SindhUniv. Res. Jour. (Sci. Ser.) Vol.50 (002) 265-268 (2018)

http://doi.org/10.26692/sujo/2018.06.0046

SINDH UNIVERSITY RESEARCH JOURNAL (SCIENCE SERIES)

Sindhi-English Bilingual Parallel Ontological Dictionary

M. K. RATTAR, M. A. RATTAR*, M. M. RIND**, M. HYDER***, W. KHAN***

Departmentof Software Engineering, Mehran University of Engineering and Technology, Jamshoro

Received 01st July 2017 and Revised 17th April 2018

Abstract: Background: Since 2001 Sindhi language has been part of the computational world with the advent of Abdul Majid Bhurgiri's Fonts technology.It is evolving technologically and becoming part of various computational areas, which includes Word Processing, Optical Character Recognition and Natural Language Processing.**Aims and Objectives:**Primarily this research enables Sindhi language to be computationally platform independent. In this regard this research takes an initiative to identify methods to develop Sindhi-English bilingual parallel Ontology to be used as bilingual dictionary.

METHODOLOGY: This research proposes a methodology to exploit two most common terminological resources i.e. **Lexicon** and **Taxonomy**. This system maps available taxonomies from one language(English) or creates the**Taxonomy** of other language(Sindhi) manually. More than a hundred different digital dictionaries and thesaurieavailable online as traditional (**lexical**) terminological sourcesin Sindhi-English language pair, these sources were accessed and analyzed to create alternative (**semantic taxonomic**) source (**Ontology**).

Findings: Mapping **lexicons** to **taxonomies**cannot be done directly, since these are fundamentally two different terminological sources. This research provides an alternative through parallel Ontologies. Ontologies are Semantic Taxonomies and proved to be universal terminological source. To the best of our knowledge this is going to be the novel attempt in the field of Semantic Web Technologies for Sindhi-English language pair.

Application: This translation system can be used to displaymulti-lingual search results, multi-lingual Education and to develop real time translation of application's user interface.

Keywords: Bi-lingual, Ontological, Dictionary, Taxonomy, Lexicon, Semantic Web.

1. INTRODUCTION

Dictionaries are lexicographic resources of language, which compares two different but equally important groups of information. In fact English language has been world's most dominant languages at least from last two centuries[1], Sindhi-English language pair has plenty of the translatedmaterial in literaturewith prominent work of computing. For a Sindhi-English pair of languages a good number of dictionaries are available in different platforms, including Android, Windows and web based.

Taxonomy of concepts uses semantics termed as Ontology[2]. Ontologies of different domain are being developed and integrated day.By day. The current version of web which is human readable (is being shifted to be machine readable), and named as semantic web[3]. Primarily Ontologies are language-neutral terminological resource for Machine Translation that makes Ontologies a universal resource for adding semantics to web with multilingual support. It is not only the web which demands language neutrality, but there are many other resources of language communication that can use this technology.

The concept of parallel and multi-lingual Ontologies is emerging as a new trend of bilingual language translation. The traditional way of bilingual language translation is to use dictionaries that have been used for educational and other related purposes successfully[4]. This research aims to map Lexicographic resources, i.e. Dictionaries to Taxonomical resources of corresponding languages[5] (Sindhi and English). Sindhi language which is now part of computational world, has lexicographic resources based on Unicode based dictionaries and Word Net, but its lakes at Semantic Lexicographic resource i.e. Ontology. English language which is computationally a rich one and have both of the above resources at a good level of accuracy. The actual challenge of this research is not only the mapping of two languages but also to produce a second resource for Sindhi language to provide some methods for mapping.

The **novelty of this research** is i) the attempt for Sindhi language which is one of the oldest and richest languages of Indo-Pak region to be the part of semantic web technologies ii) Transformation of Lexical source of information in English or Sindhi inthe bilingual Taxonomic source.

Although Sindhi language based work in linguistic computing is not rich as compared to English language but it is not ignorable too[6]. Normally a standard

++engrmaqsooda@hotmail.com, engr.arifrattar@live.com mansoor.hyder@sau.edu.pk, engineermalook@gmail.com, wafa_Laghari@hotmail.com
*Department of Computer System Engineering, Quid-e-Awam University of Engineering Science and Technology, Nawabshah
**Department of Computer Science, Sindh Madrasat-u-Islam, Karachi
***Information Technology Centre, Sindh Agriculture University Tandojam

Dictionary contains synonyms, antonyms, part of speech information and example sentences to recognize the contextual meaning of words. A number of Sindhi-English bilingual digital dictionaries are available with the required information mentioned above. It's easy for humans to read textual and visual information available in digital dictionaries, however, having a bilingual digital dictionary is not enough for associating links between lexicographic and taxonomical resources computationally.

