Noname manuscript No.
(will be inserted by the editor)

A Word Sense Disambiguation Corpus for Urdu

Ali Saeed · Rao Muhammad Adeel Nawab · Mark Stevenson · Paul Rayson

Received: date / Accepted: date

Abstract The aim of Word Sense Disambiguation (WSD) is to correctly identify the meaning of a word in context. All natural languages exhibit word sense ambiguities and these are often hard to resolve automatically. Consequently WSD is considered an important problem in Natural Language Processing (NLP). Standard evaluation resources are needed to develop, evaluate and compare WSD methods. A range of initiatives have lead to the development of benchmark WSD corpora for a wide range of languages from various language families. However, there is a lack of benchmark WSD corpora for South Asian languages including Urdu, despite there being over 300 million Urdu speakers and a large amounts of Urdu digital text available online. To address that gap, this study describes a novel benchmark corpus for the Urdu Lexical Sample WSD task. This corpus contains 50 target words (30 nouns, 11 adjectives, and 9 verbs). A standard, manually crafted dictionary called Urdu Lughat is used as a sense inventory. Four baseline WSD approaches were applied to the corpus. The results show that the best performance was obtained using a simple Bag of Words approach. To encourage NLP research on the Urdu language the corpus is freely available to the research community.

Ali Saeed
COMSATS University Islamabad, Lahore, Pakistan
E-mail: fa15-pcs-004@cuilahore.edu.pk

Rao Muhammad Adeel Nawab
COMSATS University Islamabad, Lahore, Pakistan
E-mail: adeelnawab@cuilahore.edu.pk

Mark Stevenson
University of Sheffield, UK
E-mail: mark.stevenson@sheffield.ac.uk

Paul Rayson
Lancaster University, UK
E-mail: p.rayson@lancaster.ac.uk

Keywords Word Sense Disambiguation · Lexical Sample Task · Sense Tagged Urdu Corpus

1 Introduction

WSD is the problem of identifying the correct sense of a word used in a given context [1]. All natural languages exhibit word sense ambiguity. Human speakers are generally able to resolve this ambiguity unconsciously but doing so is generally a challenging problem for machines. Despite extensive research into the WSD over several decades, the difficulty of the problem means that it is still an open challenge [2].

WSD has potential applications in many areas of language processing including Information Retrieval [3], Machine Translation [4], Information Extraction [5], Content Analysis [6], Text Summarization [7], Discourse Analysis [8] and Natural Language Generation [9].

Two variants of the WSD problem have been explored: (1) the All-Words WSD task and (2) the Lexical Sample WSD task [10]. In the first case the aim is to disambiguate all content words in a given piece of text (normally a sentence or paragraph) while in the second case a predefined set of target words are provided and the aim is to disambiguate instances of these terms.

Previous studies on Lexical Sample WSD tasks have focused on a range of languages including English, Basque, Italian, Japanese, Korean, Swedish, Spanish, Catalan, Chinese and Romanian [10][17][20][23][26][27][31][32]. However, WSD has not been widely explored for South Asian languages [33][34][35] despite the fact that these languages are spoken by a large number of people. The focus of this paper is on the Lexical Sample WSD task for Urdu, a widely spoken but under-resourced South Asian language.

Urdu is an Indo-Aryan language that inherits its vocabulary and grammatical forms from a range of languages including Arabic, Persian and South Asian languages [14]. It is morphologically rich, some of its words (verbs and nouns) can have more than 40 forms making it difficult to process automatically [16]. Urdu is one of the important international languages with more than 300 million speakers [11][12][13]. 151 million are native speakers and the remainder second language speakers[1]. Urdu is the national language of Pakistan where there are more than 11 million speakers. Other countries with a large number of Urdu speakers include India, Bangladesh, USA, UK, and Canada [12][13]. It is also spoken globally due to the large South Asian diaspora [14]. Despite its wide usage Urdu is still a poorly resourced language for NLP and efforts are being made to create Urdu computational resources [15].

Benchmark corpora are required to develop, evaluate, analyze and compare WSD systems for the Lexical Sample WSD task. The majority of corpora developed for Lexical Sample WSD tasks are for English and other European languages [1][17]. However, there is a lack of standard evaluation resources for South Asian languages, particularly Urdu. This study describes a novel

[1] `https://www.ethnologue.com/language/urd` Last visited: 23-October-2018

benchmark corpus for the Urdu Lexical Sample WSD task and reports baseline experiments using a range of approaches.

Our Urdu Lexical Sample WSD 2018 (ULS-WSD-18) corpus contains 50 target words (30 nouns, 11 adjectives, and 9 verbs). Target words are manually tagged using a sense inventory extracted from a hand crafted dictionary called Urdu Lughat [18]. The corpus contains 7,185 sense tagged instances ($75 + 15n$ sentences for each target word, where n is the number of possible senses of the target word). ULS-WSD-18 is a sense annotated corpus, split into training and testing sets using a 2:1 ratio, and also a dataset that can be used for the Lexical Sample WSD task.

Four baseline WSD algorithms were evaluated against ULS-WSD-18 to demonstrate how the corpus can be used for the development and evaluation of Urdu Lexical Sample WSD systems.

The rest of this paper is organized as follows. Section 2 presents existing evaluation resources for the WSD task. Section 3 describes the corpus generation process for our proposed corpus. Section 4 explains the WSD techniques applied on our proposed corpus and how they are evaluated. Section 5 shows results and their analysis. Finally, the paper is concluded in Section 6.

2 Related Work

Previous literature describes a number of efforts to develop standard evaluation resources for the WSD task. Broadly, existing corpora either focus on the All-Words WSD task or the Lexical Sample WSD task.

The most prominent effort in developing resources for WSD task is the series of competitions organized under the Senseval and SemEval banner. The outcome of these competitions is a set of benchmark corpora for WSD tasks, including Lexical Sample and All-Words [10][19][20][21]. Languages for which these corpora were developed include English, Basque, Italian, Japanese, Korean, Spanish, Swedish, Catalan, Chinese and Romanian. WordNet was used as a sense inventory for many of the resources with sense assignments determined by manual tagging [22].

Three other corpora for the All-Words WSD task are also worthy of mention: (1) SEMCOR WSD corpus [23], (2) Google WSD corpus[2] and (3) OMSTI (One Million Sense-Tagged Instances) corpus [24]. The SEMCOR WSD corpus is a manually annotated corpus of English. The source text was taken from Brown corpus [25]. It contains 234,000 manually sense annotated sentences. WordNet was used as a source of sense inventory. In addition, Dutch [26] and Japanese [27] versions of this corpus have also been developed. The Google WSD corpus is the largest manually annotated corpus for English. The corpus source text was taken from SEMCOR WSD corpus and MASC WSD corpus [28] (a sense annotated corpus). It is comprises of 248,000 sense annotated sentences. All sentences were manually annotated using the New Oxford

[2] `https://github.com/google-research-datasets/word_sense_disambigation_corpora` Last visited: 23-October-2018

American Dictionary (NOAD) [29]. The OMSTI corpus is another large sense annotated corpus of English. Source text for developing this corpus was obtained from the MultiUN corpus[3]. It is comprises of one million word instances semi-automatically annotated with senses from WordNet.

