Automatic Normalization of Punjabi Words

Vishal Gupta

Assistant Professor UIET, Panjab University
Sector-25, Chandigarh, India

***Abstract*—For any language in the world, automatic normalization of words is a basic linguistic resource required to develop any type of application in Natural Language Processing (NLP) with high accuracy like: machine translation, document classification, document clustering, text question answering, topic tracking, text summarization and keywords extraction etc. It is not possible to achieve high accuracy without using automatic normalization of words for NLP applications for any language. This paper concentrates on automatic normalization of Punjabi words. Punjabi is the official language for state of Punjab. But Punjabi is under resource language. There are very less number of computational-linguistic resources available for Punjabi. This is 1st in history that automatic standardization of terms related to Punjabi is implemented and this system can be very much useful in creating other applications for Punjabi having good efficiency. For example it can be applied in different NLP applications like machine translation, document association, documents clustering, topic tracking and text summarization etc.**

***Keywords*— Punjabi words normalization, normalized Punjabi words, standardized words.**

I. INTRODUCTION

For any language in the world, automatic normalization of words is a basic linguistic resource required to develop any type of application in Natural Language Processing (NLP) [1] with high accuracy like: machine translation, document classification, document clustering, text question answering [4], topic tracking, text summarization [2] and keywords extraction etc. It is not possible to achieve high accuracy without using automatic normalization of words for NLP applications for any language. For Indian languages, Normalization of words is required because of their nature which is phonetic & dialects which are multiple, names transliteration, terms of foreign languages used as such has created variations in spellings of same term. These type of variations many times can assumed as writing errors.

Different conventions related to spellings are very essential feature for a language which is used. Dictionary used by university of Cambridge addresses different spellings as forming terms having different accurate characters in proper order or it is defined as capability for doing it in which variation is considered as difference or treated as deviation of that structure. Presence of different variants is not of much important for common man while applying that language as this thing usually will not affect the manner of communication of speech properly. This thing is very essential for different NLP applications like text summarization, machine translation, document classification, clustering etc.

This paper discusses the automatic normalization of Punjabi noun words. The main difficulty for Punjabi language is spellings of Punjabi words which are not standardized. Most of nouns in Punjabi can be written in different manners like ਅੰਮ੍ਰਿਤੀ ammritī can also be typed as ਅੰਮਰਿਤੀ ammaritī and other Punjabi noun ਪੱਤਰ pattar can also be typed as ਪਤਰ patar. For eliminating this difficulty with Punjabi, we have standardized the input text typed in Punjabi & morph of Punjabi nouns [5] for overcoming many variations in spellings of nouns in Punjabi. There are 37297 nouns in morph of Punjabi nouns. Input text is standardized for different letters as: bindi at top ◌ , bindi at foot in ਫ਼ fa, ਖ਼ kẖa, ਸ਼ sha, ਜ਼ za, and ਲ਼ ḷa, ਗ਼ ġa, ◌ॅ aadak, , foot letters ◌ੁ in ਹ ha, ਵ v and ਰ ra.

II. RELATED WORK

Goyal et al. (2010) discussed normalization of different variations of spellings of text related to Hindi language while implementing system of machine translation, job for normalizing different variations in spellings for additional text processing is assumed doing very good task for enhancing efficiency of machine translation. This system used rules oriented technique in normalizing different variations in spellings for text related to Hindi while implementing translation technique from Hindi text to Punjabi text. They concluded that text of 7.45% was translated by this technique and we can say that this normalization job enhanced the efficiency of this system [3]. Alam et al. (2009) discussed normalization system which is rule oriented for Bangla and it contained semiotic classes detection, implementation of various rules used in generating tokens & verbalization, categorisation of tokens, word sense disambiguation and creation of standard terms[8]. Script of Python [9] was applied for recognizing class related to semiotic in corpus of news articles. Efficiency of this proposed rules oriented system is 90%. Yuxianget al. (2008) implemented various NSWs based on very big corpus of Chinese, and it suggested strategy of two steps NSWs disambiguation, for starting categorization automata of finite state & for disambiguation of subclass classifiers related to maximum entropy. Results of implementation proved that it attained the performance which is very good and can be applied to novel fields as well. But, there some mistakes appeared in these results of experiments like error related to number sequence [10]. Zhu et al. (2007) discussed tagging technique which is unified in nature for conducting the job by applying conditional random fields. They showed with creation of tiny tags , majority of this job

