# Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi

*Rakesh Kumar, Dr. Ravinder Khanna*
*GGS Sachdeva Engg. College, Punjab*

***Abstract :*** *Word Sense Disambiguation (WSD) is an important part of Machine Readable Dictionary (MRD) which is extensively used in Expert System/Intelligent Systems. All languages have multiple meanings of words or phrases depending on the context of their usage. WSD draws the correct (intended) meaning using a database called Machine Readable Dictionary (MRD). Some rudimentary designs of MRD have been made for some European Languages. In this paper a preliminary attempt has been made towards the formulation and design of MRD in Punjabi Language using modified Lesk Algorithm which uses a simple method for relating the appropriate word sense relative to set of dictionary meanings of the word or phrase.*

## 1. Introduction

One of the most challenging and active area of Artificial Intelligence is Natural Language Engineering (NLE) also referred as Natural Language Processing (NLP). The main goal of NLE is to design and develop software that will analyze, understand and generate languages that humans use naturally. It also determines a system of symbols, relations and conceptual information that can be used by computer logic to communicate with human. This implementation requires the system to have the capacity to translate, analyze and synthesize language. Natural language is any human "spoken or written language governed by set of rules and conventions sufficiently complex and subtle enough for there to be frequent ambiguity in syntax and meaning." The main tasks of artificial NLE are to replace the human processor with a machine processor and to get a machine to understand the natural language input and then transform it appropriately. The term natural language is used to distinguish human languages (like Punjabi, Hindi, English etc.) from formal or computer languages (like C, C++, Java,

Perl). Natural language processing requires two types of resources: linguistic data in an electronic format and computational engines. Typical electronic linguistic resources are electronic dictionaries and corpora. Computational engines process the linguistic resources and produce an output **[1], [2], [3], [4], [5]**.

## 2. Word Sense Disambiguation (WSD)

One of the most important issues in the field of natural language engineering is WSD. A lot of research has been carried out on the topic at MIT, Rutgers University, Stanford University in USA and some Indian Universities like IITs etc. In WSD, a system attempts to determine the sense of a word from contextual features. Many words in written languages have more than one meaning and we have to select the meaning which makes the most sense in context. In computational linguistics, WSD is the process of identifying which sense of a word is used in any given sentence, when the word has a number of distinct senses. For example, consider the following phrases in English:

(a) Do not tell a lie

(b) Lie down here

In the above phrases the word 'Lie' is ambiguous as it has more than one meaning. In case of (a) 'Lie' is for 'opposite of truth' and in case of (b) 'Lie' is for 'refers to a position of body'. In the same way in Punjabi language, the word **vfr** (Vaar) has six different meanings like **kiv rcnf df iek r|p ijs ivc s|rbIrqf df ijkr h|df h|, hmlf, bhq vfr, imtI afid dI icxfel ivc idqf iek gVf, idn (ijv| sinc|rvfr), krvfn krnf [8], [15]**. WSD is an intermediate task **[35]** i.e. it does not constitute an end in itself but aims at providing information that is considered essential for language understanding. WSD contributes various applications in natural language engineering for which it is potentially an issue for Machine Translation, Information Retrieval, Question Answering and Dialogues etc. The main goal of WSD is to figure out the intended or pretended meaning.

There are four conventional approaches to WSD:

- Dictionary and knowledge-based methods
- Supervised methods
- Semi Supervised methods
- Unsupervised methods

**[13], [14], [23]** explore the subject in detail. The available research primarily focuses on European languages only. No sufficient work has been done on the topic in the context of Indian languages. The present research will deal with WSD in the

context of Indian languages, in general, and with Punjabi language, in particular. An attempt for WSD in case of nouns in Hindi has been made at IIT Bombay **[27]**.

## 3. Machine Readable Dictionary (MRD)-Structure and Role

Machine Readable Dictionary (MRD) is a dictionary which is stored in a computer as database instead of being printed on paper. MRD contains a rich set of relationships between their senses, and indicate them in variety of ways. It may be single language explanatory dictionary to support translation between two or more languages or a combination of both. WSD approaches utilize external knowledge sources such as MRD, Thesauri, Tagged or Untagged corpora. LDOCE (Longman Dictionary of Contemporary English) has been the most widely used MRD in context of WSD. Such dictionaries are readily available for European languages including English. But no such MRD is define the work available for Punjabi. This work will include design and construction of such a dictionary. It work will lead to better understanding of the relationships between word meanings, which in turn will help in implementation of Lesk Algorithm by retrieving all sense definitions of the words to be disambiguated from MRD.

