22

DEVELOPING A TRILINGUAL (ENGLISH-HINDI-KASHMIRI) E-DICTIONARY: ISSUES AND SOLUTIONS[i]

Tariq Ahmad Banday
Feroz Ahmad Lone
Kaiser Ahmad Malik
Oveesa Panzoo
Shajoo Nazir
Shagufta Rasool
Sheeba Maqsood

INTRODUCTION

Dictionary describes the meaning of words, often illustrating how they are pronounced and used in context. Modern dictionaries often include information about spelling, etymology, usage, synonyms and grammar, and some may include illustrations as well. In many languages, words can appear in many different forms, but only the undeclined or unconjugated form appears as the head word in most dictionaries, rather we can say that the words are looked at from a lexeme perspective. Dictionaries can vary widely in coverage, size and scope. A maximizing dictionary lists as many words as possible from a particular speech community, whereas minimizing dictionary exclusively attempts to cover only a limited selection of words from a speech community. Corpus-based dictionary is to provide learners with relevant, idiomatic and useful information that will help them setting up native-like links between words and meanings. In a corpus based dictionary lexicographers are keen to include corpus information about lexico-semantic relations such as synonyms, antonyms, hyponyms and super ordinates.

In any natural language application, dictionary look-up plays a vital role. Natural language is inherently ambiguous. A word can have multiple meanings (or senses). Given an occurrence of a word *w* in a natural language text, task of Word Sense Disambiguation (WSD) is to determine the correct sense of word in that context. WSD is a fundamental and central open problem of Natural Language Processing (NLP).

Highly ambiguous words pose continuing problems for NLP applications. They can lead to irrelevant document retrieval translations in Machine Translation systems (Palmer et al., 2000). Lexical ambiguity is syntactic or semantic. A word's syntactic ambiguity can be resolved by applying part-of-speech taggers which predict the syntactic category of a word in texts with high levels of accuracy (Brill, 1995; Brants, 2000). The problem of resolving semantic ambiguity, which is generally known as WSD, has proved to be more difficult than syntactic disambiguation.

The only way to determine the meaning of a word in a particular usage is to examine its context. One could envisage building a WSD system using handcrafted rules or knowledge obtained from linguists. Such an approach would be highly labor-intensive, with questionable scalability. Another approach involves the use of dictionary or thesaurus to perform WSD. There are also three ways to approach WSD: a knowledge-based approach, which uses an explicit lexicon, corpus-based disambiguation, where the relevant information about word senses is gathered from training on a large corpus, or, third alternative, a hybrid approach combining aspects of aforementioned methodologies (Ide and Veronis, 1998). There are different types of dictionaries viz simple, sense based, synset based and so on, apart from being monolingual or multilingual which can be used for WSD. Here the main emphasis will be on sense based dictionary as the present paper focuses on developing a maximized sense- based English-Hindi-Kashmiri e-dictionary. Sense based dictionaries are actually concept based. A sense based dictionary includes all the possible senses that a word can have depending upon the function of a particular word in a particular context. Consider the word *pen* it has five different senses viz:

Pen — a writing implement with a point from which ink flows.

Pen — an enclosure for confining livestock.

Playpen, pen — a portable enclosure in which babies may be left to play.

Penitentiary, pen — a correctional institution for those convicted of major crimes.

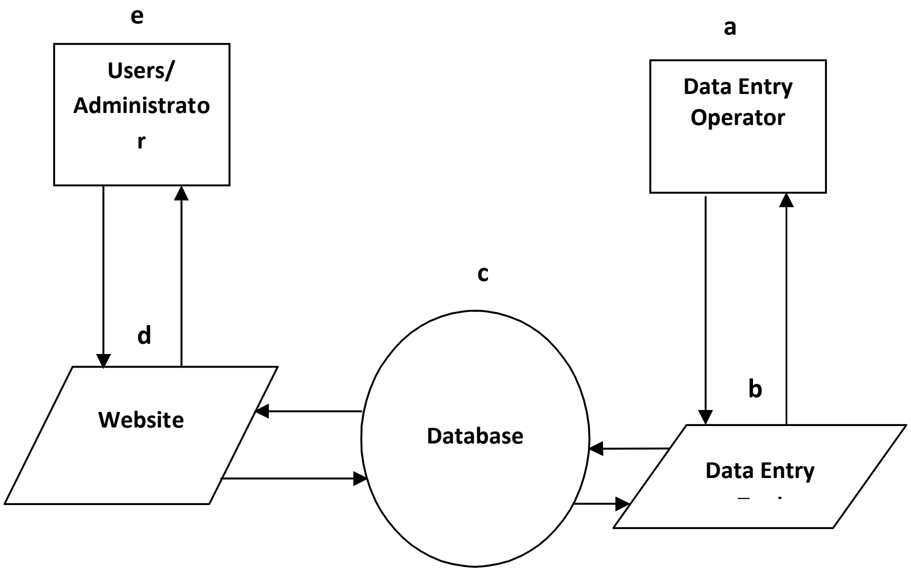
Pen — female swan.