related to normalization of text can be done in this technique. Efficiency of this proposed technique is good, as different jobs for standardization are not dependent and these should be done together with each other [11]. Filip et al. (2006) showed different hurdles for standardization of text for Polish language which is also inflected one [12]. Panchapagesan et al. (2004) discussed new technique for normalization of text, but in creation of small tokens and starting categorization of tokens are mixed to one step, and 2nd step after this related to word sense disambiguation. In generation of small tokens & starting categorization, space which is white is applied mostly as separator among tokens & it is largely used in generating tokens [13]. Xydas et al. (2004) discussed new technique, stated as text to pronunciation used in absolutely standardization of terms which are non Standard in nature for texts which are unrestricted [14]. Wong et al. (2002) discussed the technique which applies dictionary related to addresses & the system for ranking on the basis of neural network which is analog for changing addresses which are free and are formatted to form which is canonical. Normalization related to address is type recognition problems for different patterns, here different sub fields are similar to feature vector in addresses, & this method is similar to categorizer having decision rules. It used scoring technique which was on the basis of neural nets which are analog in nature, here dictionary containing addresses used a technique which is digital [15]. Adda et al. (1997) discussed investigation which is quantitative in nature for standardization of text on linguistic based systems for recognizing the speech related to French language. This approach showed 11.2% of terms error rate for AUPELF "Frenchspeaking" which is type of test for evaluating the recognition of speech by applying two big dictionaries of French dictionaries i.e. DELAF & BDLEX [16]. Atwell et al. (2004) discussed approach where particular pattern of characters can be considered as single word or sub words group which we can further divide. Its different results shown were not good because of performance of categorizer, as small rate of recall reduced its rate of accuracy [9].

III. AUTOMATIC NORMALIZATION OF PUNJABI NOUNS

Punjabi is the official language for state of Punjab. But Punjabi is under resource language. There are very less number of computational-linguistic resources available for Punjabi. But a lot of research is going on for developing NLP applications in Punjabi language. This is 1st in history that automatic standardization of terms related to Punjabi is implemented and this system can be very much useful in creating other applications for Punjabi having good efficiency. For example it can be applied in different NLP applications like machine translation, document association, documents clustering, topic tracking and text summarization etc. [7].

For applying standardization of nouns present in Punjabi morph and text present in input, change presence of different characters as bindi on top ◌ਂ, aadak ◌ੱ & bindi on foot ◌੍ with null & also change presence of Punjabi foot letters i.e. ◌੍ with any of suitable ਹ ha, ਵ v or ਰ ra letters. Table I. shows some of Punjabi nouns in input text and noun morph which need to be normalized.

TABLE I
SAMPLE LIST OF PUNJABI NOUNS FOR NORMALIZATION

Characters to be normalized	**Example Punjabi noun words**
◌ੱ aadak	ਟੱਬ ṭabb "tub", ਗੁੱਟguṭṭ "group", ਜੱਜ jajj "judge", ਪੱਤਰ pattar "letter", ਸਲਿੱਪ salipp "slip", ਸੱਭਿਅਤਾ sabbhitā "culture" and ਹੱਡ haḍḍ "bone" etc.
◌ਂ bindi at top	ਕਵਰਤੋਂ kavartōṃ "misuse", ਸਰਗਰਮੀਆਂ sargarmīāṃ "activities", ਸਰਕਾਰਾਂ sarkārāṃ "governments", ਹਾਕਮਾਂ hākmāṃ "rulers", ਹੰਟਰਾਂ haṇṭrāṃ (whips) and ਹਿੰਦੂਆਂ hindūāṃ "to follower of hinduism" etc
◌੍ foot characters	ਆਕ੍ਰਿਤੀ ākritī "shape", ਸ੍ਵਰ svar "sound", ਅੰਮ੍ਰਿਤੀ ammritī "a kind of sweetmeat", ਅੰਤਰਦ੍ਰਿਸ਼ਟੀ antradrishṭī "inner vision", ਆਲ੍ਹਣਾ ālhṇā "nest", ਸ੍ਵਸਥਤਾ svasathtā "healthiness", ਸ੍ਵੱਛਤਾsvachtā "cleanness" and ਸ੍ਵਰਗ svarag "heaven" etc.
◌ੰ bindi at foot	ਜ਼ਮਾਨਤ jaॅmānat "bail", ਖ਼ਿਆਲ khaiāl "idea" and ਜ਼ਿਲ੍ਹਾ jailhā "district" etc.