Corpora created for the Lexical Sample WSD task include those developed for Senseval/SemEval and other resources such as: (1) DSO corpus [17][30], (2) Line-Hard-Serve corpus [31], (3) Interest corpus [32], (4) Hindi Sense Tagged corpus [36]. The DSO corpus was constructed using 191 frequent and ambiguous words (121 nouns and 70 verbs). The DSO corpus was developed by selecting 192,800 sentences from the Brown corpus [25] and Wall Street Journal corpus[4]. All the sentences were sense tagged by human taggers (students of University of Singapore) using WordNet as a sense inventory. The Line-Hard-Serve corpus was constructed using three target words: line, hard and serve. Line is used as a noun, hard as an adjective, and serve as a verb. 12,000 instances were manually annotated using WordNet. Source data (for the Line-Hard-Serve corpus) was gathered from Wall Street Journal[4], American Printing House for the Blind [37], and San Jose Mercury [38]. The Interest corpus contains 1,470 hand labeled instances of a single target word, interest (used as a noun). Senses were taken from LDOCE[5] dictionary for this corpus and sentences from the Wall Street Journal. The Hindi Sense Tagged corpus was built using 40 target words (nouns). This corpus contains 2,369 manually tagged instances (senses for a target word were collected from Hindi WordNet[6]). Source data for creating this corpus was taken from India info Dainik Jagran[7], Khoj, Hindi Wikipedia[8], Webdunia[9] websites and the EMILLE corpus [39][40].

We only found one sense tagged Urdu corpus in the previous literature. The Sense Tagged CLE Urdu Digest corpus was developed for the All-Words WSD task. The source text for the development of this resource was taken from CLE Urdu Digest corpus [41]. It contains 17,006 manually sense annotated sentences (senses of tagged words were extracted from CLE Urdu WordNet [42]). All sentences in the corpus were annotated by a single annotator over a period of 10 months.

A detailed survey of Urdu language processing [13] describes resources (and their characteristics), tasks, techniques, and applications of Urdu language processing. Studies have also been carried out with a particular focus on the Urdu Named Entity Recognition (NER) task. Approaches were categorized into three groups by [43]: (1) Rule-Based, (2) Machine Learning, and (3) Hybrid. The challenges faced when processing the Urdu language and a novel algorithm for Urdu NER were also proposed [43]. A rule-based Urdu N-

[3] `http://opus.nlpl.eu/MultiUN.php` Last visited: 23-October-2018

[4] `https://catalog.ldc.upenn.edu/LDC2000T43` Last visited: 23-October-2018

[5] `http://www.ldoceonline.com/` Last visited: 23-October-2018

[6] `http://www.cfilt.iitb.ac.in/wordnet/webhwn/` Last visited: 23-October-2018

[7] `http://www.jagran.com/` Last visited: 23-October-2018

[8] `https://en.wikipedia.org/wiki/Hindi` Last visited: 22-October-2018

[9] `http://www.hindi.webdunia.com/` Last visited: 23-October-2018

ER algorithm was proposed by [12]. They also discuss the differences between Urdu and other South Asian languages, particularly Hindi, in the context of the NER task.

The literature includes some significant contributions around WSD for South Asian languages. Three studies apply supervised learning methods for Urdu WSD [16, 44, 45]. However, these approaches are not based on standard datasets, do not apply extensive feature extraction methods and do not explore a range of classifiers. Researchers also explored WSD for other South Asian languages including Hindi [46], Tamil [47] and Telugu [48].

To conclude, benchmark Lexical Sample WSD corpora have been developed for many languages but not for Urdu. As far as we are aware, the ULS-WSD-18 corpus is a sense annotated data set developed for the Urdu Lexical Sample WSD task.

3 Corpus

3.1 Source Data

The ULS-WSD-18 corpus was created using the UrMono corpus [15], the largest freely available corpus of Urdu language that can be used for non-commercial research purposes. The UrMono corpus contains 95.4 million tokens and 5.4 million sentences. All tokens were Part of Speech (PoS) tagged using the CLE PoS tagset [49] with an accuracy of 87.98%. Data in the UrMono corpus was collected from a variety of domains including news, religion, blogs, literature, science and education.

3.2 Words Selection

A predefined set of target words is needed to create the gold standard ULS-WSD-18 corpus. We selected 50 words (30 nouns, 11 adjectives and 9 verbs) which were highly frequent and polysemous in the entire UrMono corpus.

Table 1 shows the 50 selected words along with their PoS tag, frequency (in the UrMono corpus) and number of senses (in the Urdu Lughat). Stop words were ignored and only the most frequent and ambiguous content words were selected. The advantage of selecting frequent and ambiguous words is that it increases lexical coverage and makes the task more challenging [20]. The number of senses For each selected word varies from 2 to 8. The most frequent of the 50 selected words is دل (pronounced as Dil and appears 83,949 times in the UrMono corpus) and the least frequent خط (pronounced as Khat and appears 7,487 times).

3.3 Sense Inventory

We found three resources which could potentially be used to generate a sense inventory: (1) Indo WordNet[10] [50], (2) CLE Urdu WordNet[11] [42] and (3)

[10] `http://www.cfilt.iitb.ac.in/indowordnet/` Last visited: 23-October-2018

[11] `http://www.cle.org.pk/clestore/urduwordnet.htm` Last visited: 23-October-2018

Table 1 Fifty highly frequent and ambiguous words with roman Urdu transliteration selected from the UrMono corpus for the ULS-WSD-18 corpus (P represents PoS tag, Fr represents frequency in UrMono corpus, N represents the number of senses in Urdu Lughat).

Sr.	Word	P	Fr.	N	Sr.	Word	P	Fr.	N
1	پاس (Pass)	N	60780	8	26	طور (Tor)	N	68411	2
2	دور (Daur)	N	49402	8	27	ڈاکٹر (Doctor)	N	21783	2
3	پانی (Pani)	N	29235	8	28	سفر (Safar)	N	18572	2
4	سر (Sir)	N	40738	7	29	آیت (Aayat)	N	8084	2
5	حصہ (Hissa)	N	26978	7	30	خط (Khat)	N	7487	2
6	روشنی (Roshni)	N	13498	7	31	کہنا (Kehna)	V	59700	8
7	دل (Dil)	N	83949	6	32	دیکھ (Dekh)	V	53776	8
8	نظر (Nazar)	N	83332	6	33	مل (Mil)	V	42492	8
9	کتاب (Kitaab)	N	45169	6	34	لگ (Lag)	V	25438	7
10	زبان (Zabaan)	N	34732	6	35	چل (Chal)	V	24479	7
11	برس (Baras)	N	10810	6	36	سوچ (Soch)	V	12840	3
12	ملک (Mulk)	N	61292	5	37	بھول (Bhool)	V	12582	3
13	کار (Car)	N	21829	5	38	پڑھ (Parh)	V	11493	3
14	رنگ (Rang)	N	19949	5	39	سمجھ (Samajh)	V	25885	2
15	شکل (Shakal)	N	12694	5	40	خاص (Khas)	A	27611	6
16	سوال (Sawal)	N	40171	4	41	تیار (Tayyar)	A	22954	6
17	عمر (Umar)	N	37783	4	42	بند (Band)	A	27785	5
18	خون (Khoon)	N	12908	4	43	ذندہ (Zindah)	A	13005	5
19	شکر (Shukar)	N	11708	4	44	مکمل (Mukammal)	A	35939	4
20	قسم (Qisam)	N	30125	3	45	صحیح (Sahih)	A	25943	4
21	ذکر (Zikar)	N	25805	3	46	شریف (Shareef)	A	15990	3
22	درمیان (Darmiyan)	N	22836	3	47	کم (Kam)	A	63381	2
23	دین (Deen)	N	20351	3	48	شامل (Sahaamil)	A	50559	2
24	حل (Hal)	N	19340	3	49	غیر (Ghair)	A	44116	2
25	بجلی (Bijli)	N	13028	3	50	اہم (Ahem)	A	25918	2

Urdu Lughat dictionary[12]. These resources were manually inspeected to determine which is most suitable for use as a sense inventory.