In this manner all the entries for a word *pen* will be included in a sense based dictionary. Uses of sense based dictionaries have been very helpful in achieving consistent levels of accuracy on a variety of word types and ambiguities.

THE E-DICTIONARY

The project, which forms the main theme of the paper, revolves around the building of a trilingual English-Hindi-Kashmiri sense-based online dictionary. This dictionary will incorporate three scripts; the Roman script for English, the Devanagari script for Hindi and the modified Persio-Arabic script for Kashmiri, and from a computational perspective this means the use of three fonts viz: 'Arial' for English, 'Mangal' for Hindi and 'Afan Koshur' for Kashmiri. Each entry or lexeme will display its varied senses and these senses have their counterparts in the other two languages.

Below given conceptual diagram represents the overall process of this e-Dictionary:



a) DATA ENTRY OPERATOR

Performs the entries of the e-dictionary to the database through data entry tool. The entries are:

- *Insertion of new records.*
- *Modifying the existing records.*
- *Deleting the entries.*
- *Viewing the records.*

b) **DATA ENTRY TOOL**

The tool acts as an interface between the data entry operator and database. It is developed in java to perform all the operations and takes the input in three languages viz. English, Hindi and Kashmiri

c) **DATABASE**

The whole schema and data of the e-dictionary is stored in the database. All users, database administrator and data entry operators have access via data entry tool and website to the same database with different authentication. It is dynamic and saves all the events related to e-Dictionary.

d) **WEBSITE**

Interface developed for users so that they can use and share the knowledge of e-Dictionary. It acts as an intermediary between the user/administrator and database. Users can share their knowledge through feedback pages of the website, which can make the dictionary better.

e) **ADMINISTRATOR/USER**

The user of the dictionary is the person who uses or views this e-Dictionary; furthermore a feedback is also expected from the user. The feedback is then checked by an authorized person called the Administrator, after that, if the feedback or input regarding to a particular entry is accepted then the input will be reflected to the database.

The dictionary procedure involves the following main tasks:

a. Database creation standards and priorities.
b. Development of tool for entering and modifying records.
c. Input provisions.
d. Website

DATABASE CREATION STANDARDS AND PRIORITIES

The main sources of the data included the corpus collected by the Kashmiri Language Technology Development team at the Department of Linguistics, some free-source available dictionaries, etc. The database was created on the basis of required fields like head_word, lex_cat, sense, meaning_hindi, meaning_target, english_example.

SCHEMA OF THE DATABASE IS

Name	**Type**
Dic_id	Number (5)
Head_Word	Varchar2 (30)
Lex_cat	Varchar2(15)
Sense	Number (2)
Meaning_hindi	Nvarchar2(780)
Meaning_target	Nvarchar2(780)
Eng_example	Varchar2(200)

The character set of the database was changed so that it could support Hindi and Kashmiri words which are Unicode (UTF-8) based characters.

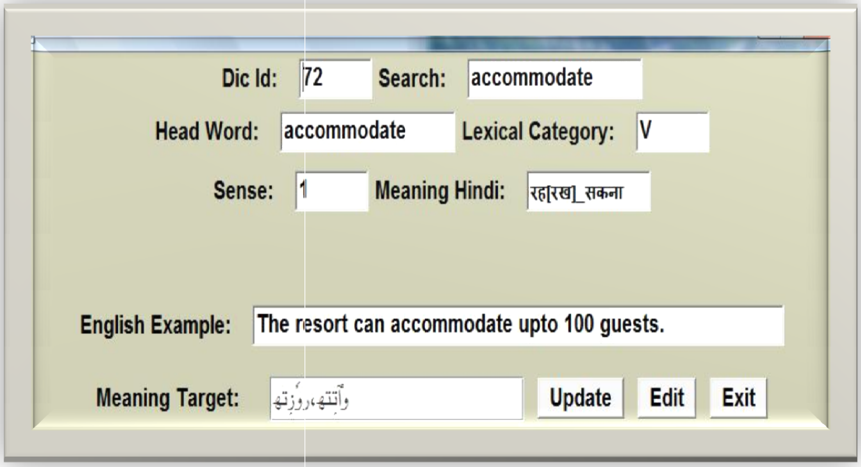
Each entry of the dictionary has a dictionary-id as a Primary key. Each head word has different senses which are differentiated by the dictionary-id. The database accepts more than one meaning in Kashmiri as well as in Hindi for a single entry. The

