A Sense Annotated Corpus for All-Words Urdu Word Sense Disambiguation

ALI SAEED, COMSATS University Islamabad, Lahore Campus, Pakistan
RAO MUHAMMAD ADEEL NAWAB, COMSATS University Islamabad, Lahore Campus, Pakistan
MARK STEVENSON, University of Sheffield, UK
PAUL RAYSON, Lancaster University, UK

Word Sense Disambiguation (WSD) aims to automatically predict the correct sense of a word used in a given context. All human languages exhibit word sense ambiguity and resolving this ambiguity can be difficult. Standard benchmark resources are required to develop, compare and evaluate WSD techniques. These are available for many languages but not for Urdu, despite this being a language with more than 300 million speakers and large volumes of text available digitally. To fill this gap, this study proposes a novel benchmark corpus for the Urdu All-Words WSD task. The corpus contains 5,042 words of Urdu running text in which all ambiguous words (856 instances) are manually tagged with senses from the Urdu Lughat dictionary. A range of baseline WSD models based on *n*-grams are applied to the corpus and the best performance (accuracy of 57.71%) is achieved using word 4-grams. The corpus is freely available to the research community to encourage further WSD research in Urdu.

CCS Concepts: • **Computing methodologies → Language resources**;

Additional Key Words and Phrases: Word Sense Disambiguation, All-Words Task, Sense Tagged Urdu Corpus

ACM Reference Format:
Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019. A Sense Annotated Corpus for All-Words Urdu Word Sense Disambiguation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 1, 1 (February 2019), 14 pages. https://doi.org/0000001.0000001

1 INTRODUCTION

Word Sense Disambiguation is the problem of identifying the appropriate sense of a word when it is used in context. WSD is a long established and widely explored task in Natural Language Processing (NLP) [45]. Work on WSD has focused on English but the problem has also been explored for other languages [3, 39, 47, 65, 70]. WSD is an important problem since it has potential to improve performance in many areas of language processing, including Information Retrieval [59], Information Extraction [31], Machine Translation [27], Natural Language Understanding [4], Natural Language Generation [7], Text Summarization [22], Question Answering [63], Content Analysis [64], Lexicography [38], Word Processing [25], Text Classification [72], Discourse Analysis [52], Genre Identification [6], Document Indexing [12], Theme Extraction [50], Cross Document

Authors' addresses: Ali Saeed, COMSATS University Islamabad, Lahore Campus, 1.5 KM off Raiwind Rd, Near Labors Colony Defence Rd, Defence Road, Lahore, Punjab, 54000, Pakistan, fa15-pcs-004@ciitlahore.edu.pk; Rao Muhammad Adeel Nawab, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan, adeelnawab@cuilahore.edu.pk; Mark Stevenson, University of Sheffield, Sheffield , UK, mark.stevenson@sheffield.ac.uk; Paul Rayson, Lancaster University, Lancaster, UK, p.rayson@lancaster.ac.uk.

2375-4699/2019/2-ART $15.00
https://doi.org/0000001.0000001

Referencing [68], Automatic Summarization [24], Sentiment Analysis [21], Complex Searching Queries [14], and Semantic Web search [9].

Previous literature on WSD has explored two main tasks [19, 39]: (1) Lexical Sample (or Targeted Words) WSD and (2) All-Words WSD. The goal of the first task is to disambiguate a set of pre-defined targeted words in a given text while the aim of the second is to disambiguate all ambiguous words that appear in a particular text. Approaches to the lexical sample task generally involve training a classifier for each target word. This method is often an effective way to develop accurate WSD systems but requires annotated training data and can only be applied to the set of target words. Solutions to the all-words task are generally considered to be more useful for downstream applications, but the task is more challenging and obtaining training data for a very large lexical sample is impractical. The focus of our research is All-Words WSD for the Urdu language.

Researchers used various approaches for both tasks[1, 45] which can be grouped into (1) Knowledge based approaches and (2) Machine Learning approaches. Knowledge based approaches need external sources such as dictionaries, thesauri and ontologies to perform disambiguation. While, Machine Learning approaches rely on some form of training data and can be further be categorized into (1) Supervised (2) Unsupervised, and (3) Semi-supervised approaches. Supervised approaches require labeled examples of ambiguous words annotated with the correct sense and use this data to train WSD systems. Unsupervised approaches do not require labeled examples and while semi-supervised approaches use a combination of labeled and unlabeled data, however they are less widely applied for WSD than supervised approaches.

Nivigli [45] noted that supervised approaches are useful for the Lexical Sample WSD task. But for the All-Words WSD task supervised systems suffer from the problem of data sparseness since labeled training examples are unlikely to be available for all words that need to be disambiguate. No large annotated datasets are available for Urdu. The dataset developed in this research is too small to train supervised machine learning approaches and is used for testing only. Consequently, knowledge based approaches are applied in this work.

The All-Words WSD task has been studied for a wide range of languages including English, Italian, Czech, Dutch, Estonian and Chinese [3, 19, 39, 65]. However, South Asian Languages, including Urdu in particular, have not been explored in previous work. The work described here addressed that oversight.

Urdu is among one of the most important international languages. Urdu has around 300 million speakers mainly spread across 20 different countries [26, 32, 55] and is also spoken across the globe due to the significant South Asian diaspora [54]. Urdu is an Indo-Aryan language widely spoken in Pakistan, India, Jammu & Kashmir and Bangladesh. Its vocabulary and grammatical form is inherited from Arabic, Persian and other South Asian languages [54]. It is morphologically rich, verbs and nouns may have more than 40 forms, which makes it complicated to process [44].

To develop, evaluate and compare WSD systems benchmark corpora are needed, however none is available for Urdu. This study introduces a novel benchmark corpus for the Urdu All-Words WSD task.

