Interdisciplinary Journal of Linguistics
Volume [5] 2012, Pp.169-184

Challenges and Opportunities in Automatically Building Bilingual Lexicon from Web Corpus

Kiran Pala*
S. V. Ganagashetty*

Introduction

Bilingual lexicons or dictionaries are an essential and invaluable resource for any language, without which the use of that language for academic, official and commercial purposes becomes difficult, and so does learning of that language for those who are not native speakers. Furthermore, in this digital age the chances of that language being used on the digital media or of building computational applications for the language reduce drastically. Since creating bilingual lexicon manually is very time consuming and effort consuming as well, some researchers have been trying to do at least a part of this job automatically or semi-automatically. This has hardly been tried so far for Indian languages, and especially for Telugu.

From our experiences as practitioners of Natural Language Processing (NLP) which deals with such problems, we describe the challenges that need to be overcome to achieve the goal of automatization , atleast, a part of the process of building bilingual lexicon for Indian languages in general and for Telugu in particular. Since one of the largest and easily available resources nowadays for natural language data is the World Wide Web (WWW), we concentrate on using it as the source from which we can try to mine the information required to build bilingual lexicon. We also try to suggest some directions along which the possible solutions of this problem may be found. The problems are related to the nature of the scripts used for Indian languages, the lack of resources and computational tools for them, the lack of awareness as well as the lack of interest among the Indian (and the Telugu speaking) community in the linguistic and

* International Institute of Information Technology, Hyderabad, India.

computational issues involved (even among the NLP practitioners), the lack of demand for such resources and tools among those who can pay for them, and also the very fact that most educated Indians are hesitant to use their own language even for communication on the digital media, whereas, ironically, linguistic chauvinism seems to be on the rise at the same time in various parts of India.

For Dravidian languages like Telugu, the problem of automatically building lexicon has the added complication that these languages are morphologically rich, which leads to both linguistic and computational problems for the purpose being described here, although this richness has its positive aspects in other respects. We survey the computational methods that can be used to solve this problem and we also suggest some language specific techniques that could make task easier. One of these is using akshara (roughly, an orthographic syllable) as an important unit for text processing for Indian languages.

The Structure of Indian Languages

All languages belong to certain groups or families; this grouping is based on the origin and long historical proximity (Emeneau, 1956, 1980). Indian languages have two large families, i.e.; Indo-Aryan (Hindi, Punjabi, Marathi, Gujarati etc.), and Dravidian languages (Tamil, Kannada, Malayalam and Telugu),apart from Austro-Asiatic and Tibeto-Burman families. Brahmi is the origin of these language scripts. Basically, spellings of Indian languages are phonetic and the scripts are syllabic; alpha-syllabaries or abugidas-- these consist of symbols for consonants and vowels.

Word Structure

The consonants each have an inherent vowel which can be changed to another vowel or muted by means of diacritics. Words in Dravidian languages, especially in Telugu are long and complex, i.e., because they also have suffixation, words are built up from many affixes that combine with one another. Telugu has been the language of choice for lyrical compositions for its vowel ending words. For example the vocabulary of Telugu is highly Sanskritized in addition to having the Persian-Arabic

borrowings కటురు /kaburu/ `story', జవాబు /javaabu/ `answer'; Urdu తరాజు /taraaju/ `balance'. It does have cognates from other Dravidan languages too such as పులి /puli/ `tiger', ఊరు /uuru/ `village'; తల /tala/ `head' (Rajendran, S. and Shivapratap, G. and Dhanlakshmi, V. and Soman, KP, 2010).

Agglutinative Morphology

Linguistic analysis of words is concerned with retrieving the structure—more specifically, the syntactic and morphological properties of a complex word (Caldwell, R, 1875). Basically, Dravidian languages are agglutinative, highly inflectional and rich in morphology. The major inflectional categories are nouns and verbs. Nominal morphology of Telugu is simple as compared to verb morphology and it also allows polyagglutination (Caldwell, R, 1875). For instance, a single Telugu verb can take at least 200 forms without including the auxiliary information whereas a noun can inflect for only 8 cases. Extremely simple paradigms can be used to categorize the root words. As of now, the current implementation of these paradigms outputs all possible known legitimate splits.

