Word Sense Alignment of Sanskrit Lexica

Dhaval Patel
Department of Sanskrit Studies,
University of Hyderabad
drdhaval2785@gmail.com

Amba Kulkarni
Department of Sanskrit Studies,
University of Hyderabad
ambakulkarni@uohyd.ac.in

Abstract

Word sense alignment is a field of study in which lexical resources or texts are aligned at the level of word sense rather than the word. The present paper tries to evaluate the possibility of mechanically aligning Sanskrit lexica at the level of word sense computationally.

1 Introduction

Sanskrit, an ancient Indian language, has been a medium of transmission of knowledge in various fields of study for centuries. Compilation of word lists in Sanskrit commenced at an early date as it was found necessary to access the old literature such as Vedic literature, while the language was undergoing some transformations with meaning shifts. The lexical resources known as kośas were developed. They are of two types – (1) Samānārthaka kośas and (2) Anekārthaka kośas. Samānārthaka kośas enlist the synonyms together. The synonyms are arranged following some theme, semantic criterion, or ontological classification scheme. For example, in the most famous samānārthaka kośa viz. Amarakośa, the words are arranged in three kāṇḍas and further within the kāṇḍas, the headwords are arranged based on either semantic or ontological properties. Anekārthaka kośas enlist different meanings of a given word. The words may or may not have any alphabetic arrangement. Both kinds of kośas were meant to be memorized, applied to texts, and cited as and when the usage of the said word in the literature was to be justified in a commentary. Therefore, the kośas were almost invariably in a verse form. Vogel (2015) has given a comprehensive coverage of these Sanskrit kośas and commentarial literature thereon.

Because of the influence of Western lexicography, a few Sanskrit-Sanskrit dictionaries like Vācaspatyam and Śabdakalpadruma were also compiled on the lines with the Western methodology of arranging headwords alphabetically and in prose form. Several bilingual Sanskrit dictionaries such as Sanskrit-English, Sanskrit-French, and Sanskrit-German were created starting from the early 19^{th} century. Almost all of the major dictionaries that are free from copyright are available on the Cologne Digital Sanskrit Dictionaries website (CDSD, 2023). This digitized data has various levels of markup. In the recent years Huet (2019) has developed a digital Sanskrit-French dictionary where the lexical items are directly linked to the inflectional and derivational morphology.

Some of these dictionaries, in addition to providing the meaning of Sanskrit words in the target language, also provide citations from Sanskrit texts. The citations in different dictionaries vary. These citations play an important role in understanding the context in which the sense is being used. Aligning the senses of different dictionaries would provide us with more than one example sentence for each sense to understand the context and the semantic criterion that decides the sense of the word in a given usage. Further, with the availability of word embeddings for words in several languages such as Hindi, English, French, German, etc. if the senses in Sanskrit bilingual dictionaries are aligned, one can take advantage of the

existing modelling of the world knowledge and the domain knowledge of other languages to disambiguate Sanskrit words. Such sense mapping would be useful in the Machine Translation system, for Information Retrieval, and even for a casual learner of the Sanskrit language. This motivated us to look at the problem of aligning various Sanskrit bilingual dictionaries according to the senses.

In what follows, we first explain the word sense alignment problem, and the challenges therein. This is followed by the discussion on the methodology followed for automatic sense alignment. In section 4 we discuss the sense alignment of two dictionaries Sanskrit-English and Sanskrit-Hindi by Apte. The results of the alignment algorithms are extended to other pairs of dictionaries, which is the topic of section 5. Finally we discuss other possible ways of alignment before concluding.