The corpus (hereafter referred to as the UAW-WSD-18 corpus) contains 5,042 words of Urdu obtained from the UrMono corpus[1]. The corpus contains 466 ambiguous types and 856 ambiguous tokens. All instances of ambiguous words are tagged manually by three annotators using senses from the Urdu Lughat dictionary (initially each sentence is tagged by two annotators and conflicting entries are annotated by third person). A range of WSD approaches (e.g. Most Frequent Sense, Jaccard Similarity, Overlap Similarity, Dice Similarity, Euclidean Distance, Cosine Similarity and

[1] A freely available POS tagged corpus that can be downloaded from https://ufal.mff.cuni.cz/urmonocorp Last Visited: 16-October-2018

Voting Based Approach) are applied to the corpus, to demonstrate how the resource can be used for the development and evaluation of WSD systems. The sense annotated corpus (UAW-WSD-18), sense inventory (manually created using Urdu Lughat), gloss inventory (manually extracted from Urdu Lughat) and code (for both WordNGram and CharNGram approaches) are freely available for public use under Creative Commons license (CC-BY-NC-SA)[2]).

The remainder of this paper is organized as follows. Section 2 reviews existing evaluation resources for the WSD task. Section 3 describes the creation process for our corpus. Section 4 explains the WSD techniques applied to the corpus. Section 5 presents the results from applying these approaches and their analysis. Finally, Section 6 concludes the paper.

2 RELATED WORK

Researchers have developed a range of resources to support both Lexcial Sample and All-words WSD tasks although the vast majority of these have focused on English and other European languages.

The series of competitions organized by SENSEVAL/SemEval has been the most important effort in the development of WSD resources. The result of these competitions is a set of WSD benchmark corpora for both Lexical Sample and All-Words WSD tasks [19, 39, 53]. Corpora were created for a range of languages including English, Spanish, Swedish, Catalan, Basque, Italian, Japanese, Korean, Chinese, Dutch and Romanian. WordNet was the most commonly used sense inventory [20]. The SENSEVAL All-Words corpora contain 5,000 words of running text. The Lexical Sample corpora contained 75+15n sentences, where n represents number of senses in WordNet for a target word. In the majority of cases, source data was been taken from Wall Street Journal news articles [61].

Other corpora for Lexical Sample WSD tasks include the DSO corpus [47], the Line-hard-Serve corpus [36], the Interest corpus [13] and the Hindi Sense Tagged corpus [40].

Other corpora developed for the All-Words WSD task include (1) SEMCOR [33], (2) the Google WSD corpus [70] (3) the Italian Syntactic-Semantic Treebank (ISST) [41], and (4) the CLE Urdu Sense Tagged corpus [67]. SEMCOR contains 234,000 manually sense annotated sentences from the Brown corpus. Versions for a range of languages have been developed including DUTCH SEMCOR [69], JAPANESE SEMCOR [11], BASQUE SEMCOR [2], BULGARIAN SEMCOR [34], and SPANISH SEMCOR [28]. The Google WSD corpus is the largest manually annotated corpus of English. Text for this corpus was taken from SEMCOR WSD corpus and MASC WSD corpus [51] (a sense annotated corpus). It comprises of 248,000 sentences manually annotated using New Oxford American Dictionary (NOAD). The Italian Syntactic-Semantic Treebank (ISST) is created manually for Italian language [41]. Source data for the creation of ISST was taken from Balanced corpus [8] and Specialized corpus [41]. It contains 305,547 tokens including 81,236 content words tagged using Italian WordNet [57].

For the Urdu Lexical Sample WSD task, [46] recently developed a benchmark corpus called the ULS-WSD-18 corpus. This corpus contains 50 target words (30 nouns, 11 adjectives, and 9 adverbs) and 75+15n sentences for each target word, where n represents the number of senses in sense inventory. A standard, manually crafted dictionary called Urdu Lughat is used as a sense inventory and annotation carried out by three human annotators.

The Sense Tagged CLE Urdu Digest corpus [67] is the only sense tagged corpus for All-Words Urdu WSD described in the literature. Source data for this resource was obtained from CLE Urdu Digest corpus [67]. It contains 17,006 sense tagged instances manually annotated by a single tagger over a period of 10 months. CLE Urdu WordNet [71] was used as a sense inventory. The corpus suffers from some serious limitations. Firstly, the coverage of CLE Urdu WordNet is very limited. It contains only 5,000 words, and even common Urdu words are not included, such as دل (Heart), دیکھ

[2] https://comsatsnlpgroup.wordpress.com/ Last Visited: 14-October-2018

(Watch), عمر (Age), and روشنی (Light). In addition, the resource contains a small number of senses, for example, it returned two senses for word نظر (View) and only a single sense is for word بات (Talk). Secondly, the corpus was tagged by a single annotator.

Previous work on Urdu WSD has focused on Lexical Sample tasks [1, 5, 44]. These approaches are based on supervised learning methods. All three approaches applied Naive Bayes, [5] also used Support Vector Machines while [1] also used Support Vector Machines and Decision Trees.

Researchers have also explored other language processing tasks for Urdu including Word Segmentation [18, 37] and Named Entity Recognition [56, 60].

The problem explored in this paper, All-Words WSD for Urdu, is a difficult and important language processing problem. The All-Words WSD task can be viewed as being more challenging and complicated than the Lexical Sample WSD task since all ambiguous words have to be disambiguated. However, progress on All-Words WSD tasks has the potential to provide a generic solution to WSD problem. The two main challenges faced in the development of All-Words WSD approaches are (1) the creation of suitable resources and (2) development of disambiguation methods. Previous work has largely focused on developing these resources and methods for English, and some other languages, but not for Urdu. The work described here addressed this gap by proposing a novel resource (sense annotated corpus) and developing methods for Urdu WSD based on text similarity.

To conclude, benchmark All-Words WSD corpora have been developed for a wide range of languages, but not Urdu. This study attempts to fill this gap by developing a benchmark Urdu All-Words WSD corpus. In addition, this study also attempts to develop various text similarity methods for WSD including a specially designed approach called Voting Based Approach. As far as we are aware no other corpus is currently available for the Urdu All-Words task.

3 CORPUS

Fig. 1 shows the steps involved in the corpus generation process and these are described in the following subsections.