Because of these structural features computational description and processing of Indian languages become a formidable problem. The problem of building bilingual lexicon is thus also multiplied given the nature of such difficulties and challenges. We can now move on to a description of the aspects of web that can be manipulated and maximized for use in building bilingual lexicon from web resources.

Web as a Resource

The Web corpus owes its popularity to its tremendous size, broad linguistic, geographical and social range, up-to-dateness, multimodality, and wide availability at a minimal cost. We shall consider each advantage briefly before discussing the limitations of the Web corpus.

Size of Data

The data on the web are constantly expanding, so its size is unknowable. In 2008 Google noted that it had identified (but not actually indexed) over a trillion (1012) distinct URLs (Web addresses), and that several billion (109) new webpages appear daily (Alpert & Hajaj, 2008).

Range of Data

The web data have increased the range towards World Wide. This keeps goes hand in hand with the availability of web space and easy accessibility which has made even lay people contribute their own linguistic experiences to the information communication society. Encompassing earlier situations, most of the virtual community around the world was within English-language users. Now the content of other world languages has increased their share in web space, but they have not lost their minority status online. Internationally-oriented web portals provide multilingual materials which can be an aid to translators and language learners. For instance, Wikipedia is one of the major contributors to the web data or corpus. It invites the public to make contributions and distributes the knowledge free of cost. Another contributor is social networking websites, these sites have redefined the notion of community and in addition, it has brought private and public interaction together in public view. As a consequence, what is quite visible is that the users in social networking websites have unintentionally supplied both linguistic and demographic data.

Data Up-to-datedness

The Web is the most effective resource for understanding the reflections of contemporary culture through the linguistic representations of the users. For instance, the new words easily coined through an interaction with various virtual communities and the use of those words if increased have a dramatic effect on the language; thus, the frequency of lexical items in different contexts can be an additional linguistic input to the language. Similarly borrowings also take place. And if that condition is not acceptable, it can be treated as a threat to the minor linguistic community and also to that language. Most of the search engines closely observe and

trace eventual developments in the web and incorporate those content topics in the emerging usage (REF: web as corpus).

Formats of Data

Web data have usually consisted of multimodality (sound, image, sight and text) and multimedia content. In general, the content should be in the form of machine-readable text or with linguistic annotations for the purpose of tracking specific paths. For instance, the content of some of the interviews or commercial broadcasting videos and audios do consist of transcriptions in different languages as subtitles. It can easily be accessible to the user or crawler. Some of the sites are even made available to the users in the transcriptions by means of the formation of audio. Some sites collect linguistic annotations of the images from the web users as captcha. Some sites offer games for language learning for the purpose of input verifications or error corrections in linguistic data. Social media and networking websites are examples of such activities.

Availability of Data

These days even in slow bandwidth networks it is possible to access the web content so easily. This was possible with constant efforts of technological developments. The recent technological developments in web browsers and content managements are brought by the machine-readable content and easy loadable formats. Internet access, in most of the countries, is fast and inexpensive. Also, the cost of hardware equipment has reduced to a large extent and the availability of linguistic input tools in local languages has extended to the local doors. This helps users upload news articles, blogs, magazines and also entire text books. A user with a home broadband connection can compile and process a multimillion word corpus in minutes (REF: web as corpus).

