I.J. Information Technology and Computer Science, **2016, 8, 75-86**
Published Online August 2016 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijitcs.2016.08.09

Typology for Linguistic Pattern in English-Hindi Journalistic Text Reuse

Aarti Kumar
Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal-462003, India
E-mail: aartikumar01@gmail.com, Mob: +919303132828

Sujoy Das
Department of Computer Applications, Maulana Azad National Institute of Technology, Bhopal-462003, India
E-mail: sujdas@gmail.com, Mob: +919826345195

Abstract—Linking and tracking news stories covering the same events written in different languages is a challenging task. In natural languages same information may be expressed in multiple ways and newspapers try to exploit this feature for making the news stories more appealing. It has been observed that the same news story is presented in same as well as in different language in different ways but normally the gist remains the same. Diversity of linguistic expressions presents a major challenge in identifying and tracking news stories covering the same events across languages, but doing so may provide rich and valuable resources as comparable and parallel corpora can be generated with this resource. In the case of Indian languages there exist limited language resources for Natural Language Processing and Information Retrieval tasks and identifying comparable and parallel documents would offer a potential source for deriving bilingual dictionaries and training statistical Machine Translation systems. Paraphrasing is the most common way of reproducing news stories and translated text is also a type of paraphrase. Prior to linking monolingual or bilingual news stories, these paraphrase types need to identified and classified to help researchers to devise techniques to solve these challenging problems. English-Hindi language pair not only differs in their scripts but also in their grammar and vocabulary. A number of paraphrase typologies have been built from the perspective of Natural Language Processing or for some or the other specific applications but as per the knowledge of the authors, no typology have been reported for English-Hindi cross language text reuse. In this paper a typology is formulated for cross lingual journalistic text reuse in English-Hindi. Typology unravels level of difficulties in English-Hindi mapping. It shall help in devising techniques for linking and tracking English-Hindi stories

Index Terms—Paraphrasing, typology, linguistic transformation, lexical, cross-lingual, journalistic text reuse.

I. INTRODUCTION

Newspapers report events that are taking place in any part of the world at more or less same time across different languages. Any news conveys same facts across language but news reporters try to incorporate their viewpoints according to their findings. Linking news stories covering the same events and with same content written in different languages may provide rich and valuable multilingual resources of both parallel and comparable text. Translation equivalents provides parallel fragments and paraphrases provides comparable fragments [2]. Guan and Yuan [29], while working with mislabeled data, have also emphasized on the importance of pattern classification in machine learning. In case of Indian languages there exist limited language resources for Natural Language Processing (NLP) and Information Retrieval (IR) tasks and identifying comparable and parallel documents would offer a potential source for deriving bilingual dictionaries and training statistical Machine Translation (MT) systems [2, 19]. Paraphrasing is the most common way of reproducing news stories. In paraphrasing, substitution for semantic equivalents and grammar, are performed over the text which make even similar contents difficult to identify. Although, linguistic transformations take place in the paraphrased sentence but meaning is still preserved. Translated text in a different language is also a type of special kind of paraphrasing [4]. In order to determine what paraphrasing types make text reuse detection harder to be revealed, analysis, identification and classification of the different types of paraphrasing strategies applied during the text re-use process is important. Typology is nothing but drawing boundaries among different paraphrase types, identifying their manifestations, going into depth to find their characterization and finally classifying them [4]. Building a Typology has been a tool for many NLP researchers to apprehend paraphrasing [24].

Knowledge of paraphrase typology will help in identifying and linking similar news stories by applying suitable techniques. It is also an important aspect in IR research which also deals with document representation languages and models, and finding similar matching contents from documents collections on the web [30]. Therefore in this paper paraphrases are identified across English-Hindi language and a Typology for English-Hindi journalist text reuse has been proposed. It is a pioneer work in context with English-Hindi journalist text reuse and unravels level of difficulties in English-Hindi mapping. The proposed typology has been built by considering other monolingual typologies given by different authors and comprises of previously defined categories in addition to many new categories to encompass the unique cases of English-Hindi cross-lingual journalistic reuse. The existing typologies have either been mapped in context to English Hindi language or are modified according to the intrinsic representation of the transformation across these two languages. The proposed typology may help in devising techniques for linking similar paraphrased contents in in English Hindi document pairs.

The rest of the paper is structured as follows: In Section II various monolingual typologies given by different authors are discussed. In Section III proposed paraphrase typology for English-Hindi journalistic text reuse is discussed. Section IV presents discussion on the typology classes in context to empirical evidence and Section V presents the conclusion.

II. Chronological Related Work

Early works on paraphrase typologies are by Culicover [9] in1968 and Honeck [17] in 1971. They divided paraphrase types into those classes which can either be formally mapped in natural language processing or cannot be.

Culicover [9] logically grouped paraphrasing into five types and separated accessible paraphrase relationships from inaccessible ones.

A taxonomy in the fields of Psychology was given by Honeck [17] which classified three types of paraphrases including transformational, lexical and formalexic.