3.1 Source Data

The UAW-WSD-18 corpus was developed using the UrMono corpus [30] see Fig. 1 (A). The UrMono corpus was selected since it is the largest available dataset for Urdu and is readily available for non-commercial research purposes[3]. It contains 95.4 million Urdu words and 5.4 million sentences. The UrMono corpus is tokenized and POS tagged using the CLE POS tagset [58] with an accuracy of 87.98%. The corpus contains documents from a range of domains including news, religion, blogs, literature, science, and education.

3.2 Text Selection

A significant amount of text is needed to develop a useful sense annotated Urdu WSD corpus. We selected 5,042 words of running text from the UrMono corpus , see Fig. 1 (B) and (C). The main motivation for the selection of this subset is that the SENSEVAL guidelines suggested at least 5,000 words of running text is required to adequately evaluate the All-Words WSD task, and all content words should be tagged [19]. Content words (nouns, verbs, adjectives and adverbs) were extracted from the running text of 5,000 words resulting in a total of 2,306 content words (1,315 nouns, 567 verbs, 328 adjectives and 96 adverbs). The set of 2,306 content words wer manually inspected to remove non-ambiguous words, which resulted in a subset of ..... ambiguous words.

We extracted content words (verbs, nouns.....) form the WSD Corpus, thereofre, stop words were automatically removed.

[3] https://ufal.mff.cuni.cz/urmonocorp Last Visited: 16-October-2018

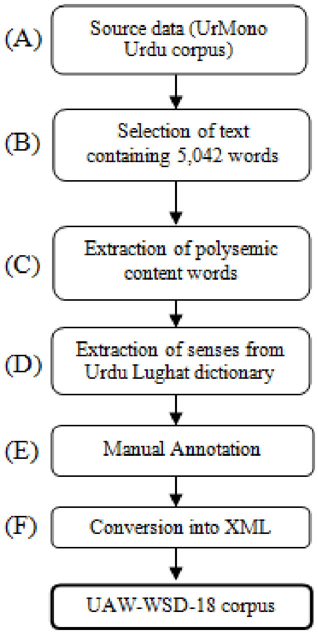

(A)
Source data (UrMono Urdu corpus)
(B)
Selection of text containing 5,042 words
(C)
Extraction of polysemic content words
(D)
Extraction of senses from Urdu Lughat dictionary
(E)
Manual Annotation
(F)
Conversion into XML
UAW-WSD-18 corpus

Fig. 1. Explaining corpus generation process

[4] Stop words are commonly occurring words in a language with no meanings [29]. These stop words can be ignored for WSD task [? ? ?]. For example in English the words “to", “on", “an" and “the" are stop words. Similarly the words “گی", “خو", “اور", and “تک", are the examples of Urdu stop words. Stop words from the on-line available link[5] were removed, leaving 1,378 words. Senses for the remaining words were manually retrieved from the Urdu Lughat dictionary, a comprehensive Urdu to Urdu dictionary and can be accessed via an online interface[6] (see Section 3.3 for details of Urdu Lughat). Urdu Lughat returned multiple senses for only 466 words (342 nouns, 14 verbs, 10 adverbs, and 100 adjectives). For the remaining 912 words, Urdu Lughat either did not return any senses or only returned a single sense.

The complete list of 466 polysemous words along with their senses and glosses (i.e. descriptive examples) is available for download[7].

3.3 Sense Inventory

The selection of the sense inventory for the creation of UAW-WSD-18 corpus is an important decision. Three resources (1) Indo WordNet [43], (2) CLE Urdu WordNet [71], and (3) Urdu Lughat [10] were considered. These resources were compared by manually inspecting the senses they return for the two most frequent words in our corpus: الله (God), which appears 22 times, and بات (talk), which appears 12 times.

[4] As we extracted the content words from our proposed WSD corpus, therefore, stop words (most frequent words) were automatically removed.
[5] https://www.ranks.nl/stopwords/urdu Last Visited:16-October-2018
[6] http://urdulughat.info/ Last Visited: 16-October-2018
[7] https://comsatsnlpgroup.wordpress.com/ Last Visited: 14-October-2018

Table 1. An example of two instances from the sense inventory

Word	POS	Fr	Sense 1	Sense 2	Sense 3	Sense 4	Sense 5	Sense 6	Sense 7	Sense 8	Sense 9
الله (God)	NN	22	معبود، خدائے تعالٰی کا اسم ذات، اسماء صفاتی کے مقابل (God)	استعجاب کے پر، موقع مترادف: حیرت ہے، تعجب ہے (Strange)	شکوئے شکایت یا دوئی موقع کے پر (Complaining)	کمال اضطراب اور کے یاس پر موقع (Wonder)	کے تمنا پر، موقع مترادف: خدا کرئے (Wish)	دعا کے وغیرہ پر موقع (Prayer)	ناجانے، خدا معلوم، الله ہی کو معلوم ہے (Unknown)	الله، کے خدا لیے، خدارا خوشامد (Welcome)	کسی شخص کا نام جیسے الله بخش (Name)
بات (Talk)	NN	12	لفظ، بول، کلمہ، فقرہ، گفتگو، قول (Word)	کہاوت (Saying)	خیال (Idea)	واقع، ماجرا، سرگزشت (Event)	منگنی، کی لڑکی شادی کا پیام سلام، بیاہ کی نسبت (Engagement)	حادثاتی کیفیت، واردات (Accident)			

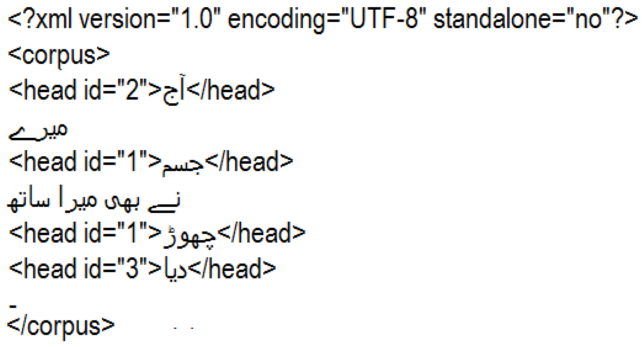
The focus of Indo WordNet project was to develop WordNets for the languages of India including Hindi, Marathi, Urdu, Tamil, Malyalam and Telugu. We applied the word الله (God) as input to Indo WordNet using its online interface[8] and it failed to return any senses. Indo WordNet only returned a single sense for the word بات (Talk).