The Indo WordNet project aimed to develop WordNets for multiple languages spoken in India including Hindi, Marathi, Konkani, Urdu, Sanskrit, Nepali, Kashimiri, Assamese, Tamil, Malyalam, Telugu, Kannad, Manipuri, Bodo, Bangla, Punjabi and Gujarati. We applied our selected target words as input to Indo WordNet using its online interface[10] and it failed to return senses for many words or the number of senses returned were very low. For example, for the target words دل (Dil) and طور (Tor), Indo WordNet returned nothing (i.e. the words were not found in the resource) and only one sense was returned for the target word نظر (Nazar).

The second available choice for generating a sense inventory was CLE Urdu WordNet. It contains only 6,000 unique words along with their senses. Again, similar to Indo WordNet, the majority of our selected words were not found or had a very small number of senses. For example, دل (Dil) was not found

Table 2 Example showing five highly frequent ambiguous words with roman Urdu transliteration along with their senses collected from Urdu Lughat.

Sr. No	Word	Sense1	Sense2	Sense3	Sense4	Sense5	Sense6
1	دل (Dil)	سینے کے اندر ایک عضو (Seenay kay andar aik uzoo)	سخاوت، فیاضی (Sakhavat, Fayazi)	مزاج، طبیعت (Mizaaj, Tabiyat)	جرات، ہمت، شجاعت، دلیری (Jurrat, Himmat, Shujaat, Dileri)	ذہن، دماغ تخیل (Zehen, Dimagh, Takhayyul)	مرکز، محور (Markaz, Mehwar)
2	نظر (Nazar)	بغور دیکھنا (Baghor Dekhna)	تیور، بصارت روشنی، چشم (Tevar, Basarat, Roshni Chasham)	آنکھ، چشم ، نین (Aankh, Chasham, Nain)	جن و پری، بھوت پریت (Jin-o-Pari, Bhoot Pret)	امید، توقع (Umeed, Tawaqqa)	نیت، ارادہ (Niyat, Iradah)
3	طور (Tor)	پہاڑ (Pahar)	انداز، طریقے (Andaaz, Tareeqay)	-	-	-	-
4	کم (Kam)	تھوڑا (Thora)	بد، برا (Bad, Bura)	-	-	-	-
5	ملک (Mulk)	وطن، دیس (Watan, Daes)	فرشتے (Farishtay)	دودھ (Doodh)	سنہرا سانپ (Sunehra Saanp)	-	-

in CLE Urdu WordNet. It returns only three senses for نظر (Nazar) and two senses for طور (Tor) (i.e. the number of senses returned is small).

The third choice was the Urdu Lughat dictionary [18]. Urdu Lughat is an Urdu to Urdu dictionary, which is manually created by the Dictionary Board, Karachi, Pakistan and is freely available for research purposes through its online interface[12]. It contains more than 120,000 unique words along with their senses, synonyms, glosses and descriptive examples. We found multiple senses for all the 50 target words. For example, it returns 6 senses for دل (Dil), 6 senses for نظر (Nazar) and 2 senses for طور (Tor).

To conclude, among all the three available resources, manual inspection showed that the best resource for sense inventory generation was Urdu Lughat and it was therefore selected for this study.

Each entry in the sense inventory comprises one of the 50 target words, its PoS tag, frequency of occurrence in the UrMono corpus and number of senses (obtained from Urdu Lughat). Table 2 shows five highly frequent and ambiguous target words along with their senses extracted from Urdu Lughat[13].

[12] `http://www.urdulughat.info/` Last visited: 23-October-2018

[13] The complete list of 50 selected words along-with their senses can be downloaded from `http://www.comsatsnlpgroup.wordpress.com` Last visited: 23-October-2018

Table 3 Example sentences for six different senses (Urdu Lughat senses) of a target word دل (Dil).

Sense No.	Sense	Sentence with roman Urdu transliteration
1	سینے کے اندر ایک عضو (Seenay kae andar aik uzoo)	آج میرئے بھائی کے دل کا آپریشن ہے ۔ Aaj meray bhai ke dil ka operation ha.
2	سخاوت، فیاضی (Sakha-vat, Fayazi)	اپنے اخلاق کے ساتھ انھوں نے میرا دل جیت لیا ۔ Apne ikhlaq ke sath unhon na mera dil jeet liya.
3	مزاج، طبیعت (Mizaaj, Tabiyat)	آپ نے میری پوسٹ نمبر ۱۳ کو ناپسند کر کے میرا دل توڑ دیا ۔ Aap na meri post number 13 ko napasand kar k mera dil toar diya.
4	جرات، ہمت شجاعت، دلیری (Jurrat, Himmat, Shujaat, Dileri)	میں نے دل پر ہاتھ رکھ کے سب کچھ لکھ دیا ۔ Mein nay dil par haath rakh ke sab kuch likh diya.
5	ذہن، دماغ، تخیل (Zehen, Dimagh, Takhayyul)	میں نے دل میں سوچا ۔ Mein nay dil mein socha.
6	مرکز، محور (Markaz, Mehwar)	لاہور پاکستان کا دل ہے ۔ Lahore Pakistan ka dil hay

Table 2 illustrates the range of senses for words in the corpus, making the WSD task more realistic and challenging.

3.4 Sentence Selection

$75 + 15n$ sentences were extracted from the UrMono corpus for each target word, where n represents the number of senses for a target word [10][20]. For instance, the target word دل (Dil) has six senses, therefore 165 $(75 + 15 \times 6)$ sentences were chosen. The number of senses of the target words varies from 2 (105 sentences) to 8 (195 sentences). Table 3 shows example sentences for six different senses of the target word دل (Dil).

