Semantic Tagging for the Urdu Language: Annotated Corpus and Multi-Target Classification Methods

JAWAD SHAFI, Department of Computer Science, COMSATS University Islamabad, Lahore Campus, and InfoLab21, Lancaster University, Lancaster, U.K.
RAO MUHAMMAD ADEEL NAWAB, Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Pakistan.
PAUL RAYSON, School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, U.K.

Extracting and analysing meaning-related information from natural language data has attracted the attention of researchers in various fields, such as natural language processing, corpus linguistics, information retrieval, and data science. An important aspect of such automatic information extraction and analysis is the annotation of language data using semantic tagging tools. Different semantic tagging tools have been designed to carry out various levels of semantic analysis, for instance, named entity recognition and disambiguation, sentiment analysis, word sense disambiguation, content analysis, and semantic role labelling. Common to all of these tasks, in the supervised setting, is the requirement for a manually semantically annotated corpus, which acts as a knowledge base from which to train and test potential word and phrase-level sense annotations. Many benchmark corpora have been developed for various semantic tagging tasks, but most are for English and other European languages. There is a dearth of semantically annotated corpora for the Urdu language, which is widely spoken and used around the world. To fill this gap, this study presents a large benchmark corpus and methods for the semantic tagging task for the Urdu language. The proposed corpus contains 8,000 tokens in the following domains or genres: news, social media, Wikipedia, and historical text (each domain having 2K tokens). The corpus has been manually annotated with 21 major semantic fields and 232 sub-fields with the USAS (UCREL Semantic Analysis System) semantic taxonomy which provides a comprehensive set of semantic fields for coarse-grained annotation. Each word in our proposed corpus has been annotated with at least one and up to nine semantic field tags to provide a detailed semantic analysis of the language data, which allowed us to treat the problem of semantic tagging as a supervised multi-target classification task. To demonstrate how our proposed corpus can be used for the development and evaluation of Urdu semantic tagging methods, we extracted local, topical and semantic features from the proposed corpus and applied seven different supervised multi-target classifiers to them. Results show an accuracy of 94% on our proposed corpus which is free and publicly available to download.

Additional Key Words and Phrases: Urdu corpus annotation, Multi-target classifiers, Semantic annotation, Semantic tagger

1 INTRODUCTION

Semantic analysis of natural language data is worthwhile for a number of research areas and practical applications, for instance, Natural Language Processing (NLP), text mining, and Human Language Technology (HLT) systems. In recent research, different types of semantic tagging tools have been suggested and developed to carry out various levels of semantic analysis. For instance, some semantic tagging tools are designed to identify topics of a given text [10]. Others are used to extract specific or partial information, for example, types of named entities or events [85, 112]. Another set of semantic tagging tools are designed to identify semantic field categories for all

Authors' addresses: Jawad Shafi, Department of Computer Science, COMSATS University Islamabad, Lahore Campus, and InfoLab21, Lancaster University, Lancaster, Lancaster, U.K., jawadshafi@cuilahore.edu.pk; Rao Muhammad Adeel Nawab, Department of Computer Science, COMSATS University Islamabad, Lahore Campus, Lahore, Pakistan., adeelnawab@ciitlahore.edu.pk; Paul Rayson, School of Computing and Communications, InfoLab21, Lancaster University, Lancaster, Lancaster, U.K., p.rayson@lancaster.ac.uk.

2375-4699/2023/2-ART
https://doi.org/10.1145/3582496

lexical units (words) using a predefined semantic taxonomy. In order to support semantic information extraction and analysis from language data, the latter types of tools require richer semantic lexical resources and provide a coarser level of sense disambiguation, and thus, are challenging to create. In this research work, our focus will be on developing a benchmark corpus and methods for a semantically rich text analytical tool.

Several semantically rich lexical resources and tagging tools are available for monolingual analysis, particularly for English e.g. WordNet [55, 61], but very few resources or tools exist that can be used to carry out semantic analysis for multilingual text, such as, EuroWordNet [109], BabelNet [61], and USAS[1] [76], which have many applications in the development of intelligent NLP and HLT systems [94]. For example, the original English USAS semantic tagging tool (or semantic tagger) has been applied in numerous research studies such as entrepreneurship [31], software engineering [102], empirical language analysis [73], requirements engineering [78], historical semantic analysis via HTST 1 [69], to train a chatbot [96], and several others [13, 93]. Moreover, USAS [76] has been ported previously to cover multiple other languages[2] (Arabic, Finnish, Russian, Chinese, Welsh, Italian, Portuguese, Czech, Dutch and Spanish) with a unified semantic annotation scheme. Following this established framework i.e. *USAS* [76] therefore, in this research work our focus will be the development of a coarse-grained *all-words* semantic analysis tool rather than annotating fine-grained word senses as in WordNet.

Originally developed for English semantic tagging, USAS [76] is a commonly used semantic field-oriented analysis system. Compared to word sense disambiguation systems, it does not disambiguate between fine-grained word sense definitions, but rather, it assigns a semantic category (or categories) to each word or phrase by employing a unified semantic annotation taxonomy. USAS is also different from those systems which extract other types of information (named entity recognition, semantic role labelling, etc), in that it assigns semantic field tag(s) to every lexical unit in a running text. The required resources and methods in the development and evaluation of the USAS [76] system are: (i) a set of semantic field tags, for major semantic field tags, (ii) single and multi-word semantic lexicons, and (iii) semantic field disambiguation methods.

The USAS was developed based on semantic lexicons acting as a knowledge base from which to select word and phrase level sense annotation(s) using a variety of disambiguation methods to select the most likely semantic tag in context. However, supervised multi-target[3] classifiers have never been applied for semantic analysis previously. Each single-word or Multi-Word Expression (MWE) in USAS output may appear with multiple possible semantic field tags to show the different meanings which can be taken in different contexts, and these are left in the output in rough likelihood order if disambiguation methods cannot resolve the correct sense (more details can be found in [69, 76]). For such systems, multi-target classifiers can be potentially beneficial, where the word(s) may be associated with multiple labels/tags, and as a pre-processing step to full disambiguation.

A multi-target semantic tagging tool is different from *one sense per discourse* [35] where a polysemous word appears two or more times in a well-written discourse, it is extremely likely that they will all share the same sense. On the other hand, in the case of *one sense per collocation* [115], a polysemous word exhibits essentially only one sense per collocation (here collocation means the co-occurrence of two words in some defined relationship).

To develop and evaluate semantic tagging methods (and thereby semantic taggers) based on supervised multi-target classifiers, we need benchmark corpora. In the previous literature, some research has been carried out to develop benchmark corpora for semantic tagging of natural language data. However, the majority of these are for English and European languages [40, 44] and there is a lack of such benchmark evaluation resources for South Asian languages, in particular Urdu. The Urdu language is widely spoken and is ranked nineteenth among the native speaking languages of the world [95]. In addition, a rapidly increasing amount of Urdu digital text is readily available on-line.

[1]USAS: the UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System, HTST: the Historical Thesaurus Semantic Tagger

[2]https://ucrel.lancaster.ac.uk/usas/ - Last visited: 11-January-2019

[3]In the semantic classification, each target variable can take multiple class values (i.e. target variables are not binary).

This study presents our research on developing a benchmark Urdu Semantically Annotated Corpus (hereafter called the USA-23 Corpus). Our corpus contains 8,000 manually annotated tokens (2,000 each for news, social media, Wikipedia, and historical text). Each word in the USA-23 Corpus is assigned from one to nine semantic tags. To demonstrate how the USA-23 Corpus can be used for the development and evaluation of supervised multi-target classification methods, we extracted local (raw words, Part-Of-Speech (POS) tags and lemmas), topical (bag-of-words context, collocations) and semantic features (domain indicators) from the proposed corpus and applied seven multi-target classifiers including Bayesian classifier chain, classifier chain, classifier chain probabilities, class relevance, nearest set replacement, RAkELd, and super class classifier.

We believe that the USA-23 Corpus and semantic tagging methods presented in this study will have potential benefits including (i) fostering research in a low-resourced language i.e. Urdu, (ii) developing and evaluating new semantic tagging tools/methods for the Urdu language, (iii) to test the lexical coverage and accuracy of Urdu semantic lexicons, (iv) our proposed semantic tagging methods can be used for performing various Urdu NLP tasks including named entity extraction, information extraction, document classification, (v) as several equivalent semantic taggers are developed based on the same USAS tagset which acts as a kind of a "meta-dictionary" between the languages, therefore, it opens the door for the development of multi-lingual and cross-lingual applications, machine translation, plagiarism detection, and information extraction as well as retrieval tasks.

The remainder of the paper is organized as follows: in Section 2 we describe the related work. Section 3 presents the corpus generation process. Section 4 explains the experimental set-up, dataset, and semantic annotation methods which we have applied to our proposed corpus, evaluation measures, and evaluation methodology. Section 5 discusses results and their analysis. Finally, Section 6 concludes the paper with future work directions.

2 LITERATURE REVIEW

The research field which is most closely related to semantic tagging is Word Sense Disambiguation (WSD) (see Section 1) [76]. Therefore, in this section, we will present the corpora and methods developed for WSD and semantic tagging tasks.

2.1 Corpora and Techniques for Word Sense Disambiguation

Corpora. Developing large-scale freely available standard evaluation resources to investigate the problem of WSD is a non-trivial task. In previous literature, efforts have been made to develop benchmark corpora for the WSD task. An in-depth discussion of all the WSD corpora will be beyond the scope of this study. Therefore, we only present some of the most prominent studies.

The most prominent effort in developing standard evaluation resources for WSD task is a series of SensEval competitions[4]. The outcome of these competitions is a set of benchmark corpora for the WSD task. The SensEval competitions on the WSD task have been organized from 1998 to 2004. The competitions focused on two main types of WSD: (i) all-words WSD task and (ii) lexical sample WSD task. The languages for which WSD corpora were developed include English, Basque, Italian, Japanese, Korean, Spanish, Swedish, Spanish, Chinese, and Romanian. The lexical resources or dictionaries that were used in the development of WSD corpora include WordNet. SensEval WSD corpora are large and freely available for research purposes [60].

In previous literature, other than SensEval, efforts have been made to develop WSD corpora for English and other languages such as the SEMCOR WSD Corpus [46], Google WSD Corpus [116], and DutchSemCor WSD corpus [110]. However, for the Urdu WSD task, only three corpora have been found in previous research, (i) an Urdu sense tagged corpus [107], (ii) Urdu Lexical Sample WSD (ULS-WSD-18) corpus [87] and (iii) all-Words WSD corpus for Urdu (UAW-WSD-18) [88].

[4]http://www.senseval.org/ - Last visited: 11-January-2019

The Urdu sense tagged corpus [107] was developed for the Urdu all-words WSD task and contains 17K manually sense annotated sentences with 2,285 unique senses by a single annotator over a period of 10 months. Whereas, the ULS-WSD-18 corpus [87] has been developed for the lexical sample WSD task and contains 7,185 manually sense tagged sentences for 50 target words (senses of tagged words were extracted from a hand crafted dictionary called Urdu Lughat Board [19]) by three different annotators. Finally, a recently released all-Words WSD corpus for Urdu [88] is also worth mentioning here, containing 5,042 Urdu words. In this corpus, all 466 ambiguous types and 856 ambiguous words have been manually annotated with senses from the Urdu Lughat dictionary by three annotators.

It can be observed from the above discussion that several WSD based corpora are available for English and Urdu languages. Such fine-grained WSD corpora are not always necessary for many NLP applications[5]. However, there is a dearth of semantically annotated standard evaluation resources for Urdu and several world languages. Therefore, this study addresses this gap by constructing a large and freely available semantically annotated multi-target corpus for the Urdu language with USAS semantic field tags.

Techniques. WSD research is closely related to our work and different WSD techniques have been used to resolve semantic tag ambiguity such as mentioned in [76]. Therefore, in this section, we provide an overview of WSD techniques.

Over the years, many different WSD techniques have been proposed, and they can be classified into the following four categories: (i) Artificial Intelligence (AI), (ii) knowledge-based, (iii), corpus-based, and (iv) hybrid techniques [60, 76].

Prominent efforts to tackle WSD based on AI techniques began in the early 1970s via large-scale language understanding [2, 43]. For example, Wilks [113] described a "preference semantics" system, using selectional restrictions and lexical semantics (*case frames*[6]) to find a set of senses for a word in a sentence.

Knowledge-based WSD techniques use lexical resources to provide contextual knowledge which is essential to determining the appropriate sense(s) of polysemous words [76]. These resources can be thesauri [86], machine-readable dictionaries [74], or computational lexicons [55, 76]. A wider survey of these resources can be found in [5].

Current state-of-the-art techniques for the resolution of word sense ambiguity stem from the field of Machine Learning (ML). These ML (or corpus-based) WSD techniques can be primarily classified into, (i) unsupervised, (ii) semi-supervised, and (iii) supervised.

Unsupervised techniques have the potential to acquire contextual information directly via knowledge acquisition [34] i.e. senses can be deduced from untagged raw text using similarity measures [53] based on the idea that occurrences of the same sense of a word will have similar neighbouring words. Example techniques for unsupervised WSD are co-occurrence and spanning tree-based graphs [4], word clustering [20], and recently developed neural network language models [67].

Semi-supervised ML WSD techniques usually train a classifier with a small set of labelled examples and then bring further improvements in the process of iterative learning i.e. a classifier is retrained, and this learning process continues until convergence. There have been a number of studies that have used semi-supervised ML WSD techniques, for instance, [63] used label propagation algorithm for WSD, whereas, [116] used sequence learning neural network to differentiate different senses.