Database is designed in such a manner that it can be further used for other applications like Spell Checker, and other different types of dictionaries (minimized, multilingual etc).

Initially the database was designed in Oracle 10g but considering the usage of memory and access time it will be converted into MySql.

DEVELOPMENT OF TOOL FOR ENTERING AND MODIFYING RECORDS

To make the required modifications like insertion of equivalent Kashmiri words, altering existing records, deleting some records or inserting some new entries, a tool was developed in Java. The snapshot of the data entry tool is given below:



Dic Id: 72 Search: accommodate

Head Word: accommodate Lexical Category: V

Sense: 1 Meaning Hindi: रह[रख]_सकना

English Example: The resort can accommodate upto 100 guests.

Meaning Target: وآتتھ،روزتھ

There are eight input boxes in the data entry tool each followed by a label. Label defines the input type of each input box, where the first one “Dic_id” is the unique id for each entry and it is not editable. The “Search” is used for searching a particular word with all corresponding entries, “Head Word” as its name depicts, describes the head word of the dictionary. Similarly there are other input boxes with different attributes like ”Lexical category”, “Sense”, “Meaning Hindi”, “English Example” and “Meaning Target”. The task or working of these fields is depicted by their names. Sense is the sense-id of a head word i.e. if the head word can be repeated with same or different category, a different sense-id is allocated. “Meaning Hindi” and “Meaning target" takes the input in Hindi and Kashmiri respectively. There are also three buttons in the tool viz. Update, Edit and Exit, where after modification or insertion the update button is pressed to reflect the changes to the database. Edit button is used to modify a particular record. The data entry operator can navigate between the records by pressing the UP or DOWN Key of the keyboard. The "Exit" button exits from the application. There are also short cuts for different operations like pressing Esc key for exit.

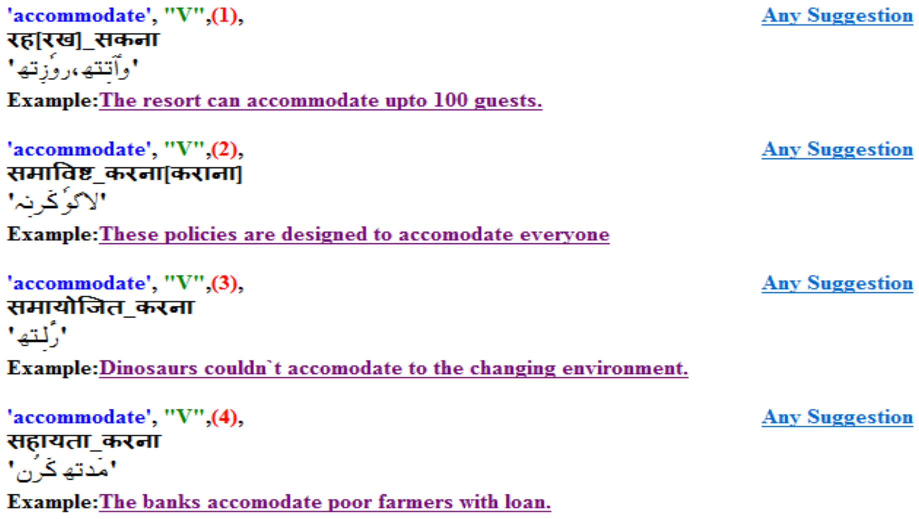
INPUT PROVISIONS

The dictionary has three types of inputs viz: English, Hindi and Kashmiri. But Hindi and Kashmiri are stored in the database as Unicode values therefore the process of conversion takes place from character to Unicode while inserting the data into the database and similarly conversion from Unicode to character while retrieving the data from the database. For this conversion processes Java function was written. Input languages were automatically changed between English, Hindi and Kashmiri while typing the data.