3.5 Urdu Annotation Tool (UAT)

A web based UAT was built using PHP and MySQL to manually annotate sentences containing one of the target words. The complete code of UAT is

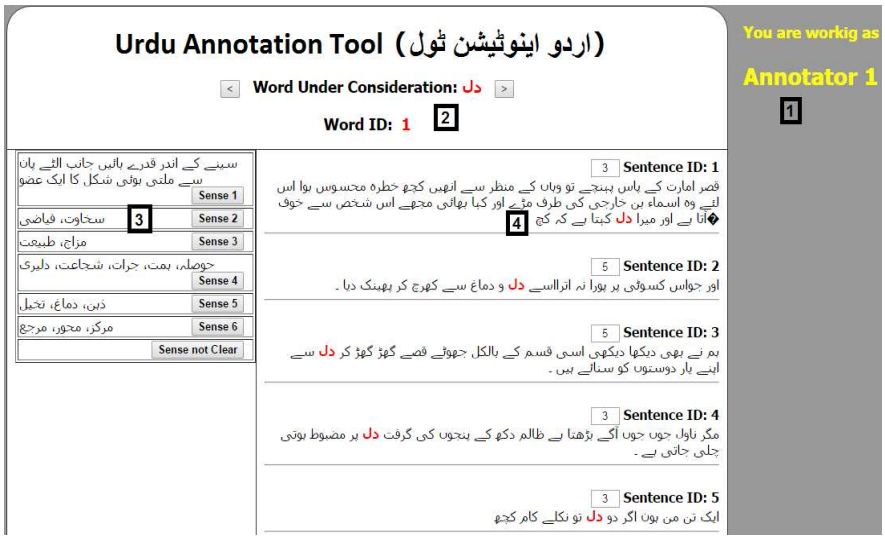

Fig. 1 Graphical User Interface of Urdu Annotation Tool shown to the first two annotators.

available on GitHub[14]. Fig. 1 shows the Graphical User Interface (GUI) of the UAT. Box 1 shows that currently annotator 1 is assigning senses to sentences. The sentences to be annotated appear in Box 4. The target word under consideration is shown in Box 2 along with the word ID. Box 3 shows all possible senses for the target word being annotated.

In Fig. 1, the word under consideration for annotation, دل (Dil) (see Box 3), has six possible senses. An additional option, "Sense not Clear", is available for when none of these senses appear appropriate (a value of -1 will be assigned when this option is selected). When an annotator assigns a sense to the target word of a sentence, it will appear in a small text box near the sentence ID (see first line in Box 4). The Save button is used to store all data in a persistent storage. Annotators can use arrow buttons near Box 2 to load the next word for annotation.

The corpus was analyzed by three annotators. The first two independently tagged senses in each sentence and conflicts were resolved by the third annotator. Fig. 2 shows the GUI of the UAT shown to the third annotator. This annotator is only shown the sentences where there was disagreement between the first two annotators. The interface shows the senses assigned by the first two annotators and allows the third annotator to make the final sense selection.

3.5.1 Annotations and Inter-annotator Agreement

The ULS-WSD-18 corpus was manually annotated by three annotators. All the annotators were native speakers of Urdu with a high level of proficiency

[14] `https://github.com/alisaeed007/Urdu-Annotation-Tool-UAT-` Last visited: 23-October-2018

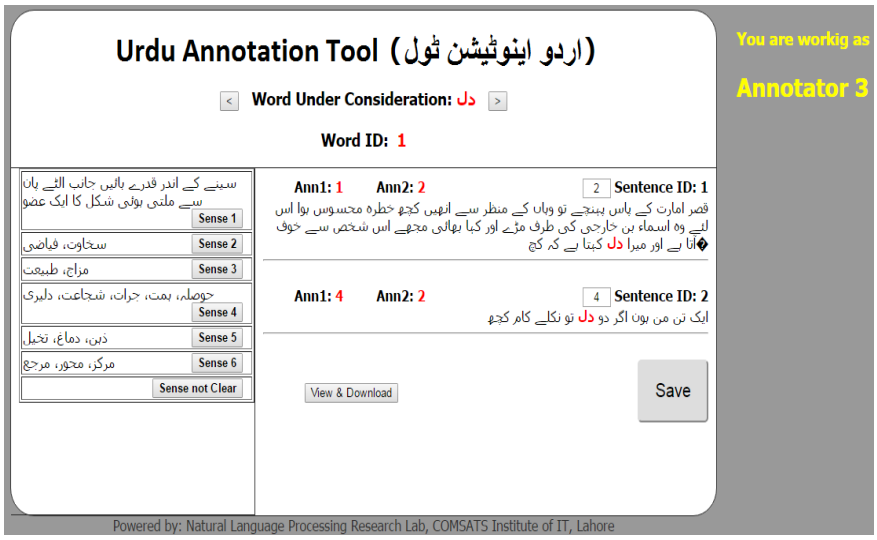

Fig. 2 Graphical User Interface of Urdu Annotation Tool shown to the third annotator to resolve conflicts.

in the language and had knowledge of the WSD task. In the first stage, annotators 1 and 2 annotated a subset of 200 sentences, computed inter-annotator agreement and discussed their annotations, particularly conflicting ones. In the second stage, the remainder of the corpus was annotated by annotators 1 and 2 and inter-annotator agreement computed for the entire corpus. Sentences where annotators 1 and 2 disagreed were then annotated by the third annotator.

The inter-annotator agreement obtained is 90.1% and the weighted kappa score is 0.82. This result shows a good agreement, considering the fact that each target word has at least two possible senses.

3.6 Corpus Characteristics

The corpus contains 222,533 tokens and 7,185 sentences. The tokens consist of 57,150 nouns, 25,719 verbs, 15,297 adjectives, 3,557 adverbs and 120,810 belonging to the other PoS categories.

The corpus is also split into training and testing sets. The training set contains two-thirds of the sentences (4,790 sense tagged instances) and the test set the remaining third (2,395 sense tagged instances). The splits take account of the number of sentences that belong to a particular sense. For example, if sense S_1 of word W_1 has 30 sentences then 20 will be used in training set and 10 for testing set. The ULS-WSD-18[15] corpus is freely and publicly available for research purposes under a Creative Commons license.

[15] `http://www.comsatsnlpgroup.wordpress.com` Last visited: 23-October-2018

```
<?xml version="1.0" encoding="utf-8"?>
 <contextfile fileno="1" filename="دل">
   <s snum="1">
      <wf pos="NN">آگے</wf>
      <wf pos="PP">آپ</wf>
      <wf pos="P">نے</wf>
      <wf pos="G">میری</wf>
      <wf pos="PN">پوسٹ</wf>
      <wf pos="NN">نمبر</wf>
      <wf pos="CA">13</wf>
      <wf pos="P">کو</wf>
      <wf pos="NN">ناپسند</wf>
      <wf pos="VB">کر</wf>
      <wf pos="P">کے</wf>
      <wf pos="G">میرا</wf>
      <wf pos="NN" ws="3">دل</wf>
      <wf pos="VB">توڑ</wf>
      <wf pos="AA">دیا</wf>
      <wf pos="TA">ہے</wf>
      <wf pos="SM">.</wf>
    </s>
</contextfile>
```

Fig. 3 An example of a sense annotated instance from the ULS-WSD-18 corpus.

3.7 Corpus Encoding

The corpus is released in a standard XML format [51] containing 50 context files (one file for each target word). Each file contains $75+15n$ sentences (where n represents the number of senses for a target word) with training and testing splits.

Fig. 3 shows a single sentence from the corpus in XML format. In this example *<contextfile fileno="file_number" filename="urduname">* indicates the beginning of the context file. The *fileno* attribute shows the file number (which ranges from 1 to 50) and the *filename* attribute the name of the file. *<s snum="sentence_number">* indicates the beginning of a sentence and *snum* unique sentence number assigned to a particular sentence in a context file. The *<wf PoS="PoS_tag" ws="Word_sense_number" >* tag shows the beginning of each tagged word while the *PoS* and *ws* attributes show (respectively) the word's PoS tag and sense number.