Supervised single-label classification techniques apply where each word is only associated with a single label or class, that is, they assign the appropriate sense to a target word. There have been a number of research studies where single-label ML classification techniques are applied for English and European language WSD tasks, for example, [3] used decision lists [84], whereas, [56] used C4.5 (decision tree) and concluded that it outperformed

[5]Several NLP problems can be solved without access to a full set of dictionary/WordNet definitions.
[6]These contain information about words, their relation to other words, and their roles in individual sentences

all the other single-label ML techniques, simple Naïve Bayes is applied in [22], [104] (based on neural networks), k-nearest neighbour [28]. A complete overview and discussion of all single-label classification techniques are beyond the scope of this section. Therefore, we will present the single-label classification studies adopted for Urdu. We have found two such studies in the previous literature: (i) machine learning based WSD [1], and (ii) Bayesian classifier based WSD [59].

The authors in [1] developed a lexical-sample based WSD system using single-label classifiers including, Naïve Bayes, Decision Tree, and Support Vector Machines with POS tags and bag-of-words as features. Twenty named entities were used to evaluate the system performance. The reported F_1 scores for Naïve Bayes, decision tree, and support vector machines are: 71%, 34%, and 34% respectively. Another study was conducted using Naïve Bayes classifiers for the development of lexical-sample WSD system [59]. The authors resolved the ambiguity in four words including three verbs and one noun. Bag-of-words and POS tags were used as features and the reported highest F_1 score was 95.15%.

The final of our four categories of techniques is the hybrid technique, representing studies using a combination of the various above-mentioned techniques. A number of research studies have been carried out using hybrid ensemble techniques, for instance, [99] used LDOCE with information derived from corpora.

It can be observed from the above discussion that a number of WSD techniques have been used for sense resolution. However, these techniques have several shortcomings as follows: (i) AI techniques are not to be practical for large-scale language understanding [76], (ii) knowledge-based methods are a useful way to represent linguistic or lexicographic knowledge of word sense ambiguity, and they have produced good results. However, they are not very robust as natural language is a dynamic phenomenon i.e. new words and senses are added and old ones become archaic or outdated, thus, they lack complete coverage as new words or senses may not exist in these resources, (iii) moreover, for knowledge-based systems lexical resources are readily available for English and other European languages, but not for the under-resourced Urdu[7] language, (iv) semi-supervised ML techniques have a major drawback in that they lack a method for selecting optimal values for classifiers i.e. the number of iterations and labelled examples [62]. Further, these types of techniques are tested on small corpora [60], (v) unsupervised ML techniques automatically acquire contextual information and are often erroneous and noisy and alone [1], thus are unlikely to solve large-scale problems, (vi) ML based hybrid techniques require several resources (lexicons and corpora), which is difficult for resource-poor languages, and (vii) supervised single-label classifiers can assign only one tag/label. These shortcomings imply that these techniques are not a promising basis for Urdu semantic tagging.

In the multi-target classification task, word(s) may be associated with multiple semantic labels/tags [111]. Multi-target classifiers have been applied in a number of research studies; text classification [27], bio-informatics [25], scene classification [21], shape detection in ultrasound images [117]. However, to the best of our knowledge, multi-target classifiers had never been explored for the WSD task in general and particularly in the context of the Urdu language. Therefore, this research study addresses another gap in the existing research by extracting features and then applying various off-the-shelf multi-target classifiers to deal with the WSD problem by employing a broad semantic taxonomy rather than fine-grained word sense definitions. This provides a practical means of coping with the semantic disambiguation task and can be seen as an important step for a more robust wide coverage candidate semantic tag assignment before final disambiguation.

2.2 Corpora and Techniques for Semantic Tagging

Corpora. A number of studies in the literature have devoted a great deal of research effort to the development of semantic annotation, such as Semantic Role Labelling, Named Entity Recognition, Content Analysis, and several others. Usually, these semantic annotation systems have used annotated corpora or WordNet to induce or cluster

[7]A recent study [92] involved Urdu semantic lexicons (both single and multi-words) of 2K entries, however, it is lacking wide lexical coverage.

different meanings or senses [60]. The most prominent effort in developing standard evaluation resources for various semantic annotation tasks are the series of SemEval competitions for English and other languages [60].

The outcome of these competitions (from 2007 to date) are a set of benchmark corpora with semantic annotations for various NLP tasks, Information Extraction, Sentiment Analysis, Textual Semantic Similarity, Word Semantic Similarity, Question Answering, Labeling of MWEs and Supersenses (SemEval-2012[8], SemEval-2013[9], SemEval-2014[10], SemEval-2015[11], SemEval-2016[12], SemEval-2017[13], and SemEval-2018[14]) for a variety of languages including English, French, Italian, Dutch, Chinese, Arabic and several others.

In the case of Urdu, only four semantically annotated corpora for semantic role labelling have been found in the past research as: an Urdu dependency Treebank (UDT) [17], (ii) a Hindi/Urdu Treebank (HUTb) [39], (iii) a Proposition Bank for Urdu (PBU) [11] and (iv) a multilayered Urdu Treebank [6].

The UDT corpus [17] has been built following the computational paninian grammar [16]. This Treebank contains morphological, POS, chunking information, and dependency relations for newspaper articles manually annotated by a team of linguistic experts. Around 200K words, (7,000 sentences) have been annotated with the previously mentioned annotations, where each sentence contains an average of 29 words and an average of 13.7 chunks of average length 2. Moreover, the tagset involved in this research contains 43 tags. However, this Treebank does not tag words that a verb can take depending upon the subject and object pertaining to the verb.

The HUTb [39] has been annotated for the deep analysis of the language by integrating the functional structure of lexical functional grammar. This treebank encodes traditional syntactic notions i.e. subject, direct as well as an indirect object, complement, adjunct, functional, and morphological information. This treebank does not provide a complete dependency bank, but rather provides an argument that developing such a resource will be beneficial for several NLP applications. Therefore, the authors of this treebank have annotated several randomly selected sentences for the annotation, the detailed statistics are not given.

The Proposition Bank for Urdu (PBU) corpus [11] contains the text of the Urdu dependency treebank and adds a further semantic layer (add argument structures of both simple and complex predicates) into the UDT corpus. This Treebank corpus has been annotated manually by two annotators, for simple predicate 180K words have been tagged whereas, for complex verb predicate 100K token are annotated. The tagset used to tag simple and complex predicate argument structures of verbs in this research work contains a total of 28 tags.

The recently constructed multilayered Urdu treebank corpus [6] contains 1,300 sentences of the centre for language engineering Urdu digest corpus [106]. A small set of tags (12) have been used to annotate the phrase, grammatical functions, semantic roles, demonstrative phrase, interjection grammatical functions, and discontinuous phrases. The purpose of this treebank corpus is to provide parse trees for the text to speech applications.

It can be observed from the above discussion that semantically annotated corpora are widely available in English and other European languages. However, there is a dearth of semantically annotated standard evaluation resources for Urdu. Therefore, this study addresses this gap by constructing a large and freely available semantically annotated multi-target corpus for the Urdu language, which is annotated with USAS semantic field tags. As far as we are aware, a USAS based semantically annotated multi-target corpus for the Urdu language has not previously been developed.

Techniques. Tagged corpora are used to induce or cluster different senses or meanings, aiming to identify and assign certain types of semantic information required by specific tasks. These types of semantic annotations

[8]https://www.cs.york.ac.uk/semeval-2012/index.html - Last visited: 11-January-2019
[9]https://www.cs.york.ac.uk/semeval-2013/ - Last visited: 11-January-2019
[10]http://alt.qcri.org/semeval2014/ - Last visited: 11-January-2019
[11]http://alt.qcri.org/semeval2015/ - Last visited: 11-January-2019
[12]http://alt.qcri.org/semeval2016/ - Last visited: 11-January-2019
[13]http://alt.qcri.org/semeval2017/ - Last visited: 11-January-2019
[14]http://alt.qcri.org/semeval2018/ - Last visited: 11-January-2019

have been researched in [26] and [10] to identify the topic or themes of a given text. There are yet further studies [66, 85, 112] which are conducted to extract specific or partial information, such as named entities, categories of relations between the specific named entities, and/or types of events.

Via another group of knowledge-based sense inventories (WordNet, BabelNet), a semantic annotation can be used to assign fine-grained word senses [55]. However, WordNet does not readily generalise to OOV words [89]. Moreover, WordNets have been developed for English, European, and several Asian languages. These resources have also been ported to provide multilingual word sense inventories [60].

Other semantic annotation research aims to assign each content word with a semantic category using a component-based semantic classification scheme, for instance, tagging the word "mother" as [HUMAN, FEMALE, ADULT] and "paprika" as [NON-HUMAN, VEGETABLE]. A number of research studies based on this concept have been reported previously, including [50]. Other knowledge and information management systems provide general purpose semantic annotations based on ontologies [72].

In addition, a similar semantic tagging approach to the one proposed here is STREUSLE[15], which integrates comprehensive annotations of MWEs and semantic supersenses[16] for lexical expressions with a unified tagset [91]. The supersenses tags or labels apply to both single and mulitwords but only for noun plus verb categories, and to prepositional/possessive expressions. This method of annotation has been used to classify lexical MWEs of English web reviews [90].

Directly related to our research presented here is the development of coarse-grained semantic tagging tools, such as USAS [76] and several others cited in [29] and [15]. USAS is different from other WSD systems as it assigns tags from a pre-defined coarse-grained semantic field taxonomy rather than fine-grained word meaning. Furthermore, USAS is also different from LaSIE (a named entity identification system) [42], in that it does not just focus on a small number of specific classes of words, rather, it assigns a tag(s) to every word in a running text.

Recently, the systems based on USAS semantic fields have been ported to support fine-grained semantic annotation [69] for historic English text. Moreover, the coarse-grained semantic analysis system has been ported to Finnish [49], Russian [57], and to several other European and world languages using a common semantic taxonomy [32, 68, 92].

From the above discussion, it can be observed that a number of semantic annotation or tagging tools and resources have been previously developed. However, WSD and WordNet based semantic tagging provide fine-grained word senses, thus, which are not always required in many NLP tasks [76]. Moreover, the STREUSLE based semantic tagging provides supersense for only verbs and nouns. Furthermore, these systems and resources have mainly been created for European languages. However, the USAS semantic annotation framework is a knowledge-based system, which therefore poses more difficulties when creating similar resources for poorly-resourced languages. Still, it is a worthwhile task, since if we can design similar semantic tagging tools for multiple languages, they can potentially provide a bridge for multilingual Machine Translation and WSD systems.

The semantic annotation report in this paper falls under the category of a coarse-grained but *all-words* based semantic tagger for Urdu text but employs supervised multi-target classifiers. In USAS, one word may have one or more semantic tag(s). To handle such an annotation process there exists a supervised multi-target classifier, where, words may be associated with multiple labels. To the best of our knowledge, this combination has not previously been incorporated into any language. Therefore, this research work extends the capability of the existing USAS (knowledge-based system) in terms of porting USAS for Urdu using supervised multi-target classifiers.

[15]Supersense Tagged Repository of English with a Unified Semantics for Lexical Expressions
[16]Synsets or word senses of WordNet

3 CORPUS CONSTRUCTION

In USAS (see Section 1) not all words fall into one predefined semantic category, rather, some words can belong to two or more semantic categories. For instance, the word "officer" can be tagged with G3/S7.1/S2, since it can be considered to belong to the semantic category "Warfare, defence and the army; Weapons" (G3), as well as to the category "Power, organizing" (S7.1), and to the category "People" (S2). These multiple memberships of categories have been indicated with "slash tag (/)" separating tags in USAS. Furthermore, USAS is a concept-driven tagging tool rather than content driven, in that it provides a general conceptual structure of the world, instead of trying to offer a semantic taxonomy for specific domains [68]. Therefore, our proposed multi-target Urdu Semantically Annotated (USA-23) Corpus has been annotated with multiple potential USAS semantic tags (up to nine[17], if required). This section describes the USAS semantic tagset, its importance, and the creation of our proposed gold standard USA-23 Corpus, including raw data collection, development of an annotation tool, the annotation process, corpus statistics, and standardization of the corpus.

3.1 USAS Semantic Fields/Tags

The USAS semantic tagset was loosely based on Tom McArthur's Longman Lexicon of Contemporary English (LLOCE) [52]. This tagset adopts a general ontological approach and has proved to be a most appropriate thesaurus type classification of word senses or for a semantic field kind of analysis [70, 71, 76]. Furthermore, this classification scheme has been considerably revised in the light of practical tagging problems met in the course of ongoing research [68–71, 76, 92, 94]. The revised tagset has been arranged in a hierarchy with a 21 major semantic fields (see Table 1), which, are expanded into 232 sub-fields[18]. In the USAS tagset capital letters have been used to denote major semantic field tags, while numbers are used to indicate subdivisions of the fields/tags. The grouping of these tags are related by the virtue of their being connected at some level of generality with the same logical concept[19] [12]. The USAS semantic field tags group words in more general or coarse-grained senses rather than fine-grained senses. For instance, the word 'bank' in some dictionaries may differentiate the conceptual categories of the physical branches of the bank and the type of financial institute. However, the USAS tags consider both of these senses as related to one semantic field tag i.e. 'I: Money and Commerce'.

The reason for selecting the USAS tagset for the annotation of the USA-23 corpus is many fold: (i) it has been revised in the light of problems met in the course of applied research, (ii) successfully applied to the following research studies[20]: market research analysis [114], software engineering [77], deep semantic analysis [69], historical semantic tagging [9], analysis of Weblogs [65], analysis and standardisation of SMS spelling variation and detecting gender differences in Twitter [14, 101], discourse analysis [8, 64], ontology learning [33], phraseology [37], political science research [47], social networks child Protection [75], psychological profiling [51], sentiment analysis [97], to train chatbots [96], deception detection [51], are a representative selection, (iii) equivalent semantic taggers have been designed based on these semantic fields[21], which enables the development of multi-lingual NLP, HLT, text mining, translation, and other types of information and communication technology systems, (iv) many of the semantic taggers developed so far using this tagset are for resource-poor languages from Asia [94].