2 Word Sense Alignment

Word sense alignment, also known as sense alignment or sense mapping matches two entries from two lexical resources based on the sense the two entries express. Word sense alignment emerged out of various efforts toward the word sense disambiguation (WSD) problem. WSD is an important task for several NLP applications such as Machine Translation, Information Retrieval, Question Answering, Summarisation, and so on. At the same time, it is one of the most difficult problems in the field of Natural Language Processing (NLP). It is considered as being an AI-complete problem (Agirre and Edmonds, 2007). The difficulties arise due to poor understanding of the process involved. Various factors such as linguistic, contextual, domain-specific, cultural, and world knowledge contribute to the process of manual word sense disambiguation. In the case of resource-rich languages such as English, there are several lexical resources with varied granularity such as WordNet (Miller, 1995), FrameNet (Baker et al., 1998), ConceptNet (Speer et al., 2017), VerbNet (Schuler, 2005) etc., and sense-tagged corpora available in digital media. This resulted in several efforts aiming at the alignment of such resources known as Word Sense Aligned (WSA) resources. The development of Euro-WordNet (Vossen, 1998) and Indo-WordNet (Bhattacharyya, 2010) are also steps towards generating Word sense-aligned lexical resources so that the sense-tagged corpus in one language can become available in another with minimum effort. In the recent years, Word Sense Alignment has gained importance. Languages with low resources would like to take advantage of the resources available in resource-rich languages, by aligning their resources to those of the rich languages. For example Salgado et al. (2020) describes the challenges of word sense alignment of Portuguese Language Resources. Joshi et al. (2012) present a heuristic approach to link English and Hindi WordNets by linking their senses. Two closely related Czech lexical resources VALLEX[1] and PDT-VALLEX[2] were aligned fully automatically (Bejcek et al., 2014). Johansson and Pina (2015) used word sense embeddings to automatically link the Swedish language banks.

During his post-doctoral fellowship in 2012 at Inria, Pawan Goyal aligned the Sanskrit Heritage dictionary with an XML version of Monier-Williams available at CDSD (Goyal et al., 2012). Goyal used the online google translator to translate the French entries into English and then aligned them with the entries in Monnier Williams' Sanskrit-English dictionary, by manually aligning the entries wherever there were ambiguities/multiple choices available. The alignment process is incremental and thus may be iterated on successive versions of the Sanskrit Heritage dictionary.

[1] http://ufal.mff.cuni.cz/vallex/2.6/
[2] http://lindat.mff.cuni.cz/services/PDT-Vallex/

2.1 Challenges

The conceptual space is a continuum that is divided into discrete units by the lexicon of a language. Since the lexicon is denumerably finite, a word represents a piece of continuous conceptual space and not a discrete point. This sometimes leads to one word representing a spectrum of meanings. Such words are termed polysemous words. Sometimes, more than one lexical unit produces the same word form. Such word forms are called homonyms. Among the homonymous and polysemous meanings, typically the homonyms are provided with different headword entries, while the polysemous meanings are clubbed under a single head. Within polysemous meanings, the granularity is decided by the lexicographer. Deciding the granularity of the meaning is not trivial. It is not at all clear when a sense of the word should be treated as a separate meaning and when it should be subsumed within an already existing meaning. Further deciding between a polysemy and homonymy is subjective due to the fuzzy boundary between them. Another factor is the inclusion of metaphoric meanings in the dictionary. Indian tradition discusses three types of meanings viz. abhidhā (literal), lakṣaṇā (metaphoric or secondary) and vyañjanā (suggestive). While it is impossible to provide the suggestive meanings, which are subjective in nature, and also depend on the context, the lexicographers do consider the secondary or suggestive meanings for inclusion in the dictionaries. Even in the case of dictionaries from the same lexicographer, the intended audience, printing or economic considerations may force the lexicographer to deal with sense granularity in different ways across different dictionaries. Therefore, the choice of sense granularity is mostly left to the discretion of the lexicographer, as has been observed through various lexical resources. Because of these reasons, the word sense mapping between different lexical resources is not trivial.

3 Methodology

Manually aligning lexical resources at the word sense level is a very laborious task. It would also require the person to be well versed in two languages e.g. aligning a Sanskrit-English and Sanskrit-Hindi dictionary would require the person to know at least English and Hindi, and preferably Sanskrit too. For a resource-starved languages like Sanskrit, this may be very costly and time-consuming.

The present work focuses on finding out the similarity between different meanings of a given word and present the human annotator with a similarity score or a confidence score, so that the annotator may devote more time to the places where the machine performs with a low confidence level. We also aim at finding a more or less language-agnostic way of automatically or semi-automatically aligning lexical resources at the word sense level so that it can be extended to other language pairs.

For any mechanical mapping between entries from two dictionaries to be successful, either both the target languages need to be the same or a model trained on both languages to identify similar concepts across both languages is needed. The first approach is simpler. Because of the advancement in machine translation technologies and publicly available resources such as Google Translate[3], it has become possible to translate various texts from one language to another. Thus, in the absence of models trained in two different languages, one can still use Google Translate to identify similar concepts across languages. The task of finding out the similarity between two documents (in our case, the meaning of the word for a given sense) is a common theme in information retrieval (IR), topic modeling, ontology matching, etc. There are various algorithms which have already been tested for the same.