4 WSD Experiments

We developed four baseline WSD approaches to demonstrate how the ULS-WSD-18 corpus can be used for the development and evaluation of Lexical Sample WSD systems for Urdu: (1) Most Frequent Sense, (2) Part of Speech tags based method, (3) Bag of Words and (4) Word Embeddings. The following sections describe these approaches, the data set used for the experiments, our evaluation methodology and the evaluation measures employed.

4.1 Approaches for Lexical Sample Word Sense Disambiguation

4.1.1 Most Frequent Sense (MFS) Approach

It is common for one sense of a polysemous word to occur more frequently than the others [52][53] and this sense is commonly known as MFS. A very simple disambiguation approach can be designed that assigns to each word its most frequent meaning. We applied the MFS approach for each of the 50 target words separately and reported the averaged accuracy in Table 7.

4.1.2 Part of Speech Tags based Approach

PoS tags of neighboring words are a useful feature for WSD [1]. Let w_i be a target word then a PoS based feature vector (2x2 widow size) can be written as: $[PoS_{i-2}, PoS_{i-1}, PoS_{i+1}, PoS_{i+2}, Labeled\_Sense]$ in which PoS_{i-1} and PoS_{i+1} indicate the PoS tags of the previous and next words respectively. Similarly PoS_{i-2} and PoS_{i+2} represent the PoS tags of the preceding and following two words of a target word. $Labeled\_Sense$ indicates the manually assigned sense.

Table 4 Examples of different context window sizes.

Context Size	Sentence targeted on word دل (Dil) with roman Urdu transliteration
Unigram	… **دل** کا … نہیں (... ka **dil** nahi ...)
BiGram	… کس کا **دل** … نہیں کرتا (... kis ka **dil** nahi karta...)
TriGram	… تو کس کا **دل** … نہیں کرتا کے (... to kis ka **dil** nahi karta kay ...)
Sentence	معلوماتی چیزیں ہیں تو کس کا **دل** نہیں کرتا کے فائدہ اٹھائے ۔ (Malomati cheezen hain to kis ka **dil** nahi karta kay faida uthaye.)

Table 5 Examples of three PoS tags based feature vectors.

Sentence	PoS_{i-2}	PoS_{i-1}	PoS_{i+1}	PoS_{i+2}	Labeled Sense
(a)	ADJ	NN	PP	P	Sense1
(b)	-	VB	TA	SC	Sense2
(c)	ADJ	CC	NN	P	Sense3

Table 4 shows example sentences of a target word دل (Dil) with different context window sizes. We applied three PoS tags based approaches i.e. PoS-1x1, PoS-2x2, and PoS-3x3 on the ULS-WSD-18 corpus. PoS-1x1 means to construct a feature vector PoS tags of one word from the right and the left of target word is considered and for PoS-2x2 PoS tags of two words from the right and the left are considered and so on. Table 5 shows an example of three

feature vectors (2x2 of word دل (Dil)) with their labeled senses. We converted all 50 data files (training and testing) within ULS-WSD-18 corpus into PoS based feature vectors, which were used to train and test a range of machine learning algorithms.

4.1.3 Bag of Words (BoW) Approach

BoW is another method extensively used for supervised WSD algorithms [54]. In this approach, the words surrounding the target are used as features. The positions of words in the context are ignored and the features simply encode whether a particular word appears in the context or not. All of the words surrounding words a target word were used to construct the feature.

4.1.4 Word Embeddings (WE) Approach

WE are low dimensional vectors designed to represent word semantics. They do not produce sparse vectors unlike some alternative approaches to distributional semantics. The widely applied Word2Vec embeddings [55] were used to create another WSD method. There are two variants of this model: (1) Continuous Bag Of Words (CBOW) and (2) Skip-gram. In the CBOW architecture, a system aims to predict the nearest word on the basis of provided context words, in contrast, in the skip-gram model a system aim to predict nearest words on the basis of a given target word. The skip-gram architecture was used for this study.

To accurately train a WE model, a large amount of training data is required [56]. For these experiments, a WE model (with skip-gram architecture) was trained on the entire UrMono corpus, which contains 94.5 million tokens of Urdu. We used Word2Vec with the *deeplearning4java* library[16] [57]. The trained WE model was used to extract nearest word embeddings for all 50 target words in the ULS-WSD-18 corpus. The number of nearest words extracted for each target word were: 100, 200, 300, 400 and 500. Feature vectors (based on nearest neighbors) were used to train and test the various machine learning algorithms used in this study.

4.2 Evaluation Methodology

Following standard practice in WSD research, the problem of correctly tagging the sense of a target word in a given sentence is treated as a supervised classification task.

We applied three feature extraction techniques i.e. PoS tags based approaches, BoW approaches and WE approaches (see Section 4.1). Five different machine learning algorithms were explored: Naive Bayes [58], Support

[16] A Java based library to implement WE models - `https://deeplearning4j.org/` Last visited: 23-October-2018

Vector Machine (SVM) [59], K-Nearest Neighbor [60], ID3 [61], and Multilayer Perceptron [62]. Each classifier was separately trained and tested for each target word using the training and test splits. Results are averaged over all 50 words in the corpus.

4.3 Evaluation Measures

Measures are borrowed from Machine Learning and Information Retrieval are conventionally used to evaluate WSD [1]. In this study, the evaluation was carried out using the most widely used evaluation measures i.e., Accuracy, Precision, Recall and F_1 measures. Consider the confusion matrix shown in Table 6 [63]. On the basis of this confusion matrix, evaluation measures can be defined as follows:

Table 6 Confusion matrix for binary classification problem.

Data Classes	Classified as Positive	Classified as Negative
Positive	True Positive (TP)	False Negative (FN)
Negative	False Positive (FP)	True Negative (TN)

The Accuracy (A) of a WSD system is defined as the proportion of the total number of predictions that were correct.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

The Precision (P) of a WSD system is the proportion of the predicted positive cases that were correct.

$$P = \frac{TP}{TP + FP} \tag{2}$$

The Recall (R) of a WSD system is defined as the proportion of positive cases that were correctly identified.

$$R = \frac{TP}{TP + FN} \tag{3}$$

F_1 measure is a specific relationship (harmonic mean) between precision (P) and recall (R).

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{4}$$

Averaged Accuracy, Precision, Recall and F_1 scores are computed using all the 50 target words in the proposed ULS-WSD-18 corpus and are reported in this study.

5 Results and Analysis

Table 7 shows averaged accuracy, Precision, Recall and F_1-measure scores for various WSD techniques used in this study. Table 8 shows averaged accuracy scores obtained on three PoS categories (nouns, verbs, and adjectives) when four different WSD approaches are applied. In these tables, "Technique" refers to the WSD feature extraction approach used to identify the proper sense of a word in a particular sentence. "ML Algorithm" refers to the Machine Learning algorithms used for training and testing models based on features extracted using WSD approaches. PoS-1x1 means that a feature vector is formed by considering the PoS tag of one word from the left, and one word from right side of target word. Similarly, PoS-2x2 means that feature vectors are formed by considering the PoS tags of two words from the left, and two words from right side of target word and so on. "WE-100" means the 100 neighboring words (from trained WE model) of target word are used to form a feature vector, "WE-200" means the 200 words (from the trained WE model) neighboring the target word are used and so on. "SVM" refers to Support Vector Machine classifier. "KNN" refers to K-Nearest Neighbors classifier. "ID3" refers to Iterative Dichotomiser 3 classifier.