[17] In a separate study we tagged words and have found that a word can be tagged with up to nine tags.
[18] For the full tagset visit https://ucrel.lancaster.ac.uk/usas/USASSemanticTagset.pdf- Last visited: 26-March-2020
[19] Through a process of synonymy, antonymy, hypernymy and/or hyponymy.
[20] A complete list of publications and applications using Wmatrix (in which USAS is embedded) can be found at https://ucrel.lancs.ac.uk/usas/ andhttps://ucrel.lancs.ac.uk/wmatrix/ - Last visited: 19-March-2020
[21] Currently, there are sixteen non-English semantic taggers or lexicons available for Arabic, Chinese, Czech, Dutch, French, Italian, Malaysian, Portuguese, Spanish, Urdu, Indonesian, Turkish, Swedish, Finnish, Russian, and Welsh languages. More details can be accessed through the following URL: https://ucrel.lancs.ac.uk/usas/ - Last visited: 17-March-2020

Table 1. USAS major semantic fields

Domain: Description
A: General and abstract terms
B: The body and the individual
C: Arts and crafts
E: Emotional actions, states and process
F: Food and farming
G: Government and the public domain
H: Architecture, buildings, housing and the home
I: Money and commerce
K: Entertainment, sports and games
L: Life and living things
M: Movement, location, travel and transport
N: Numbers and measurement
O: Substances, materials, objects and equipment
P: Education
Q: Linguistic actions, states and process
S: Social actions, states and processes
T: Time
W: The World and our environment
X: Psychological actions, states and processes
Y: Science and technology
Z: Names and grammatical words

3.2 Data collection

To train and test supervised multi-target machine learning algorithms, an Urdu annotated corpus is required based on the USAS semantic taxonomy. Therefore, to develop a corpus with realistic examples, we have collected data from different domains. For example, social media texts are short and informal, whereas, newspaper articles are formally written and of moderate length. To develop the USA-23 Corpus, raw data is collected from the following domains: (i) news articles, (ii) social media (Twitter[22], Facebook[23], and Blogs), (iii) literary magazines, and (iv) Wikipedia[24] articles.

The reasons for collecting data from these domains are, firstly, they contain data that are significantly different from one another. Secondly, variation in data poses different types of challenges for the semantic annotation task, which makes our proposed corpus more realistic and challenging. Thirdly, data from these sources are free and readily available in digital format for research purposes. Fourthly, they enable the evaluation of semantic annotation tools (or methods) on a variety of writing styles and publication times. Fifthly, to make sure that our vocabulary inventory is of sufficient coverage. Finally, to produce a more robust semantic field annotated corpus.

[22]https://twitter.com/ - Last visited: 11-January-2019
[23]https://facebook.com/ - Last visited: 11-January-2019
[24]https://ur.wikipedia.org/wiki/ - Last visited: 11-January-2019

The raw text of news articles is collected from various sources including BBC Urdu[25], Express news[26], Urdu Library[27], and Minhaj Library[28] using a Web crawler[29]. The newspaper text is useful as it is written in continuous prose and purports to be a mainly factual report of events that have taken place. The news articles collected were from different genres including Sports, Politics, Showbiz, Science and Technology, Business, Health and Religion. There are in total 2,100 word tokens in the collected text (for each genre there are 250-300 tokens). We call this sub-corpus the USA-23-raw-news corpus.

To form a sub-corpus from social media, raw data is collected from the following four sources: Twitter[30], Facebook[31], Blogs, and Reviews. These sources serve monthly around 2,375 million active users[32]. We manually collected publicly available data (user generated content) on different topics to make sure that the collected data is genuine, realistic, diverse and of high quality. From each source, we collected Urdu texts of 600 tokens (a total of 2,400 tokens). We call this sub-corpus the USA-23-raw-smedia corpus. It has been shown [30] that social media text poses additional challenges to automatic NLP methods, as text from these sources tends to be less grammatical. Thus, forming a corpus from social media sources provides the challenging text for the Urdu semantic annotation task.

To form a third sub-corpus, Urdu text is collected from the following Wikipedia[33] articles: Culture, History, Geography and Areas, Personalities, Science and Technology. A passage of size 300-350 words is excerpted from each of these Wikipedia articles (giving a total of around 2,300 words). The sub-corpus is called USA-23-raw-wiki corpus. The reason for using Wikipedia as a text collection source is that it is large, reliable, freely available, contains texts on a variety of topics and articles written by different authors exhibiting language variation.

The last and fourth type of collected Urdu text consists of words from old Urdu literature (fiction and non-fiction short stories). The raw text of Urdu literature of the early 1940s is collected from HamariWeb[34]. We collected Urdu text of approximately 2,200 words. This sub-corpus is called the USA-23-raw-historic corpus and contains Urdu text with a variety of writing styles and time periods.

3.3 Pre-processing

In this study, four different raw sub-corpora (USA-23-raw-news, USA-23-raw-smedia, USA-23-raw-wiki, and USA-23-raw-historic) have been used to form the gold standard USA-23 Corpus. All four sub-corpora are pre-processed as follows. Text in a sub-corpus is cleaned by removing multiple spaces, duplicated text, diacritics as they are optional (only used for altering pronunciation), HTML tags, hashtags, links, URLs, and emoticons. Only sentences with five or more words were kept (as the empirical analysis of another study [94] has shown that sentences with a length of less than five words are typically incorrectly tagged). A language detection tool[35] was used to discard foreign words, which resulted into the removal of 957 tokens. After pre-processing, the four cleaned sub-corpora contain the raw text of 8,000 tokens (2,000 tokens in each sub-corpus).

In the next step of pre-processing, the raw text of 8,000 tokens is tokenized, lemmatized and POS tagged. The tokenization and POS tagging are carried out by using the Urdu natural language tools [94]. These tools use an

[25]BBC terms of use are available at this link: https://www.bbc.com/urdu/institutional-37588278 - Last visited: 27-January-2019
[26]https://www.express.pk/ - Last visited: 11-January-2019
[27]http://www.urdulibrary.org/ - Last visited: 11-January-2019
[28]http://www.minhajbooks.com/urdu/control/ - Last visited: 11-January-2019
[29]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-65A9-5 - Last visited: 11-January-2019
[30]To address privacy issues, we asked users for their permission to use the tweets, https://twitter.com/en/privacy - Last visited, 27-January-2019
[31]Under its privacy policy we can ask Facebook users to share their data, https://www.facebook.com/about/privacy/ - Last visited: 27-January-2019.
[32]https://www.statista.com - Last visited: 11-January-2019
[33]Its terms of use are available via this link: https://foundation.wikimedia.org/wiki/Terms_of_Use/en - Last visited: 27-January-2019
[34]http://www.hamariweb.com/ - Last visited: 11-January-2019
[35]https://lindat.mff.cuni.cz/repository/xmlui/handle/11858/00-097C-0000-0023-65A9-5 - Last visited: 21-January-2019

Urdu CLE POS tagset consisting of 35 tags [100]. This POS tagset is simple but based on the critical analysis of several previous iterations of Urdu POS tagset[36]. Furthermore, simplification of the POS tagset generally does not affect USAS semantic annotation system accuracy, as a single semantic category tends to span similar POS categories, for instance, present, past, and progressive tense of verbs [68, 94]. Lemmatization is carried out using an online Urdu tool[37]. Finally, the 8,000 tokens with automatically assigned POS tags and lemmas are stored in txt files (called USA-23-pp-news, USA-23-pp-smedia, USA-23-pp-wiki, and USA-23-pp-historic) given the name USA-pp-20.

3.4 Annotation Guidelines

Once raw texts are pre-processed, the next step is to manually annotate the extracted text with the USAS semantic field tags. Given a single- or multi-word, it can be annotated with the 232 categories of USAS semantic fields (see Section 3.1) depending on the text. Annotation conventions have been recorded as the annotation progressed. The annotation guidelines describe general issues and considerations which have been taken (Urdu inflectional and derivational morphology, MWE types and examples, foreign fragments, POS tags along with examples, common and punctuation symbols, quantity/numbering/date/time conventions), then briefly discusses 232 semantic field tags.

The further general guidelines followed by the annotators for the multi-target classification task are as follows: (I) read each sentence and annotate each individual word by understanding its local/surrounding context, (II) assign at least one and up to nine semantic field tags which best describes the single- and multi-word along with its POS tag, (III) annotate words with more relevant tags and in order of preference, domain of discourse can be use to alter rank ordering of semantic tags, (IV) if annotators faced difficulties deciding on the order of preference or tags then the USAS tagset provides a brief explanation of each tag along with its prototypical examples so read and understand these, (V) collected text must be carefully read and any ambiguity found in it can be discussed with the first author of this paper, (VI) proper/geographical names are considered as multi-word expressions, likewise for the case for abbreviations, (VII) misspelled or unconventionally spelled tokens in a text are interpreted according to their understanding and its context and must be tagged, otherwise tag them with Z99 (Unmatched) tag, (VIII) improperly tokenized[38] words must be joined as multi-word and tagged accordingly with POS and semantic field tags, (IX) MWE's take priority over single word tagging, and for the case of MWEs, try to form a complete MWEs, and finally (X) avoid as much as possible tagging words with Z99– unmatched tag.

3.5 Annotation Tool for Urdu Semantic Annotations

To facilitate the annotation of Urdu text with semantic field tags, we developed a user-friendly Java based Graphical User Semantic Annotation Interface (henceforth called GUSAI). Figure 1 shows the GUSAI screen-shot for a sample word "بات" ("Talk") (see Label 3 of screenshot) for the sentence "ایشرساں کیا بات ہیِ؟" ("Easher what's the matter?") (see Label 2) along with other information (this information was loaded from a file, see Section 3.3) including POS tag (see Label 4), a lemma (see Label 5), and semantic field tags[39] (see Label 6). Annotators were asked to attach as many (up to nine and at least one) USAS semantic field tag(s), as they deem appropriate for all senses of a word[40] and place them in descending order of importance. We asked annotators to edit the POS tag,

[36] http://www.cle.org.pk/Downloads/langproc/UrduPOStagger/UrduPOStagset.pdf - Last visited: 11-January-2019
[37] http://lemmatization.herokuapp.com/ - Last visited: 11-January-2019
[38] i.e. Split into two different words, but where they are a single word.
[39] For the process of semantic field tags assignment, a word along its POS tag information were looked up in the Urdu semantic lexicons (developed in another research project - available at the URL mentioned in Section 6), resulting in 7,461 semantically annotated tokens. The remaining 539 tokens which are not found in the Urdu semantic lexicons were manually annotated.
[40] The GUSAI has been developed to annotated only single words, however, MWEs are tagged manually (more details are mentioned in forthcoming sections).

lemma, and semantic field tags(s), if the pre-assigned information is incorrect, inappropriate, or incomplete. For words whose information is missing, they must add POS tag, lemma, and semantic field tag(s) information using GUSAI.

To assign semantic field tag(s) (if the assigned tag(s) is/are incomplete), an annotator needs to click on the "مزید ٹیگز منتخب کریں" ("add more tags") button (see Figure 1). Furthermore, to understand appropriate and common senses of a word, "بات" ("Talk") in our case (see Figure 1, Label 3), the references (of dictionaries, and thesauri) are displayed alongside the GUSAI. However, annotators were free to use any other resources as they wished.

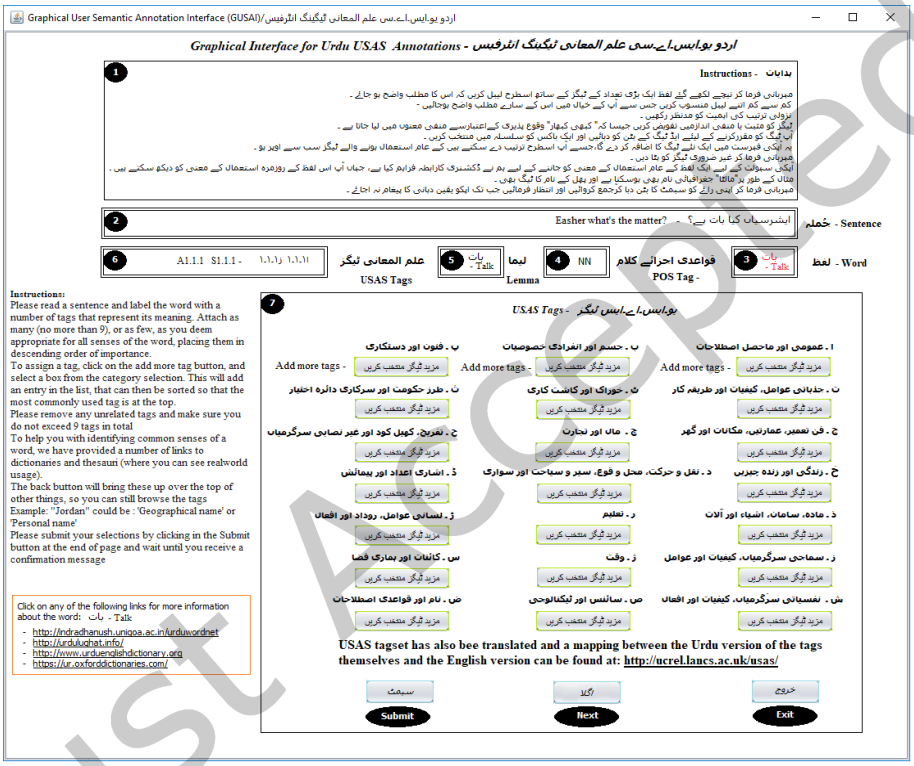

Fig. 1. Graphical User Semantic Annotation Interface (GUSAI) developed for the semantic annotations of our proposed USA-23 Corpus

By clicking "مزید ٹیگز منتخب کریں" ("add more tags") button (see Figure 1), an annotator is directed to sub-GUSAI (see Figure 2) in order to attach more semantic field tag(s) (see Section 3.1) or to remove irrelevant, incorrect ones by selecting or deselecting the check-boxes respectively. Furthermore, by clicking *go back*, it redirects to the main-GUSAI (see Figure 1), where the annotator may complete the remaining (add/remove relevant/irrelevant tag(s)) annotation process. However, by clicking the *submit* button it finalizes the annotation process for a word and then stores annotated information i.e., word, POS tag, lemma, and semantic field tag(s), in persistent storage. *Next* button will load the following word along with its POS tag, lemma, and semantic field

tag(s). When annotations are completed for the entire corpus, an annotator is prompted with an “annotation completion message” and (s)he can use the *Exit* button to close the annotation tool.