Histories for acquiring semantic information of words or phrases natural language processing techniques are used. A remarkable work donefor Sindhi Word Segmentation Model, that helps to tokenize individual words by recognizing from a collection of text and then validating against a pre-built terminological resource[7, 8].This model was used for spell and grammar checking and other natural language processing operations.A properly collected and maintained text corpus is necessary for morphological, syntactic and semantic analysis, information retrieval and extraction and machine translation. Later a corpus construction modelwas proposed, which addresses the issues, including corpus acquisition, pre-processing, and tokenization[9]. Preliminary results and observations are also included for letter unigram, bigram and trigram frequencies in which 368 word Net structures were developed especially for non-diacritic words of Sindhi language. This model collectsa group of analogical words of one specific word by using rule based semantic POS(Part of Speech) Tagging.

Traditionword Net is a lexical database used for natural language processing applications .Computationally English language work has been evolved enough to use word Net as one of the automatic or semi-automatic Ontology construction resource[10].

2. METHODOLOGY

2.1. Lexical Source

To start transforming word meaning from one language to another, the very fundamental source can be a lexicon irrespective of its size. The size of the lexicon is only important whenit concerns accuracy, but at this stage only successful translation is demanded. Mapping structure of meaning of words shown below in **(Table 1).** Any word from Sindhi languagesw_ican have words of similar meaning in Sindhi language as well. To associate Sindhi meaning of each word$sw_i$is labeled as sw_{i-01}, sw_{i-02},sw_{i-k}, $sw_{i-n}$and furthermore English meaning to each Sindhi meaningssw_{i-01}, sw_{i-02}, ….., sw_{i-k}, $sw_{i-n}$is labeled as ew_{01-1}, ew_{01-2} …., ew_{02-1}, ew_{02-2} …….. ew_{k-1}, ew_{k-2}, …… ew_{n-1}, ew_{n-2}, …….

Table 1. Meaning structure of Lexical source

Sindhi Word	Sindhi Meanings	English Meanings of each Sindhi Meaning
sw_i	sw_{i-01}	ew_{01-1}, ew_{01-2} ….
	sw_{i-02}	ew_{02-1}, ew_{02-2} ……
	……..	……..
	sw_{i-k}	ew_{k-1}, ew_{k-2}, ……
	sw_{i-n}	ew_{n-1}, ew_{n-2}, ……

Words of similar meaning called synonyms and set of synonyms is simply named as synset in word Net terminology[11]. Synset is not the only concern of this structure; along with synset other properties of words will be extracted in a similar way and kept labeledrespectively.

2.2. Taxonomic Source

Taxonomies are set of concepts logically associates other concepts, Ontologies in this regard are a great source of extracting information and transformed into knowledge by using wisdom in terms of logic.Ontology can be classified into terminological ontology, information ontology and knowledge ontology[12]. Basically Ontologies can be represented as taxonomic trees of related terms, but practically it is more complicated. The representational primitives of ontology often include the information about the meaning and constraints on the logical consistent concepts.

Before constructing a bilingual Ontology, Ontology of each language is developed separately. Ontologies provide language-neutrality, by this feature mapping of the multilingual Ontology pair is becomingconvenient as illustrated in **(Fig.1).**

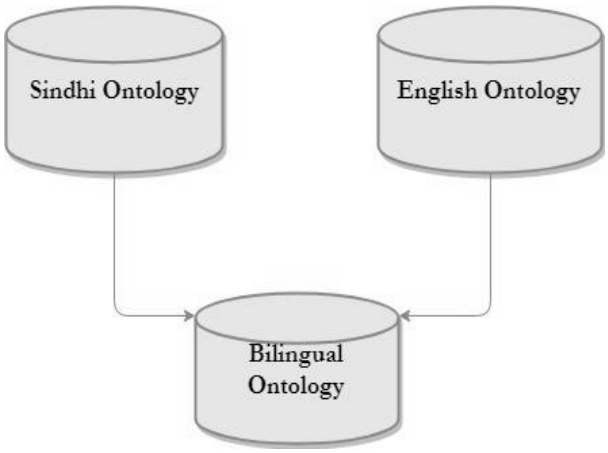

Fig. 1. Structure of Taxonomic source

2.3. *Mapping Of Concepts*

On the primitive stage this research presents and tests only three types of mapping, from lexical sources of taxonomic sources. i) Mapping synonyms ii) Mapping antonyms ii) Mapping other morphological properties, e.g. part of speech information.