The following Table gives the basic idea or description of MRD in Punjabi. In this table we are using some words in Punjabi which has more than one meaning for WSD:

**Word Synset**

**sr qlf, srvr, qfs d Kz jf ck pq, srkVf, kfnf, sl mfn jl, ApfDl, hvf d qj clx dl avfj, iks jfnvr d clx dl avfj, ijq, Azd pqg dl zr n iZlf krn df Bfv**

**bd bnx dl ikiraf jF Bfv, srlr d afgF df jV, iesqrlaF d hQ df gihxf, kivqf ivc d cfr jF pj qkF dl iekfel, BliVaf hieaf,Zikaf hieaf,KVf**

**pql Bfievfll, abfdl df ihsf, cfh d pq, gn d sk pqr, Pl dl plKVl, lh dl pqll kfqr**

**sfvDfn hisafr, ckknf,Prqllf, jfgdf, cks, scq, cqnqf**

**kcf ijs ivc aj imTfs nhl bxl (axpk), axiriJaf, j riVaf nf jF cgl qrF isikaf nf hv, axAbilaf, anfVl, kmjlr, nrm, nf BrsXg, \grsrkfrl, atklpc, iCmfhl-iqmfhl(iemiqhfn), iPkf p jfn vflf, j Bq nf rlK sk, j sQfel nhl, afrjl, ijsdl sjf df \Pslf nhl hieaf, kcf mfl, grB ivclf QV mihinaf df bcf, pgzzl, Ctrl Amr df mzf, j pkl vhl ivc drj nf hv, srmsfr, qfjf**

Table 1.1: MRD structure in Punjabi **[16]**

## 4. Word Sense Disambiguation in Punjabi

The Punjabi language is morphologically rich. As already mentioned, WSD is defined as the task of finding the correct sense of the word in the context. The task needs large amount of word and word knowledge. Words may have different meaning in different contexts. Let us consider the word **vt** (Vatt). This word is used normally in eight different senses in Punjabi.

i. **vt - pll df bnf, Kfl aid df Acf kzf: Example "pfxl rknf h qF vt Acl kr lv"**

ii. **vt - gsf, nrfjgl: Ex. "m As d mh q scl gl kih idql! Asn vVf vt ciVaf"**

iii. **vt - jrnfl: Ex. "aiDafpk n mlz d vt k cpV mfr idql"**

iv. **vt - JrVl, ivg, vl, Bg: Ex. "bZfp ivc cmVl q vt px lg jfd hn"**

v. **vt - iqAVl: Ex. "Gr afe mihmfan vl dK k mQ vt nhl pfAnf cfhldf"**

vi. **vt - iZz ivc pn vflf mrV: Ex. "m kc amrd Kf igaf hf! hn iZz ivc vt p irhf h"**

vii. **vt - pslnf, hm, hmV, BVfs, BVdfa: Ex. "grml bhq h! mn vt qg kr irhf h"**

viii. **vt - mCf n vt cfVn: Ex. "srdfr jl svr rj pg bnx q bfad mCf n vt cfVd hn"**

In the above sentences, the word **vt** has different meanings **[8], [15]**. There are hundreds of examples of such words in Punjabi. In English language Yarowsky proposed a solution to WSD using the thesaurus and supervised learning approach **[14]**. The main idea is to compare the context of the word in a sentence with the context constructed from the database and assign correct sense to the word. In this research we will use Punjabi Lexical Resources, Gurmukhi OCR, Harkirat Singh's Shabad Kosh, Punjabi Thesaursus and Online Technical English Punjabi glosses. The work involves the design & construction of a Machine Readable Dictionary (MRD) and its use in implementation of Lesk Algorithm or its modified form for WSD in Punjabi. The study will be a pioneering work on the chosen topic in the context of Indian languages, in general, and Punjabi language, in particular.