CLE Urdu WordNet[9] contains 5,000 unique words along with their senses. Again, like Indo WordNet, sense coverage of the terms contained in our corpus was too low in this resource for it to be useful. الله (God) was not found in CLE and only a single sense is returned for بات (Talk).

The third choice, the Urdu Lughat [10], is a comprehensive Urdu dictionary created by the Dictionary Board, Karachi, Pakistan and is freely available for research purposes through an online interface[10]. The dictionary contains approximately 120,000 unique words with multiple senses, synonyms, glosses, and descriptive examples. We found multiple senses for large number of selected ambiguous words available in the proposed corpus. As Table 1 shows, Urdu Lughat returns nine senses for الله (God) and six for بات (Talk).

To conclude, manual inspection highlights the fact that the most suitable resource for generation of sense inventory was Urdu Lughat and therefore this resource was used to construct the corpus.

[8] http://www.cfilt.iitb.ac.in/indowordnet/ Last Visited: 16-October-2018
[9] http://www.cle.org.pk/clestore/urduwordnet.htm Last Visited: 16-October-2018
[10] http://urdulughat.info/



```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<corpus>
<head id="2">آج</head>
میرے
<head id="1">جسم</head>
نے بھی میرا ساتھ
<head id="1">چھوڑ</head>
<head id="3">دیا</head>
-
</corpus>
```

Fig. 2. An example sentence obtained from proposed corpus, in XML format

3.4 Annotations and Inter Annotator Agreement

All ambiguous words in UAW-WSD-18 corpus were manually annotated by three human annotators (see Fig. 1 (E)). All annotators were native Urdu speakers with understanding of the WSD task. In the first stage two annotators (A and B) annotated a subset of 200 ambiguous words. Annotations were compared and reasons for conflict discussed. In the second stage the complete corpus was annotated by A and B. Disagreements were resolved by a third annotator (annotator C). The Inter-Annotator Agreement (IAA) achieved on our proposed UAW-WSD-18 corpus is 90.18%. This result shows a good agreement, highlighting the fact that annotators produced annotations of high quality.

3.5 Corpus Standardization

The corpus (UAW-WSD-18 corpus) is formatted using a standard XML format from SENSEVAL[11] [19] (see 1 (F)). The corpus is stored in a single file in which all ambiguous words are tagged. Fig. 2 shows an example of a single sentence from the UAW-WSD-18 corpus. The <corpus> tag appears as the root element of the entire corpus and all polysemous words are enclosed in a <head> tag. This particular sentence contains four polysemous words, آج (Today), جسم (Body), چھوڑ (Leave), and دیا (Burner). “ID" is an attribute of the <head> tag and contains the sense number, corresponding to the sense from the sense inventory manually assigned by human taggers.

3.6 Corpus Characteristics

The corpus contains 5042 words and 252 sentences with an average length of 20 tokens. There are 856 sense annotated tokens and 466 unique types. The distribution of sense across types is shown in Table 2.

4 EXPERIMENTAL SETUP

We carried out a set of experiments to demonstrate how our proposed dataset (UAW-WSD-18 corpus) can be used for the development and evaluation of All-Words WSD systems for the Urdu language. This section describes the All-Words Word Sense Disambiguation approaches, the dataset used for experiments, evaluation methodology and evaluation measures.

[11] http://www.hipposmond.com/senseval2/ Last Visited: 17-October-2018

Table 2. Number of senses for types in corpus

No. senses	Total Count	% of Total
2	119	(26.68%)
3	104	(23.31%)
4	74	(15.87%)
5	65	(13.94%)
6	45	(9.65%)
7	23	(4.93%)
8	26	(5.57%)
9	5	(1.07%)
10	4	(0.85%)
11	1	(0.21%)

4.1 Approaches

Most Frequent Sense. For polysemouns words a single sense often occurs more frequently than the others [3] and is referred to as the Most Frequent Sense (MFS). We applied MFS approach on the entire corpus by assuming the first sense in the dictionary is the most frequent and reported average accuracy score (see Table 4). MFS is considered a strong baseline for All-Words WSD systems [45].

N-gram Models. The similarity between a pair of sentences can be computed by counting the *n*-grams that they have in common, an approach commonly used for WSD [17, 39]. An ambiguous word is disambiguated by comparing its context against the gloss of each possible sense and choosing the one with the highest similarity. A gloss is a textual description of a meaning and may also contain usage examples. Table 3 shows an example of two words (اللہ (God) and بات (Talk)) with their glosses obtained from the Urdu Lughat dictionary.

Similarity Coefficients. Given a target sentence TS (which contains the ambiguous content word) and the gloss of a particular sense G_i (where i represents the sense number). The similarity score using Jaccard Similarity Coefficient (JSC) [49, 66], Overlap Similarity Coefficient (OSC) [16, 35], Dice Similarity Coefficient (DSC) [42], Euclidean Distance Measure (EDM) [23], and Cosine Similarity Coefficient (CSC) [15, 48] are computed using the following equations and the sense number with highest similarity is selected.

Let $T_1 = S(TS, n)$ and $T_2 = S(G_i, n)$ represent the set of (word or character) n-grams of length n for TS and G_i respectively, then

$$JSC(T_1, T_2) = \frac{|T_1 \cap T_2|}{|T_1 \cup T_2|} \tag{1}$$

$$OSC(T_1, T_2) = \frac{|T_1 \cap T_2|}{min(|T_1|, |T_2|)} \tag{2}$$

$$DSC(T_1, T_2) = \frac{2|T_1 \cap T_2|}{|T_1| + |T_2|} \tag{3}$$

$$EDM(T_1, T_2) = \sqrt{\sum_{i=1}^{n} (T_{1i} - T_{2i})} \tag{4}$$

$$CSC(T_1, T_2) = \frac{T_1.T_2}{|T_1|^2.|T_2|^2} = \frac{\sum_{i=1}^{n}(T_{1i}.T_{2i})}{\sqrt{\sum_{i=1}^{n} T_{1i}^2}\sqrt{\sum_{i=1}^{n} T_{2i}^2}} \tag{5}$$

It can also be noted that, to apply EDM and CSC we first converted sentences T_1 and T_2 from to numeric representation (“one hot" encoding).