Fig. 2. Sub-GUSAI to add/remove semantic field tag(s).

3.6 Annotation process

Our proposed USA-23 Corpus (containing 8,000 tokens) was semi-automatically annotated by three annotators (A, B, and C), over the course of four months. All three annotators were Urdu native speakers and had a very good understanding of the USAS semantic tagset (see Section 1). All the annotators were graduate NLP students, experienced in text annotations, and had a high level of proficiency in Urdu. The USA-23 Corpus has been annotated at the word level with 21 major semantic fields and 232 sub domains of the USAS semantic tagset (see Section 3.1). The complete annotations were carried out in three phases: (i) training phase, (ii) annotations, and (iii) conflict resolution.

In the beginning, training consensus annotation sessions were held every other day (bi-diurnal). The rate of these sessions was decreased to weekly as the agreement score improved between the annotators. Although consensus annotations average scores are 3/6 for sentences (only for single words), every sentence was at least reviewed independently and jointly. Furthermore, in the annotation training process annotators are trained for the GUSAI uses (see Section 3.5) and asked to report any difficulty or error they have faced. For each sentence, the time of annotation is recorded. This process helped us to update the experimental interface and provided us with more information to make the task as well as GUSAI more efficient, reliable, and convenient. However, we observed in training sessions that all annotators were able to complete the task for a sentence within five minutes. In response, to the feedback of annotators, we have provided the ‘instruction set’ written in the GUSAI main interface (see Figure 1) to efficiently complete the task and a short description to complete the task. Furthermore,

these annotators were also observed for spamming annotations[41] and we have observed they are not submitting automatically assigned annotations or neither random results.

It has been observed that the training phase is more simple and easy for single words (using GUSAI), but, it is more difficult as annotators have less consensus i.e. 1/6 for MWEs. This is due to the fact that the Urdu language is highly inflected, morphologically rich, and poorly resourced and less research effort has been reported in past literature for MWEs [58]. There exists a study [38] which only defines named entities i.e. location and person name. For all other MWEs, Urdu words lead to an ambiguity problem because there is no clear agreement as to when to classify them as single words or MWEs [94]. For instance, the MWE, وزیر اعلی ('chief minister'), بہن بھائی ('sibling', literally 'brother sister'). The same is the case for reduplications, فر فر ('fluent'), and affixation, بد اخلاق ('depravedly'). These distinct forms may be perceived as single- or multi-words even by Urdu native speakers.

To train annotators, so that they can identify and annotate MWEs correctly, we have coordinated with Urdu linguistic experts. These experts have divided Urdu MWEs into seven types as follows: phrasal verbs, proverbs, collocations, idioms, noun phrases, proper names, and abbreviations. In addition, training sessions[42] for MWEs have been conducted.

In the training phase, two annotators (A and B) manually annotated single words of a subset of 62 sentences from the USA-pp-20 Corpus (see Section 3.3) using GUSAI. Afterward, both annotators were asked to identify MWEs from these single-word annotated sentences and assigned them POS tags and semantic field tag(s). These identified MWEs are then grouped into USAS semantic field tags (232) regardless of their morpho-syntactic patterns i.e. classifying MWEs in terms of semantic field tag(s) they represent. For instance, Table 2 shows five sample entries of various types of MWEs. Each row of the MWEs column stores expressions in indexing format. As an example, the idiom type of MWE is composed of three words (see Index 1, 1.1, and 1.2) along with its POS tags (see Section 3.3) and semantic tag(s) (see Section 3.1), listed on the left of brackets ([]). The term [MWE-1] in the semantic tag(s) column represents that it is the first MWE which consists of three parts/words ([MWE-1.1], [MWE-1.2], and [MWE-1.3]). The labels after the brackets ([]:) show semantic tag(s) for each individual word. It can be observed that some MWEs have only one semantic tag associated with them (see MWEs 3rd, 4th, and 5th start from the index 3, 4, and 5, respectively). However, the second reduplication type of MWE (see Index 2 and 2.1) combines three semantic tags (M1– Moving, coming and going), (M2– Putting, taking, pulling, pushing, transporting) and (M3– Movement/transportation: land) fields into one tag.

In addition, annotators A and B discussed the annotations both for single- and multi-words (both those agreed and conflicting pairs) on the initial subset of 62 sentences to further improve the quality of annotations. After that, the remaining corpus comprising 461 sentences was either semi-automatically (for single words) or manually (for MWEs) annotated by annotators A and B. After the annotation process, the Inter-Annotator Agreement (IAA) was computed for the entire corpus. In the third and last phase, the conflicting tokens were annotated by a third annotator (C – the first author of this paper), which resulted in a gold-standard semantically annotated corpus for the Urdu language.

The IAA on the entire USA-23 Corpus was calculated by using three approaches: (i) first correct – check whether the first semantic field tag selected by the annotator A matches with the first semantic field tag of annotator B, (ii) fuzzy-order – check whether semantic field tags selected by an annotator A are contained within

[41]Not taking the task seriously or were attempting to manipulate the system for personal gain by injecting some wrong tags or by just submitting the already assigned tags.

[42]All annotators undertook a practical training session on annotation tasks of MWEs and their types. Each annotator was given an annotation assignment of 150 random sentences of the UMC corpus [45] and requested to extract MWEs and identify their types. These tasks were marked and each annotator was awarded a score. Annotators are considered trained for MWEs annotation tasks when they scored 80% or above

Table 2. Annotated Sample of MWEs with USAS Semantic Tags

Index	MWEs	POS Tag	Semantic Tag(s)
1	آستین (arm)	NN	A5.1 G2.2 X9.1 [MWE-1.1]: B1 O2 B5
1.1	کا (is)	PSP	A5.1 G2.2 X9.1 [MWE-1.2]: A3 Z5
1.2	سانپ (snake)	NN	A5.1 G2.2 X9.1 [MWE-1.3]: L2
Translation of MWE: آستین کا سانپ (Devious)			
2	آنا (come)	VBI	M1 M2 M6 [MWE-2.1]: N3.1 A2.2
2.1	جانا (go)	AUXA	M1 M2 M6 [MWE-2.2]: A1.1.1
آنا جانا (Pop in)			
3	یو (U)	NNP	Z2 [MWE-3.1]: Z5 Z8
3.1	۔	PU	Z2 [MWE-3.2]:
3.2	کے (K)	NNP	Z2 [MWE-3.2]: N3.3 N3.7 Z5 Z8
Translation of MWE: یو ۔ کے (U.K)			
4	عمر (Umer)	NNP	Z1 [MWE-4.1]: Z3 T3 T1.3
4.1	فاروق (Farooq)	NNP	Z1 [MWE-4.2]: Z3 G2.2
Translation of MWE: عمر فاروق (Umer Farooq)			
5	تعلیم (Education)	NN	P1 [MWE-5.1]: X2.3
5.1	بالغاں (Adult)	NN	P1 [MWE-5.2]: T3 S2
Translation of MWE: تعلیم بالغاں (Adult education)			

the tags annotated by B in any order, (iii) strict-order – check whether annotator A semantic field tag(s) is/are identical to B in terms of semantic field tag(s) selection and order.

On the entire USA-23 Corpus, we obtained an IAA of 79.88% (first-correct), 81.61% (fuzzy-order), and 26.56% (strict-order) (see Table 3). It is important to note that annotators had an agreement on 6,390, 6,529, and 2,125 words for first-correct, fuzzy-order, and strict-order approaches, respectively. The IAA scores of first-order and fuzzy-order are considered as good, considering the difficulty of the Urdu semantic annotation task. However, strict-order shows low IAA results (26.56%). The Kappa Coefficient [54] computed for the entire USA-23 Corpus was 77.01%, 74.96%, and 21.07% using first-correct, fuzzy-order, and strict order semantic tagging approaches, respectively.

The details of IAA for four domain-wise sub-corpora (USA-23-News, USA-23-SMedia, USA-23-Wiki, and USA-23-Historic) are also shown in Table 3. It shows that the highest IAA score is obtained on the USA-23-News sub-corpus using the first-correct semantic tagging approach (84.65%). IAA scores of 83.76% and 81.05% are obtained for USA-23-SMedia and USA-23-Wiki sub-corpora respectively. The lowest IAA score of 70.07% is obtained for the USA-23-Historic sub-corpus. The possible reason for a low IAA score on the USA-23-Historic sub-corpus is that text in this sub-corpus is from older Urdu literature and annotators would have faced difficulty in correctly understanding the meanings of words from old Urdu. For the fuzzy-order semantic tagging approach, the USA-23-News sub-corpus has obtained the highest IAA score (86.06%), followed by USA-23-Wiki (82.42%), and USA-23-SMedia (81.97%) sub-corpora. The lowest score is 75.98% for the USA-23-Historic sub-corpus. Finally, for the strict-order semantic tagging approach, the highest IAA score is obtained by USA-23-News sub-corpus

i.e. 31.86%. The USA-23-SMedia, USA-23-Wiki, and USA-23-Historic sub-corpora have obtained IAA of 28.78%, 25.95%, and 19.63%, respectively.

The above discussion highlights the fact that in the case of first-order and fuzzy-order, the annotators are consistent, however, the strict-order annotators have huge variability. It also shows that the nature of text has an impact on the quality of semantic annotations as the USA-23-Historic sub-corpus obtained the lowest IAA compared to the other three sub-corpora on all three semantic tagging approaches i.e. first-order, fuzzy-order and strict-order. Finally, it is worth noting here that in the majority of cases, annotators have annotated the first tag correctly, it shows that on most important or core tags, annotators have good IAA scores.

Table 3. Inter-Annotator Agreement scores for USA-23 corpus and domain wise sub-corpora.

IAA approach Corpus/Sub-corpus	First-correct	Fuzzy-order	Strict-order
USA-23	79.88%	81.61%	26.56%
USA-23-News	84.65%	86.06%	31.86%
USA-23-SMedia	83.76%	81.97%	28.78%
USA-23-Wiki	81.05%	82.42%	25.95%
USA-23-Historic	70.07%	75.98%	19.63%

3.7 Corpus statistics

Table 4 shows the detailed statistics of the USA-23 Corpus. The gold standard USA-23 Corpus consists of 8,000 words (tokens), 2,213 unique tokens, and 523 sentences. The average number of words per sentence is approximately 15. In the proposed Corpus, there are 2,442 nouns, 1,529 verbs, 814 adjectives, 636 pronouns, and 161 adverbs. Furthermore, the total count for all semantic field tags in the corpus is 15,624.

To characterize the properties of any multi-targeted Corpus (USA-23 in our case), several useful multi-label indicators have been used in the recent past [118]. The primary and natural way to measure the multi-labeledness of the entire USA-23 Corpus is label cardinality. *Label cardinality* is a standard measure to calculate the average number of tags or labels per example present in the USA-23 Corpus. For a given multi-target corpus (USA-23), the label cardinality can be computed using the following equation.

$$Label\ cardinality = \frac{1}{N}\sum_{j=1}^{N}\sum_{k=1}^{L}(USA-23)_{k}^{j} \tag{1}$$

Where N means the number of examples and L is the number of labels. If the label cardinality score is greater than 1 then it means the corpus is a multi-targeted corpus (note that when L=1 the corpus is a single-label corpus). On the other hand, a label cardinality score of less than 2 means it is low multi-targeted. On our proposed USA-23 Corpus, we obtained a label cardinality score of 2.09. This high number shows that our corpus has a good label frequency.

The USA-23 Corpus contains 16 words with typos (spelling errors), which are annotated with Foreign Fragment "FF" POS tag and "Z99" (unmatched token) semantic field tag. Note that these typos are carried inherently from sources mentioned in Section 3.2. Typos were not replaced with correct words because it would be interesting to see the behaviour of semantic annotation methods (see Section 2.1) on such typographical words. Our proposed USA-23 Corpus is free and publicly available for research purposes (see Section 6).