As reported by Vila et al.[23] in 2011 Apresjan (1973) mainly dealt with lexical paraphrases and Martin (1976) focused mainly on connotation, opposition and synonymy based paraphrases.

An editing taxonomy has been given by Faigley & Witte [15] in 1981 which divides revisions into two major categories; surface changes and meaning changes, each of which have 2 subcategories finally culminating into 23 types at the lowest level.

Dras [13] in 1999 studied syntactic paraphrases using Synchronous Tree Adjoining Grammars and classified paraphrasing types into classes based either on the formal change observed in the paraphrase pair or according to the paraphrase effect which makes them not mutually exclusive. The five classes of paraphrase that he identified are Change of Perspective, Change of Emphasis, Change of Relation, Deletion, and Clause Movement which are further divided into 51 sub-types.

Barzilay et al. [5] in 1999, Dolan et al. [11] in 2004 and Dutrey et al. [14] in 2011 gave an NLP typology of the most frequent types in a corpus whereas Kozlowski et al. [18] in 2003, Dorr et al. [12] in 2004 and Boonthum [7] in 2004 concentrated on the paraphrases that NLP addresses. Rinaldi et al. [21] in 2003 focused on classic paraphrases with illustrative purposes.

16 obfuscation types were reported by Clough [8] in 2003 in his paraphrase typology which dealt with text reuse.

Conversives, non-literal language use and extended paraphrases were studied by Dorr et al. [12] in 2004 while dealing with paraphrases with equivalent meanings. He focused on the syntax, lexicon, and grammatical features of the paraphrases.

Based on the type of linguistic units or the range of difference between the original and paraphrased sentences Shimohata [22] in 2004 has classified the paraphrase into three types only-Sentential, Phrasal and Lexical. Each paraphrasing type requires a different kind of knowledge to deal with. Sentential paraphrasing requires pragmatic knowledge, phrasal paraphrasing requires syntactic knowledge, and lexical paraphrasing requires lexical knowledge

Fujita [16] in 2005 analyzed a variety of linguistic phenomena in Japanese and provided a more detailed classification of paraphrases than in Shimohata [22]. He classified them on the basis of their similarities and differences in syntactic characteristics. He presented a classification of lexical and structural paraphrases grouped into six classes including paraphrases of single content words, function-expressional paraphrases, paraphrases of compound expressions, clause-structural paraphrases, multi-clausal paraphrases, and paraphrases of idiosyncratic expressions. These where further subdivided into 24 types.

Barreiro [3] in 2008 divided paraphrases into 5 classes- referential, lexical, phrasal, syntactic, lexical-syntactic and paraphrasing of multiword expression. The typology is based on the extent of paraphrasing within a sentence ranging from a single lexicon to a phrase to more than one phrase or more than one level of paraphrasing.

Clough and Gaizauskas [25] in 2009 studied journalistic text reuse and gathered three recurrently applied operations which are analogous to some entries of their typology: deletion, lexical substitution, changes in syntax and summarization.

A general typology of quasi-paraphrases together with their relative frequencies has been given by Bhagat [6] in 2009. The basis of classification of paraphrases is lexical and each of the types of paraphrase is linked to the compositional alterations involved.

Marta Vila et al. [23] in 2011 hypothesize that there exists a correlation between the differences in propositional content and the differences in wording on the one hand, and the degree of sameness of meaning or paraphrasability on the other, both being gradual properties. The typology they have presented classifies paraphrases according to the linguistic nature of their difference in wording and consists of a two-level typology of 2 paraphrasing types grouped into 5 classes. Paraphrasing types reflect a general paraphrase mechanism and classes represent the level of language where this mechanism takes place.

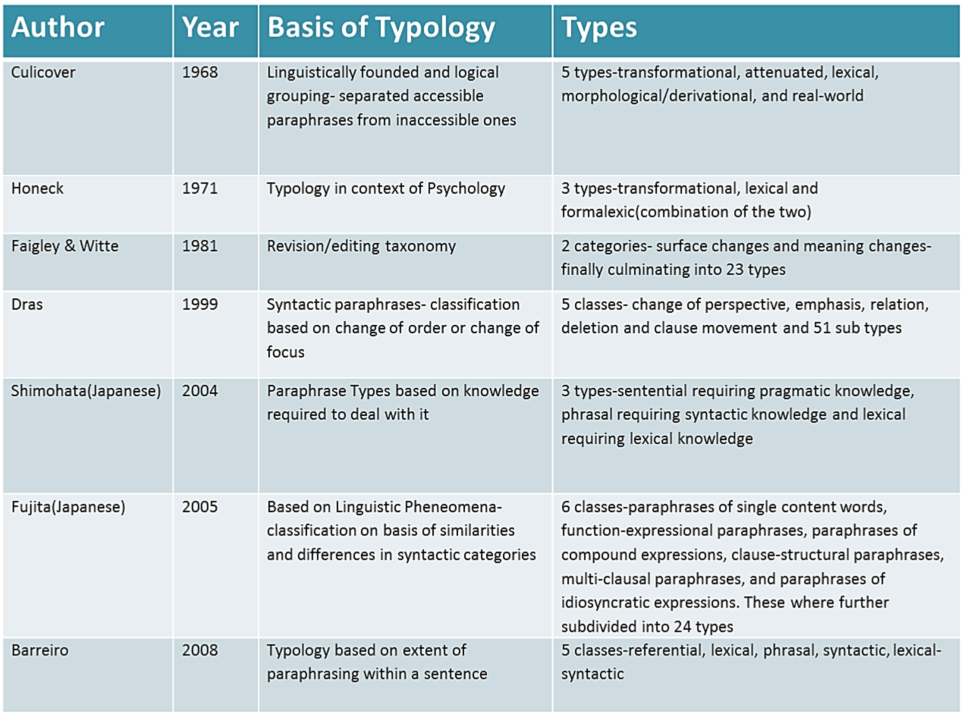
The paraphrase typology given by Barron et al. [4] in