Various normalization rules for Punjabi nouns have been shown in Table II.

TABLE I I

NORMALIZATION RULES FOR PUNJABI NOUNS

Characters to be replaced	Characters replaced with	Example normalization rules for Punjabi noun words
ੱ aadak	null character	ਟੱਬ *ṭabb* →ਟਬ *ṭab* , ਗੁੱਟ*guṭṭ* →ਗੁਟ *guṭ*, ਜੱਜ *jajj* →ਜਜ jaj, ਪੱਤਰ *pattra*→ਪਤਰ *ptar*, ਸਲਿੱਪ *salipp* →ਸਲਿਪ *salip* and ਸੱਭਿਅਤਾ *sabbhitā*→ਸਭਿਅਤਾ *sabhitā* etc.
ੰ bindi at top	null character	ਕਵਰਤੋਂ *kavratōṃ*→ਕਵਰਤੋ *kvartō*,ਸਰਗਰਮੀਆਂ *sargarmīāṃ*→ਸਰਗਰਮੀਆ *sragramīā* and ਸਰਕਾਰਾਂ *sarkārāṃ*→ਸਰਕਾਰਾ *srakārā* etc
੍ Punjabi foot characters	ਰ or ਵ or ਹ	ਆਕ੍ਰਿਤੀ *ākritī*→ਆਕਰਿਤੀ *ākritī*, ਸ੍ਵਰ *svar*→ਸਵਰ *savar*, ਅੰਮ੍ਰਿਤੀ *ammritī*→ਅੰਮਰਿਤੀ *ammaritī*, ਆਲ੍ਹਣਾ *ālhṇā*→ਆਲਹਣਾ *ālhaṇā*, and ਸ੍ਵਸਥਤਾ *svasathtā*→ਸਵਸਥਤਾ *savsathtā* etc.
੍ bindi at foot	null character	ਜ਼ਮਾਨਤ *ja◌्mānta*→ਜਮਾਨਤ *jmānat*, ਖ਼ਿਆਲ *khiāl*→ਖਿਆਲ *khiāl* and ਜ਼ਿਲ੍ਹਾ *jailhā* → ਜਿਲ੍ਹਾ *jilhā* etc.

Standardization Procedure for Nouns in Punjabi language:
This procedure for standardization of nouns in Punjabi works after storing complete morph of Punjabi nouns to some other table called morph_standardized_nouns. For every noun term present in morph_standardized_nouns we can proceed as [17]:

Step 1 : Change presence of all letters of aadak ੱ with letter null.

Step 2: Change presence of all letters of bindi on top ੰ with null letter.

Step 3: Change presence of all Punjabi foot letters ੍ with suitable ਵ (v) or ਹ (ha) or ਰ (ra) letters.

Step 4: Change presence of all letters of bindi on foot ◌੍ with letter null.

Step5: morph_standardized_nouns is normalized

Step 6: End of this procedure

Algorithm Input: ਸੱਭਿਅਤਾ sabbhitā, ਸ੍ਵਰ svar, ਜ਼ਿਲ੍ਹਾ jailhā and ਆਲ੍ਹਣਾ ālhṇā

Algorithm Output: ਸਭਿਅਤਾ sabhitā ਸਵਰ savar, ਜਿਲ੍ਹਾ jilhā and ਆਲਹਣਾ ālhaṇā

IV. IMPLEMANTATION AND ANALYSIS OF RESULTS

We have developed and implemented this automatic normalization of Punjabi noun using access as database and at front end ASP.NET. we have implemented a module which is offline for creating normalization database. This standardization system uses different normalization rules, on 50 news articles of Punjabi. In this way it stores nouns which are non standard and standard terms retrieved while processing and analysing news articles of Punjabi with their count in the database. After this, Punjabi nouns with highest count in its variants of spellings are treated as standard. Then standard term can also be changed with its other variant when count of this variant increases the count of present standard term. In this way, normalization database is generated.