Table 4. Detailed statistics of USA-23 Corpus

Complete Urdu semantically annotated corpus	
Sentence count	434
Word count	8,000
Unique words	2,213
Words with Z99	16
Tagged words	7,477
Untagged words	523
MWEs	872
Semantic tags	15,624
Named entities	590
Average no of words per sentence	15
Label cardinality	2.09

3.8 Corpus encoding

Our proposed USA-23 Corpus is encoded in XML format. Figure 3 shows an example (اشر سان ؟ پهلي تهي روز آٹھ سي آج جو هو نهن وهِ تم هي بات ڪا- Asher Sayan, what's the matter you are not what you were eight days ago?) of a semantically annotated sentence from the USA-23 Corpus in standard XML format. In this sentence, *<contextfile fileno="1" filename="USA-23 Corpus">*, indicates the beginning of a context file. The *fileno* and *filename* attributes show file number and file name, respectively. The attribute *<s snum=350>* indicates the beginning of a sentence, with unique IDs, i.e. *snum*. The tag *<wf pos="POS_tag" lemma="Lemma_of_Word" stags="USAS_Semantic_Tags" MWE="[MWE-no_of_MWE. component_no _of_MWE]: USAS_Semantic_Tags_for_single_word_of_a_MWE>*, indicates the beginning of a word in a particular sentence. The *pos* attribute shows the POS tag for a word, and *lemma* represents the lemma of a word (i.e. the dictionary headword), and *stags* shows USAS based semantic field tag(s) for a target word, *[MWE:Index.Its part no]* represents an index no of a MWE in the corpus along with its part no. Whereas, *:* is used to represent semantic tag(s) for a single word part of a MWE.

4 SEMANTIC ANNOTATION METHODS

In our proposed multi-target USA-23 Corpus, a tagged word can have one to nine Urdu semantic field tags associated with it. These tags have been used to indicate multiple membership categories from the USAS semantic taxonomy. i.e. different components of one sense (see Section 3). Therefore, we treated the Urdu semantic tagging problem as a multi-target classification problem. The following sections will describe the baseline and machine learning based approaches used for Urdu semantic tagging task, dataset, evaluation methodology, and evaluation measures.

4.1 Approaches

Most frequent sense approach. The Most Frequent Sense (MFS) heuristic is a simple but primary baseline for any supervised semantic annotation task [36]. To handle multi-target classification, we have adopted the most frequent sense in a way that it always predicts the most frequent *set* of senses (semantic tags - up to nine tags, if available) in the entire USA-23 Corpus.

Fig. 3. A semantically annotated sentence in standard XML format from our proposed USA-23 Corpus.

```
<?xml version=``1.0''encoding=``UTF-8''?>
<contextfile fileno=``1'' filename=``USA-23 Corpus''>
<body>
<s snum=350>
    <wf pos=``NNP'' lemma=``انسان'' stags=``Z1 S3.2 S3'' MWE=``[MWE-800.1]: Z3''>ایشر</wf>
    <wf pos=``NNP'' lemma=``انسان'' stags=``Z1 S3.2 S3'' MWE=``[MWE-800.2]: Z3 S9''>سیان</wf>
    <wf pos=``VBF'' lemma=``کیا'' stags=``A2.2 A2''>کیا</wf>
    <wf pos=``NN'' lemma=``بات'' stags=``A1.1.1 S1.1.1''>بات</wf>
    <wf pos=``VBF'' lemma=``ہے'' stags=``A3 Z5''>ہے</wf>
    <wf pos=``PRP'' lemma=``تم'' stags=``Z8''>تم</wf>
    <wf pos=``PRP'' lemma=``وہ'' stags=``A8''>وہ</wf>
    <wf pos=``NEG'' lemma=``نہ'' stags=``Z6''>نہیں</wf>
    <wf pos=``VBF'' lemma=``ہو'' stags=``Z5''>ہو</wf>
    <wf pos=``PRR'' lemma=``جو'' stags=``Z5''>جو</wf>
    <wf pos=``NN'' lemma=``آج'' stags=``T1.1.2 T1 T1.1''>آج</wf>
    <wf pos=``PSP'' lemma=``سے'' stags=``Z5''>سے</wf>
    <wf pos=``JJ'' lemma=``آٹھ'' stags=``T1.1.1 N1'' MWE=``[MWE-801.1]: N3.2 T1.2 T3 T1.3''>آٹھ</wf>
    <wf pos=``NN'' lemma=``روز'' stags=``T1.1.1 N6'' MWE=``[MWE-801.2]: T1.1.2 T1.3''>روز</wf>
    <wf pos=``RB'' lemma=``پہل'' stags=``T1.1.1 N4'' MWE=``[MWE-801.3]: T1 T1.1 T1.1.1 T1.2 T3''>پہلے</wf>
    <wf pos=``VBF'' lemma=``تھا'' stags=``T1.1.2 T1.3 N6 Z5''>تھے</wf>
    <punc pos=``PU''>؟</punc>
</s>
</body>
</contextfile>
</xml>
```

Machine learning approach. As discussed earlier the problem of Urdu semantic tagging is treated as a multi-target classification task. For this purpose, we first extracted three different types of features from each input word, (i) local, (ii) topical, and (iii) semantic features.

Local features. It is comprised of token form, POS tags– POS tags of a token itself "w_{p_0}", for three previous tokens "w_{p-1}, w_{p-2}, w_{p-3}" and the next three tokens "w_{p+1}, w_{p+2}, w_{p+3}". However, if there are fewer tokens (before or after) in the same sentence I_u, then we denote the corresponding feature as NIL. It is worth mentioning here that a token can be a word or a punctuation symbol and each of the previously mentioned tokens must be in the same sentence as w. We have used Urdu sentence tokenizer and POS tagger [94] to segment the tokens surrounding w into sentences and assign POS tags to these tokens. Furthermore, followed by lemmas– the lemma of a target word. The exemplary feature vector for the token کامیابی (success) in the sentence "دنیا/NN [7] کا/PSP ہر/JJ فرد/NN کامیابی/NN کا/PSP آرزومند/NN ہے/VBF ۔/PU" (Everyone in the world wants success.) is as follow: < کامیابی, NN, PSP, NN, NN, VBF, PU,∅, ∅, کامیاب > Where ∅ denotes a null POS tag for a token in a sentence (I_u) followed by a lemma (کامیاب: successful) for a token (w).

Topical features. Which consists of a bag-of-words (For each training/testing word, the vocabulary of words can be used as a feature(s). We have counted the word frequencies of all target words of the USA-23 Corpus. However, it has been shown [60] that this feature is position insensitive i.e. it counts the frequency of all words and disregards the grammatical details, the word order as well as its context) followed by a number of positional features i.e. collocations (we adopted the same 11 collocations features as cited in [23] i.e. $C_{-1,-1}$, $C_{1,1}$, $C_{-2,-2}$, $C_{2,2}$, $C_{-2,-1}$, $C_{-1,1}$, $C_{1,2}$, $C_{-3,-1}$, $C_{-2,1}$, $C_{-1,2}$, and $C_{1,3}$. Collocation $C_{j,k}$ means the ordered sequence of words and punctuation characters surrounding the target word. Furthermore, j and k refer to the starting and ending positions of the sequence, respectively. A negative (positive) value refers to the word position before (after) a target word). For instance, $C_{-2,-1}$ and $C_{-1,2}$ feature vector for the word کرتا (perform) in the sentence "دنیا/NN کا/PSP ہر/JJ فرد/NN کامیابی/NN کی/PSP تمنا/NN کرتا/VBF ۔/PU" (Everyone in the world wants success.) is as follow:

تمنا . کی and ∅ . ـ . کی. As we have performed in the POS feature a collocation does not consider any word after the sentence boundary marker, in this case, it will a null (∅). It is worth mentioning here that $C_{i,j}$ is represented with one feature however, it may contain several possible feature values (it takes binary values– indicating the presence or absence of that word). For instance, for a word w کرتا the set of selected collocations for $C_{-2,-1}$ is (محفل . کی, فطرت . کی, تمنا . کی). Then the feature value for $C_{-2,-1}$ (collocation) in the previously mentioned sentence is تمنا . کی.

Semantic feature. This type of feature consists of a domain indicator (cluster of texts regarding similar topics/subjects). In our case, we have used four main domains i.e. News, Social Media, Wikipedia, and Literature (see Section 3.2). For this feature, we have used the unlabelled UNLTool-ST-Test dataset [94] and have extracted individual words from this corpus after excluding stop words. Thereafter, we have selected the number of domains i.e., T, for the Urdu unlabelled UNLTool-ST-Test dataset and then we have applied the probabilistic LDA model [18] to obtain α (conditional probability of a word $w = i$ given a domain $z = j, p(w = i|z = j) = \alpha_{ij}$). Using this model, it clusters words that occurred in the Urdu unlabelled UNLTool-ST-Test dataset according to the T domains (in our case News, Social Media, Wikipedia, and Literature). This conditional probability $p(w = i|z = j) = \alpha_{ij}$ is then used to tag the words in the corpus with the probability of each topic. Moreover, a word with the highest probability is tagged with that particular domain. For instance, a word "میچ" ("match") belongs to all four domains, therefore assigned probability score for all four domains i.e. *News: 0.53*, *Social Media: 0.21, Wikipedia: 0.48*, and *Literature: 0.01*. However, as the News domain has the highest score so this word in the feature vector is represented with the same decimal code i.e. 1 (which indicates, this word belongs to the news domain).

All the above mentioned extracted features (word form, POS tags, lemma, bag-of-words, collocation, and semantic) are used to train different multi-target classifiers. After extracting the local, topical, and semantic set of features from the entire USA-23 Corpus, we applied seven different multi-target classifiers to them. The next section discusses these multi-target classifiers in more detail.

Multi-target Classifiers. In contrast to single-label ML algorithms (see Section 2.1), in supervised multi-target settings, each target variable can take multiple class values. This type of classification is performed using two main approaches: (i) Problem Transformation, and (ii) Algorithm Adaptation [105, 118].

Problem Transformation is primarily used for multi-target classifiers – a multi-target problem is transformed into one or more single-label problems. Doing so, single-label ML algorithms are employed in such a way, that their single-label predictions are transformed into multi-label predictions. On the other hand, Algorithm Adaptation is an alternative to problem transformation, where internal modification is required in existing classifiers to handle multi-target data directly (off-the-shelf approaches include Decision Tree [108], MLRF (Multi-Label Random Forest) [48]). However, Algorithm Adaptation approaches are usually domain specific, for instance, a decision tree is popular in bioinformatics [82]. Consequently, problem transformation provides flexibility and scalability: any state-of-the-art single-label ML algorithms (K-Nearest Neighbour [98]) can be used to suit requirements. Problem transformation can be primarily sub-classified into two categories: (i) Binary Relevance [105], and (ii) Label Combination [81] classifiers.

Binary Relevance (BR) is the most common and baseline multi-target problem transformation classifier [105]. It transforms a multi-target problem into multiple independent binary classification problems, where each binary classifier is trained to predict the relevance of one of the labels. The common binary relevance label prediction is calculated by the following function:

$$\hat{y} = h_j(x) = \underset{y_j \in \{Y_j\}}{argmax}\ \rho(y_j|x), \quad j = 1, ..., L \tag{2}$$

Where, for each j, a state-of-the-art single-label ML algorithm h_j is employed to map a data instance to the relevance of the j_{th} label.

There are several families of Binary Relevance classifiers in the literature. However, an in-depth study and comparison of all these classifiers are beyond the scope of this research article. Therefore, for Urdu semantic tagging task, we will use the four most common and popular classifiers: (i) Bayesian Classifier Chains, (ii) Classifiers Chains, (iii) Classifiers Probabilities Chains, and (iv) Class Relevance [24, 80].

Another well-known Problem Transformation approach to handle the supervised multi-target classification task is Label Combination (LC). It also transforms a multi-label problem into a multi-class problem by treating all label sets as atomic labels, that is, each label set is treated as a single label in a single-label multi-class problem. Label probability in LC can be expressed by:

$$\hat{y} = \underset{y \in Y}{argmax}\ \rho(y|x), \quad |Y| \ll 2^L \tag{3}$$

For this study, we have selected three Label Combination algorithms, (i) Nearest Set Replacement, (ii) Random -labEL Disjoint Pruned Sets (RAkELd), and (iii) Super Class Classifiers [80, 82], as these have proven to be effective in previous literature [103].

These multi-target classifiers have been used for various text classification tasks (see Section 2.1). However, to the best of our knowledge, multi-target classifiers have never been explored for a semantic tagging task in general and particularly in the context of the Urdu language. Therefore, another contribution of this research study is that we have extracted various features (see Section 4.1) from our proposed USA-23 Corpus and applied seven different multi-target classifiers to them.

4.2 Evaluation measures

The performance of a multi-target classifier can be measured using two approaches: (i) *label-based* – evaluated on a per-label basis, and (ii) *instance-based* – used to carry out an evaluation on label sets [24]. In this study, we used three evaluation measures to evaluate the performance of our Machine Learning based approaches: (i) Exact Match (an instance-based evaluation measure), (ii) Hamming Loss (an instance-based evaluation measure), and (iii) Accuracy (a label-based evaluation measure).

Exact match computes the percentage of instances whose predicted set of labels ($\hat{y}$) are exactly the same as their corresponding true set of labels (y), this measure is also known as 0/1 subset or classification accuracy (see Equation 4). Where $\mathbb{I}$ is the indicator function.

$$Exact\ match = \frac{1}{N}\sum_{j=1}^{N} \mathbb{I}(\hat{y}^{(j)} \neq y^{(j)}) \tag{4}$$

Hamming loss is used to evaluate how many times, on average, an example-label pair is misclassified (see Equation 5). This is a loss function, therefore, the lower the value means the higher the performance of the classifier.

$$Hamming\ loss = \frac{1}{NL}\sum_{i=1}^{N}\sum_{j=1}^{L} \mathbb{I}[\hat{y}_j^{(i)} \neq y_j^{(i)}] \tag{5}$$

Accuracy is the proportion of label values correctly classified of the total number of labels for that instance averaged over all instances (predicted ($\hat{y}$) and true (y)), for a set of N test examples (see Equation 6).

$$Accuracy = \frac{1}{N}\sum_{j=1}^{N}\frac{\left|\hat{y}^{(j)} \wedge y^{(j)}\right|}{\left|\hat{y}^{(j)} \vee y^{(j)}\right|} \tag{6}$$

4.3 Corpus

For the set of experiments presented in this study, the entire USA-23 Corpus is used (see Section 3.7). There is a total of 8,000 tokens in the corpus (2,000 for each of USA-23-News, USA-23-SMedia, USA-23-Wiki, and USA-23-Historic).