Table 7 shows that all three approaches outperform the MFS baseline approach (Accuracy = 68.763). The best results are obtained using the BoW approach (Accuracy = 81.495 and F_1= 0.784). For the PoS tags based approach, the highest score is obtained using PoS-3x3 approach (Accuracy = 73.437 and F_1 =0.674). However, this performance is almost the same as obtained for PoS-1x1 and PoS-2x2 approaches, demonstrating that variation in the context window does not have an impact on WSD performance on our proposed corpus. For the WE approach, the highest score is obtained using WE-200 approach (Accuracy=72.361 and F_1= 0.628). Varying the size of the embedding does not greatly affect the results, and accuracy is broadly similar across all settings. Although accuracy is lower (71.842%) for WE-500.

Comparing machine learning algorithms, the highest results are obtained using Naive Bayes in most cases. Also, the overall highest results are obtained when Naive Bayes is used with BoW features.

Table 8 shows the averaged accuracy broken down by PoS catgeory. Overall, these results show that the WSD system performs better when disambiguating nouns than verbs and adjectives. A possible reason is nouns can be disambiguated more accurately using the MFS approach (Accuracy= 71.564%). The best results are obtained using the BOW approach with averaged accuracy score of 82.939%, 77.055% and 81.192% for nouns, verbs and adjectives respectively.

The highest performance for the PoS tag based approaches is obtained by using PoS-3x3 for nouns (Accuracy = 75.260%), PoS-2x2 for verbs (Accuracy = 71.620%) and PoS-1x1 for adjectives (Accuracy = 71.344%). Results for nouns are higher than other PoS categories, similar to the pattern of results observed for the BoW approach. Previous work on WSD has also found the most useful feature to vary by PoS.

Table 7 Results obtained using various WSD approaches with machine learning algorithms on ULS-WSD-18 corpus.

Technique	ML Algorithm	Accuracy	Precision	Recall	F_1
BoW	Naive Bayes	**81.495**	**0.790**	**0.814**	**0.784**
	SVM	78.824	0.767	0.788	0.760
	ID3	73.591	0.683	0.735	0.694
	KNN	73.953	0.690	0.739	0.696
	Multilayer Perceptron	76.8892	0.750	0.768	0.746
PoS-3x3	Naive Bayes	73.437	0.647	0.734	0.674
	SVM	69.414	0.672	0.694	0.673
	ID3	72.511	0.610	0.725	0.653
	KNN	67.322	0.673	0.673	0.663
	Multilayer Perceptron	67.818	0.661	0.678	0.662
PoS-2x2	Naive Bayes	73.417	0.647	0.734	0.673
	SVM	70.932	0.657	0.709	0.675
	ID3	72.560	0.609	0.725	0.653
	KNN	69.101	0.676	0.691	0.674
	Multilayer Perceptron	68.358	0.661	0.683	0.665
PoS-1x1	Naive Bayes	73.173	0.631	0.731	0.662
	SVM	72.619	0.655	0.726	0.675
	ID3	60.429	0.600	0.723	0.642
	KNN	70.301	0.648	0.703	0.663
	Multilayer Perceptron	69.743	0.653	0.697	0.664
WE-100	NaÃŕve Bayes	71.836	0.585	0.718	0.623
	SVM	72.032	0.608	0.720	0.640
	ID3	71.738	0.593	0.717	0.632
	KNN	71.488	0.622	0.714	0.638
	Multilayer Perceptron	70.507	0.628	0.705	0.634
WE-200	Naive Bayes	72.361	0.589	0.723	0.628
	SVM	72.348	0.629	0.723	0.644
	ID3	71.855	0.603	0.718	0.636
	KNN	71.522	0.631	0.715	0.643
	Multilayer Perceptron	71.430	0.643	0.714	0.648
WE-300	Naive Bayes	72.233	0.625	0.722	0.648
	SVM	72.233	0.625	0.722	0.648
	ID3	72.171	0.610	0.721	0.642
	KNN	71.664	0.651	0.716	0.653
	Multilayer Perceptron	70.748	0.641	0.707	0.647
WE-400	NaÃŕve Bayes	71.888	0.572	0.718	0.619
	SVM	72.081	0.627	0.720	0.650
	ID3	71.847	0.609	0.718	0.641
	KNN	71.444	0.644	0.714	0.653
	Multilayer Perceptron	70.535	0.640	0.705	0.650
WE-500	Naive Bayes	71.842	0.574	0.718	0.618
	SMO	71.554	0.620	0.715	0.647
	ID3	71.434	0.606	0.714	0.636
	KNN	70.978	0.639	0.709	0.651
	Multilayer Perceptron	69.996	0.639	0.699	0.650
MFS (Baseline)	-	68.763	-	-	-

Table 8 Results obtained using various WSD approaches and machine learning algorithms with PoS categorization.

Technique	ML Algorithm	Nouns	Verbs	Adjectives
BoW	Naive Bayes	**82.939**	**77.055**	**81.192**
	SVM	79.888	73.528	80.256
	ID3	75.065	68.772	73.515
	KNN	76.657	67.682	71.708
	Multilayer Perceptron	78.944	70.830	76.242
PoS-3x3	Naive Bayes	75.260	70.158	71.149
	SVM	71.179	67.698	66.005
	ID3	74.781	67.906	70.089
	KNN	69.057	65.654	63.957
	Multilayer Perceptron	69.943	64.498	64.740
PoS-2x2	Naive Bayes	75.127	71.620	70.225
	SVM	73.257	64.906	68.636
	ID3	74.842	67.759	70.263
	KNN	71.439	65.576	65.608
	Multilayer Perceptron	70.497	64.113	65.996
PoS-1x1	Naive Bayes	74.388	71.357	71.344
	SVM	74.274	69.876	70.349
	ID3	61.421	55.274	61.941
	KNN	71.279	69.188	68.542
	Multilayer Perceptron	70.436	68.432	68.926
WE-100	Naive Bayes	73.860	66.839	70.405
	SVM	73.928	66.760	71.171
	ID3	73.841	65.821	70.846
	KNN	73.478	65.468	70.986
	Multilayer Perceptron	72.638	62.797	71.002
WE-200	Naive Bayes	73.897	68.890	71.011
	SVM	73.822	68.374	71.580
	ID3	73.826	66.301	71.027
	KNN	73.273	64.851	72.203
	Multilayer Perceptron	73.118	65.549	71.637
WE-300	Naive Bayes	73.625	67.718	72.130
	SVM	73.625	67.718	72.130
	ID3	74.071	66.529	71.605
	KNN	73.440	64.314	72.836
	Multilayer Perceptron	72.929	62.187	71.804
WE-400	Naive Bayes	71.804	67.623	70.203
	SVM	73.924	67.027	71.189
	ID3	74.126	66.643	69.888
	KNN	72.963	65.497	72.167
	Multilayer Perceptron	72.775	61.959	71.440
WE-500	Naive Bayes	73.658	67.794	70.203
	SVM	73.381	66.368	70.813
	ID3	73.982	65.076	69.686
	KNN	73.015	64.403	70.802
	Multilayer Perceptron	72.685	60.616	70.335
MFS (Baseline)	-	71.564	63.284	65.605

Table 9 Five sentences, which were wrongly classified by BoW approach and Naive Bayes machine learning algorithm for the target word مکمل (Mukammal).