Upon analysing corpus of Punjabi articles [6] it is found that there are small variations in spelling of nouns in Punjabi. It is found that nouns of 1.562% are with spelling variations. Table III represents in 1.562% terms, %age of terms with three, two and one variations from corpus of news articles in Punjabi.

TABLE III

PERCENTAGE OF WORD OCCURRENCES WITH SPELLING VARIATION COUNT

Number of Variants	Words Frequency (%)	Example
1	99.95	ਪੰਜਾਲ਼ੀ *pañjālī* “yoke for a pair of bullocks” ਪੰਜਾਲੀ *pañjālī* “yoke for a pair of bullocks”
2	0.046	ਉੱਖਲ਼ੀ *ukkhlaī* “mortar for pounding grain” ਉਖਲ਼ੀ *ukhlaī* “mortar for pounding grain” ਉੱਖਲੀ *ukkhlī* “mortar for pounding grain”
3	0.004	ਅੰਗਰੇਜੀ *aṅgrējī* “English” ਅੰਗਰੇਜ਼ੀ *aṅgrēzī* “English” ਅੰਗ੍ਰੇਜੀ *aṅgrējī* “English” ਅੰਗ੍ਰੇਜ਼ੀ *aṅgrēzī* “English”

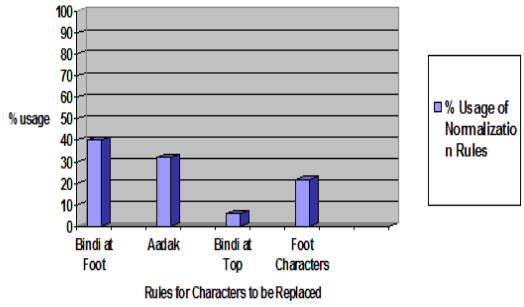

Fig. 1 Analysis of usage of various normalization rules

The above graph (Fig. 1) shows that rules for bindi on foot & letter aadak are having highest usage. The rule which is used very less is bindi on top. Foot letters rule is with maximum applicability i.e. 22% usage for normalization of nouns in Punjabi.

Table IV. carries the design of the database used for storing information about text normalization.

TABLE IV
TEXT STANDARDIZATION DESIGN OF DATABASE

Field Name	Description
nounTerm_ Non standard	It carries noun terms in Punjabi which are non standard.
freq_ Non standard_Term	Count of noun terms in Punjabi which are non standard nalysed in fifty Punjabi news documents
nounTerm_ Standard	Stores Punjabi noun words with standard spellings
freq_ standard_Term	The frequency of the standard Punjabi noun words analysed in fifty Punjabi news documents

Table V shows the sample entries for Text Normalization Database.

TABLE V

SAMPLE ENTRIES OF TEXT NORMALIZATION DATABASE

Non Standard Noun Word	Non Standard Noun Freq	Standard Noun Word	Standard Noun Freq
ਖ਼ਿਆਲ khaiāl “idea”	44	ਖਿਆਲ khiāl “idea”	118
ਆਕਰਿਤੀ ākritī “shape”	3	ਆਕ੍ਰਿਤੀ ākritī “shape”	85
ਪੰਜਾ਼ਲੀ pañjālī “yoke for a pair of bullocks”	33	ਪੰਜਾਲੀ pañjālī “yoke for a pair of bullocks”	97
ਕਵਰਤੋ kvartō “misue”	11	ਕਵਰਤੋਂ kavratōm̐ “misuse”	92
ਜਜ jaj “judge”	5	ਜੱਜ jajj “judge”	37