4.4 Evaluation methodology

The task of Urdu semantic tagging is treated as a multi-target classification task, as one word can have one or more semantic field tags. Features extracted using local, topical, and semantic approaches (see Section 4.1) are used as input to multi-target classifiers. We applied nine different multi-target classifiers (Bayesian Classifiers Chain, Classifier Chain, Classifier Chain Probabilities, Class Relevance, Nearest Set Replacement, Random -labEL Disjoint Pruned Sets (RAkELd), Super Class Classifiers, Deep Interpretation of Classifier Chains (DeepML) [41], and Deep Back-Propagation Neural Network) [79]. To better evaluate the performance of Machine Learning based Urdu semantic tagging methods, we applied 10-fold cross-validation. The *MEKA*[43] [83] implementation of the multi-target classifiers, with its default parameter settings (except RAkELd – where the following parameters are selected empirically: subset size is varied from 2 to 5, number of models selected 1 to 100, and the threshold is set to 0.1 to 0.9 with a 0.1 step and DeepML – with 20 hidden units per layer, a learning rate of 0.2, the momentum of 0.3, and 4000 iterations), is used for the supervised classification task. Furthermore, all experiments were run on a 64-bit computing machine, with 8 GB RAM.

5 RESULTS AND ANALYSIS

In Table 5, we present the Exact Match (EM), Hamming Loss (HL) and Accuracy scores obtained for Urdu semantic annotation tasks using Most Frequent Sense (MFS) and Machine Learning (ML) based approaches applied on our proposed USA-23 Corpus. "Classifiers" in the table refers to the Problem Transformation (PT) based Multi-target classifiers that produced the highest results among all the three single-label algorithms used in this research. "NB", and "RF" means Naïve Bayes and Random Forest, respectively. "RAkELd" is used as a short form of Random k-labEL Disjoint Pruned Sets. "BR" and "LC" refers to Binary Relevance and Label Combination which are problem transformation classifiers. The NN refers to the Neural Network.

Overall, for Hamming Loss and Accuracy evaluation measures, the best results are obtained using the RAkELd (Hamming Loss = 0.36 and Accuracy = 0.54). However, for the Exact Match measure, the highest scores are obtained using the Nearest Set Replacement i.e. 0.47. Thus, we can say that when we have considered all three evaluation measures the RAkELd (Exact Match = 0.46, Hamming Loss = 0.36, and Accuracy = 0.54) classifiers outperform all other multi-target classifiers. As far as the case of Deep learning methods are concerned, the results are almost comparable. However, it can be seen that the best results are obtained using deep back propagation neural network (DBPNN), exact match = 0.26, hamming loss = 0.41, and accuracy of 0.39. Also, these results are significantly higher than the baseline approach i.e. Most Frequent Sense (Accuracy = 0.31) (see Section 4.1). As can be noted that very promising results are obtained for the Urdu semantic annotation task indicating that the multi-target classifiers are effective in assigning semantic field tag(s) to Urdu words in our proposed corpus. Furthermore, it can be observed that deep learning methods, DeepML (Deep Interpretation of Classifier Chains) and Deep Back Propagation Neural Network (DBPNN) are also lower than other multi-target classifiers. The

[43] http://waikato.github.io/meka/ - Last visited: 17-January-2022

Table 5. Results obtained on USA-23 Corpus using Most Frequent Sense and Machine Learning based approaches

	Classifiers		**Evaluation Measures**		
	PT based Multi-target	**Single-label**	**EM**	**HL**	**Accuracy**
Approach	**Type: Name**				
MFS					
	–	–	–	–	0.31
ML					
	BR: Bayesian Classifier Chain	NB	0.28	0.43	0.48
	BR: Classifier Chain	*NB*	*0.40*	*0.37*	*0.53*
	BR: Classifier Chain Probabilities	NB	0.33	0.45	0.45
	BR: Class Relevance	NB	0.31	0.39	0.51
	BR: DeepML	NB	0.23	0.39	0.32
	LC: Nearest Set Replacement	NB	0.33	0.40	0.51
	LC: RAkELd	NB	0.31	0.40	0.50
	LC: Super Class Classifier	NB	0.33	0.40	0.50
	LC: Deep Back-Propagation NN	NB	0.21	0.40	0.38
ML					
	BR: Bayesian Classifier Chain	RF	0.45	0.37	0.53
	BR: Classifier Chain	RF	0.45	0.36	0.53
	BR: Classifier Chain Probabilities	RF	*0.45*	*0.37*	*0.53*
	BR: Class Relevance	RF	0.45	0.36	0.53
	BR: DeepML	RF	0.25	0.41	0.38
	LC: Nearest Set Replacement	RF	***0.47***	0.37	0.54
	LC: RAkELd	***RF***	0.46	***0.36***	***0.54***
	LC: Super Class Classifier	RF	0.42	0.43	0.51
	LC: Deep Back-Propagation NN	RF	0.26	0.41	0.39
ML					
	BR: Bayesian Classifier Chain	J48	0.39	0.37	0.50
	BR: Classifier Chain	J48	0.42	0.37	0.53
	BR: Classifier Chain Probabilities	J48	0.42	0.28	0.53
	BR: Class Relevance	J48	0.49	0.41	0.51
	BR: DeepML	J48	0.25	0.40	0.52
	LC: Nearest Set Replacement	*J48*	*0.45*	*0.36*	*0.53*
	LC: RAkELd	J48	0.30	0.40	0.51
	LC: Super Class Classifier	J48	0.46	0.38	0.53
	LC: Deep Back-Propagation NN	J48	0.25	0.40	0.38

reason for such low results is the size of our proposed dataset (see Section 3.7), as deep learning techniques are data hungry and this is not feasible in our case.

Among BR and LC sub-classifiers, although the best results (based on average) are obtained using Label Combination considering all three evaluation measures (Exact Match, Hamming Loss, and Accuracy), however,

the difference in performance is small. The possible reason for this might be its classification style. Where each class is considered as a random subset of labels and thus learned a single-label classifier for prediction of each label in the powerset. This highlights the fact that both BR and LC type of Problem Transformation based multi-target classifiers are effective in Urdu semantic annotations on our proposed corpus as compared to the deep learning methods.

Regarding single-label ML algorithms (Naïve Bayes, Random Forest, and J48) which are used in combination with multi-target classifiers, the best results are obtained using Random Forest on both BR (Classifier Chain) and LC (RAkELd) sub-classifiers. The possible reason for obtaining good results using Random Forest is that it is considered the best ensemble learning algorithm for the single-label classification task, thus when combined with multi-target classifiers (RAkELd and Classifier Chain) it constructs multiple single-label training sets from the multi-targeted USA-23 Corpus.

Table 6 presents the Exact Match (EM), Hamming Loss (HL), and Accuracy scores obtained for Urdu semantic annotation tasks using Machine Learning (ML) based approaches applied on our various sub corpora (USA-23-News, USA-23-SMedia, USA-23-Wiki, and USA-23-Historic). For the set of experiments presented here single-label Random Forest algorithm has been used (selected as this has produced better results (see Table 5) as compared to two others, NB and J48). All other terms of the table are the same as described previously. The best average results obtained overall on the sub corpus is presented in bold, whereas, the second highest average results on sub corpus is presented in italic.

It can be observed, the best average results are obtained on the USA-23-Historic sub corpus. Where the average EM, HL, and Accuracy have the following scores, 0.30, 0.51, and 0.46, respectively. The lowest average results are observed for the USA-23-Wiki sub corpus (EM = 0.26, HL = 0.49, and Accuracy = 0.43). The average results on USA-23-SMedia sub corpus have EM score of 0.27, HL score of 0.48, and an Accuracy of 0.45. On the USA-23-News sub corpus obtained average results are as, EM: 0.26 HL: 0.48, and Accuracy: 0.45.

Table 7 presents some more detailed results (using Exact Match (EM), Hamming Loss (HL), and Accuracy scores) of local, topical, and semantic features (see Section 4.1) which has been used to train and test different multi-target classifiers on the proposed USA-23 Corpus. This analysis is also based on the Random Forest single-label algorithm. All others terminologies of the table are the same as described previously.

The average results are as expected. The best average results on the USA-23 Corpus are obtained using Local features (EM = 0.24, HL = 0.55, and Accuracy = 0.40). The lowest results are obtained using the Semantic feature i.e. EM = 0.22, HL = 0.58, and Accuracy = 0.38. However, the last Topical feature has also produced the similar type of results i.e. EM = 0.20, HL = 0.60, and Accuracy = 0.37.

To conclude, the best results on the USA-23 Corpus are obtained using RAkELd and Classifier Chain when considering all three evaluation measures. However, when several sub corpora are evaluated using different ML based techniques, the best results are obtained for the USA-23-Historic sub-corpus, which reflects that for the historic type of text, multi-target classifiers are more appropriate. However, the best highest average weighted features for the USA-23 Corpus are Local whereas, the second highest feature for Urdu semantic tagging task is Topical. It also shows that semantic features are less useful for the multi-target semantic tagging task for the Urdu text. Moreover, it can be observed from the above discussion that deep learning methods are not appropriate for our proposed dataset. However, among deep learning classifiers, deep back propagation neural network has produced a good result as compared to DeepML.

6 CONCLUSION AND FUTURE WORK

This study presents a benchmark corpus for the Urdu semantic tagging task. Our proposed USA-23 Corpus contains 8,000 tokens (2,000 tokens each from News, Social Media, Wikipedia, and Historic articles). Each word in the USA-23 Corpus is annotated with one to nine semantic field tag(s) using the USAS semantic taxonomy

Table 6. Results obtained on various sub corpora using Machine Learning approaches

Corpus	**Multi-target Classifiers Type: Name**	**Evaluation Measures EM**	**HL**	**Accuracy**
USA-23-News				
	BR: Bayesian Classifier Chain	0.27	0.48	0.49
	BR: Classifier Chain	0.27	0.48	0.49
	BR: Classifier Chain Probabilities	0.27	0.48	0.49
	BR: Class Relevance	0.27	0.47	0.49
	BR: DeepML	0.20	0.49	0.32
	LC: Nearest Set Replacement	0.30	0.47	0.49
	LC: RAkELd	0.27	0.47	0.49
	LC: Super Class Classifier	0.25	0.48	0.48
	LC: Deep Back-Propagation NN	0.21	0.48	0.32
	Average score of all classifiers	0.26	0.48	0.45
USA-23-SMedia				
	BR: Bayesian Classifier Chain	0.29	0.47	0.49
	BR: Classifier Chain	0.29	0.47	0.49
	BR: Classifier Chain Probabilities	0.29	0.47	0.49
	BR: Class Relevance	0.29	0.47	0.49
	BR: DeepML	0.20	0.51	0.31
	LC: Nearest Set Replacement	0.30	0.47	0.49
	LC: RAkELd	0.30	0.46	0.49
	LC: Super Class Classifier	0.29	0.47	0.49
	LC: Deep Back-Propagation NN	0.21	0.49	0.31
	Average score of all classifiers	0.27	0.48	0.45
USA-23-Wiki				
	BR: Bayesian Classifier Chain	0.28	0.48	0.48
	BR: Classifier Chain	0.28	0.48	0.48
	BR: Classifier Chain Probabilities	0.28	0.48	0.48
	BR: Class Relevance	0.28	0.48	0.48
	BR: DeepML	0.20	0.52	0.30
	LC: Nearest Set Replacement	0.29	0.48	0.48
	LC: RAkELd	0.28	0.48	0.48
	LC: Super Class Classifier	0.22	0.55	0.41
	LC: Deep Back-Propagation NN	0.20	0.49	0.31
	Average score of all classifiers	*0.26*	*0.49*	*0.43*
USA-23-Historic				
	BR: Bayesian Classifier Chain	0.34	0.4	0.52
	BR: Classifier Chain	0.34	0.46	0.52
	BR: Classifier Chain Probabilities	0.34	0.46	0.52
	BR: Class Relevance	0.34	0.46	0.52
	BR: DeepML	0.18	0.67	0.23
	LC: Nearest Set Replacement	0.35	0.46	0.52
	LC: RAkELd	0.34	0.46	0.52
	LC: Super Class Classifier	0.25	0.53	0.46
	LC: Deep Back-Propagation NN	0.18	0.63	0.24
	Average score of all classifiers	**0.30**	**0.51**	**0.46**

Table 7. Results obtained on the USA-23 Corpus using local, topical and semantic features

	PT based Multi-target Classifiers	**Evaluation Measures**		
Features	**Type: Name**	**EM**	**HL**	**Accuracy**
Local				
	BR: Bayesian Classifier Chain	0.26	0.52	0.44
	BR: Classifier Chain	0.26	0.52	0.44
	BR: Classifier Chain Probabilities	0.26	0.52	0.44
	BR: Class Relevance	0.26	0.52	0.44
	BR: DeepML	0.18	0.67	0.23
	LC: Nearest Set Replacement	0.27	0.53	0.45
	LC: RAkELd	0.27	0.51	0.45
	LC: Super Class Classifier	0.25	0.53	0.44
	LC: Deep Back-Propagation NN	0.18	0.63	0.24
	Average score of all classifiers	**0.24**	**0.55**	**0.40**
Topical				
	BR: Bayesian Classifier Chain	0.24	0.54	0.43
	BR: Classifier Chain	0.24	0.54	0.42
	BR: Classifier Chain Probabilities	0.24	0.55	0.43
	BR: Class Relevance	0.24	0.54	0.43
	BR: DeepML	0.17	0.70	0.22
	LC: Nearest Set Replacement	0.24	0.54	0.43
	LC: RAkELd	0.26	0.55	0.42
	LC: Super Class Classifier	0.19	0.57	0.40
	LC: Deep Back-Propagation NN	0.17	0.69	0.22
	Average score of all classifiers	*0.22*	*0.58*	*0.38*
Semantic				
	BR: Bayesian Classifier Chain	0.21	0.57	0.41
	BR: Classifier Chain	0.21	0.57	0.41
	BR: Classifier Chain Probabilities	0.21	0.57	0.41
	BR: Class Relevance	0.21	0.57	0.41
	BR: DeepML	0.17	0.71	0.21
	LC: Nearest Set Replacement	0.22	0.56	0.42
	LC: RAkELd	0.22	0.55	0.42
	LC: Super Class Classifier	0.21	0.57	0.41
	LC: Deep Back-Propagation NN	0.17	0.70	0.22
	Average score of all classifiers	0.20	0.60	0.37

(21 major semantic fields and 232 sub-fields). To demonstrate how our proposed corpus can be used for the development and evaluation of an Urdu semantic tagging method(s) we explored local, topical, and semantic feature extraction approaches and applied seven multi-target classifiers. Our results show that RAkELd and Classifier Chain multi-target classifiers outperform all other classifiers. The USA-23 Corpus, Graphical Urdu Semantic Annotation Interface tool and other supporting resources are publicly available for research purposes at

https://github.com/UCREL/USA-23Corpus under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License[44].