2012 relies on the paraphrase concept defined by Recasens and Vila [20] in 2010 and Vila et al. [23] in 2011. It consists of an upgraded version of the one presented in the latter. Their typology also consists of a two-level typology but of 20 paraphrase types instead of 9 there grouped into six classes instead of 5.

Marta Vila et al. [24] in 2014 in their recent work refined their former typology and have given a new three level typology of 24 paraphrase types grouped in 5 classes.

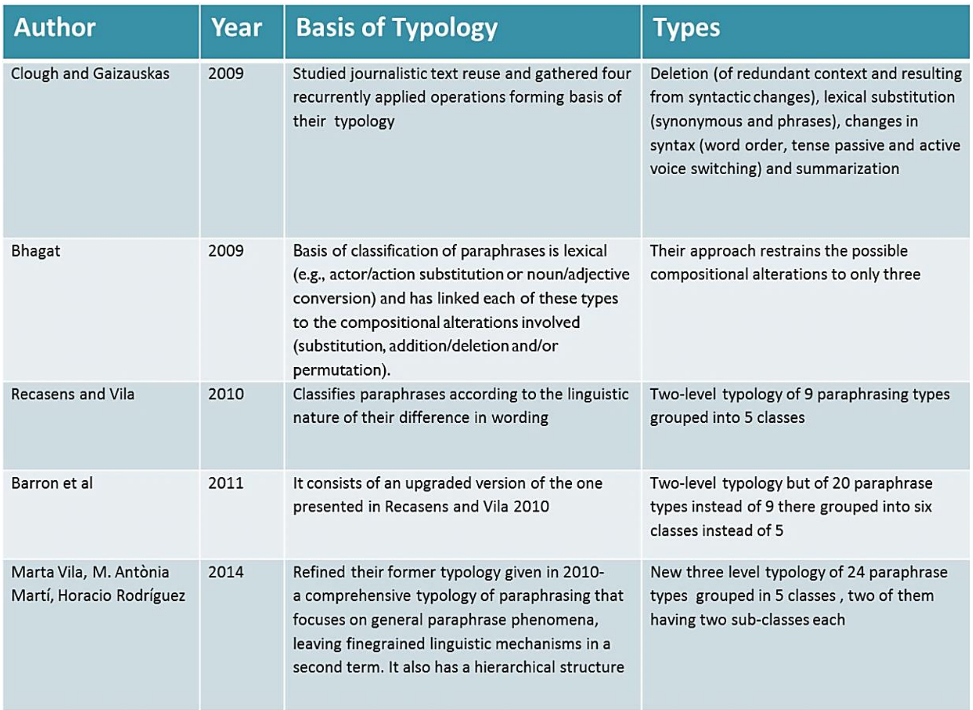
The paraphrase typologies and their basis are compiled in Fig. 1a and Fig. 1b.

Although some work has been done towards finding text reuse or linking news stories in English-Hindi but as per the knowledge of the authors, no paraphrase typology for these two language pairs has been reported so far. Also, the work done in these two language pair is directly proportional to the tasks defined by FIRE since 2009.



Author	Year	Basis of Typology	Types
Culicover	1968	Linguistically founded and logical grouping- separated accessible paraphrases from inaccessible ones	5 types-transformational, attenuated, lexical, morphological/derivational, and real-world
Honeck	1971	Typology in context of Psychology	3 types-transformational, lexical and formalexic(combination of the two)
Faigley & Witte	1981	Revision/editing taxonomy	2 categories- surface changes and meaning changes- finally culminating into 23 types
Dras	1999	Syntactic paraphrases- classification based on change of order or change of focus	5 classes- change of perspective, emphasis, relation, deletion and clause movement and 51 sub types
Shimohata(Japanese)	2004	Paraphrase Types based on knowledge required to deal with it	3 types-sentential requiring pragmatic knowledge, phrasal requiring syntactic knowledge and lexical requiring lexical knowledge
Fujita(Japanese)	2005	Based on Linguistic Pheneomena- classification on basis of similarities and differences in syntactic categories	6 classes-paraphrases of single content words, function-expressional paraphrases, paraphrases of compound expressions, clause-structural paraphrases, multi-clausal paraphrases, and paraphrases of idiosyncratic expressions. These where further subdivided into 24 types
Barreiro	2008	Typology based on extent of paraphrasing within a sentence	5 classes-referential, lexical, phrasal, syntactic, lexical-syntactic