Bär et al. (2013) have enumerated and implemented the following similarity algorithms in their software: Longest Common Substring, Greedy String Tiling (Wise, 1996), Jaro (1989),

[3]https://translate.google.com/ accessed on 20 September 2023

Jaro-Winkler (Winkler, 1990), Monge and Charles (1997), Levenshtein (1966), Jiang and Conrath (1997), Resnik (1995), Latent Semantic Analysis (Landauer et al., 1998) and Explicit Semantic Analysis (Gabrilovich and Markovitch, 2007). As the dictionary meanings are relatively small chunks of text, with sizes ranging upto two or three sentences at maximum, structural and stylistic similarity measures mentioned in the said paper are not of much relevance to the task at hand. Semantic similarity measures presume some graph-like structure and use the structure of those graphs to find out the similarity or nearness between two nodes. These work best when there is some ontological representation of the world knowledge or some hierarchy of the word/word senses and their relationship is explicitly coded. In an alphabetically arranged dictionary, such a relationship is almost non-existent. Therefore, these measures were not tried. Latent Semantic Analysis (LSA) and explicit semantic analysis (ESA) require a lot of computational resources. Training and running these algorithms on a large corpus like two full-fledged dictionaries will be computationally too heavy. LSA may be able to identify similarities between 'child' and 'offspring', which a normal text-based scorer may miss. However, due to limited computational resources, we have not tried them either. Other than these measures, there are following string-based measures implemented by rapidfuzz library[4] - Damerau (1964), Hamming (1950), Indel, OSA, Prefix and Postfix.

The present paper focuses on the usability of string-based similarity measures for finding out the mapping of word senses. We present here our efforts towards the word sense alignment of Sanskrit lexical resources. We present three case studies. The first one is with two different target languages but the same compiler. Here we have chosen Sanskrit-English[5] and Sanskrit-Hindi dictionaries of Apte[6]. Since the second dictionary is based on the first one, the assumption is there would be a good chance of getting one-one mapping. The second pair is with the same target language but different compilers. Here we have chosen Sanskrit-English by two different compilers – Wilson (1832)[7] and Yates (1846)[8]. The third one was a pair of monolingual English lexical resources viz. Webster's Unabridged Dictionary of the English Language (Webster, 1900)[9] and English Wordnet (Miller et al., 1990)[10].

4 WSA of Sanskrit-English and Sanskrit-Hindi of Apte

Apte Sanskrit-English (AP90) dictionary (Apte, 1890) has been used in this experiment. The later 1957 version (AP) of the dictionary (Apte et al., 1957) is still under copyright. Therefore CDSD does not have its data for open usage. AP90 is not fully marked up to show different word senses separately. It has some rudimentary markup or patterns by the help of which crude parsing was done and word meanings were separated. Apte Sanskrit-Hindi (ASH) dictionary (Apte, 2007) is a Hindi translation of Apte's Sanskrit-English dictionary. It is not an exact translation. Many of the words have been omitted, and many meanings have been merged, deleted, or separated. It seems that ASH had the advantage of using the data of the 1957 edition too. Therefore, the new words or meanings added in that edition are also used in ASH. At the same time, ASH has been made more concise. Therefore, multiple meanings have been combined together. Rare meanings have been dropped altogether too. Therefore, it was not trivial to align the word senses in these two dictionaries, and hence, these were taken up to attempt word sense level alignment between them.

[4]https://pypi.org/project/rapidfuzz/

[5]https://www.sanskrit-lexicon.uni-koeln.de/scans/AP90Scan/2020/web/webtc/download.html

[6]Developed by the SHMT (Sanskrit-Hindi Machine Translation) consortium during 2008-2011, now a part of Samsaadhanii Platform at https://sanskrit.uohyd.ac.in/scl/

[7]https://www.sanskrit-lexicon.uni-koeln.de/scans/WILScan/2020/web/webtc/download.html

[8]https://www.sanskrit-lexicon.uni-koeln.de/scans/YATScan/2020/web/webtc/download.html

[9]https://www.gutenberg.org/files/29765/29765-0.txt

[10]https://github.com/fluhus/wordnet-to-json/releases/download/v1.0/wordnet.json.gz

4.1 Gold Standard Data

As there is no previously existing gold standard data regarding word sense alignment of unstructured lexical resources like a dictionary pair, a manual gold standard data was created by selecting a random starting point and taking roughly 1000 ASH entries starting therefrom (See Table 1). Corresponding entries of AP90 were also taken up (See Table 2).