WEBSITE

Finally the website will be developed for users to get access to the e-dictionary. The website will be developed by using different web technologies. The proposed website will use three languages viz. English, Hindi and Kashmiri as discussed above, the problem is how to display and store the Kashmiri language as there is no Unicode based Kashmiri font developed so far. And this issue is resolved by developing a Unicode based font for Kashmiri. Now again the issue is how to display Kashmiri language on webpage because the font developed is not released yet. This issue will be resolved by embedding fonts in the web page.

User will search the meaning of words in any of the three languages. As it is sense



'accommodate', "V",(1),
रह[रख]_सकना
'وآتنتھ،روزِتھ'
Example:The resort can accommodate upto 100 guests.

'accommodate', "V",(2),
समाविष्ट_करना[कराना]
'لاگوٚ کَرِنہ'
Example:These policies are designed to accomodate everyone

'accommodate', "V",(3),
समायोजित_करना
'رٔلِتھ'
Example:Dinosaurs couldn`t accomodate to the changing environment.

'accommodate', "V",(4),
सहायता_करना
'مَدتھ کَرُن'
Example:The banks accomodate poor farmers with loan.

based dictionary the user will get the outcome for head word “accommodate” as:

It will display all the entries of a head word with different senses. Where first line of each entry contains the head word in single quotes’’, lexical category in double quotes ”” and sense-id in parenthesis ( ). Second and third line contains meaning in Hindi and meaning in Kashmiri respectively. Fourth line shows the English example.

Link “Any Suggestion” on right side of each entry, takes the user to the feedback page where the user can give feedback about the particular word. The user may have suggestions with the above examples in a number of ways:

a. Spellings in
   i. English
   ii. Hindi
   iii. Kashmiri
b. Grammatical categories in
   i. English
   ii. Hindi
   iii. Kashmiri
c. Example sentence
d. Need of more alternatives
e. Redundancy of some alternatives or no need of some alternatives
f. Adding a new entry.

It will be the first e-Dictionary which will be using Kashmiri.

CONCLUSION

The present paper discussed the overall making of an e-dictionary and some issues related to it but further work is required to extend this work. Once this set begins to grow in size, it should be possible to apply it in the analysis of the syntax/semantics interface. The dictionary can become a good resource for future research and language maintenance measures.

REFERENCES

Agirre, Eneko and Philip, Edmonds. (eds.). 2006. *Word Sense Disambiguation: Algorithms and Applications*. Dordrecht: Springer.

Bar-Hillel, Yehoshua. 1964. *Language and Information*. New York: Addison-Wesley.

Coward, David, F. and Charles, E. Grimes. 1995, 2000. *Making dictionaries: A Guide to Lexicography and the Multi-Dictionary Formatter.* Waxhaw: Summer Institute of Linguistics.

Edmonds, Philip and Adam, Kilgarriff. 2002. "Introduction to the Special Issue on Evaluating Word Sense Disambiguation Systems". *Journal of Natural Language Engineering*. 8(4):279-291.

Edmonds, Philip. 2005. "Lexical Disambiguation". Keith, Brown. (ed.). *The Elsevier Encyclopedia of Language and Linguistics*. (2nd ed.): 607-23. Oxford: Elsevier.

Herbert, Schildt. 2007.*Java*: *The Complete Reference,* 7th Ed. Delhi: Tata McGraw-Hill.

Ide, Nancy and Jean, Véronis. 1998. "Word Sense Disambiguation: The state of the art". *Computational Linguistics*. 24(1):1-40.

Jurafsky, Daniel and James, H. Martin. 2000. *Speech and Language Processing*. New Jersey, USA: Prentice Hall.

Lesk, Michael. 1986. "Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone". *Proceedings of SIGDOC-86: 5th International Conference on Systems Documentation*, Toronto, Canada. 24-26.

Manning, Christopher, D. and Hinrich, Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

Mihalcea, Rada. 2007. Word Sense Disambiguation. *Encyclopedia of Machine Learning*. Springer-Verlag.

<www.unicode.org>

<www.google.com>

<www.javaforum.com>

<www.oracle.com>

[i] The authors wish to acknowledge that this work has been financed by the Ministry of Communication and Information Technology, Govt. of India.