Our novel research contributions are as follows: (i) the development of the first large semantically annotated corpus for Urdu, made freely available to the research community, (ii) the application of various multi-target machine learning classifiers to the semantic tagging task for the first time in any language, and (iii) the development of different supporting resources (such as the annotation tool, and the single word semantic lexicon).

In the future, we plan to explore other feature extraction approaches and multi-label classifiers. Increasing the size of our proposed corpus by adding several other genres of Urdu literature is another avenue for future work, further to this it will be interesting to see the classification of various POS categories.

ACKNOWLEDGEMENT

This work has been supported by COMSATS University Islamabad, Lahore Campus, Pakistan, and Lancaster University, UK, under the split site Ph.D. programme.

REFERENCES

[1] Muhammad Abid, Asad Habib, Jawad Ashraf, and Abdul Shahid. 2017. Urdu word sense disambiguation using machine learning approach. *Cluster Computing* (2017), 1–8. https://doi.org/10.1007/s10586-017-0918-0

[2] Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*. Vol. 33. Springer Science & Business Media.

[3] Eneko Agirre and David Martinez. 2000. Exploring automatic word sense disambiguation with decision lists and the web. In *Proceedings of the COLING-2000 Workshop on Semantic Annotation and Intelligent Content, Luxembourg*. Association for Computational Linguistics, 11–19.

[4] Eneko Agirre, David Martínez, Oier López de Lacalle, and Aitor Soroa. 2006. Two graph-based algorithms for state-of-the-art WSD. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'04), Sydney, Australia*. Association for Computational Linguistics, 585–593.

[5] Eneko Agirre and Mark Stevenson. 2006. Knowledge sources for WSD. *Word Sense Disambiguation* 33 (2006), 217–251.

[6] Tafseer Ahmed, Toqeer Ehsan, Almas Ashraf, Mutee u Rahman, Sarmad Hussain, and Miriam Butt. 2020. A Multilayered Urdu Treebank. In *Proceedings of the Conference on Language & Technology, (CLT'20), Lahore, Pakistan*. Centre for Language Engineering Al-Khawarizmi Institute of Computer Science University of Engineering and Technology, 1–7.

[7] Tafseer Ahmed, Saba Urooj, Sarmad Hussain, Asad Mustafa, Rahila Parveen, Farah Adeeba, Annette Hautli, and Miriam Butt. 2015. The CLE Urdu POS tagset. In *LREC 2014, Ninth International Conference on Language Resources and Evaluation, Reykjavik, Iceland*. 2920–2925.

[8] Bandar Al-Hejin. 2015. Covering Muslim women: Semantic macrostructures in BBC news. *Discourse & Communication* 9, 1 (2015), 19–46.

[9] Marc Alexander, Fraser Dallachy, Scott Piao, Alistair Baron, and Paul Rayson. 2015. Metaphor, popular science, and semantic tagging: Distant reading with the Historical Thesaurus of English. *Digital Scholarship in the Humanities (DSH)* 30, suppl_1 (2015), i16–i27.

[10] James Allan. 2012. *Topic detection and tracking: event-based information organization*. Vol. 12. Springer Science & Business Media.

[11] Maaz Anwar, Riyaz Ahmad Bhat, Dipti Misra Sharma, Ashwini Vaidya, Martha Palmer, and Tafseer Ahmed Khan. [n.d.]. A Proposition Bank of Urdu. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*.

[12] DE Archer, Paul Rayson, Scott Piao, and AM McEnery. 2004. Comparing the UCREL semantic annotation scheme with lexicographical taxonomies. In *Proceedings of the Eleventh EURALEX International Congress (EURALEX'04), Lorient, France*. European Association for Lexicography, 817–827.

[13] Giuseppina Balossi. 2014. *A corpus linguistic approach to literary language and characterization: Virginia Woolf's The Waves*. Vol. 18. John Benjamins Publishing Company.

[14] Alistair Baron, Caroline Tagg, Paul Rayson, Philip Greenwood, James Walkerdine, and Awais Rashid. 2011. Using verifiable author data: Gender and spelling differences in Twitter and SMS. In *International Computer Archive of Modern and Medieval English (ICAME 31), Oslo, Norway*. 61–73.

[15] Roberto Basili, Michelangelo Della Rocca, and Maria Teresa Pazienza. 1997. Towards a bootstrapping framework for corpus semantic tagging. *Tagging Text with Lexical Semantics: Why, What, and How?* (1997).

[44] https://creativecommons.org/licenses/by-nc/4.0/- Last checked: 26-January-2023

[16] Rafiya Begum, Samar Husain, Arun Dhwaj, Dipti Misra Sharma, Lakshmi Bai, and Rajeev Sangal. 2008. Dependency annotation scheme for Indian languages. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II (IJCNLP'08), Hyderabad, India*. Asian Federation of Natural Language Processing (AFNLP), 721–726.

[17] Riyaz Ahmad Bhat and Dipti Misra Sharma. 2012. A dependency Treebank of Urdu and its evaluation. In *Proceedings of the Sixth Linguistic Annotation Workshop (The LAW'12), Jeju, Republic of Korea*. Association for Computational Linguistics Special Interest Group for Annotation (ACL SIGANN), 157–165.

[18] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3 (2003), 993–1022.

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

[20] Stefan Bordag. 2006. Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation.. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL'06), Trento, Italy*. Association for Computational Linguistics, 137–144.

[21] Matthew Boutell, Xipeng Shen, Jiebo Luo, and Chris Brown. 2003. *Multi-label semantic scene classification.* Technical Report. technical report, department of computer sciences. u. Rochester.

[22] Rebecca F Bruce and Janyce M Wiebe. 1999. Decomposable modeling in natural language processing. *Computational Linguistics* 25, 2 (1999), 195–207.

[23] Jun Fu Cai, Wee Sun Lee, and Yee Whye Teh. 2007. NUS-ML: Improving word sense disambiguation using topic features. In *Proceedings of the 4th International Workshop on Semantic Evaluations, Prague, Czech Republic*. Association for Computational Linguistics, 249–252.

[24] F Charte, AJ Rivera, MJ del Jesus, and F Herrera. 2018. *Multilabel Classification: Problem Analysis, Metrics and Techniques.* Springer.

[25] Amanda Clare and Ross King. 2001. Knowledge discovery in multi-label phenotype data. *Principles of data mining and knowledge discovery* (2001), 42–53.

[26] Hamish Cunningham, Diana Maynard, and Kalina Bontcheva. 2011. *Text processing with GATE.* Gateway Press CA.

[27] André de Carvalho and Alex Freitas. 2009. A tutorial on multi-label classification techniques. *Foundations of Computational Intelligence* 5 (2009), 177–195.

[28] Bart Decadt, Véronique Hoste, Walter Daelemans, and Antal Van den Bosch. 2004. GAMBL, genetic algorithm optimization of memory-based WSD. In *3rd International workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3); held in conjunction with the 42nd Annual meeting of the Association for Computational Linguistics (ACL'04), Barcelona, Spain*. Association for Computational Linguistics, 108–112.

[29] George Demetriou and Eric Steven Atwell. 2001. A domain-independent semantic tagger for the study of meaning associations in English text. In *Proceedings of the Fourth International Workshop on Computational Semantics (IWCS'4), Prague, Czech Republic*. Association for Computational Linguistics (ACL) Special Interest Group in Computational Semantics (SIGSEM), 67–80.

[30] Leon Derczynski, Diana Maynard, Niraj Aswani, and Kalina Bontcheva. 2013. Microblog-genre noise and impact on semantic annotation accuracy. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media, (HT '13), Paris, France*. ACM, 21–30.

[31] Neil Doherty, Nigel Lockett, Paul Rayson, and Stuart Riley. 2006. Electronic-CRM: a simple sales tool or facilitator of relationship marketing?. In *29th Institute for Small Business & Entrepreneurship Conference. International Entrepreneurship-from local to global enterprise creation and development (ISBE'06), Cardiff, Wales, United Kingdom.*

[32] Mahmoud El-Haj, Paul Rayson, Scott Piao, and Stephen Wattam. 2017. Creating and validating multilingual semantic representations for six languages: expert versus non-expert crowds. In *Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications (SENSE'17), Valencia, Spain*. Association for Computational Linguistics, 61–71.

[33] Ricardo Gacitua, Pete Sawyer, and Paul Rayson. 2008. A flexible framework to experiment with ontology learning techniques. In *Research and Development in Intelligent Systems XXIV*. Springer, 153–166.

[34] William A Gale, Kenneth W Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities* 26, 5 (1992), 415–439.

[35] William A Gale, Kenneth W Church, and David Yarowsky. 1992. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language (HLT'91), PA, USA*. Association for Computational Linguistics, 233–237.

[36] Eva Gibaja and Sebastián Ventura. 2015. A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)* 47, 3 (2015), 52.

[37] Sylviane Granger, Magali Paquot, and Paul Rayson. 2006. Extraction of multi-word units from EFL and native English corpora: The phraseology of the verb 'make'. *Phraseology in motion I: Methoden und Kritik* (2006), 57–68.

[38] Annette Hautli and Sebastian Sulger. 2011. Extracting and classifying Urdu multiword expressions. In *Proceedings of the the ACL-HLT Student Session (ACL-HLT'11), Portland, OR, USA*. Association for Computational Linguistics, 24–29.

[39] Annette Hautli, Sebastian Sulger, and Miriam Butt. 2012. Adding an annotation layer to the Hindi/Urdu treebank. *Linguistic Issues in Language Technology* 7, 1 (2012).

[40] Verena Henrich, Erhard Hinrichs, and Tatiana Vodolazova. 2012. Webcage: a web-harvested corpus annotated with germanet senses. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12), Avignon, France*. Association for Computational Linguistics, 387–396.