Head word	Hindi sense_id (ASH)	Hindi Meaning (ASH)
आकल्पः (Ākalpaḥ)	247	आभूषण, अलंकार
आकल्पः (Ākalpaḥ)	248	वेशभूषा
आकल्पः (Ākalpaḥ)	249	रोग, बीमारी

Table 1: sample ASH entries

Head word	English sense_id (AP90)	English Meaning (AP90)
आकल्पः (Ākalpaḥ)	440	An ornament, decoration
आकल्पः (Ākalpaḥ)	441	Dress (in general), accoutrement
आकल्पः (Ākalpaḥ)	442	Sickness, disease
आकल्पः (Ākalpaḥ)	443	Adding to, increasing

Table 2: sample AP90 entries

Every word sense had been given a unique identifier for both dictionaries. A manual examination of the data was done and a manual mapping was created. As and when some parsing error was detected in the data, the same was manually corrected. Sample entries of the sense alignment of entries from ASH and AP90 are shown in Table 3.

Head word	Hindi sense_id (ASH)	English sense_id (AP90)
आकल्पः (Ākalpaḥ)	247	440
आकल्पः (Ākalpaḥ)	248	441
आकल्पः (Ākalpaḥ)	249	442
आकल्पः (Ākalpaḥ)	-	443

Table 3: Gold Standard Data for Alignment of ASH and AP90

Since the two dictionaries selected had different target languages, for aligning the entries, we decided to use Google Translate to translate the meanings of AP90 into Hindi. In Table 4, column GSH shows the Google translation of the entries in AP90 into Hindi. The task at hand is to map the English sense_id to Hindi sense_id using GSH. Please note that sometimes Google Translate does not translate some difficult words like 'accoutrement' and leave them as they are, when processed via bulk upload.

Head word	English sense_id (AP90)	English Meaning (AP90)	GSH
आकल्पः (Ākalpaḥ)	440	An ornament, decoration	एक आभूषण, सजावट
आकल्पः (Ākalpaḥ)	441	Dress (in general), accoutrement	पोशाक (सामान्य रूप से), accoutrement
आकल्पः (Ākalpaḥ)	442	Sickness, disease	रोग, रोग
आकल्पः (Ākalpaḥ)	443	Adding to, increasing	जोड़ना, बढ़ाना

Table 4: sample AP90 entries along with their Hindi Translations

Similarly, entries of ASH were translated into English with the help of Google translate. Please see column GSE of Table 5.

Head word	Hindi sense_id (ASH)	Hindi Meaning (ASH)	GSE
आकल्पः (Ākalpaḥ)	247	आभूषण, अलंकार	jewelery , Ornament
आकल्पः (Ākalpaḥ)	248	वेशभूषा	Costumes
आकल्पः (Ākalpaḥ)	249	रोग, बीमारी	Disease , Disease

Table 5: sample ASH entries along with their English Translations

In the next section, we present various algorithms, and their performance on the gold standard data.

4.2 Algorithms

Our algorithms are based on simple string-level similarity measures. We define four different units of comparison, and four different units of measure for comparison resulting in 16 different algorithms. We describe them below.

4.2.1 unit of comparison

The three basic units we propose are words, shingles (n-grams at character levels), and syllables. While glancing at the ASH entries with GSH manually, we also realized that in the case of languages like Hindi, depending upon the presence of post-positions, the last character of the word is changed as in 'baharā' (बहरा) versus 'bahare' (बहरे). Hence we decided to consider a word with the last character trimmed also as a unit of comparison.