Fig.1a. Paraphrase Typologies



Author	Year	Basis of Typology	Types
Clough and Gaizauskas	2009	Studied journalistic text reuse and gathered four recurrently applied operations forming basis of their typology	Deletion (of redundant context and resulting from syntactic changes), lexical substitution (synonymous and phrases), changes in syntax (word order, tense passive and active voice switching) and summarization
Bhagat	2009	Basis of classification of paraphrases is lexical (e.g., actor/action substitution or noun/adjective conversion) and has linked each of these types to the compositional alterations involved (substitution, addition/deletion and/or permutation).	Their approach restrains the possible compositional alterations to only three
Recasens and Vila	2010	Classifies paraphrases according to the linguistic nature of their difference in wording	Two-level typology of 9 paraphrasing types grouped into 5 classes
Barron et al	2011	It consists of an upgraded version of the one presented in Recasens and Vila 2010	Two-level typology but of 20 paraphrase types instead of 9 there grouped into six classes instead of 5
Marta Vila, M. Antònia Martí, Horacio Rodríguez	2014	Refined their former typology given in 2010- a comprehensive typology of paraphrasing that focuses on general paraphrase phenomena, leaving finegrained linguistic mechanisms in a second term. It also has a hierarchical structure	New three level typology of 24 paraphrase types grouped in 5 classes , two of them having two sub-classes each

Fig.1b. Paraphrase Typologies

English and Hindi languages not only differ in scripts but also in their grammar and vocabulary. English stores the meaning of the words in positions whereas Hindi, in morphemes. Identifying equivalent translated text across language becomes a challenging task as this category of text can be treated as obfuscation as well as paraphrasing. Identifying parallel contents in cross language news becomes even more complex if too much of alternation has been done to the translated news stories.

III. Proposed Typology

Although a pioneer work in the field of English-Hindi language text reuse, the typology has been built by considering other monolingual typologies covered in the related work section. It aims to cover most of the phenomena described in these typologies. As the works of other authors, referred in this research paper, are primarily based on monolingual paraphrasing, therefore some classes that are not finding relevance across the language are dropped here. In the proposed typology, apart from inclusion of some of the previously defined categories, some new categories are introduced by us to signify their importance for cross language text reuse. The previously defined categories are followed by citations of the authors who have proposed them. Categories without any citation are the new categories proposed by us.

The typology is strictly formulated for cross lingual news stories covering English-Hindi language. Cross Language Indian News Story Search (CLINSS) corpus[1] of FIRE 2012 and 2013 with 50691 files, English newspaper Hindustan Times and Hindi newspaper Dainik Bhaskar has been used as the corpus for the study and for inferring a typology for cross language news story. The parallel stories have been extracted from these newspapers manually and have been retrieved from CLINSS corpus using relevance judgment file provided by them. Text alignment was done manually by the authors themselves. The categories are classified to be in isolation but some of them overlap i.e. two classes can co-exist. For example, if there is a sentence split, there is addition of words also. Any paraphrased parallel sentence in majority of the reported news is a combination of more than one such category. Still, while discussing any particular category of typology, only that category of paraphrasing is emphasized at that point.

The classification has been done on the basis of extent of words in the sentences which are paraphrased and on the basis of difficulty in automatic identification of cross lingual news stories. Five difficulty levels have been identified and each level describes the extent of paraphrasing.

The Hindi words/phrases/sentences which have been used as examples under each level also have their transliterated English versions following them, within brackets, for the ease of understanding by those who are not the native speakers of Hindi language.

As the following examples have been taken from original news stories, some names have been changed/hidden wherever found necessary, in view of keeping work purely for the purpose of research and not to hurt any sentiments.

A. Level I

News stories that are **almost exact translations** of their English counterpart fall under this category. 1(b) and 2(b) are nearly exact Hindi translation of 1(a) and 2(a). In such cases simple dictionary based cross language approach may be fruitful to retrieve same news story for text reuse

1 a). Palson owner convicted in attempt-to-murder case

1 b). हत्या की कोशिश केस में पालसन के मालिक दोषी

(*hatya ki koshish mein palson ke maalik doshi*)

2 a). We wanted to know where all were the camps, who were in charge.

2 b). हम जानना जाहते थे कि कैंप कहां-कहां लगाए गए थे और उनका इंचार्ज कौन था।

(*hum janana chahte the ki camp kahan kahan lagaye gaye the aur unka incharge kaun tha*)

B. Level II

In this level key content words in Hindi are unambiguously mapped from E_1 to H_1 set but sentences may have a few additions/deletions of words or trivial modifications in one language or other. Linking Cross language news stories needs to map these words. Gist or meaning in this level is preserved. Categories identified under this level have also been reported in monolingual text reuse.