[41] Geoffrey Hinton and Ruslan Salakhutdinov. 2006. Reducing the Dimensionality of Data with Neural Networks. *Science* 313, 5786 (2006), 504–507.
[42] Kevin Humphreys, Robert Gaizauskas, Saliha Azzam, Chris Huyck, Brian Mitchell, Hamish Cunningham, and Yorick Wilks. 1998. University of Sheffield: Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the 7th Message Understanding Conferences (MUC-7), Fairfax, Virginia*. Association for Computational Linguistics, 1–20.
[43] Nancy Ide and Jean Véronis. 1998. Word sense disambiguation: the state of the art. *Computational linguistics* 24, 1 (1998), 1–41.
[44] Jerker Järborg, Dimitrios Kokkinakis, and Maria Toporowska Gronostaj. 2002. Lexical and Textual Resources for Sense Recognition and Description.. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands, Spain*. European Language Resources Association (ELRA), 1492–1497.
[45] Bushra Jawaid, Amir Kamran, and Ondrej Bojar. 2014. A Tagged Corpus and a Tagger for Urdu.. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'09), Reykjavík, Iceland*. European Language Resources Association (ELRA), 2938–2943.
[46] Adam Kilgarriff. 2004. How dominant is the commonest sense of a word?. In *International Conference on Text, Speech and Dialogue (TSD'04), Brno, Czech Republic*. Springer, 103–111.
[47] Beata Beigman Klebanov, Daniel Diermeier, and Eyal Beigman. 2008. Automatic annotation of semantic fields for political science research. *Journal of Information Technology & Politics* 5, 1 (2008), 95–120.
[48] Feng Liu, Xiaofeng Zhang, Yunming Ye, Yahong Zhao, and Yan Li. 2015. MLRF: Multi-label classification through random forest with label-set partition. In *International Conference on Intelligent Computing (ICIC'15) Fuzhou, China*. Springer. 407–418.
[49] Laura Löfberg, Dawn Archer, Scott Piao, Paul Rayson, Tony McEnery, Krista Varantola, and Jukka-Pekka Juntunen. 2005. Porting an English semantic tagger to the Finnish language. In *Proceedings of the Corpus Linguistics 2005 conference, Birmingham, UK*. 457–464.
[50] John B Lowe. 1997. A frame-semantic approach to semantic annotation. In *In Proceedings of the SIGLEX workshop "Tagging Text with Lexical Semantics, (SIGLEX'97), Washington D.C., USA*. 18–24.
[51] David M Markowitz and Jeffrey T Hancock. 2014. Linguistic traces of a scientific fraud: The case of Diederik Stapel. *PloS one* 9, 8 (2014), e105937.
[52] Tom McArthur. 1986. *Longman lexicon of contemporary English*. Longman.
[53] Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Computational Linguistics* 33, 4 (2007), 553–590.
[54] Mary L McHugh. 2012. Interrater reliability: the Kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.
[55] G.A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K.J. Miller. 1990. WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3, 4 (1990), 235–312.
[56] Raymond J Mooney. 1996. Comparative experiments on disambiguating word senses: An illustration of the role of bias in machine learning. *arXiv preprint cmp-lg/9612001* (1996).
[57] Olga Mudraya, Bogdan Babych, Scott Piao, Paul Rayson, and Andrew Wilson. 2006. Developing a Russian semantic tagger for automatic semantic annotation. *Corpus Linguistics 2006* (2006), 290–297.
[58] Smruthi Mukund, Rohini Srihari, and Erik Peterson. 2010. An information-extraction system for Urdu—A resource-poor language. *ACM Transactions on Asian Language Information Processing (TALIP)* 9, 4 (2010), 43.
[59] Asma Naseer and Sarmad Hussain. [n.d.]. Supervised Word Sense Disambiguation for Urdu Using Bayesian Classification.
[60] Roberto Navigli. 2009. Word sense disambiguation: a survey. *Comput. Surveys* 41, 2 (2009), 1–69.
[61] Roberto Navigli and Simone Paolo Ponzetto. 2012. BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence* 193 (2012), 217–250.
[62] Vincent Ng and Claire Cardie. 2003. Weakly supervised natural language learning without redundant views. In *In Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (NAACL'03), Edmonton, Canada*. 173–180.
[63] Zheng-Yu Niu, Dong-Hong Ji, and Chew Lim Tan. 2005. Word sense disambiguation using label propagation based semi-supervised learning. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05), Michigan, USA*. Association for Computational Linguistics, 395–402.
[64] Kieran O'Halloran. 2011. Limitations of the logico-rhetorical module: Inconsistency in argument, online discussion forums and Electronic Deconstruction. *Discourse Studies* 13, 6 (2011), 797–806.
[65] Beng Yeow Vincent Ooi, Kok Wan Peter Tan, and Kok Leong Andy Chiang. 2007. Analyzing personal weblogs in Singapore English: The Wmatrix approach. (2007).
[66] Lluís Padró and Evgeny Stanilovsky. 2012. Freeling 3.0: Towards wider multilinguality. In *LREC2012*.
[67] Maria Pelevina, Nikolay Arefyev, Chris Biemann, and Alexander Panchenko. 2017. Making sense of word embeddings. In *Proceedings of the 1st Workshop on Representation Learning for NLP (RepL4NLP'16), Berlin, Germany*. Association for Computational Linguistics (ACL), 174–183.

[68] Scott Piao, Francesca Bianchi, Carmen Dayrell, Angela D'egidio, and Paul Rayson. 2015. Development of the multilingual semantic annotation system. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL-HLT'15), Denver, Colorado, USA*. Association for Computational Linguistics (ACL), 1268–1274.

[69] Scott Piao, Fraser Dallachy, Alistair Baron, Jane Demmen, Steve Wattam, Philip Durkin, James McCracken, Paul Rayson, and Marc Alexander. 2017. A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation. *Computer Speech & Language* 46, 2017 (2017), 113–135.

[70] Scott Songlin Piao, Paul Rayson, Dawn Archer, and Tony McEnery. 2005. Comparing and combining a semantic tagger and a statistical tool for MWE extraction. *Computer Speech & Language* 19, 4 (2005), 378–397.

[71] Scott SL Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18 (SIGLEX'03), Sapporo, Japan*. Association for Computational Linguistics (ACL), 49–56.

[72] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. 2003. KIM–semantic annotation platform. In *Proceedings of the 2nd International Semantic Web Conference (ISWC'03), Sanibel Island, FL, USA*. Springer, 834–849.

[73] Amanda Potts and Paul Baker. 2012. Does semantic tagging identify cultural change in British and American English? *International journal of corpus linguistics* 17, 3 (2012), 295–324.

[74] Paul Procter. 1978. Longman dictionary of contemporary English.

[75] Awais Rashid, Philip Greenwood, James Walkerdine, Alistair Baron, and Paul Rayson. 2012. Technological solutions to offending. In *Understanding and preventing online sexual exploitation of children*. Routledge, 244–259.

[76] Paul Rayson, Dawn Archer, Scott Piao, and AM McEnery. 2004. The UCREL Semantic Analysis System. In *proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC'04), Lisbon, Portugal*. European Language Resources Association (ELRA), 7–12.

[77] Paul Rayson, Luke Emmet, Roger Garside, and Pete Sawyer. 2000. The REVERE Project: Experiments with the application of probabilistic NLP to Systems Engineering. In *Proceedings of the 5th International Conference on Applications of Natural Language to Information Systems (NLDB'00), Versailles, France*. Springer, 288–300.

[78] Paul Rayson, Roger Garside, and Pete Sawyer. 1999. Recovering legacy requirements. In *Proceedings of the of the 5th International Workshop on Requirements Engineering: Foundations of Software Quality (REFSQ'99), Heidelberg, Germany*. Foundation for Software Quality, 49–54.

[79] Jesse Read and Jaako Hollmen. 2014. A Deep Interpretation of Classifier Chains. In *Advances in Intelligent Data Analysis XIII - 13th International Symposium, IDA 2014, Leuven, Belgium*. 251–262.

[80] Jesse Read, Luca Martino, and David Luengo. 2014. Efficient monte carlo methods for multi-dimensional learning with classifier chains. *Pattern Recognition* 47, 3 (2014), 1535–1546.

[81] Jesse Read, Bernhard Pfahringer, and Geoff Holmes. 2008. Multi-label classification using ensembles of pruned sets. In *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM'08), Pisa, Italy*. IEEE, 995–1000.

[82] Jesse Read, Bernhard Pfahringer, Geoff Holmes, and Eibe Frank. 2011. Classifier chains for multi-label classification. *Machine learning* 85, 3 (2011), 333–359.

[83] Jesse Read, Peter Reutemann, Bernhard Pfahringer, and Geoff Holmes. 2016. MEKA: a multi-label/multi-target extension to WEKA. *The Journal of Machine Learning Research* 17, 1 (2016), 667–671.

[84] Ronald L Rivest. 1987. Learning decision lists. *Machine learning* 2, 3 (1987), 229–246.

[85] Giuseppe Rizzo and Raphaël Troncy. 2012. NERD: a framework for unifying Named Entity Recognition and Disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12), Avignon, France*. Association for Computational Linguistics (ACL), 73–76.

[86] Peter Mark Roget. 2008. *Roget'S International Thesaurus, 3/E*. Oxford and IBH Publishing.

[87] Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2018. A word sense disambiguation corpus for Urdu. *Language Resources and Evaluation* (2018), 1–22.

[88] Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019. A Sense Annotated Corpus for All-Words Urdu Word Sense Disambiguation. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 18, 4 (2019), 1–14.

[89] Nathan Schneider, Dirk Hovy, Anders Johannsen, and Marine Carpuat. 2016. Semeval-2016 task 10: Detecting minimal semantic units and their meanings (dimsum). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16), San Diego, California*. Association for Computational Linguistics (ACL), 546–559.

[90] Nathan Schneider, Spencer Onuffer, Nora Kazour, Emily Danchik, Michael T Mordowanec, Henrietta Conrad, and Noah A Smith. 2014. Comprehensive annotation of multiword expressions in a social Web corpus. (2014), 1–7.

[91] Nathan Schneider and Noah A Smith. 2015. A corpus and model integrating multiword expressions and supersenses. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

(NAACL-HLT'15), Denver, Colorado. Association for Computational Linguistics (ACL), 1537–1547.

[92] Piao Scott, Rayson Paul, Archer Dawn, Bianchi Francesca, Dayrell Carmen, El-Haj Mahmoud, Jiménez Ricardo-María, Knight Dawn, Křen Michal, Löfberg Laura, Adeel Nawab Rao Muhammad, Shafi Jawad, Lee Teh Phoey, and Mudraya Olga. 2016. Lexical coverage evaluation of large-scale multilingual semantic lexicons for twelve languages. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16), Portorož, Slovenia*. European Language Resources Association (ELRA), 2614–2619.

[93] Elena Semino, Zsófia Demjén, Jane Demmen, Veronika Koller, Sheila Payne, Andrew Hardie, and Paul Rayson. 2017. The online use of Violence and Journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study. *BMJ supportive & palliative care* 7, 1 (2017), 60–66.

[94] Jawad Shafi. 2019. *An Urdu semantic tagger– lexicons, corpora, methods and tools*. Ph. D. Dissertation. The UCREL and Data Science Group, School of Computing and Communication (Infolab21), Lancaster University, Lancaster, U.K.

[95] Jawad Shafi, Hafiz Rizwan Iqbal, Rao Muhammad Adeel Nawab, and Paul Rayson. 2022. UNLT: Urdu Natural Language Toolkit. *Natural Language Engineering* (2022), 1–36.

[96] Bayan Abu Shawar and Eric Atwell. 2003. Using dialogue corpora to train a chatbot. In *Proceedings of the Corpus Linguistics 2003 conference, Lancaster, UK*. Lancaster University, UK, 681–690.

[97] William Simm, Maria-Angela Ferrario, Scott Piao, Jon Whittle, and Paul Rayson. 2010. Classification of short text comments by sentiment and actionability for voiceyourview. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*. IEEE, 552–557.

[98] Eleftherios Spyromitros, Grigorios Tsoumakas, and Ioannis Vlahavas. 2008. An empirical study of lazy multilabel classification algorithms. In *Proceedings of the 5th Hellenic conference on artificial intelligence (SETN'08), Berlin, Germany*. Springer, 401–406.

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

[100] Ahmed Tafseer, Saba Urooj, Sarmad Hussain, Asad Mustafa, Rahila Parveen, Farah Adeeba, Annette Hautli, and Miriam Butt. 2014. The CLE Urdu POS Tagset. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland*. European Language Resources Association (ELRA), 2920–2925.

[101] Caroline Tagg, Alistair Baron, and Paul Rayson. 2014. "i didn't spel that wrong did i. Oops": Analysis and normalisation of SMS spelling variation. *In L.-A. Cougnon & C. Fairon (Eds.), SMS Communication: A Linguistic Approach, Lingvisticæ Investigationes* 35, 2 (2014), 217–237.

[102] François Taïani, Paul Grace, Geoff Coulson, and Gordon Blair. 2008. Past and future of reflective middleware: Towards a corpus-based impact analysis. In *Proceedings of the 7th workshop on Reflective and adaptive middleware (Middleware'08), Leuven, Belgium*. ACM, 41–46.

[103] Konstantinos Trohidis, Grigorios Tsoumakas, George Kalliris, and Ioannis P Vlahavas. 2008. Multi-Label Classification of Music into Emotions.. In *Proceedings of the 9th International Conference of Music Information Retrieval (ISMIR'08), Philadelphia, USA*, Vol. 8. 325–330.

[104] George Tsatsaronis, Michalis Vazirgiannis, and Ion Androutsopoulos. 2007. Word Sense Disambiguation with Spreading Activation Networks Generated from Thesauri.. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07), Hyderabad, India*, Vol. 7. 1725–1730.

[105] Grigorios Tsoumakas and Ioannis Katakis. 2006. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining* 3, 3 (2006).

[106] Saba Urooj, Sarmad Hussain, Farah Adeeba, Farhat Jabeen, and Rahila Parveen. 2012. The CLE Urdu digest corpus. In *Proceedings of the Conference on Language & Technology, (CLT'12), Lahore, Pakistan*. Centre for Language Engineering Al-Khawarizmi Institute of Computer Science University of Engineering and Technology, 47–53.

[107] Saba Urooj, Sana Shams, Sarmad Hussain, and Farah Adeeba. 2014. Sense Tagged CLE Urdu Digest Corpus. In *Proceedings of the Conference on Language and Technology (CLT'14), Karachi, Pakistan*. Centre for Language Engineering (CLE), 1–8.

[108] Celine Vens, Jan Struyf, Leander Schietgat, Sašo Džeroski, and Hendrik Blockeel. 2008. Decision trees for hierarchical multi-label classification. *Machine Learning* 73, 2 (2008), 185–214.

[109] Piek Vossen. 1998. *A multilingual database with lexical semantic networks*. Springer.

[110] 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'13), Hissar, Bulgaria*. 710–718.

[111] Willem Waegeman, Krzysztof Dembczyński, and Eyke Hüllermeier. 2018. Multi-target prediction: a unifying view on problems and methods. *Data Mining and Knowledge Discovery* (2018), 1–32.

[112] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13), Washington, USA*. ACL-SIGDAT (the Association for Computational Linguistics special interest Group for linguistic data and corpus-based approaches to natural language processing), 133–137.

[113] Yorick Wilks. 1973. *Preference semantics*. Technical Report. STANFORD UNIV CA DEPT OF COMPUTER SCIENCE.

[114] Andrew Wilson and Paul Rayson. 1993. Automatic content analysis of spoken discourse: a report on work in progress. *Corpus based computational linguistics* (1993), 215–226.
[115] David Yarowsky. 1993. *One sense per collocation.* Technical Report. Pennsylvania Univ Philadelphia Dept. of Computer and Information Science.
[116] Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING'16), Osaka, Japan.* Association of Natural Language Processing (ANLP), 1374–1385.
[117] Zeng Zeng, Nanying Liang, Xulei Yang, and Steven Hoi. 2018. Multi-target deep neural networks: Theoretical analysis and implementation. *Neurocomputing* 273 (2018), 634–642.
[118] Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering* 26, 8 (2014), 1819–1837.