4.2.2 measure of comparison

We have identified four different measures for calculating the similarity. Suppose the meanings from two dictionaries are stored as a list of words L_1 and L_2. l_1 and l_2 are sets of unique words amongst L_1 and L_2 respectively. In the following notation, $|A|$ denotes the cardinality of set A. The four different measures are defined as

$$m_1 = \frac{|L_1 \cap L_2|}{|L_1 \cup L_2|}$$

$$m_2 = \frac{|l_1 \cap l_2|}{|l_1 \cup l_2|}$$

$$m_3 = \frac{|L_1 \cap L_2|}{|L_1|}$$

$$m_4 = \frac{|l_1 \cap l_2|}{|l_1|}$$

The phrases describing the senses are tokenised and stop-words are removed. In the case of English, all the words are converted to lower case. Let us assume the two senses that need to be aligned are 'space, place in general' and 'free space or vacuum'. As a first step the phrases are tokenised and the stop words are removed, and the words are changed to lower case. This results into two word lists

L_1 = [“space”, “place”, “general”], and
L_2 = [“free”, “space”, “vacuum”]

As there are no duplicate words in any of these two word lists, $l_1 = L_1$ and $l_2 = L_2$. Thus, for the above lists $m_1 = \frac{1}{6}$; $m_2 = \frac{1}{5}$; $m_3 = \frac{1}{3}$; $m_4 = \frac{1}{3}$

The shingles for each word are the all possible n-grams of characters. Thus,
shingles(“space”) = [“s”, “sp”, “spa”, “spac”, “space”, “p”, “pa”, “pac”, “pace”, “a”, “ac”, “ace”, “c”, “ce”, “e”].

The trimmed words are obtained by trimming the last character of the word. So the trimmed word list for L_1 is [“spac”, “plac”, “genera”]. While, we do not see this trimmed word list of any advantage in the case of English, for languages like Hindi these are useful. For example, in the word mapping the words ‘baharā’ (**बहरा**) and ‘bahare’ (**बहरे**) will not match, but after trimming the last phoneme, both the words will match ‘bahar’.

With the 4 units of comparison and 4 units of measures of similarity, there are 16 different measures for judging the similarity between the two senses. These 16 measures are shown in Table 6.

unit	m_1	m_2	m_3	m_4
word	CR1	CR2	CR3	CR4
shingles	CR5	CR6	CR7	CR8
trimmed word	CR9	CR10	CR11	CR12
syllable	CR13	CR14	CR15	CR16

Table 6: Metrics used for evaluation

A threshold of 0.2 was defined to ignore the mappings with low similarity score. Another measure delta was also calculated. It is the difference between the word sense pair across dictionaries with the highest similarity score and the second best pair. If delta is high, it means that the pair at the first rank is ahead of the second rank comfortably. A threshold value of delta was kept at 0.1.

4.3 Evaluation on Gold Data

After setting these thresholds, the comparison of the results of all algorithms was made. The gold standard comprises of 2022 word sense pairs manually validated. The results of some of the standard algorithms implemented by rapidfuzz library are shown in Table 7.

Algorithm	Pairs identified	Percentage	Algorithm	Pairs identified	Percentage
Levenshtein	1793	88.67%	Damerau	1794	88.73%
Hamming	1688	83.48%	Indel	1822	90.11%
Jaro	1816	89.91%	Jaro-Winkler	1818	89.91%
OSA	1793	88.67%	Prefix	1749	86.50%
Postfix	1690	83.58%			

Table 7: Percentage of word sense pairs correctly identified from gold standard data by already existing algorithms

With the same thresholds, the results obtained from algorithms CR1 to CR16 are shown in Table 8.

Algorithm	Pairs identified	Percentage	Algorithm	Pairs identified	Percentage
CR1	1866	92.28%	CR9	1861	92.04%
CR2	1871	92.53%	CR10	1865	92.24%
CR3	1855	91.74%	CR11	1848	91.39%
CR4	1856	91.79%	CR12	1847	91.35%
CR5	1856	91.79%	CR13	1847	91.35%
CR6	1887	**93.32%**	CR14	1878	92.87%
CR7	1858	91.89%	CR15	1841	91.04%
CR8	1863	92.14%	CR16	1851	91.54%

Table 8: Percentage of word sense pairs correctly identified from gold standard data by various algorithms

As can be seen from the results, CR6 gave the best result of all the algorithms. Therefore, the algorithm CR6 was selected out of these algorithms. CR6 makes use of shingles and hence captures various features like terminal case removal, textual similarity between tatsama words and tadbhava words, common verb or common noun in compounds etc. This may be the reason why CR6 gives better result than other algorithms.

4.4 Evaluation on complete dictionaries

CR6 was applied to the complete dictionaries ASH (D1) and AP90 (D2). As AP90 definitions are in English language and ASH definitions are in Hindi language, both were translated with the help of Google Translate and an English version of ASH (E) and a Hindi version of AP90 (H) were created. D1 and H were compared against each other (both with Hindi definitions) and D2 and E were compared against each other (both with English definitions). Having comparisons with both the languages helped in a big way. There are cases where one language is insufficient to map satisfactorily, but the other language could map without any difficulty. Let us see such a case with an example.