B.1. Word Insertion/Deletion

New information is **added** to a sentence by adding or deleting words [4, 23], leading to a paraphrase at the time of cross language text reuse (3(a) & 3(b) and 4(a) & 4(b)). It may have minor syntactic transformation or lexical replacement. Robin [26] in 1994 introduced the term 'Information adding' paraphrases for such type of paraphrase.

3 a). The base price of $225-million remains the same.

3 b). उन्होंने कहा कि टीमों का बेस प्राइस 22.5 करोड़ डॉलर यानी करीब 10 अरब रुपये ही रहेगा

(*unhone kaha ki teamon ka base price 22.5 karod dollar yani kareeb 10 arab rupye hi rahega*)

4 a). This has remained secret until now.

4 b). दस्तावेजों के इस तरह नष्ट या गुम हो जाने का राज आज तक बना हुआ था।

[1] http://users.dsic.upv.es/grupos/nle/clinss.html

(*dastavejon ke is tarah nasht ya gum ho jaane ka raaj aaj tak bana hua tha*)

In 3(b) and 4(b) underlined lexical units are added but same information of 3 (a) and 4 (a) is communicated. In 3 (b) few lexical units are transliterated instead of translation such as "*base price*" is transliterated as "*बेस प्राइस*".

In case of **deletion** of lexical unit normally words in a sentence that are superfluous or peripheral in sentence are removed. The constituents deleted are: hedging verbs, relative pronouns etc. [13]. In 5 (b) "*will completely*" and "*from circulation*" are removed while translating 5 (a). This may be done to shorten the news stories.

5 a). {..} will completely withdraw from circulation {..}

5 b). {..} वापस लेगा

(*{..}wapas lega*)

B.2. Sentence Split/Join

The information may be spread over more than one sentence or may be combined in single sentence. These types of paraphrases have two components text units and connective between clauses which is normally altered [13]. The sentence in 6 (a) has been split and translated into two sentences in Hindi 6 (b).

6 a). Ramesh, 50, who was serving his life imprisonment, is survived by his wife and two children Rakesh and Karuna both of whom are college students.

6 b). रमेश के परिवार में उसकी पत्नी और दो बच्चे राकेश तथा करुणा हैं। दोनों कॉलेज के विद्यार्थी हैं।

(*Ramesh ke parivar mein uski patni aur do bachche rakesh tatha karuna hain. Dono college ke vidyarthi hain*)

B.3. Change in Modality

The modality of the sentence may also be changed (7(a) & 7(b) and 8(a) & 8(b)). Normally they may also be considered in discourse based change in which structure of the sentence is normally changed [4, 24].

7 a). He won't be able to make {..} into {..}

7 b). उनकी हैसियत नहीं {..} को {..} बनाने की

(*unki haisiyat nahin{..}ko {..} banana ki*)

8 a). The meaning of {..} is 24 hours electricity

8 b). जानते हैं {..} बनाने का मतलब? {..} बनाने का मतलब होता है 24 घंटे बिजली.

(jante hain {..} banane ka matlab? {..} banana ka matlab hota hai 24 ghante bijli)

B.4. Passive vs. Active/Direct-Indirect Style Alteration/Voice Alternation/Change of Emphasis [4, 13, 23, 24]

It involves syntactic reorganization and contains those diathesis alternations where the meaning is preserved but the voice or style is changed at the time of translation (9 (a) & 9(b) and 10 (a) & 10 (b)).

9 a). Water released from the dam completely submerged the fields

9 b). बांध से जारी पानी में खेत पूरी तरह जलमग्न

(*baandh se jaari paani mein khet poori tarah jalmagn*)

10 a). "Of course I am disappointed. But it was the decision of the governing council," he said

10 b). आईपीएल कमिश्नर ने बाद में कहा कि टेंडर रद्द होने से हालांकि वह निराश हैं लेकिन यह गवर्निंग काउंसिल का फैसला है।

(*IPL commissioner ne baad mein kaha ki tender radd hone se halanki who nirash hain lekin yeh governing council ka faisla hai*)

B.5. Representational Change

Many times category of noun may be changed at the time of translation. In 12 (b) natives are represented by country at the time of translation from 12 (a).

11 a). Indians showed great restraint after the last {..} attack

11 b). पिछले {..} हमले के बाद भारत ने जबर्दस्त धैर्य का परिचय दिया

(*pichhle {..} hamle ke baad bharat ne jabardast dhairya ka parichay diya*)

C. Level III

In this level, normally translated sentence contains few content lexical units that are not proper translation across the language. Linking such news stories is a challenging task as news stories communicate same information but words may not have direct mapping. Such cases may fall under the category of low obfuscation or lexical paraphrasing.