2.3.1. Mapping synonyms

Synonyms from lexical sources are easy to map for taxonomical source as property 'equivalent'. Terms equivalent in nature are simply synonyms, by enabling this association almost every property or constrain of a term from one language will be considered true for the term from the other corresponding language.

2.3.2. Mapping antonyms

Terms labeled as antonyms are words of opposite meaning. The concept having antonym property true for any other term will be mapped as antonyms of synonyms for that word respectively. The property mutually exclusive specifies set of concepts that cannot overlap each other's boundary of properties and constraints. The set of synonyms for each concept must be made mutually exclusive to set of antonyms of the same concept.

2.3.3. Mapping other morphological properties

Properties like part of speech are made simple by adding annotations in the Ontology. Annotations are like comments in other computational platforms that contain some extra information related to any concept in Ontologies. Data and Object properties in Ontologies helps to add constrains etc.

3. RESEARCH ASSUMPTIONS

Due to the novelty this research is in its preliminary phase, which of courseencompasses some of the limitations.

i) The transformation of two (Lexical and Taxonomy) terminological sources for this research entirely done by hand.

ii) This transformation involves only noun phrases because Noun phrases are very easy to translate from English to Sindhi.

4. RESULTS AND DISCUSSION

This research proposes a methodology for creating a terminological resource named as BI-LINGUAL PARALLEL ONTOLOGICAL DICTIONARY in short BPO on pair of languages i.e. Sindhi and English. It takes traditionally used lexicon based terminology resources as input and create a taxonomic resources based on characteristics from lexicon, then embed and transform the sources. The transformed new source BPO tended to be lexical in nature which benefits it from the latest technologies of the semantic web. For understanding the effects of transformation from lexical sources of terminology to taxonomic source i.e. Ontology, some of the measures has been taken into account.

More than a 1000 Noun, Verb and Adjective phrases have been taken into consideration. Consideringthe methodology proposed above, it has been found that lexicons of single to single association can be transformed into its equivalent texonomic structure efficiently. Mapping different parts of taxonomy to proper Ontology takes place on behlf of following constraints:

i- Noun Phrases are suitable to translate as ***Name of Entities*** in class hierarchy

ii- Verb and Adjective can collectively be categorized through **DataProperty** and **ObjectProperty** parts of Ontology.

iii- Other characteristic like synonymes, antonyms can be associated using various types of object properties associate Ontology (Equivalent Object, Disjoint Object Properties, Inverse Object Propertiesetc...).

Accuracy of translation found between 90 and 100% , but it should not ignore the fact that given approach is manual. As much it is concerned to moves from manual to automated process of translation using this methodology, more efforts shall be needed.

REFERENCES:

Adult and youth literacy, (1990-2015). Institute for Statistics; Montreal UNISCO, 2012.

Bhatti Z., I. A. Ismaili, SW. J. oomro, D. N. Hakro (2014). **Word segmentation model for Sindhi text. *American Journal of Computing Research Repository*.** 2(1), 1-7.

Cimermanová I. (2012) Corpus vs dictionary in EFL classes. English Matters III. 65-73.
Article title. http://www.indjst.org/index.php/vision. Date accessed: 01/01/2015.

Dootio M A. and A. I. Wagan (2017) Syntactic parsing and supervised analysis of Sindhi text. *Journal of King Saud University-Computer and Information Sciences.*

Enabling Pakistani Languages through Unicode.http://download.microsoft.com/download/1/4/2/142aef9f-1a74-4a24-b1f4-782d48d41a6d/PakLang.pdf. Date accessed: 01/04/2018.

Feigenbaum L, I. Herman, T. Hongsermeier, F. E. Neumann, S. Stephens(2007). The Semantic Web in Action. *Scientific American*. 297(6), 90-97.

Gonzalo J., F. Verdejo I. Chugur, J. Cigarran 1998 Indexing with WordNet synsets can improve text retrieval.arXiv preprint cmp-lg/9808002.. Aug 5.

Liu L. (2009) Encyclopedia of database systems. New York, NY: Springer;.

Okumura A, E. Hovy (1994) Lexicon-to-ontology concept association using a bilingual dictionary. *First Conference of the Association for Machine Translation in the Americans*, 177-184.

Rahman M. U. (2010) Towards Sindhi corpus construction.*Conference on Language and Technology, Lahore, Pakistan*..

Schierholz S. (2015) Methods in Lexicography and Dictionary Research. Lexikos. 323- 352.

Wang, H. I. (2009). The framework of an ontology based bilingual terminology system on college campus of Taiwan-the terminological theory approach. *InPervasive Computing (JCPC), Joint Conferences, IEEE*. 2009 Dec 3, 737-742