D1.91	जो चुराये जाने के योग्य न हो, या हटाये जाने अथवा दूर ले जाये जाने के योग्य न हो
D1.92	श्रद्धालु, निष्ठावान्
D1.93	दृढ,अविचल,अननुनेय
D1.94	पहाड़

Table 9: Entry of the word 'अहार्य' in the ASH (D1) dictionary

D2.1916	not to be stolen, removed or carried
D2.1917	not to win (by fraud), devoted, loyal
D2.1918	firm, steadfast, hard
D2.1919	a mountain

Table 10: Entry of the word 'अहार्य' in the AP90 (D2) dictionary.

As can be seen from the contents the four senses of D1 correspond sequentially to the four entries of D2.

Now, let us look at the Google translations of D1 into English and D2 into Hindi.

Had we used only translation of AP90 into Hindi through Google translator, and compared it with the entries in ASH, the words 'पहाड़' (D1.94) and 'एक पर्वत' (H.1919) will not get good similarity score. However, the same words when translated to English will be highly similar viz. 'a mountain' (D2.1919) and 'Mountain' (E.94). Therefore, the similarity score with English as the destination language will be very high. Similarly, 'श्रद्धालु, निष्ठावान्' will not match 'समर्पित,

E.91	Unstealable, or not capable of being removed or taken away
E.92	Devotees , loyal
E.93	Strong , motionless , irresistible
E.94	Mountain

Table 11: Entry of ASH translated to English via Google Translate (E)

H.1916	चोरी, हटाया या ले जाने के लिए नहीं
H.1917	जीतने के लिए नहीं (धोखाधड़ी से), समर्पित , वफादार
H.1918	दृढ, अडिग, कठोर
H.1919	एक पर्वत

Table 12: Entry of AP90 translated to Hindi via Google Translate (H)

वफादार' much at character level, but 'devoted, loyal' will match 'Devotees, loyal' at character level. There are also cases where Hindi fares better. Mapping '**फँसा हुआ**' with '**फँसा हुआ**' is easier than mapping 'trapped' and 'entangled' (D1.342 and D2.556). Thus, using two languages helps us to take care of some cases where one language uses different synonyms and the other language has only one word for the concept or may have used the same word out of available synonyms.

Creating mapping with two languages also gives us some more benefits. It give us more confidence about a given mapping if both the languages give the same mapping. Based on these insights, an analysis of the mappings of gold standard data and full dictionary data was carried out. We classify the confidence levels of machine into 7 different categories. These categories are shown in Table 13 with their correspondence confidence levels.

Category	Description	Confidence
A	Both languages give above sim_threshold, and both languages give the same first match	High
B	(Language1 above sim_threshold, and Language2 gives lower similarity score) or (Language2 above sim_threshold, and Language1 gives lower similarity score)	High
C	Headword present in only one dictionary, and absent in the other	High
D	Both languages give below sim_threshold, and both languages give the same first match	Low
E	(Language1 below sim_threshold, but better than Language2) or (Language2 below sim_threshold, but better than Language1)	Low
F	Headword present in both dictionaries, but all entries of dictionary1 have already been assigned to other entries of dictionary2 or vice versa. Hence, there is no mapping	High
G	Force mapped, as this is the only remaining match	High

Table 13: Categorization of various mappings along with their confidence level

Analysis of gold standard data with these codes yielded the following results (See Table 14).

It is worth noting that the machine generated a total of 2096 mappings. The gold standard data has 2022 mappings. It is because the machine does not know what are the number of mappings present in the gold standard data. Word senses may have one-one, one-many and many-one mappings. Therefore, it is not possible to determine in advance how many word sense

Category	Pairs in the category	Percentage
A	636	30.34 %
B	221	10.54 %
C	738	35.21 %
D	60	02.86 %
E	59	02.81 %
F	360	17.18 %
G	22	01.05 %
Total	2096	100 %

Table 14: Categorization of gold standard mapping generated via algorithm CR6

mappings are to be generated. Therefore, the machine generated a total of 2096 mappings. Among these, the entries falling in the category of D and E have low confidence. Thus roughly 5-6% cases are such which are of low quality and would improve with human intervention. Rest of 94-95% cases can be mechanically aligned, saving precious resources. As the gold standard data was corrected as and when some parsing error or typographic error was seen, the error rate in gold standard data is much less. Whereas, there was no attempt made to clear these kind of errors in full dictionaries. Therefore, the error rates in the full dictionaries is more than the gold standard data.