C.1. Localization Related Issues

In this class cross language text reuse is dominated by localization related issues. It is observed that such types of usages are among the most common ones in new story text reuse. [23] has considered this class as a case of change of format. Date (12 (a)) currency (12 (b) and 12 (c)) are the most common examples of this class (Table 1). Date can be written in any applicable format in English and gives rise to many permutations and combinations of Hindi translations formats. Likewise in natural language, currency can be expressed in users own ways and may not always conform to the dictionary equivalents. Like in 12 (b) “*ten million”* is translated in Hindi as “*दस लाख*” which is wrong, and not as “एक करोड” when given to a Machine Translation system. But a person can intrerpret it as “एक करोड” or as transliterated version of its English counterpart.

Many such examples are present in FIRE corpus. Some of the examples are shown in Table 1

Table 1. Localization related issues in English-Hindi Cross Language Text Reuse

Example No.	English	Hindi(which is not translated by MT system)
8 a)	June 12, 2009	बारह जून 2009 (*barah june 2009*)
8 b)	$10 million	एक करोड़ डॉलर / 10 मिलियन डॉलर (*ek karod dollar/10 million dollar*)
8 c)	$5 million	50 लाख डॉलर (*50 lakh dollar*)

C.2. Partial Improper Translation

In this class linguistically the situation is communicated as per the syntax of the respective language but if one tries to map the text reuse then few lexical units may not map due to partial improper translation (13 (a) & 13 (b) and 14 (a) & 14 (b)). “Crashed on him” will never map to “टकरा गए थे” and “amended” does not mean “कटौती”.

13 a). {..} door crashed on him

13 b). वे दरवाजे से टकरा गए थे.

(*ve darwaze se takra gaye the*)

14 a). {..} Can be amended.

14 b). {..} कटौती हो सकती है

(*katauti ho sakti hai*)

Automatic mapping for text reuse under this class is quite challenging. In 15(a) “*boy”* the actor is referred as “*व्यक्ति*” in 15(b) which is not proper. Table 2 shows some of the words present in the new stories that shall never mapped properly.

15 a). {..}an affair with a boy from a different community

15 b). {..}किसी और बिरादरी के व्यक्ति से प्रेम{..}

(*{..} kisi aur biradari ke vyakti se prem{..}*)

Table 2. Partial Improper Translation

English word	Hindi wordfound in News Articles	Exact dictionary mapping in Hindi
Boy	व्यक्ति (*vyakti*)	लड़के (*ladke*)
Girl	महिला (*mahila*)	लड़की (*ladki*)
Ordered	फैसला सुनाया (*faisla sunaya*)	आदेश दिया (*aadesh diya*)
Police	टीम (*team*)	पुलिस (*pulis*)
Bill	कोटा (*kota*)	विधेयक (*vidheyak*)
Calculation	समीकरण (*samikaran*)	गणना (*ganana*)
Chief	कमिश्नर (*commisioner*)	प्रमुख/मुख्य (*pramukh/mukhya*)
Chairman	कमिश्नर (*commisioner*)	अध्यक्ष (*adhyaksh*)

C.3. OOV words substitutions

Socio-cultural influence across globe results in acceptability of some of the lexical units that are normally treated as Out of Vocabulary (OOV) for native language. In such cases although the Hindi equivalents of the words are available, but instead of taking exact word translations, transliterated words are accepted at the time of news reporting because such transliterated versions are more in use than the translation equivalents. Hindi has adopted many such words in its day to day writings and conversations but such words do not find any place in the dictionary as Hindi meanings of English words. These words also create problems if we go for Dictionary based approaches for mapping these words. Some of the words of FIRE corpus are shown in Table 3.

Table 3. OOV words substitutions

English	Accepted transliterated form	Exact Hindi Translation
Net worth	नेट वर्थ (*net worth*)	निवल मूल्य (*nival mulya*)
South Asia	साउथ एशिया (*south asia*)	दक्षिण एशिया (*dakshin asia*)
Counter-terrorism Strategy Initiative Co-Director	काउंटर टेररिजम स्ट्रैटिजी इनीशिएटिव के को-डायरेक्टर (*counter terrorism strategy initiative ke co director*)	आतंकवाद विरोधी रणनीति पहल सह-निदेशक (*aatankwad virodhi rananiti pahal sah-nideshak*)
Criminal cases	क्रिमिनल केस (*criminal case*)	आपराधिक मामला (*aapradhik mamla*)
neither factual nor legal	न कानूनी और ना ही फैक्चुअल (*na hi kanuni aur na hi factual*)	न कानूनी और ना ही तथ्यात्मक/तथ्यपूर्ण (*na hi kanuni aur na hi tathyatmak/tathyapurna*)
House Homeland and Security Committee	हाउस होमलैंड एंड सिक्युरिटी कमिटी (*house homeland and security committee*)	गृह मातृभूमि और सुरक्षा समिति (*grih matribhumi aur suraksha samiti*)

C.4. Role and Thought Shifting

One news may depict thought and news in other language depicts other's role. 16 (a) shows what a person is thinking about himself and the person's own decision but its Hindi equivalent 16(b) leaves the decision on others.