Voting Based Approach (VBA). The Voting Based Approach (VBA) is a simple approach which assigns a sense based on the maximum number of votes it has received from the previous approaches (JSC, OSC, DSC, EDM, and CSC). For example, for a particular sentence S_1, if the first three methods JSC, OSC and DSC predict sense number 2, and remaining two methods EDM and CSC predict sense number 1, then VBA will assign sense 2. However, if there is a tie then, VBA assigns the sense with the lowest number.

Table 3. An example of two instances with their corresponding glosses obtained from the source sense inventory

Word	Sense No.	Sense	Gloss obtained from Urdu Lughat dictionary
الله	Sense 1	معبود، خدائے تعالٰی کا اسم ذات، اسماء صفاتی کے مقابل	آپ کے جانے کے بعد سے گھر بھائیں بھائیں کر رہا ہے الله اس دم کو رکھے -
	Sense 2	استعجاب کے موقع پر، تعجب ہے مترادف: حیرت ہے،	کیوں گر پڑی ہو، خیر تو ہے کیا ہوا اسےالله! میرے ہونٹ بھی زہریلے ہوگئے -
	Sense 3	شفوے شکایت یا دوئی کے موقع پر-	الله اختر تم اتنے کٹھور بھی ہو سکتے ہو آخر کس لیے -
	Sense 4	کمال اضطراب اور یاس کے موقع پر-	تسکین مرگ بھول گیا اضطراب میں الله پڑ گیا مرا دل کس عذاب میں -
	Sense 5	تمنا کے موقع پر، مترادف: خدا کرے -	جنون عشق نے سودائے شوق نے کھویا دماغ سے کہیں الله خلل جائے -
	Sense 6	دعا وغیرہ کے موقع پر	بیٹھے بٹھائے عشق کا آزار ہو گیا الله کس بلا میں گرفتار ہو گیا -
	Sense 7	نا جانے، خدا معلوم، الله ہی کو معلوم ہے -	الله کدھر ہے در کا شانۂ ساقی مدہوش ہے وارفتۂ پیمانۂ ساقی -
	Sense 8	الله ، خدا کے لیے، خدارا، خوشامد	الله خرم بھائی آپ ڈرائیے نہیں -
	Sense 9	کسی شخص کا نام جیسے الله بخش	بیت الله محسود سرحد کے علاقے میں پیدا ہوئے -
بات	Sense 1	گفتگو، قول ـ لفظ، بول، فقرہ،	چلا جب کہہ کر پکارا اسے بات اور پیغام بر رہ گئی -
	Sense 2	کہاوت	زلف رخ پر جو چھٹی رات ہوئی یا نا ہوئی کیوں، نا کہتا تھا میری بات ہوئی یا نا ہوئی -
	Sense 3	خیال	میرے دل میں اک بات آئی ہے -
	Sense 4	واقع، ماجرا، سرگزشت	بات اک یاد آئی ہے مجھ کو میری آنکھوں کےآگے گزری جو -
	Sense 5	منگنی، لڑکی کی شادی کا پیام سلام، بیاہ کی نسبت	بچی کو دیکھ دیکھ کر اس کے ہوش اڑے جاتے ہے مگر کوئی بات ڈھنگ کی نہیں ملتی تھی -
	Sense 6	حادثاتی کیفیت، واردات	کل سے تم اداس اداس ہو کچھ شاید کوئی بات ہو گئی ہے -

4.2 Evaluation Methodology

The entire UAW-WSD-18 corpus was used for the experiments. The text similarity based methods described above were used to carry out WSD. If a particular text similarity approach returns the same value for multiple senses, then the system tags a particular ambiguous word with lowest sense number, obtained from Urdu Lughat. For example, if a text similarity coefficient returns same similarity scores for Sense 1, Sense 3 and Sense 5 then the WSD system will assign Sense 1 to the ambiguous word because it has the lowest sense number in Urdu Lughat.

Overall, we can divide the experiments into two broad categories: (1) word *n*-gram and (2) character *n*-gram approaches. In the word *n*-gram approach, the value of *n* varies from 1 to 5. However, in the character *n*-gram, value of *n* ranges from 2 to 10.

4.3 Evaluation Measures

The accuracy measure, borrowed from Machine Learning, is commonly used to evaluate the performance of WSD systems [45]. Accuracy figures in our experiments are calculated according to Equation 6 [62].

$$Accuracy = \frac{Words\ Disambiguated\ Correctly}{All\ Words\ Disambiguated} \times 100 \tag{6}$$

5 RESULTS AND ANALYSIS

Table 4 shows accuracy scores achieved when a range of WSD algorithms are applied to the UAW-WSD-18 corpus. In this table, “*n*-gram Model" refers to the token (i.e. word or character) selection methods from a given piece of text. “Parameters" refers to the description of “*n*-gram Model" on the basis of the value of *n*. Terms “JSC", “OSC", “DSC", “EDM", “CSC" and “VBA" indicate the various textual similarity based methods “JSC" refers to Jaccard Similarity Coefficient, “OSC" too Overlap Similarity Coefficient, “DSC" to Dice Similarity Coefficient, “EDM" to Euclidean Distance Measure, “CSC" to Cosine Similarity Coefficient and “VBA" to Voting Based Approach.

Overall, we carried out fourteen experiments, of which five use word based approaches and nine use character based approach.

The value of *n* is varied from 1 to 5 for word based approaches and from 2 to 10 for character based approaches. The results indicate that word based methods perform better than character based methods. Generally, JSC, OSC, DCS, and VBA produce better results with character 10-gram, word 4-gram, and word 5-gram for the Urdu WSD task.