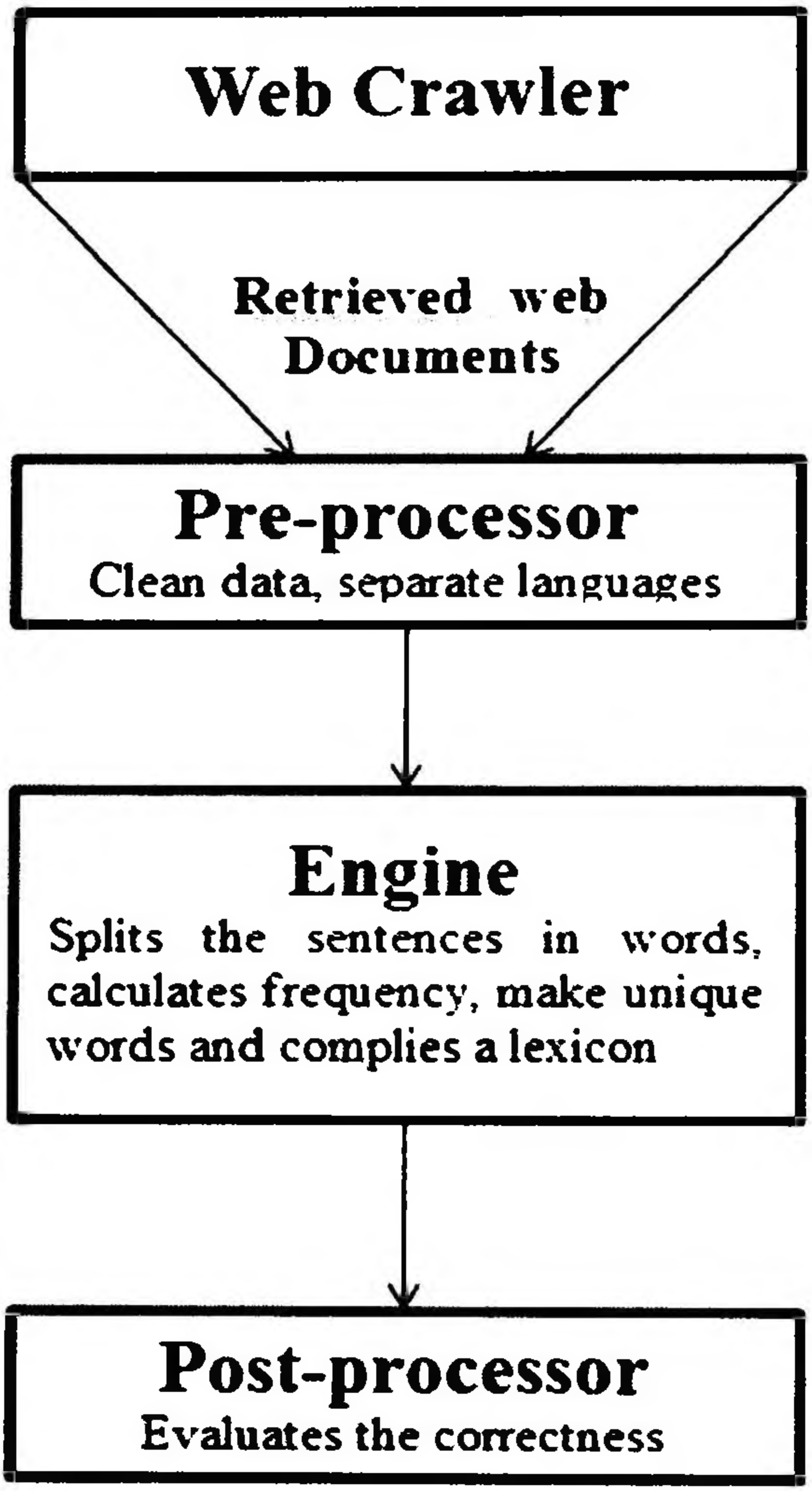
Stages in Data Collection from Web Corpus

Web content as a language resource has drawbacks and limitations too. What purpose is the text intended for? What target audience does it represent? Was it written by a native speaker, or is it translation? Is the content of the document represented in authoritative linguistic form, encoding and tagset? Does the content of the document contain any copyright?