16 a). Will consider PM job if we win
16 b). सांसद चाहेंगे तभी बनूँगा प्रधानमंत्री

(*saansad chahenge tabhi banunga pradhanmantri*)

C.5. Syntax/Discourse Structure Changes [4, 23]

While translating 17(a) interjection is converted to assertion in 17 (b) along with same polarity substitution. Whereas 17(a) expresses surprise, 17 (b) asserts that it can never be true.

17 a). Kapoor {..} he didn't believe that the minister was capable of harming her.
17 b).साथ उन्होंने {..} कपूर उनकी माँ को शारीरिक नुकसान नहीं पहुंचा सकते थे

(*saath unhone{..} kapoor unki maa ko sharirik nuksaan nahin pahuncha sakte the*)

C.6. Contextual Related Word

The contextual related word may be used in place of exact translation. Let word be e_1, its exact translation be h_1 and contextual words related to E in H be h_{c1}, h_{c2}, h_{c3}. The contextual words h_{c1}, h_{c2}, h_{c3} may be used in place of h_1 (18(a) & 18(b)). In a simpler way, these are those translations, where an English lexicon can be represented by any of its Hindi synonyms.

18 a) I told them that the bill could be amended to address their concerns in respect of OBC and Muslim women.
18 b) मैंने उनसे कहा है कि ओबीसी और अल्पसंख्यक महिलाओं को लेकर उनकी जो चिंता है, उसे खत्म करने के लिए बिल में संशोधन किया जा सकता है।

(*maine unse kaha hai ki OBC aur alpsankhyak mahilaon ko lekar unki jo chinta hai, use khatm karne ke liye bill mein sanshidhan kiya ja sakta hai*)

C.7. Transliteration of Synonym

In this class lexical mapping across the language is present but one uses transliterated synonym of lexical unit at the time of news reporting (19 and 20). The synonyms for hired and plea are contract and appeal respectively and these English synonyms only have been transliterated for using in the Hindi stories.

19). hired killer कॉन्ट्रैक्ट किलर (*contract killer*)
20). plea अपील (*appeal*)

C.8. Abbreviation vs. Polysemy

In this class abbreviation is either transliterated or its expanded form is translated or transliterated. There can also be more than one translation equivalents of the same word and it is difficult to map these words across language (Table 4). Fujita [16] and Barron et al. [4] referred this class as "Altering notational variants, abbreviations, and acronyms" and "Lexicon based spelling and format changes" respectively.

Table 4. Abbreviation vs. Polysemy

Acronym	**Expanded forms can be either of these**
SP	सपा, समता पार्टी, पुलिसअधीक्षक (*sapa,samta party, police adhikshak*)
BJP	भाजपा, भारतीय जनता पार्टी (*bhajpa, bhartiya janata party*)
SC	सुप्रीम कोर्ट, अनुसूचित जाति (*supreme court, anushuchit jaati*)
Central Bureau of Investigation	सीबीआई, केंद्रीय अन्वेषण ब्यूरो (*CBI, kendriya anveshan bureau*)
CBI	सीबीआई, केंद्रीय अन्वेषण ब्यूरो (*CBI, kendriya anveshan bureau*)
RBI	रिज़र्व बैंक, आर बी आई (*reserve bank, RBI*)
Sgt	सार्जेंट (*seargent*)
Lt	लेफ्टिनेंट (*leutinent*)
Gen	जनरल (*general*)

C.9. Sentimental Outburst to Add Sensation

In this category some phrase, idioms and words arousing emotional outburst may be added across the language (21 (a) & 21 (b)). Here "सामूहिक दुष्कर्म (*saamuhik dushkarm*)" is not the exact translation of rape but has simply been used to arouse sensation.

21 a). 12 rape girl on panchayat order
21 b). पंचायत के आदेश पर १२ लोगो ने किया सामूहिक दुष्कर्म

(*panchayat ke aadesh par 12 logon ne kiya saamuhik dushkarm*)

D. Level IV

Translations falling under this category come under pragmatic paraphrasing which have been dealt by several researchers. As special types of paraphrases it goes beyond pure semantic similarity to fall within the field of pragmatics [24]. Paraphrasing extends to a group of words. Linking and tracking news stories under this class becomes quite challenging.

D.1. Action vs. Consequence

This category has been referred to as Textual Entailment [1, 16, 24]. In this class meaning of one expression can be inferred from the other [10]. Newspaper may report action or decision taken in one language, but its translation in other language may report the consequences of the action or decision (22 (a) & 22(b))

22 a). All pre-2005 notes go out of currency

22 b). वापस करने होंगे 2005 से पहले के सभी नोट

(*wapas karne honge 2005 se pahle ke sabhi note*)

D.2. Change in Reference

In this category referencing of time period may be changed across language (23 (a) & 23(b)).

23 a). Anybody who has such notes can get these exchanged in any bank after April 1.

23 b). {..} के मुताबिक 30 जून तक कोई भी व्यक्ति कितना भी नोट बैंक में जाकर बदल सकता है

(*{..} ke mutabik 30 june tak koi bhi vyakti kitna bhi note bank mein jakar badal sakta hai*)

D.3. Focus Shifting

The focus may be shifted at the time of translation while preserving the gist of the sentence. In 24(a) the focus is on reason but in 24 (b) it is on relation.