The following are the results of application of the above methodology to full dictionaries Apte Sanskrit–Hindi (ASH) and Apte Sanskrit–English (AP90) (See Table 15).

Category	Pairs in given category	Percentage
A	51964	32.59 %
B	16655	10.44 %
C	47131	29.56 %
D	9730	06.10 %
E	6258	03.92 %
F	25710	16.12 %
G	2023	01.27 %
Total	159471	100%

Table 15: Categorization of word sense mappings generated by algorithm CR6 for Apte Sanskrit-Hindi and Apte Sanskrit-English dictionaries

Roughly 10% of cases fall under low confidence zone, which may require human intervention. Three random numbers were selected and 100 entries starting therefrom were examined for false positives. The following is the result. (See Table 16)

A	B	C	D	E	F	G	False Positives/total pairs
01	14	00	01	10	00	04	30/300

Table 16: False positives from randomly selected mappings

Thus, manual examination also yields around 10% error rate.

5 Mapping other dictionaries

Similar exercise was also tried for different dictionary pairs like (1) Apte Sanskrit-English and Monnier Williams Sanskrit-English dictionary[11] (2) Wilson Sanskrit-English and Yates Sanskrit-

[11] https://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/2020/web/webtc/download.html

English dictionary and (3) English WordNet and Webster's English dictionary. The results are shown in Table 17.

Category	Apte – MW	Wilson – Yates	WordNet – Webster
A	17.84 %	53.22 %	09.43 %
B	04.53 %	11.40 %	03.90 %
C	53.12 %	13.44 %	61.83 %
D	08.39 %	07.99 %	06.92 %
E	02.78 %	04.80 %	05.01 %
F	12.70 %	08.03 %	11.68 %
G	00.64 %	01.13 %	01.23 %

Table 17: Categorization of mappings for various dictionary pairs

Thus, in almost all dictionary pairs studied, the error rate (D+E) is roughly to the tune of 11-13%. These are the places where human annotators can make maximum impact by manual examination and correction.

6 Way ahead

We are exploring the possibility of using graph based similarity scores or semantic measures such as LSA or ESA to find out similarity in cases where text based similarity scores are below threshold. These approaches are computationally heavy and may require more computational resources. In the present case, the thresholds of similarity scores were chosen empirically or rather arbitrarily. It may be possible to learn these thresholds by optimizing its F-scores. As the gold standard (training data) is quite small, this exercise is not yet tried. Once we have large manually validated data, it will be worthwhile to find out the optimum thresholds with statistical methods.

The present methodology can be expanded to other language pairs and check whether findings in different language pairs are similar or otherwise. Effect of quality of translation services like Google Translate between different language pairs may add a cascading effect on the performance.

7 Conclusion

Undertaking the task of mapping of dictionaries at the level of word sense seems daunting at first, but after experimenting with a few dictionary pairs, it was only 11-13% of word senses that required manual examination by human expert. Once a quick implementation having an accuracy of 87-90% is created by machine, human annotators / users can be given an option to change the mapping if they feel that the mapping generated by the machine is incorrect. It holds immense potential to expand sense-mapped text resources from one language to another. It will particularly help the users of languages which are having scarce resources e.g. a Sanskrit work which has been disambiguated and sense-mapped in English with help of Sanskrit-English dictionary can be extended to the users of, say, French language by mapping Sanskrit-English dictionary to Sanskrit-French dictionary at word sense level.

References

Eneko Agirre and Philip Edmonds. 2007. *Word sense disambiguation: Algorithms and applications*, volume 33. Springer Science & Business Media.

Vāmana Śivarāma Apte, Paraśurāma Kriṣṇa Gode, Cintāmaṇa Gaṇeśa Karve, and Kaśināṭha Vāsudeva Abhyankara. 1957. *Revised and Enlarged edition of Prin. V. S. Apte's The Practical Sanskrit-English Dictionary.* Prasad Prakashan, Poona.

