology and Globalization in Multilingual Context" Delhi, June 23-26, 2012.

[17]. P.Pavan Kumar, J.Sreedhar "Language teaching and MLE in the context of the third revolution" DLA Conference on Dravidian Languages and Translation Technology" HCU, Hyderabad, June 18-20, 2012.

[18]. Hyuk-Chul Kwon, Minho Kim, Youngim Jung, "Hybrid word sense disambiguation using language resources for transliteration of Arabic numerals in Korean", International Conference on Hybrid Information Technology (ICHIT '09), ISBN: 978-1-60558-662-5, pp: 314-321,2009.