The highest accuracy for word *n*-gram approaches is achieved with Word 4-grams (Accuracy = 57.71%) and lowest accuracy achieved with word 1-grams (Accuracy = 20.50%). For word based approaches, our results indicate that the value of *n* has a significant impact on the accuracy of the WSD system. As the value of *n* increases from 2 to 4, the accuracy almost doubles from 28.5% to 51.71%.

For character *n*-gram approaches, the highest results are achieved with character 10-grams (Accuracy = 56.89 %) and the lowest for character 2-grams (Accuracy = 34.81%). The accuracy of the WSD system is also highly dependent on *n* for character based approaches. Generally, we can observe that, a single increment in the value of *n* gives rise to almost a 2% increase accuracy of Urdu WSD system in many cases.

VBA was a specially designed approach to increase the accuracy of Urdu WSD system. The outcome of this approach is largely dependent on other approaches. This approach shows better performance for variety of experiments. For instance, the performance of VBA was higher than other methods for character 2- and 10-grams and also for word 3-, 4- and 5-grams.

Interestingly, no technique performs better than the baseline approach (Most Frequent Sense), demonstrating the difficulty of the WSD task in this corpus. A possible reason for this is the variation in number of senses for each word (2 to 11).

Fig. 3 shows the detailed results for the word 4-gram approach (on which we achieved highest accuracy). It shows that EDM is not a suitable measure for Urdu WSD. However, the other approaches

Table 4. Results obtained using various WSD approaches on the UAW-WSD-18 corpus

N-gram Model	Parameters	JSC[a]	OSC[b]	DSC[c]	EDM[d]	CSC[e]	VBA[f]
Word *n*-gram	Word 1-gram	28.50	29.90	29.79	40.07	29.20	30.17
	Word 2-gram	52.92	53.27	52.92	49.64	52.92	53.03
	Word 3-gram	57.59	57.59	57.59	49.88	57.59	57.59
	Word 4-gram	**57.71**	**57.71**	**57.71**	49.88	**57.71**	**57.71**
	Word 5-gram	57.59	57.59	57.59	49.88	57.59	57.59
Character *n*-gram	Character 2-gram	36.91	33.52	37.50	35.51	34.81	38.55
	Character 3-gram	38.20	38.78	38.66	44.27	38.55	40.53
	Character 4-gram	37.61	39.71	37.61	47.66	38.43	39.71
	Character 5-gram	36.09	37.73	36.56	48.36	35.98	36.33
	Character 6-gram	39.60	40.42	39.71	48.83	39.71	40.30
	Character 7-gram	48.59	48.83	49.41	48.83	49.29	48.59
	Character 8-gram	51.63	51.86	51.63	48.94	51.51	51.63
	Character 9-gram	55.72	55.84	55.72	49.06	55.72	55.72
	Character 10-gram	56.89	56.89	56.89	49.06	56.89	56.89
Most Frequent Sense	58.20%	-	-	-	-	-	-

[a] Jaccard Similarity Coefficient(%)
[b] Overlap Similarity Coefficient(%)
[c] Dice Similarity Coefficient(%)
[d] Euclidean Distance Measure(%)
[e] Cosine Similarity Coefficient(%)
[f] Voting Based Approach(%)

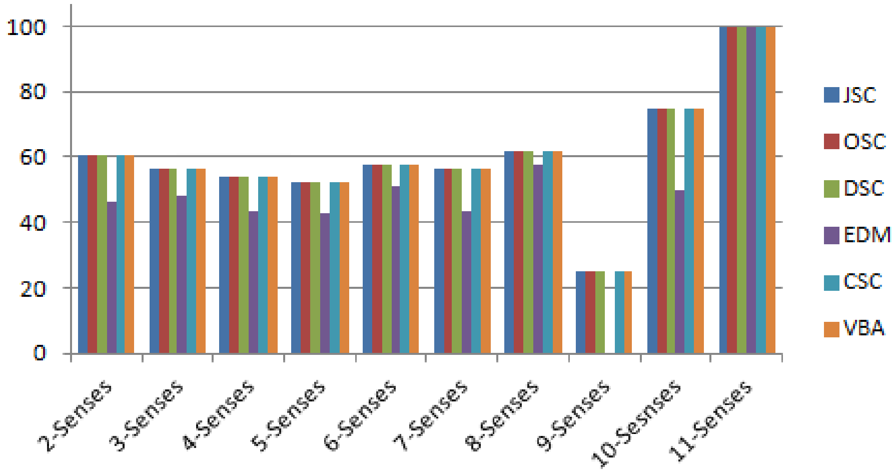


Fig. 3. Explaining sense wise detail of word 4-gram approach

perform equally well for different numbers of senses. For 2, 8, 10 and 11 senses, the system outperforms the baseline accuracy (MFS).

6 CONCLUSION

This paper described a new, freely available All-Words Word Sense Disambiguation corpus for Urdu, a widely spoken language which is critically under-resourced for natural language processing research. The paper's main contribution is the UAW-WSD-18 corpus which contains 5,042 words of running text and all ambiguous words manually annotated. An additional contribution is the application of six approaches which illustrate the use of our corpus for WSD experiments. Our results showed that Jaccard, Overlap, Dice, and Cosine similarity coefficients show the highest accuracy with Word 4 gram (57.71%). We have made the corpus and experimental data freely available in order to encourage research in this fledgling area of NLP for Urdu. In the future, we plan to apply other approaches to further address the All-Words Urdu WSD problem.

REFERENCES

[1] Muhammad Abid, Asad Habib, Jawad Ashraf, and Abdul Shahid. 2017. Urdu word sense disambiguation using machine learnaseer2009supervisedning approach. *Cluster Computing* (2017), 1–8.

[2] E Agirre, I Aldezabal, J Etxeberria, E Izagirre, K Mendizabal, E Pociello, and M Quintian. 2005. *EUSEMCOR: euskarako corpusa semantikoki etiketatzeko eskuliburua; editatze-, etiketatze-eta epaitze-lanak.* Technical Report. Internal report.

[3] E Agirre, O Lopez de Lacalle, C Fellbaum, A Marchetti, A Toral, PTJM Vossen, L Màrques, and R Wicentowski. 2009. All-words Word Sense Disambiguation on a Specific Domain (SemEval-2010 Task 17). Association for Computational Linguistics (ACL).

