Noise Issues in Sentence Structure for Morphological Analysis of English Language Sentences for Hindi Language Users

Seema Shukla and Usha Sinha

***Abstract*—This paper identifies some issues in English Language sentences which are interpreted by Hindi speakers. Sentences may seem grammatically correct but since they may not have equivalent constructs in Hindi Language, it may be difficult for NLP processes to interpret as correctly as human mind. This gap of knowledge transfer from a language to another by NLP processes would need additional knowledge base. Often, NLP systems need to use such knowledge base either as rule base or empirical formulations identified out of statistical methods on large set of bilingual corpus. Bilingual parallel corpus, though essential, is not easily available. Grammar mapping of a language to another is also difficult. The structures in a sentence which may not have proper mapping can be viewed as noise. 1000 unique English Language sentences from a 460000 word corpus were identified as representative sentences. These sentences were translated manually as well as using Machine Translation System. The outputs were compared to find out most common issues wherein MT did not interpret as correctly as human being. This misinterpretation by NLP system has been marked as noise. This paper identifies ten categories of such noises.**

***Index Terms*—NLP processes, knowledge base, bilingual corpus, grammar mapping, noise, machine translation, recursive transition networks (RTN), finite state transducers (FST).**

I. Introduction

Research in NLP, over decades, can be overviewed to conclude that efficacy of NLP systems such as Machine Translation, Auto-summarization, Auto-tagging, etc. can never be perfect for general domain. However, significant amount of efficacy can be brought out by "domainizing" the approach [1]. However, domainizing does not often solve the problem since general domain part continues to be integral part of the corpus within a specific domain. Therefore, scientific studies need be carried out for sentence structure analysis and word level morphology together.

Construction of sentences are affected not only by culture, but also by creator's mother tongue, particularly by what person has learnt as a language during his/her childhood. It also gets affected by the way emphasis is laid down in a sentence through set of words. This is so because normally people do not create the sentence but translate what they "think" in their native language(s). India being a multilingual country, people speak and write sentences of mixed forms. For example, Northern Indian would know Hindi, Punjabi and English. The sentences get created in one language with mix of these three languages, not only at word level but at construction of the sentence level too. E.g. ट्रेन लेट है. मेल सेंड कर दो, etc. This type of influence while creating sentence can be seen as "noise" [2], so that correct language sentence could be derived after identifying this noise and not only removing it at word or phrase level but also by removing its impact on other words in a sentence, which will result in formulation of a correct sentence. Computer algorithms, being static, do not have enough knowledge base, to understand ill framed sentences. Sentence structure and/or word level morphological analysis done by these algorithms may not produce correct information for the main program to support the objective of the system (such as MT). Hence, the identification and categorization of noise is necessary for improving knowledge base of algorithms. This paper proposes one methodology for categorization of such noise. To support this methodology for noise categorization an empirical architect is also proposed.

Manuscript received September 19, 2014; revised January 28, 2015.

The authors are with the Linguistics Dept., Lucknow University, Lucknow, India (e-mail: seemashukla@jssaten.ac.in).

II. Methodology for Noise Categorization

Fig. 1 illustrates the methodology used for noise categorization.

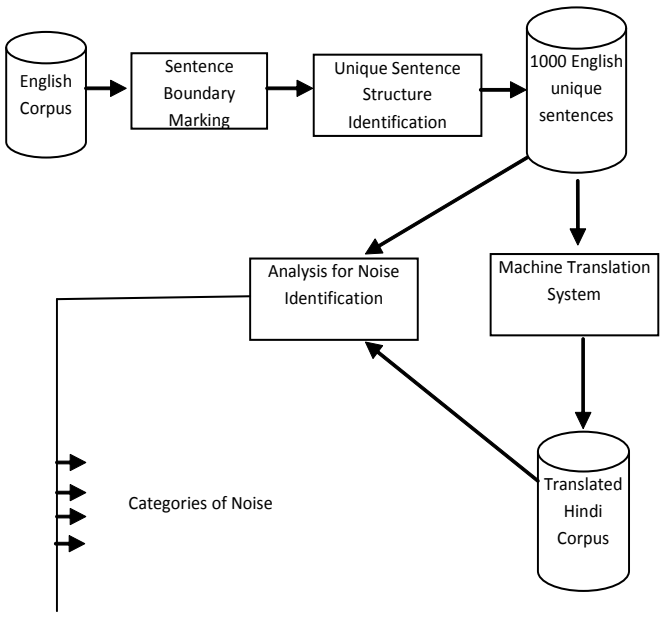

Fig. 1. Methodology for noise categorization.

Broad tourism related English Language corpus of about 460000 words was collected from various sources. It was

analyzed in a semi-automatic way with the help of specifically written computer programs to identify similar structure sentences. About 1000 sentences of unique structure were identified. Hindi translation of these sentences was obtained using Machine Translation System [3]. The incorrectly translated sentences were analyzed to identify different categories of noise.