Sr No	Sentence with roman Urdu transliteration	Correct	Predicted
1	ان کے مطابق اہلکاروں نے اس شرط پر ہتھیار ڈالے ہیں کے انہیں **مکمل** تحفظ فراہم کیا جائے گا ۔ Un kay mutabiq ehalkaron nay is shart par hathyaar dale hain kay inhen **mukammal** tahaffuz frahem kya jaye ga.	Sense2	Sense1
2	(**مکمل** اردو پیکج انسٹال کرنے کے لیے اچھا فری ہوسٹ ہے) لہذا یہاں میں نے اردو پی ایچ پی فیوژن کی کمپریس فائل اپلوڈ کر کے انسٹال کی ہے ۔ (**Mukammal** Urdu package install karne kay liye acha free host hay) lehaza yahan mein nay Urdu PHP fusion ki compress file upload kar kay install ki hay.	Sense4	Sense1
3	علی الصبح بازار سے کتابوں کی خریداری **مکمل** کرنے کے بعد ہم لذت کام و دہن کی آزمائش کی خاطر کراچی صدر کی مشہور صابری نہاری کا قصد کیے چند احباب کی معیت میں نکلا ہی چاہتے تھے کے ایک کرم فرما نے رستے میں روکا اور سرگوشی کے انداز میں کہا چپکے سے چلے آئیے ۔ Ali-alsubah bazaar say kitabon ki kharidari **mukammal** karnae kay baad hum lazzat kaam-o-dahan ki azmaish ki khatir Karachi saddar ki mashhoor Sabri nehari ka qasad kiye chand ahbaab ki mayt mein nikla hi chahte thay kay aik karam farma nay rastae mein roka aur sargoshi kay andaaz mein kaha chupkay say chalay aayea.	Sense1	Sense4
4	یعنی **مکمل** ووٹنگ کی بنیاد پر جمہوریت ہو ۔ Yani **mukammal** voting ki bunyaad par jamhoriat ho.	Sense4	Sesne1
5	دوسرا ہاف **مکمل** طور پر آسٹریلیا کے نام رہا اور اس نے مزید چار گول کیے ۔ Dosra half **mukammal** tor par Australia kay naam raha aur us nay mazeed chaar goal kyae.	Sense4	Sense1

Simlarly, the best performance for the WE approach is obtained for nouns using the WE–300 approach (Accuracy = 74.071%). The highest results for verbs and adjectives are obtained using WE-200 and WE-300 (68.890% and 72.836% respectively). Again different context sizes are shown to produce different results for three PoS categories (nouns, verbs and adjectives).

Finally, accuracy scores using the MFS approach are 71.564% for nouns, 63.284% for verbs and 65.605% for adjectives. This method produces the lowest performance among all of the four WSD approaches that were applied.

Table 9 shows the correct senses and wrongly classified senses (using BoW approach with Naive Bayes machine learning algorithm) for the target word مکمل (Mukammal). In this table, "Correct" means the actual sense manually assigned by human annotators and "Predicted" means the sense predicted by the algorithm. The word مکمل (Mukammal) has four senses in the sense inventory,: تکمیل کیا گیا (Takmeel kya gaya), بھرپور (Bharpoor), خالص (Khalis), and

کلی، بے کم (Kulie, Bekam). The majority of mistakes for the examples in Table 9 are between Sense1 and Sense4, reflecting the algorithm's behaviour.

6 Conclusion

Urdu is a widely spoken but severely under-resourced language in terms of corpora suitable for NLP purposes. Our novel contribution as described in this paper is a newly developed and freely available benchmark corpus for the Urdu Lexical Sample WSD task. The corpus contains 50 target words (30 nouns, 11 adjectives, and 9 verbs) and 7,185 sentences. In addition to the creation of the dataset, we applied four baseline WSD approaches to the corpus in order to evaluate their suitability. The results of our WSD experiments show that the BoW approach gives the highest performance. In the future, we plan to continue the work by applying other WSD approaches to the Lexical Sample WSD corpus.

References

1. Navigli, R., 2009. Word sense disambiguation: A survey. ACM Computing Surveys (CSUR), 41(2), p.10.
2. Iacobacci, I., Pilehvar, M.T. and Navigli, R., 2016. Embeddings for word sense disambiguation: An evaluation study. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Vol. 1, pp. 897-907).
3. Schütze, H., Manning, C.D. and Raghavan, P., 2008. Introduction to information retrieval (Vol. 39). Cambridge University Press.
4. Hutchins, W.J., 1995. Machine translation: A brief history. In Concise history of the language sciences (pp. 431-445).
5. Jiang, J., 2012. Information extraction from text. In Mining text data (pp. 11-41). Springer, Boston, MA.
6. Prasad, B.D., 2008. Content analysis. Research methods for social work, 5, pp.1-20.
7. Lin, C.Y. and Hovy, E., 2000, July. The automated acquisition of topic signatures for text summarization. In Proceedings of the 18th conference on Computational linguistics-Volume 1 (pp. 495-501). Association for Computational Linguistics.
8. Gee, J.P. and Green, J.L., 1998. Chapter 4: Discourse analysis, learning, and social practice: A methodological study. Review of research in education, 23(1), pp.119-169.
9. DiMarco, C., Covvey, H., Cowan, D., DiCiccio, V., Hovy, E., Lipa, J. and Mulholland, D., 2007. The development of a natural language generation system for personalized e-health information. In Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems (p. 2339). IOS Press.
10. Edmonds, P. and Cotton, S., 2001, July. SENSEVAL-2: overview. In The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (pp. 1-5). Association for Computational Linguistics.
11. Hussain, S., 2008, January. Resources for Urdu Language Processing. In IJCNLP (pp. 99-100).
12. Riaz, K., 2010, July. Rule-based named entity recognition in Urdu. In Proceedings of the 2010 named entities workshop (pp. 126-135). Association for Computational Linguistics.
13. Daud, A., Khan, W. , and Che, D. 2016. Urdu language processing: a survey. Artificial Intelligence Review: 1-33. doi: 10.1007/s10462-016-9482-x
14. Rahman, T., 2004, January. Language policy and localization in Pakistan: proposal for a paradigmatic shift. In SCALLA Conference on Computational Linguistics (Vol. 99, p. 100).
15. Jawaid, B., Kamran, A. and Bojar, O., 2014, May. A Tagged Corpus and a Tagger for Urdu. In LREC (pp. 2938-2943).