[4] James Allen. 1995. *Natural language understanding.* Pearson.

[5] Syed Zulqarnain Arif, Muhammad Mateen Yaqoob, Atif Rehman, and Fuzel Jamil. 2016. Word sense disambiguation for Urdu text by machine learning. *International Journal of Computer Science and Information Security* 14, 5 (2016), 738.

[6] Inger Askehave and John M Swales. 2001. Genre identification and communicative purpose: A problem and a possible solution. *Applied linguistics* 22, 2 (2001), 195–212.

[7] John Bateman and Michael Zock. 2003. Natural language generation. In *The Oxford Handbook of Computational Linguistics 2nd edition.*

[8] Luisa Bentivogli, Christian Girardi, and Emanuele Pianta. 2003. The MEANING Italian Corpus. In *Proceedings of the Corpus Linguistics 2003 conference.* Citeseer, 103–112.

[9] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The semantic web. *Scientific american* 284, 5 (2001), 34–43.

[10] Urdu Dictionary Board. 2008. Urdu Lughat. *Urdu Lughat Board, Karachi, Pakistan* (2008).

[11] Francis Bond, Timothy Baldwin, Richard Fothergill, and Kiyotaka Uchimoto. 2012. Japanese SemCor: A sense-tagged corpus of Japanese. In *Proceedings of the 6th Global WordNet Conference (GWC 2012).* Citeseer, 56–63.

[12] Abraham Bookstein and Don Kraft. 1977. Operations research applied to document indexing and retrieval decisions. *Journal of the ACM (JACM)* 24, 3 (1977), 418–427.

[13] Rebecca Bruce and Janyce Wiebe. 1994. Word-sense disambiguation using decomposable models. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics.* Association for Computational Linguistics, 139–146.

[14] Stefano Ceri, Adnan Abid, Mamoun Abu Helou, Davide Barbieri, Alessandro Bozzon, Daniele Braga, Marco Brambilla, Alessandro Campi, Francesco Corcoglioniti, Emanuele Della Valle, et al. 2010. Search Computing: Managing complex search queries. *IEEE Internet Computing* 14, 6 (2010), 14–22.

[15] Sung-Hyuk Cha. 2007. Comprehensive survey on distance/similarity measures between probability density functions. *City* 1, 2 (2007), 1.

[16] Surajit Chaudhuri, Venkatesh Ganti, and Raghav Kaushik. 2006. A primitive operator for similarity joins in data cleaning. In *Data Engineering, 2006. ICDE'06. proceedings of the 22nd International Conference on.* IEEE, 5–5.

[17] Ido Dagan, Lillian Lee, and Fernando Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 56–63.

[18] Nadir Durrani and Sarmad Hussain. 2010. Urdu word segmentation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics.* Association for Computational Linguistics, 528–536.

[19] Philip Edmonds and Scott Cotton. 2001. SENSEVAL-2: overview. In *The Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems.* Association for Computational Linguistics, 1–5.

[20] Philip Edmonds and Adam Kilgarriff. 2002. Introduction to the special issue on evaluating word sense disambiguation systems. *Natural Language Engineering* 8, 4 (2002), 279–291.

[21] Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion* (1999), 45–60.

[22] Mohamed Abdel Fattah and Fuji Ren. 2008. Automatic text summarization. *World Academy of Science, Engineering and Technology* 37 (2008), 2008.

[23] Wael H Gomaa and Aly A Fahmy. 2013. A survey of text similarity approaches. *International Journal of Computer Applications* 68, 13 (2013), 13–18.

[24] Udo Hahn and Inderjeet Mani. 2000. The challenges of automatic summarization. *Computer* 33, 11 (2000), 29–36.

[25] Nina Heck and Bettina Mohr. 2017. Response Hand Differentially Affects Action Word Processing. *Frontiers in psychology* 8 (2017), 2223.

[26] Sarmad Hussain. 2008. Resources for Urdu language processing. In *Proceedings of the 6th workshop on Asian Language Resources.*

[27] W John Hutchins. 1995. Machine translation: A brief history. In *Concise history of the language sciences.* Elsevier, 431–445.

[28] Rubén Izquierdo-Beviá, Lorenza Moreno-Monteagudo, Borja Navarro, and Armando Suárez. 2006. Spanish all-words semantic class disambiguation using cast3lb corpus. In *Mexican International Conference on Artificial Intelligence.* Springer, 879–888.

[29] Abdul Jabbar, Sajid Iqbal, and Muhammad Usman Ghani Khan. 2016. Analysis and development of resources for urdu text stemming. *LANGUAGE & TECHNOLOGY* (2016), 1.

[30] Bushra Jawaid, Amir Kamran, and Ondrej Bojar. 2014. A Tagged Corpus and a Tagger for Urdu.. In *LREC.* 2938–2943.

[31] Jing Jiang. 2012. Information extraction from text. In *Mining text data.* Springer, 11–41.

[32] Wahab Khan, Ali Daud, Jamal A Nasir, and Tehmina Amjad. 2016. A survey on the state-of-the-art machine learning models in the context of NLP. *Kuwait journal of Science* 43, 4 (2016).

[33] Adam Kilgarriff. 2004. How dominant is the commonest sense of a word?. In *International Conference on Text, Speech and Dialogue.* Springer, 103–111.

[34] Svetla Koeva, Sv Leseva, and Maria Todorova. 2006. Bulgarian sense tagged corpus. In *Proceedings of the 5th SALTMIL Workshop on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, Italy.* 79–87.

[35] Lawrence R Lawlor. 1980. Overlap, similarity, and competition coefficients. *Ecology* 61, 2 (1980), 245–251.

[36] Claudia Leacock, Geoffrey Towell, and Ellen Voorhees. 1993. Corpus-based statistical sense resolution. In *Proceedings of the workshop on Human Language Technology.* Association for Computational Linguistics, 260–265.

[37] Gurpreet Lehal. 2010. A word segmentation system for handling space omission problem in Urdu script. In *Proceedings of the 1st Workshop on South and Southeast Asian Natural Language Processing.* 43–50.