***Figure 1.1.*** *described the pictorial representation of WSD form Punjabi MRD and given context [28].*

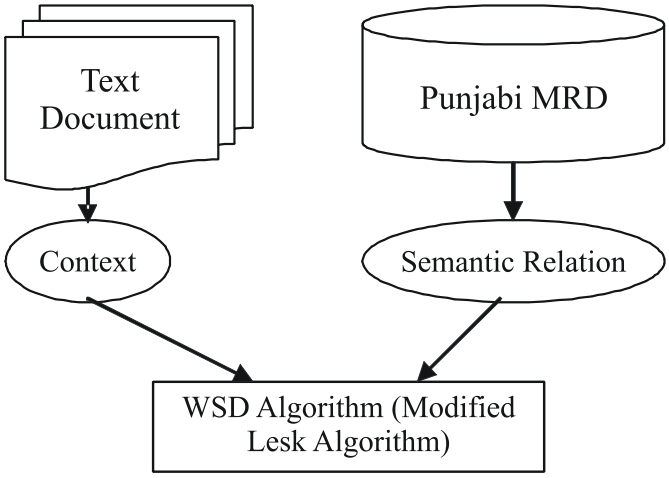


***Figure 1.1:*** *To Extraction of synonymy of given context from Punjabi MRD for WSD*

## 5. Lesk Algorithm and its implementation

The original Lesk algorithm **[23]** proposed a simple algorithm for selecting the appropriate word sense relative to a set of dictionary senses. The **Lesk** algorithm disambiguates a target word by comparing its gloss with those of its surrounding words. The target word is assigned the sense whose gloss has the most overlapping or shared words with the glosses of its neighboring words. **Lesk** demonstrates this algorithm on the words *pine cone*. Using the **Oxford Advanced Learner's Dictionary**, it finds that the word *pine* has two senses:

Sense 1: kind of **evergreen tree** with needle–shaped leaves

Sense 2: waste away through sorrow or illness.

The word *cone* has three senses:

Sense 1: solid body which narrows to a point

Sense 2: something of this shape whether solid or hollow

Sense 3: fruit of certain **evergreen tree**

Each of the two senses of the word *pine* is compared with each of the three senses of the word *cone* and it is found that the words *evergreen tree* occurs in

one sense each of the two words. These two senses are then declared to be the most appropriate senses when the words *pine* and *cone* are used together.

There are two hypothesis that underly this approach. The first is that words appears together in a sentence can be disambiguated by assigning to them the senses that are most closely related to their neighboring words. The second hypothesis is that related senses can be identified by finding overlapping words in their definitions. In this research we will implement Lesk algorithm or modified version of this, keeping in view the linguistic feature of Punjabi to disambiguate the word in Punjabi language.

## References

[1] Adam Kilgrriff, [1998], " Gold Standard Data-sets for Evaluating Word Sense Disambiguation Programs" *In Computer Speech and Languages* 12(4), Special Issues on Evaluation

[2] Eneko Agirre and German Rigau, [1996], "Word sense disambiguation using conceptual density" In *Proceedings of the 16th conference on Computational linguistics*, Association for Computational Linguistics, pages 16–22, Morristown, NJ, USA.

[3] Klavans, Judith L. and Philip Resnik (editors), [1996], The Balancing Act: Combining Symbolic and Statistical Approaches to Language. Language, Speech, and Communication. MIT Press, Cambridge, Massachusetts.

[4] Manning, Christopher D. and Hinrich Sch¨utze, [1999],. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, Massachusetts.

[5] Stevenson, Mark and Yorick Wilks, [2001], "The interaction of knowledge sources in word sense disambiguation" *Computational Linguistics*, 27(3):321–349.

[6] Adriaens, Geert, [1986], "Word expert parsing: a natural language analysis program revised and applied to Dutch." *Leuvensche Bijdragen,* pp 75(1), 73-154.

[7] Alshawi, Hiyan and Carter, David [1994], "Training and scaling preference functions for disambiguation." *Computational Linguistics,* 20(4), 635-648.

[8] Berger, Adam L., Stephen A. Della Pietra, and Vincent J. Della Pietra, [1996]. "A maximum entropy approach to natural language processing" *Computational Linguistics,* 22(1):39–71.

[9] Bhai Kahan Singh, [1981], *"Mahan Kosh (In Punjabi) Encyclopedia of Sikh Literature",* Basha Bhibag, Punjab pp 1081

[10] Brill, Eric and Philip Resnik, [1994], "A rule-based approach to prepositional phrase attachment disambiguation" *In Proceedings of the 15th International Conference on Computational Linguistics (COLING 94),* pp. 1198–1204.

[11] Br¨oker, Norbert, Michael Strube, Susanne Schacht, and Udo Hahn, [1997], Coarsegrained parallelism in natural language understanding: Parsing as message passing. In Jones and Somers, pp. 301–317.

[12] Cottrell, Garrison W., Small, Steven L, [1983], "A connectionist scheme for modeli

[13] C. Manning , Schutza "Foundations of statistical Natural Language Processing" *Word Sense Disambiguation Chapter,* The MIT Press, Cambridge, Massachusetts, London, England.