16. Naseer, A. and Hussain, S., 2009. Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification. Center for Research in Urdu Language Processing, Lahore, Pakistan.
17. Ng, H.T., Lim, C.Y. and Foo, S.K., Ng1. A case study on inter-annotator agreement for word sense disambiguation. SIGLEX99: Standardizing Lexical Resources.
18. Board, U.D., 2008. Urdu Lughat. Urdu Lughat Board, Karachi, Pakistan.
19. Palmer, M., Fellbaum, C., Cotton, S., Delfs, L. and Dang, H.T., 2001, July. English tasks: All-words and verb lexical sample. In The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (pp. 21-24). Association for Computational Linguistics.
20. Mihalcea, R., Chklovski, T. and Kilgarriff, A., 2004. The Senseval-3 English lexical sample task. In Proceedings of SENSEVAL-3, the third international workshop on the evaluation of systems for the semantic analysis of text.
21. Chklovski, T.A., Mihalcea, R., Pedersen, T. and Purandare, A., 2004, July. The Senseval-3 multilingual English-Hindi lexical sample task. Association for Computational Linguistics.
22. Edmonds, P., 2002. SENSEVAL: The evaluation of word sense disambiguation systems. ELRA newsletter, 7(3), pp.5-14.
23. Kilgarriff, A., 2004, September. How dominant is the commonest sense of a word?. In International Conference on Text, Speech and Dialogue (pp. 103-111). Springer Berlin Heidelberg.
24. Taghipour, K. and Ng, H.T., 2015. One million sense-tagged instances for word sense disambiguation and induction. In Proceedings of the Nineteenth Conference on Computational Natural Language Learning (pp. 338-344).
25. Francis, W.N. and Kucera, H., 1979. Brown corpus manual. Brown University.
26. Vossen, P., Izquierdo, R. and Görög, A., 2013. DutchSemCor: in quest of the ideal sense-tagged corpus. In Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013 (pp. 710-718).
27. Bond, F., Baldwin, T., Fothergill, R. and Uchimoto, K., 2012, January. Japanese SemCor: A sense-tagged corpus of Japanese. In Proceedings of the 6th Global WordNet Conference (GWC 2012) (pp. 56-63).
28. Passonneau, R.J., Baker, C., Fellbaum, C. and Ide, N., 2012, May. The MASC word sense sentence corpus. In Proceedings of LREC.
29. McKean, E., 2005. The new oxford American dictionary (Vol. 2). New York: Oxford University Press.
30. Landes, S., Leacock, C. and Tengi, R., 1998. Building a semantic concordance of english. WordNet: An electronic lexical database and some applications. MIT Press, Cambridge, MA.
31. Leacock, C., Towell, G. and Voorhees, E., 1993, March. Corpus-based statistical sense resolution. In Proceedings of the workshop on Human Language Technology (pp. 260-265). Association for Computational Linguistics.
32. Bruce, R. and Wiebe, J., 1994, June. Word-sense disambiguation using decomposable models. In Proceedings of the 32nd annual meeting on Association for Computational Linguistics (pp. 139-146). Association for Computational Linguistics.
33. Khan, S. N., Khan, K., Khan, A., Khan, A., Khan, A. U., & Ullah, B. (2018). Urdu Word Segmentation using Machine Learning Approaches. INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS, 9(6), 193-200.
34. Becker, D., Riaz, K., Bennett, B. H., Davis, E., & Panton, D. (2002, June). Named Entity Recognition in Urdu: A Progress Report. In International Conference on Internet Computing (pp. 757-761).
35. Sharjeel, M., Nawab, R. M. A., & Rayson, P. (2017). COUNTER: corpus of Urdu news text reuse. Language Resources and Evaluation, 51(3), 777-803.
36. Mishra, N. and Siddiqui, T.J., 2012. An Investigation to Semi supervised approach for HINDI Word sense disambiguation. Trends in Innovative Computing 2012-Intelligent Systems Design.
37. Appropriation, F.Y.A.F.Y., American Printing House for the Blind.
38. Arieff, A.I., 1912. Veterans Administration Medical Center in San Francisco. San Jose Mercury, 12.

39. McEnery, A., Baker, P., Gaizauskas, R. and Cunningham, H., 2000. EMILLE: Building a corpus of South Asian languages. VIVEK-BOMBAY-, 13(3), pp.22-28.
40. Baker, P., Hardie, A., McEnery, T., Cunningham, H. and Gaizauskas, R.J., 2002, May. EMILLE, A 67-Million Word Corpus of Indic Languages: Data Collection, Mark-up and Harmonisation. In LREC.
41. Urooj, S., Shams, S., Hussain, S. and Adeeba, F., 2014. Sense Tagged CLE Urdu Digest Corpus. In Proc. Conf. on Language and Technology, Karachi.
42. Zafar, A., Mahmood, A., Abdullah, F., Zahid, S., Hussain, S. and Mustafa, A., 2012. Developing urdu wordnet using the merge approach. In Proceedings of the Conference on Language and Technology (pp. 55-59).
43. Singh, U., Goyal, V. and Lehal, G.S., 2012. Named entity recognition system for Urdu. Proceedings of COLING 2012, pp.2507-2518.
44. Abid, M., Habib, A., Ashraf, J., & Shahid, A. (2017). Urdu word sense disambiguation using machine learning approach. Cluster Computing, 1-8.
45. Arif, S. Z., Yaqoob, M. M., Rehman, A., & Jamil, F. (2016). Word sense disambiguation for Urdu text by machine learning. International Journal of Computer Science and Information Security, 14(5), 738.
46. Sinha, M., Kumar, M., Pande, P., Kashyap, L., & Bhattacharyya, P. (2004, November). Hindi word sense disambiguation. In International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India.
47. Anand Kumar, M., Rajendran, S., & Soman, K. P. (2014). Tamil word sense disambiguation using Support Vector Machines with rich features. International Journal of Applied Engineering Research, 9(20), 7609-20.
48. Sreeganesh, T. (2006). Telugu parts of speech tagging in WSD. Language of India, 6.
49. Ahmed, T., Urooj, S., Hussain, S., Mustafa, A., Parveen, R., Adeeba, F., Hautli, A. and Butt, M., 2014. The CLE urdu POS tagset. In Poster presentation in Language Resources and Evaluation Conference (LERC 14).
50. Narayan, D., Chakrabarti, D., Pande, P. and Bhattacharyya, P., 2002, January. An experience in building the indo wordnet-a wordnet for hindi. In First International Conference on Global WordNet, Mysore, India.
51. Ide, N., 1998, May. Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In Proceedings of the First International Language Resources and Evaluation Conference (pp. 463-70).
52. Agirre, E. and Edmonds, P. eds., 2007. Word sense disambiguation: Algorithms and applications (Vol. 33). Springer Science & Business Media.
53. Bhingardive, S., Singh, D., Rudramurthy, V., Redkar, H. and Bhattacharyya, P., 2015. Unsupervised most frequent sense detection using word embeddings. In Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1238-1243).
54. Cai, J., Lee, W.S. and Teh, Y.W., 2007. Improving word sense disambiguation using topic features. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).
55. Mikolov, T., Chen, K., Corrado, G. and Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.
56. Liu, Y., Liu, Z., Chua, T.S. and Sun, M., 2015, January. Topical Word Embeddings. In AAAI (pp. 2418-2424).
57. Rong, X., 2014. word2vec parameter learning explained. arXiv preprint arXiv:1411.2738.
58. Pedersen, T., 2000, April. A simple approach to building ensembles of Naive Bayesian classifiers for word sense disambiguation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 63-69). Association for Computational Linguistics.
59. Huang, P.S., Damarla, T. and Hasegawa-Johnson, M., 2011, July. Multi-sensory features for personnel detection at border crossings. In Information Fusion (FUSION), 2011 Proceedings of the 14th International Conference on (pp. 1-8). IEEE.
60. Altman, N.S., 1992. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician, 46(3), pp.175-185.

61. Lior, R., 2014. Data mining with decision trees: theory and applications (Vol. 81). World scientific.
62. McCulloch, W.S. and Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), pp.115-133.
63. Sokolova, M. and Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), pp.427-437.