Automatic development of a bilingual dictionary from the web corpus is a sequence of processes of language resource acquisition using a web crawler. In this sequence of processes, we need to setup a web crawler for the acquisition of language resource from the web archive. The web crawler retrieves the web documents and files. These documents are in Hyper Text Markup Language (HTML) format within a specified range. The crawling range has to be set up as input for the crawler; this range could be date, keywords etc. The web crawler generates the Universal Resource Locator (URL) address for the index (first) page of any particular date. The index page of a website contains the actual content and links to some other pages, say, advertisements, videos and images etc, for example. This kind of non-text documents does not contribute to the corpus generation. It is necessary to identify the text in those HTML documents and the rest of the files are not considered further. Building or creation of language resources from the web involves a great deal of pre-processing that includes cleaning, code conversion, language pair separation and annotation. The HTML files that contain news documents are identified by the web crawler and require cleaning for the extraction of text to be stored in the corpus along with the relevant details. HTML files consist of a set of tagged data that include different texts of languages. Every HTML file has to be scanned from the beginning to end of the document for tags like font FACE, encoding range and dynamic fonts, i.e, the text on pages generated only when the system is connected to web. The web archives generally uses graphemic coding, whereas orthographic coding is required for text processing tasks (Chaudary, S Pala, K et al, 2008).

Here what the engine does is split into words the sentences from the cleaned corpus using language specific tokens without any disturbance made to the rendered information of the text. The engine filters distinct words from the words thus split and calculates the frequency of them. The engine at later stages starts to compile the lexicon using a specific computational mapping technique. These techniques have been discussed in later sections. The complied lexicon needs verification to ensure the accuracy and reliability of the used techniques. The entire process takes

place in the post-processing stage. In this stage, the data will be evaluated with the help of language-specific native speakers or by building some collaborative as well as cooperative games.



There are a diverse range of algorithms and techniques used for the data collection from the web resources based on different criteria of search and matching techniques of varying complexity. Let's now look at them.

Existing Computational Methods

Compiling monolingual lexicons has been a problem the solution to which has been devised and for this there exist various well developed and accurate techniques. But in the case of bilingual lexicons, the problem is not of the same kind as what is found in the case of monolingual lexicon compilation. This is precisely because of the fact that the development of bilingual lexical resources deals with two different languages and with huge variations like differences in scripts, variations in phonetic system, morphological variations and socio-cultural variations involving borrowed words in the lexicon etc. These are all crucial factors that need to be taken care of while one aims at building and producing the semantic resemblance of the lexemes. Hence, here we discuss some existing techniques for automatic building bilingual lexicons.The existing methods in automatizing the process of building bilingual lexicon have used different types of corpus in that various types of corpus such as comparable corpus, parallel corpus, translated data (words and continuous text) etc. are often used.

Below we describe different types of computational techniques which have been developed recently on various languages and linguistic families in the world. These techniques have been applied on cleaned and manually annotated corpus, not crawled from web servers.

Search Methods

These methods are used in parallel and translational corpora to extract the bilingual lexicons. They are very standard search methods using lexicon building techniques through Transliteration Rule, measure matching, i.e., identifying distance between two words based on their orthographic and phonemic characteristics, one-to-one correspondence, introducing heuristics related to linguistic structure or language models (unigrams, bi-grams... n-grams). And calculating the frequency is involved in the main

stages in this search method. In this search method the techniques varies according to the structure of a language family. These methods are experimented on Abugida scripts (Indian languages) and Japanese Katakana to extract the translations from the target text (Tsuji, K. 2002).

Fuzzy Text Search Method

It is a method on the identification of the notion of surface similarity which (basically for Abugida scripts) can be roughly defined as combined orthographic and phonetic similarity. A method based on a measure of surface similarity can give better results (Singh, A.K, Surana, H. and Gali, K , 2007).

Root Matching Method

It is used on a news corpus to identify several orthographic word-level features; in this approach the data have been tagged with the parts-of-speech. (Asif E and Sivaji B, 2008)

Context Heterogeneity Similarity Measure

This method consists in capitalizing on the similarity between words and their translations in compiling bilingual lexicon entries from a non-parallel corpus, say, English-Chinese corpus. Current algorithms for bilingual lexicon compilation rely on occurrence frequencies, length or positional statistics derived from parallel texts. In fact, there is little correlation between such statistics of a word and its translation in non-parallel corpora (Fung, P., 1995). Words with a productive context in one language are translated into words with a productive context in another language. Thus, context heterogeneity measures of how productive the context of a word is in a given domain are estimated independent of its absolute occurrence frequency in the text. The ensuing results can be used to bootstrap or refine a bilingual lexicon compilation algorithm.