[14] D. Narayan and P. Bhattacharyya, [2002], "Using Verb Noun Association for Word Sense Disambiguation" *International Conference on Natural Language Processing (ICON 2002),* Mumbai, India

[15] David Yarowsky, [1992], "Word Sense Disambiguation using statistical model of Roget's categories trained on large corpora" *In proceeding of 14th International Confrence on Computational Linguistics (COLING-92),* pages 454-460, Nantes, France.

[16] Dr. Madan Lal Aseeja(Chief Editor), [2001], "Parmanik Punjabi Kosh" Basha Bibagh Punjab pp 408, pp 415

[17] Franz, Alexander, [1996a], "Automatic Ambiguity Resolution in Natural Language Processing" volume 1171 of LNAI. Springer, Berlin.

[18] Fujii, Atsushi, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka, [1998], "Selective sampling for example-based word sense disambiguation" *Computational Linguistics,* 24(4):573–597.

[19] Helbig, Hermann, [1986], "Syntactic-semantic analysis of natural language by a new word-class controlled functional analysis" *Computers and Artificial Intelligence,* 5(1):53–59.

[20] Hirst, Graeme (editor), [1987], Semantic Interpretation and the Resolution of Ambiguity. Studies in Natural Language Processing. Cambridge University Press, Cambridge, England.

[21] Jones, Daniel and Harold Somers (editors), [1997], New Methods in Language Processing. University College Press, London.

[22] Jurafsky D., Martin J., [2000], "Speech and language processing" Prentice Hall

[23] Kostas Fragos, [2005], "Distributional Analysis of related synsets in WordBet-For a Word sense Disambiguation Task", Artificial Intelligence tools , vol no 14, pp 919-934.

[24] Lesk, Michael, [1986], "Automatic sense disambiguation: How to tell a pine cone from an ice cream cone" *In Proceedings of the 1986 SIGDOC Conference,* pages 24"26, New York. Association for Computing Machinery.

[25] Luisa Bentivogli and Emanuele Pianta, [2005], "Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor corpus" *Natural Language Engineering*, 11(3):247–261

[26] Mangu, Lidia and Eric Brill, [1997], Automatic Rule Acquisition for Spelling Correction. In *Proceedings of The Fourteenth International Conference on Machine Learning*, ICML 97, Morgan Kaufmann.

[27] McRoy, Susan W., [1992], Using multiple knowledge sources for word sense discrimination. Computational Linguistics, 18(1):1–30.

[28] Mahesh Kumar Reddy .R, Manish Sinha, Pushak Battacharyya, Prabhakar Panday, Laxmi Kashyap. "Hindi Word Sense Disambiguation" www.cse.iitb.ac.in/~pb/papers/HindiWSD.pdf, Department of Computer Science and engineering IIT Bombay

[29] Noam Ordan and Shuly Wintner, [2007], "Hebrew WordNet: a test case of aligning lexical databases across languages" *International Journal of Translation, special issue on Lexical Resources for Machine Translation* 19(1):39-58

[30] **P. F. Brown, S. A. Della Pietra, V. J. Della Pietra, and R. L. Mercer,** [1993], The mathematics of statistical machine translation. *Computational Linguistics*, 19(2).

[31] **P. Stathopoulou-Zois**, [2005], "A Grapheme-to-Phoneme Translator for TTS Synthesis in greek", Artificial Intelligence tools , Vol No 14, pp 901-918.

[32] **Resnik P. and D. Yarowsky,** [2000], "Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation" *Natural Language Engineering*, 5(2), pp. 113-133.

[33] **Sanda Harabagiu, editor,** [1998], "*Usage of WordNet in Natural Language Processing Systems: Proceedings of the Coling-ACL 1998 Workshop" Association for Computational Linguistics*, Montreal, Canada

[34] **Small, Steven,** [1987], **"**A distributed word-based approach to parsing" *In Natural Language Parsing Systems (edited by Leonard Bolc)*, pp. 161–201. Springer, Berlin.

[35] **Winograd, Terry,** [1972], "Understanding Natural Language" Academic Press, New York.

[36] **Wilks, Y. and M. Stevenson,** [1996], **"**The Grammar of Sense: Is Word Sense Tagging Much More Than Part of Speech Tagging" *Technical Report CS-96-05, University of Sheffield, Sheffield, UK*.

[37] **Yaacov Choueka and Serge Lusignan,** [1985], "Disambiguation by short context" *Computers and the Humanities*, 19:147–157.