[38] John B MacArthur. 1988. An analysis of the content of corporate submissions on proposed accounting standards in the UK. *Accounting and Business Research* 18, 71 (1988), 213–226.

[39] Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 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.*

[40] Neetu Mishra and Tanveer J Siddiqui. 2012. An Investigation to Semi supervised approach for HINDI Word sense disambiguation. *Trends in Innovative Computing 2012-Intelligent Systems Design* (2012).

[41] Simonetta Montemagni, Francesco Barsotti, Marco Battista, Nicoletta Calzolari, Ornella Corazzari, Alessandro Lenci, Antonio Zampolli, Francesca Fanciulli, Maria Massetani, Remo Raffaelli, et al. 2003. Building the Italian syntactic-semantic treebank. In *Treebanks.* Springer, 189–210.

[42] Miguel Murguía and José Luis Villaseñor. 2003. Estimating the effect of the similarity coefficient and the cluster algorithm on biogeographic classifications. In *Annales Botanici Fennici.* JSTOR, 415–421.

[43] Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande, and Pushpak Bhattacharyya. 2002. An experience in building the indo wordnet-a wordnet for hindi. In *First International Conference on Global WordNet, Mysore, India.*

[44] Asma Naseer and Sarmad Hussain. 2009. Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification. *Center for Research in Urdu Language Processing, Lahore, Pakistan* (2009).

[45] Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM computing surveys (CSUR)* 41, 2 (2009), 10.

[46] Saeed A Nawab RMA, Stevenson M and Rayson P. [n. d.]. A Word Sense Disambiguation Corpus for Urdu (To appear). *Language Resources and Evaluation* ([n. d.]).

[47] Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. *SIGLEX99: Standardizing Lexical Resources* (1999).

[48] Hieu V Nguyen and Li Bai. 2010. Cosine similarity metric learning for face verification. In *Asian conference on computer vision.* Springer, 709–720.

[49] Suphakit Niwattanakul, Jatsada Singthongchai, Ekkachai Naenudorn, and Supachanun Wanapu. 2013. Using of Jaccard coefficient for keywords similarity. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, Vol. 1.

[50] Francois Paradis and Catherine Berrut. 1996. Experiments with theme extraction in explanatory texts. In *Proceedings of the Second International Conference on Conceptions of Library and Information Science, CoLIS, Copenhagen, Denmark.*

13–16.

[51] Rebecca J Passonneau, Collin Baker, Christiane Fellbaum, and Nancy Ide. 2012. The MASC word sense sentence corpus. In *Proceedings of LREC*.

[52] Michel Pêcheux. 1995. *Automatic discourse analysis*. Vol. 5. Rodopi.

[53] Sameer S Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task 17: English lexical sample, SRL and all words. In *Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics, 87–92.

[54] Tariq Rahman. 2004. Language policy and localization in Pakistan: proposal for a paradigmatic shift. In *SCALLA Conference on computational linguistics*, Vol. 99. 100.

[55] Kashif Riaz. 2010. Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop*. Association for Computational Linguistics, 126–135.

[56] Kashif Riaz. 2010. Rule-based named entity recognition in Urdu. In *Proceedings of the 2010 named entities workshop*. Association for Computational Linguistics, 126–135.

[57] Adriana Roventini, Alone Antonietta, Francesca Bertagna, Nicoletta Calzolari, Cacila Jessica, Girardi Christian, Magnini Bernardo, R Marinelli, Speranza Manuela, and A Zampolli. 2003. ItalWordNet: building a large semantic database for the automatic treatment of Italian. *Linguistica Computazionale* 18 (2003), 745–791.

[58] Hassan Sajid. 2007. Urdu Part of Speech Tagset. *Center for Research in Urdu Language Processing, National University of Computer & Emerging Sciences, Lahore Pakistan, available at www. crulp. org* (2007).

[59] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. 2008. *Introduction to information retrieval*. Vol. 39. Cambridge University Press.

[60] UmrinderPal Singh, Vishal Goyal, and Gurpreet Singh Lehal. 2012. Named entity recognition system for Urdu. *Proceedings of COLING 2012* (2012), 2507–2518.

[61] Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*.

[62] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427–437.

[63] Radu Soricut and Eric Brill. 2004. Automatic question answering: Beyond the factoid. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*.

[64] Steve Stemler. 2001. An overview of content analysis. *Practical assessment, research & evaluation* 7, 17 (2001), 137–146.

[65] Xue-Ren Sun, Shao-He Lv, Xiao-Dong Wang, and Dong Wang. 2017. Chinese Word Sense Disambiguation using a LSTM. In *ITM Web of Conferences*, Vol. 12. EDP Sciences, 01027.

[66] Vikas Thada and Vivek Jaglan. 2013. Comparison of jaccard, dice, cosine similarity coefficient to find best fitness value for web retrieved documents using genetic algorithm. *International Journal of Innovations in Engineering and Technology* 2, 4 (2013), 202–205.

[67] Saba Urooj, Sana Shams, Sarmad Hussain, and Farah Adeeba. 2014. Sense Tagged CLE Urdu Digest Corpus. *Centre for Language Engineering, Al-Khawarizmi Institute of Compute Science, University of Engineering and Technology, Lahore* (2014).

[68] Arthur A Van Hoff. 1998. System for adding requested document cross references to a document by annotation proxy configured to merge and a directory generator and annotation server. US Patent 5,822,539.

[69] Piek Vossen, Rubén Izquierdo, and Attila Görög. 2013. Dutchsemcor: in quest of the ideal sense-tagged corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*. 710–718.

[70] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012* (2016).

[71] Ayesha Zafar, Afia Mahmood, Farhat Abdullah, Saira Zahid, Sarmad Hussain, and Asad Mustafa. 2012. Developing urdu wordnet using the merge approach. In *Proceedings of the Conference on Language and Technology*. 55–59.

[72] Xiang Zhang and Yann LeCun. 2017. Which Encoding is the Best for Text Classification in Chinese, English, Japanese and Korean? *arXiv preprint arXiv:1708.02657* (2017).