Using Multilingual Thesauri

Different search strategies based on multilingual thesauri are investigated in this method. The results indicate that the combination of the models significantly improves results and that the use of the hierarchical information contained in the relevant thesaurus, UMLS/MeSH, is of

primary importance. Lastly, methods for bilingual terminology extraction and thesaurus enrichment are worth looking into. (D'ejean, H., Gaussier, 'E. and Sadat, F., 2002)

Geometrical Search

Another approach that has been applied to extract a bilingual lexicon from comparable corpora is the geometrical view. In this method, the algorithm allows a re-interpretation of the existing methods. Generally, it is more of an extended method as one could expect to improve performance in recall using standard approach. So in this approach, all the dictionary words, present or not in the context vector of a given word, can be used to translate it (Gaussier, E. and Renders, J.M. and et al., 2004). The extended method uses basically two major methods-- they are multilingual PLSA and canonical correlation. Multilingual PLSA is below the standard method but when it is used with the extended method, it performs as a standard method. The complexity of the final vector space will spread over a longer processing time since it has several millions of dimensions in comparison with each other. Such cases warrant that this method has less weightage than the standard or extended methods. It is also theoretically well proven that PLSA does not lead to improved performance (Gaussier, E. and Renders, J.M. and et al., 2004).

Canonical Correlation

Under geometrical method, another method is known as canonical correlation analysis (CCA). The results we obtain with CCA and its kernel version are disappointing. As already noted, CCA does not directly solve the problems we mentioned, and results show that CCA does not provide a good alternative to the standard method. Here again, we may suffer from a noise problem, since each canonical direction is defined by a linear combination that can involve many different vocabulary words. This leads to a noise problem since spurious relations are bound to be detected. The restriction imposed on the translation pairs to be used (N nearest neighbors) directly aims at selecting only the translation pairs which are in true relation with the word to be translated. Empirical evaluation shows the

strengths and weaknesses of these methods, as well as a significant gain in the accuracy of the extracted lexicons.

Bootstrapping

Bootstrapping is one of the major techniques and approaches used in the extraction of bilingual lexicon from the corpus and different types of documents which are specifically subjective or domain-specific, in a broad sense. To build lexicons this approach uses rule-based heuristics and hand-picked annotated data as seeds. Taggers are also used for parts-of-speech information; WordNet is used for the semantic basis of sense clusters according to the requirements and demands of the method and type of resources like, small set of subjective words, raw corpus, online dictionaries and ranked algorithms etc. which are used as seeds (Liu, Y. and Yu, S. and Yu, J., 2002). This method is very useful for building and developing resources for scarce resourced languages (Banea, C. and Mihalcea, R. and Wiebe, J., 2008). In some of the deep processing, machine translation system is also used to extract data from the translated, parallel corpus and individual words like bilingual lexical templates in order to match the terminal symbols in the parses of the aligned sentences from parse-parse-match approach.

This is helpful in the automatic building of lexicons. Further processing by humans and their intervention are limited only to the verification of the output data, i.e., post-editing is also required (Turcato, D, 1998). Post-editing developers need to facilitate the use of platforms like visualized tools or gaming platforms for the lexicographers to interactively operate on them for the expression of the bilingual semantics to be put in the natural process (Pala.K, Singh. A.K, Ganagashetty., S.V., 2011). The dictionary entries are in the form of Universal Words (UWs) which are language words (primarily English) concatenated with the relevant disambiguation information. The entries are associated with syntactic and semantic properties – most of which are also generated automatically. The WordNet system uses a word sense disambiguator, an inferencer and a knowledge base (KB) of the Universal Networking Language which is a recently proposed as an interlingua (Verma, N. and Bhattacharyya, P., 2004). The facts that "appropriate, robust, monolingual grammars may not be

available" and "the grammars may be incompatible across languages" (Turcato. D. 1998) now come to the front as emerging realities.