V. CONCLUSIONS

Most of languages in India are having many variations in spellings of their terms. This variation is caused by different factors and hence it is very difficult to implement such type of systems which are having any variations in their spellings. So this system has been implemented for resolving this issue by standardizing variations in spellings for Punjabi language. This system may be combined with other system of NLP for solving their job of pre-processing. Moreover, already It is combined with automatic text summarizer in Punjabi. This system is much beneficial for increasing efficiency of this Punjabi summarizer. There are very less number of computational-linguistic resources available for Punjabi. This is 1st in history that automatic standardization of terms related to Punjabi is implemented and this system can be very much useful in creating other applications for Punjabi having good efficiency. For example it can be applied in different NLP applications like machine translation, document association, documents clustering, topic tracking and text summarization etc.

REFERENCES

[1] M. W. Berry, “Survey of Text Mining: Clustering, Classification and Retrieval,” Springer Verlag, LLC, New York, 2004.
[2] Farshad Kyoomarsi, Hamid Khosravi, Esfandiar Eslami and Pooya Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic,” In proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman, UK, pp. 347-352, 2008.
[3] Vishal Goyal and Gurpreet Singh Lehal, “Automatic standardization of spelling variations of Hindi Text,” In Proceedings of international conference IEEE ICCCT’10, pp. 764-767, 2010.
[4] Praveen Kumar, Ankush Mittal and Sumit Gupta, “A query answering system for E-learning Hindi documents,” South Asian Language Review, vol. 13, 2003.
[5] Gurmukh Singh, Mukhtiar S. Gill and S.S. Joshi, “Punjabi to English Bilingual Dictionary,” Punjabi University Patiala, 1999.
[6] Punjabi Unique word Corpus.
[7] Joel Neto, “Document Clustering and Text Summarization,” Proceedings of 4th International Conference on Practical Applications of Knowledge Discovery and Data Mining, pp. 41-55, London, 2000.
[8] F. Alam, M. Khan, S.M. Habib, and Murtoza,, “Text Normalization system for Bangla,” In: Conference on Language and Technology, 2009.
[9] H. Koo, L. Moran, S. Atwell & Tae-Jin Yoon, “Text Normalization in Python” www.linguistics.uiuc.edu/grads/moran/papers/TextNorm.pdf.
[10] D. Huang, H. Wang, S. Yu, Wu Liu, Y. Jia, D. Yuan, “Text Normalization in Mandarin Text-To-Speech System,” In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP ’08, pp. 4693 – 4696, IEEE Press, New York, 2008.
[11] C. Zhu, H. Li, Hwee Tou Ng, Tie-Jun Zhao and Jie Tang, “A Unified Tagging Approach to Text Normalization,” In Proceedings of ACL'07, pp.688-695, 2007.
[12] G. Filip, W. Agnieszka, W. Mikołaj, J. Krzysztof, “Text Normalization as a Special Case of Machine Translation,” In Proceedings of

International Multi Conference on Computer Science and Information Technology, pp.51–56, 2006.

[13] A.G. Ramakrishnan, K. Panchapagesan, N.S. Krishna, P.P. Talukdar and K. Bali, "Hindi Text Normalization," In Proceedings of fifth International Conference on Knowledge Base Computer Systems, 2004.

[14] G. Xydas, G. Karberis and G. Kouroupertroglou, "Text Normalization for the Pronunciation of Non-Standard Words in an Inflected Language," In 3rd Hellenic Conference on Artificial Intelligence SETN '04, Samos, Greece, pp.390-399, 2004.

[15] M.C. Chuah and W.S. Wong, "A Hybrid Approach to Address Normalization, " IEEE Intelligent Systems, vol.9, pp. 38-45, 1994.

[16] G. Adda, M.D. Adda ., J.L. Gauvain, and L. Lamel, "Text Normalization and Speech Recognition in French," In Proceedings of ESCA Eurospeech , 1997.

[17] V. Gupta and G.S. Lehal, " Automatic Text Summarization System for Punjabi Language," Journal of Emerging Technologies in Web Intelligence, vol. 5, pp. 257-271, 2013