III. Noise Categorization

The Categorization is quite general and is specific to chosen English Hindi language pair. That means, the mother tongue of creator is considered as Hindi (that too not perfect for its grammar) and sentence creations are considered in English. It is not necessary that the created English sentence may be wrong by its grammar but may be difficult for NLP systems to understand and comprehend correctly. Some categories identified are discussed below. Use the "Body text" style for all paragraphs.

A. Absence of Subject in Second Part of the Sentence Which Is a Partial Sentence

'And' is a conjunct connecting different entities such as nouns, noun phrasals, verb, verb phrasals, etc. It can also connect two complete sentences together. If so, often, the subject part of the second sentence is not explicitly mentioned as it is usually (indirectly) considered well connected with the subject of the first sentence. It would result in and conjunct being used between a complete sentence (SVO) with partial sentence (VO). E.g. "Temple at Somnathpur is well preserved and is not to be missed." Although sentence seems correct grammatically for NLP systems it should be "Temple at Somnathpur is well preserved and it is not to be missed." Usually, in Hindi language, constructs in which subject in second part is missing are not allowed grammatically.The ISBN assigned: 978-1-84626-xxx-x, etc.

B. Absence of Subject and Verb in the Second Part of the Sentence

In this category besides subject, verb is also missing from the second part of the sentence. E.g. "Temple at Somnathpur is well preserved and not to be missed". Although sentence seems correct grammatically for NLP systems it should be "Temple at Somnathpur is well preserved and it is not to be missed."

C. Occurrence of Adverb before Verb Phrase

English language grammar permits occurrence of adverb after verb, whereas, in Hindi, adverb, since it emphasizes verbs, appears before verb. Therefore, in Indian English it is seen that adverbs appear before verbs. This leads to identifying the adverb as adjective or modified form of the noun by NLP systems. e.g. "He quickly left" since in Hindi, one might be thinking "वो जल्दी चला गया".

D. Missing Comma to Represent Some wh- Form such as Which, Who etc.

The sentence "Pune 180 km from Mumbai is a culturally rich city" should be written as "Pune, 180 km from Mumbai, is a culturally rich city" or "Pune which is 180 km from Mumbai is a culturally rich city." The original sentence needs semantic knowledge for its clear understanding. Human mind may do so easily but NLP systems may not have enough structured knowledge for understanding such semantics. Even if the knowledge is added by some way to the NLP system, it may be misused for some other cases/occurrences. The mid-path is sometimes chosen to add additional semantic knowledge to the "comma" which may not be misused by NLP systems.

E. Use of Plural Nouns as Singular (Could be a Proper Noun) Which Needs Appropriate Form of a Verb in the Sentence

Proper noun may end with an "s" making it look like the plural form of a noun. In such cases sentence may be formulated using incorrect form of verb. E.g "Indian Airlines connect Mumbai with Delhi, Kolkata, Chennai and Bangalore" whereas the correct form is "Indian Airlines connects Mumbai with Delhi, kolkata, Chennai and Bangalore". Another example is "National Parks and Monuments include Big Cypress Reserve, Biscayne, Dry Tortugas, Everglades, Castillo de San Marcos, Fort Mantazas." The correct form is "National Parks and Monuments includes Big Cypress Reserve, Biscayne, Dry Tortugas, Everglades, Castillo de San Marcos, Fort Mantazas." NLP system is capable of understanding only correct form of the sentence.

F. Missing Which Is/That Is/Who Is/etc. as a Connector between the Complete Sentence and Partial Sentence

In the sentence, "In the Southeast corner is a small Hindu shrine honoring Laxmi, the Godess of wealth", the "the Godess of wealth" is a partial sentence emphasizing object of the previous complete sentence. NP system may not understand, if "Laxmi" or "the Godess of wealth" is the correct object of the given sentence while analyzing on SVO pattern. NLP system may not understand such SVOO pattern. There is no equivalent construct in Hindi for this sentence. The correct semantics is conveyed if the sentence is written as "In the Southeast corner is a small Hindu shrine honoring Laxmi who is the Godess of wealth". It can be interpreted as a noise of missing wh form as a connector. Another example is "In all tourist destination areas English is number 1 foreign language fairly spoken and written". Correct sentence should be "In all tourist destination areas English is number 1 foreign language which is fairly spoken and written."

G. Issues with "to-Noun" Phrase

In English sentences, constructs "to-noun" often are used e.g. "to MG Road" in a sentence "No trip to MG road is completed without a bite at the Pai Dosa Shop." NLP Systems may interpret construct "trip to MG road" like "Bombay to Goa", because "trip" is also a noun. The semantics totally gets messed up when this happens. To avoid such occurrence, dictionaries are required to include "to-noun" phrasals as explicit category in lexical resource.