Statistical Approaches

This method involves an application of statistical techniques in tandem with sense discrimination optimized for the automatic building of translation lexicons from parallel corpora in order for one/us to build multilingual lexicon. It has been experimented on "1984" parallel corpus (Erjavec et al (2001). For the sense clustering criterion, they took English words as used in the "1984" parallel corpus which were available in another 6 languages. In this approach, a technique called interlingual index (ILI) mapping is used for cross-lingual validation of basic concepts of sysnsets on running texts for the refinement of word sense discrimination and sense cluster labeling (Tufis, D. 2002). One more approach for finding domain specific words (DSW's) in particular domains is also available under the rubrics of statistical models (Chang, J.S., 2005). DKvec is a method for extracting bilingual lexicons from noisy parallel corpora based on arrival distances of words in noisy parallel corpora. Notably, DKvec was used on noisy parallel corpora in English/Japanese and English/Chinese (Fung, P., 1998).

Such results in the area of statistical models can also be derived from a new method called Convec. Convec is based on the context information of a word to be translated. Since non-parallel corpora contain a lot more polysemous words, many-to-many translations and different lexical items in the two languages, it can be safely conclude that the output from Convec is reasonable and useful (Fung, P., 1998).

Word Translations Extraction

K-vec algorithm is used for the alignment of bilingual data. It estimates the proximity of the lexical items in the word lists for example: fisheries --~ p~ches. The algorithm will draw the fact by noting that the distribution of fisheries in the English text is similar to the distribution of p~ches in French (Fung, P. and Church, K.W., 1994). 2) Parenthetical translation technique allows us to identify translation pairs which appear in the entire web page. Even if it appears once in the entire webpage, it will be identified through this technique (Lin, D., Zhao, S and et al 2008).

Gao, Z.M., 1998. Automatic Acquisition of a High-Precision Translation Lexicon from Parallel Chinese-English Corpora. *Language*.Pp. *248*-254.

Hardie, A., etal.2006. Corpus-building for South Asian Languages. *Trends in Linguistic Studies and Monographs.75*, 211.Mouton de Gruyter

Kaji, N. and Kitsuregawa, M.2007.Building Lexicon for Sentiment Analysis from Massive Collection of HTML documents. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL).Pp.*1075-1083.

Kilgarriff, A., etal. 2009. Corpus Factory. *Proceedings of Asialex.*

Lin, D., et al. 2008. Mining Parenthetical Translations from the Web by Word Alignment, *ACL08.* Pp.994-1002.

Liu, Y.2002. Building a Bilingual WordNet-like Lexicon: The New Approach and Algorithms. *Proceedings of the 19th International Conference on Computational linguistics-2.*Association for Computational Linguistics.

Morin, E., etal.2011. Bilingual Lexicon Extraction from Comparable Corpora as Metasearch, *ACL HLT 2011*.Pp.35-43.

Morin, E., etal.2011. Bilingual Lexicon Extraction from Comparable Corpora Enhanced with Parallel Corpora, *ACL HLT 2011*.Pp.27–34.

Resnik, P. and Smith, N.A.2003. The Web as a Parallel Corpus. *Computational Linguistics* 29(3).349-380. MIT Press

Singh, A.K.,etal.2007.More Accurate Fuzzy Text Search for Languages using Abugida Scripts. *Proceedings of ACM SIGIR Workshop on Improving Web Retrieval for Non-English Queries, Citeseer.*

Tsunakawa,T.,etal.2008. Building a Bilingual Lexicon Using Phrase-based Statistical Machine Translation via a Pivot Language.*The Proceedings of the 22nd COLING.*

Tufis, D.2002. A Cheap and Fast Way to Build Useful Translation Lexicons. *Proceedings of the 19th international conference on Computational linguistics-Volume 1*.1—7.Association for Computational Linguistics

Turcato,D.1998.Automatically Creating Bilingual Lexicons for Machine Translation from Bilingual Text. *Proceedings of the 17th International Conference on Computational Linguistics- 2*.Pp.1299-1306. Association for Computational Linguistics.

Verma, N. and Bhattacharyya, P.2004. Automatic Lexicon Generation through WordNet. *GWC 2004*.

Vikas,O.2005. Multilingualism for Cultural Diversity and Universal Access in Cyberspace: An Asian Perspective. *Thematic Meeting for the World Summit on the Information Society*.Pp.1-49.

❑❑❑