Vaman Shivram Apte. 1890. *The Practical Sanskrit-English Dictionary, containing Appendices on Sanskrit Prosody and important Literary & Geographical names in the ancient history of India*. Shiralkar & Co. Book-sellers, Budhwar Peth, Poona.

Vaman Shivram Apte. 2007. *Sanskrit-Hindi Kośa*. Motilal Banarsidass, Delhi.

Collin F Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Eduard Bejcek, Václava Kettnerová, and Markéta Lopatková. 2014. Automatic mapping lexical resources: A lexical unit as the keystone. In *LREC*, pages 2826–2832.

Pushpak Bhattacharyya. 2010. Indowordnet. lexical resources engineering conference 2010 (lrec 2010). *Malta, May*.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2013. Dkpro similarity: An open source framework for text similarity. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 121–126.

CDSD. 2023. *Cologne Digital Sanskrit Dictionaries*. Cologne University, Cologne. version 2.4.123, accessed on 20 September 2023, at https://www.sanskrit-lexicon.uni-koeln.de.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing semantic relatedness using Wikipedia-based explicit semantic analysis. In *IJCAI*, volume 7, pages 1606–1611.

Pawan Goyal, Gérard Huet, Amba Kulkarni, Peter Scharf, and Ralph Bunker. 2012. A distributed platform for Sanskrit processing. In Martin Kay and Christian Boitet, editors, *Proceedings of COLING 2012*, pages 1011–1028, Mumbai, India, December. The COLING 2012 Organizing Committee.

Richard W Hamming. 1950. Error detecting and error correcting codes. *The Bell System Technical Journal*, 29(2):147–160.

Gérard Huet. 2019. Sanskrit lexicography, past and future. In Li Wei, editor, *Research on the Language and Script in Buddhist Sutras*. Hangzhou Buddhist Academy.

Matthew A Jaro. 1989. Advances in record-linkage methodology as applied to matching the 1985 census of tampa, florida. *Journal of the American Statistical Association*, 84(406):414–420.

Jay J. Jiang and David W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *arXiv preprint cmp-lg/9709008*.

Richard Johansson and Luis Nieto Pina. 2015. Embedding a semantic network in a word space. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1428–1433.

Salil Joshi, Arindam Chatterjee, Arun Karthikeyan Karra, and Pushpak Bhattacharyya. 2012. Eating your own cooking: automatically linking WordNet synsets of two languages. In *Proceedings of COLING 2012: Demonstration Papers*, pages 239–246.

Thomas K Landauer, Peter W Foltz, and Darrell Laham. 1998. An introduction to latent semantic analysis. *Discourse processes*, 25(2-3):259–284.

Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3.4:235–244.

George A Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Alvaro Monge and Elkan Charles. 1997. An efficient domain-independent algorithm for detecting approximately duplicate database records. In *Proc. of the ACM-SIGMOD Workshop on Research Issues on Knowledge Discovery and Data Mining*.

Philip Resnik. 1995. Using information content to evaluate semantic similarity in a taxonomy. *arXiv preprint cmp-lg/9511007*.

Ana Salgado, Sina Ahmadi, Alberto Simoes, John McCrae, and Rute Costa. 2020. Challenges of word sense alignment. In *Proceedings of the LREC 2020 7th Workshop on Linked Data in Linguistics (LDL-2020)*, pages 45–51. European Language Resources Association (ELRA).

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31(1).

Claus Vogel. 2015. *Indian Lexicography*. Motilal Banarsidass, Delhi.

Piek Vossen. 1998. A multilingual database with lexical semantic networks. *Dordrecht: Kluwer Academic Publishers. doi*, 10:978–94.

Noah Webster. 1900. *Webster's unabridged dictionary of the English language*. Kikwansha.

H. H. Wilson. 1832. *A Dictionary in Sanscrit and English; Translated, Amended, and Enlarged from an Original Compilation, Prepared by Learned Natives for The College of Fort William*. Parbury, Allen & Co., London, second edition.

William E. Winkler. 1990. String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage. *The Educational Resource Information Center (ERIC)*.

Michael J Wise. 1996. Yap3: Improved detection of similarities in computer program and other texts. In *Proceedings of the twenty-seventh SIGCSE technical symposium on Computer science education*, pages 130–134.

W. Yates. 1846. *A Dictionary in Sanscrit and English, Designed for the Use of Private Students and of Indian Colleges and Schools*. Baptist Mission Press, Calcutta.