H. Issues with "If + Adjective"

This construct is imperfect partial sentence and usually is used at the start of a sentence e.g. "If possible you should extend your leave." Since imperfect partial constructs do not

represent a complete sentence, it makes sense only human beings. NLP systems are often not capable of interpreting partial sentence constructs. The form of the sentence for NLP systems to interpret correctly would be "If it is possible you should extend your leave." Another such example is "Highly recommended for those who don't mind roughing it a bit" which should be "It is highly recommended for those who don't mind roughing it a bit."

I. Issues with "Verb + to" Constructs

In English there are some verbs which often act as noun e.g. March, May, etc. Normally NLP constructs and tools are strong enough to interpret them with their appropriate usage in a sentence. But when these verbs are prefixed or suffixed with certain prepositions (e.g. to), then it becomes difficult for NLP tools to interpret them correctly. E.g. in a sentence "April to March is a reverse order period", "to March" is often misinterpreted by NLP tools as a verb, whereas March in the sentence is a noun. Same is true for certain Hindi words to be interpreted back for absorption of knowledge to create English sentence e.g. "Bhai" in Hindi is noun as well as verb.

J. Issues with Sentence Initiators/Terminators

Sentence initiator or terminators often disturb language tools e.g. "all in all" (in a sentence, "All in all, Bangalore is a lovely city to visit"). Another example is "Of all, you have to say so." Usually, single word as initiator or terminator in a sentence can be handled by NLP tools, but if it becomes group of words (which may resemble, but may not be, a phrasal), it is difficult for the tool to interpret correctly. This is so because the emphasis in the meaning of the sentence brought in by initiator or terminator, may not be brought in in the same fashion in the target language. Placements may be changed. In the first example, "all in all" means "कुल मिला के". It may effectively be an initiator in Hindi language. But this is not true for the second example. "Of all" may not be translated correctly as an initiator for Hindi sentence.

IV. Proposed Empirical Architecture

The proposed architecture for empowering NLP tools to handle noise is shown in Fig. 2. Only word level morphology, grammatical information and syntactic structure are used for knowledge representation of a natural language sentence in current scenario [4]. The existing techniques of knowledge representation concentrate upon the knowledge at word level only without co-relating it with other words in a sentence. This work aims at finding an architecture for the representation of knowledge, which is more efficient in terms of expressing co-relation between word and its relation with other words in a sentence using techniques useful for AI systems such as Machine Translation, knowledge extraction, etc. The proposed architecture consists of An RTN [5] which is able to parse all the sentences within the corpus successfully and trace back the rules based on which each sentence is parsed.

An environment that works around FST [6], [7] to facilitate it to find out the knowledge area of each sentence in the corpus robustly.

Lex resource [8] containing grammatical/dictionary meaning and information of words and valid NPs and VPs.

Extended lex containing meaning /information of set of words (which are not chunks).

N-grams [9] of occurrence of words from large corpus.

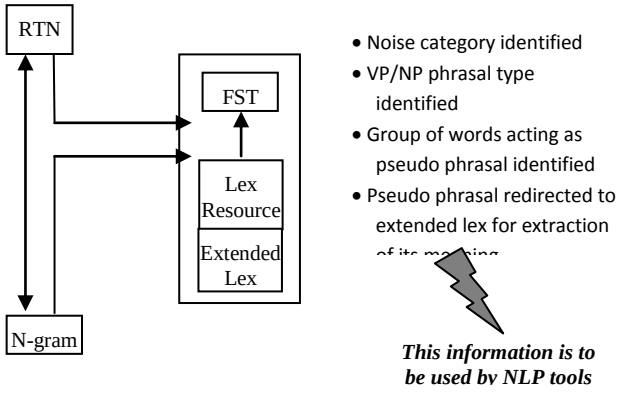

Fig. 2. Proposed empirical architecture for empowering NLP tools to handle noise.

The proposed architecture is suitable for English and Indian Languages particularly Indo-Aryan family.

For testing the architecture, rule based machine translation for English to Hindi translation was used to provide multiple outputs for representative 1000 English sentences. The multiple outputs were studied for occurrence of different noises. Word Level Morphological Analyzer and POS Tagger for "Hindi" language [10] have been successfully implemented for 91930 root words containing 50520 nouns, 81 pronouns, 33006 adjectives, 8513 verbs, 1559 base words 12 particles. The POS Tagger used was based on Maximum Entropy approach. The Morphological Analyzer gives about 95% accurate results and the POS Tagger gives the 87% accuracy for known sentences and about 80% accurate results for unknown sentences.

The RTN has been implemented for the selected 1000 English sentences and verified. Out of the set of corresponding 1000 Hindi sentences, some are used to create RTN. The two sample RTNs were compared for structural commonality to identify noise. The lex resource was used to create corresponding RTNs for set of nouns, verbs and phrasals. The information about n-grams was also used to refine observations and to authenticate application of appropriate rules.

The developed RTN and the n-gram statistics directs FST for producing appropriate information as given in Fig. 2. The lex resource provides regular meaning of the word and appropriate morphological information, if it is not part and parcel of noise. If, it is part and parcel of noise the produced results will not match with different information coming from n-gram, lex resource and RTN. Thus, the category of the noise will be visible. In extended lex, such set of words which formulate the noise component can be searched and certified to be noise. The NLP modules thus can pick up necessary information and meaning from either lex resource or extended lex resource, effectively handling the noise component.

V. Conclusion and Future Work

The corpus, though tourism related, represents, significantly, a general corpus. The corpus was scientifically

analyzed to generate lingual categories of noise which is specifically defined for bringing out clarity of divergence [11] between two languages i.e. English and Hindi. The identified categories are strong enough to provide necessary knowledge to NLP tools for syntactical and semantic understanding to help higher level systems to provide better efficacy.

This work could be progressed logically for

- Refining identified categories further for Hindi and other Indian Languages
- Finding more categories
- Structuring categories computationally for creating knowledge base for NLP systems.

REFERENCES

[1] R. Mahesh, K. Sinha, and A. Thaku, "How to get best results out of a machine translation system: A case study of English to Hindi Translation," *CSI Journal*, vol. 38, no. 4, Oct.-Dec. 2008.

[2] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, "A survey of types of text noise and techniques to handle noisy text," in *Proc. the Third Workshop on Analytics for Noisy Unstructured Text Data*, ACM, 2009, pp. 115-122.

[3] TDIL. Machine translation. Indian Language Technology Proliferation and Deployment Center. [Online]. Available: http://tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php

[4] H. Trost, X2MORF: A morphological component based on augmented two-level morphology, Research Report, 1991.

[5] A. James, *Natural Language Understanding*, 2nd ed. Pearson Education, 2008.

[6] H. Schmid, "A Programming language for finite state transducers," in *Proc. the 5th International Workshop on Finite State Methods in Natural Language Processing*, Helsinki, Finland, July 13, 2005, pp. 308-314.

[7] H. Schmid. Developing computational morphologies with the SFST tools. *Tutorial SFST Tool*. [Online]. Available: http://www.cis.uni-muenchen.de/~schmid/tools/SFST/data/SFST-Tutorial.pdf

[8] R. M. K. Sinha and A. Jain, "Angla Hindi: an English to Hindi machine-aided translation system," *MT Summit IX*, New Orleans, USA, pp. 494-497, 2003.

[9] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 2nd ed. Prentice Hall Inc., 2002.

[10] D. Kumar, M. Singh, and S. Shukla, "FST based morphological analyzer for Hindi language," *International Journal of Computer Science Issues*, vol. 9, issue 4, pp. 349-363, July 2012.

[11] V. Shukla and R. M. K. Sinha, "Divergence patterns for Urdu to English and English to Urdu translation," in *Proc. 8th International NLPCS Workshop*, Copenhagen Business School, Denmark, 20-21 August, 2011, pp. 21-28.

Seema Shukla was born in Moradabad, UP, India on 26 October, 1968. She completed her bachelor's degree in computer engineering from University of Pune, Maharashtra, India in 1993 and master degree of technology in computer science from Banasthali Vidyapeeth, Rajasthan, India in 2003. Currently, she is pursuing the PhD degree from Lucknow University, Lucknow, India in the area of computational linguistics.

She started her career as a computer programmer in New Bombay. Thereafter, she has worked as a project associate, systems officer and senior faculty in various organizations before joining JSS Academy of Technical Education, Noida (UP), India with which she has been associated for the past ten years.

Ms. Shukla is a member of IEEE and a life member of IETE.

Usha Sinha was born in Lucknow on 11 March, 1948. Her educational qualifications are masters in history, masters in linguistics and PhD in linguistics degrees with a proficiency in Sanskrit.

She joined Lucknow University, India in 1976 and retired in 2010. She served as the head of Linguistics Department since 1998. She published four books and a number of research papers. She has delivered lectures on a variety of topics such as general linguistics, dialectology, sociolinguistics, stylistics, technical terms, language teaching, translation and Hindi literature. She has organized many national conferences and seminars. She established the 'Shabdavali' group in Linguitics Department of Lucknow University with the help of Commission for Scientific and Technical Terminology, Government of India.

Dr. Sinha is a member of Board of Studies and Research Committees of many Indian University and Kelaniya University, Sri Lanka. She is a member of many linguistic societies and literary and social organizations. She is the recipient of Best Teacher award and many other awards.