24 a). Shyam Gupta dies of brain hemorrhage

24 b). राम गुप्ता के भाई श्याम गुप्ता की मौत

(*ram gupta ke bhai shyam gupta ki maut*)

D.4. Actor/Action Substitution [6]

Action may be replaced by actor to highlight actor in news stories (25 (a) & 25 (b)).

25 a). {..} admitted to hospital here.

25 b). {..} सिर और पैर में मामूली चोट आयी. उपचार के बाद वे शूटिंग पर लौट भी आये.

(*{..} sir aur pair mein maamuli chot aayi. Upchaar ke baad ve shooting par laut bhi aaye*)

D.5. Specification vs. Generalization

Group of lexical units i.e. noun phrase in one language may be replaced by some other word or by anaphora ((26 (a) & 26 (b) and 27 (a) & 27 (b) and 28 (a) & 28 (b)) [11]. Use of hyponyms and hypernyms represent this category.

26 a). {..} two-day national conclave at Karla

26 b). पुणे के पास पार्टी के दो दिन के सम्मेलन{..}

(*pune ke paas party ke do din ke sammelan{..}*)

27 a). A combination of Guptas has emerged to oppose the Bill

27 b). अमोल, मनीष और तुषार गुप्ता ने बिल का विरोध किया है

(amol, manish aur tushar gupta ne bill ka virodh kiya hai)

28 a). This has remained secret until now.

28 b). दस्तावेजों के इस तरह नष्ट या गुम हो जाने का राज आज तक बना हुआ था।

(*dastavezon ke is tarah nasht ya gum ho jaane ka raaj aaj tak bana hua tha*)

D.6. Lexicon based Opposite Polarity Substitution

Marta vila et al. [23, 24] and Barron et al. [4] referred this class as Lexicon based opposite polarity substitution but few other authors has referred it as Inter-clausal negative-affirmative paraphrasing [16]. In this class polarity may be changed twice. In this lexical unit is changed by its antonym or complementary and then in order to maintain the meaning, another change of polarity occurs within the same sentence (29 (a) & 29 (b)).

29 a). {..} "too strong" to commit suicide

29 b). {..} इतनी कमजोर नहीं थी कि ख़ुदकुशी जैसा कदम उठा ले.

(*{..} itni kamzor nahin thi ki khudkushi jaisa kadam utha le*)

D.7. Referential [16] or cognitive [27]

The cross language news stories may comprise of co reference that needs attention as it may be difficult to identify reuse (30 (a) & 30 (b) and 31 (a) & 31(b)). Referential and cognitive are to be treated as co reference rather than paraphrase but for retrieving news stories co-reference i.e. referential and cognitive might be quite useful [16, 23].

30 a).{..} photograph addressing a meeting in February 2011

30 b). {..} तीन साल पुरानी तस्वीर {..}

(*{..} teen saal purani tasveer {..}*)

31 a). He was released on 14-day parole on August 2, 1999

31 b). उन्हें खराब सेहत के आधार पर परोल पर छोड़ गया था। गत अगस्त में उसे विशेष छूट दी गई थी।

(*unhe kharab sehat ke aadhar par parol par chhoda gaya tha. Gat august mein use visheh chhoot di gayi thi*)

D.8. Handling of Phrase across Language

In this category few lexical units that are phrase needs to have exact phrasal mapping at the time of handling text reuse across language. Sometimes there might not be exact phrase but same may be mapped to nearby equivalent at time of news reporting (32 (a) & 32 (b) and 33 (a) & 33 (b)).

32 a). Mrinalini Devi {..} fighting against it tooth and nail{..}

32 b). मृणालिनी देवी {..} एडी-चोटी का जोर लगाकर लड रही हूं

(*mrinalini devi {..} edi-choti ka jor lagakar lad rahi hoon*)

33 a). "Realising {..} indulging in theatrics {..}

33 b). हर्षा राय {..} नाटक पर उतारू {..}

(*harsha rai {..} naatak par utaru {..}*)

D.9. Synthetic/Analytic Substitutions

In this category a there is replacement of word by its nearby meaning instead of exact dictionary word. These are those substitutions that have single-pieced for multiple-pieced lexical units (34 (a) & 34 (b) and 35 (a) & 35 (b)) that have same meaning [23].

34 a). {..} popular confectionery chain {..}

34 b). मिठाइयों और नमकीन प्रॉडक्ट्स की दुकानों की जानी-मानी चेन

(*mithaiyon aur namkeen products ki dukanon ki jaani maani chain*)

35 a). {..} the residence and other properties {..}

35 b). {..} ठिकानों {..}

(*{..} thikanon{..}*)

E. Level V

The categories belonging to this level are the toughest of all to link news stories across language as it comprises of news stories that are equivalent in propositional concept but words used to express news stories are completely different [23].

E.1. Modification of Action

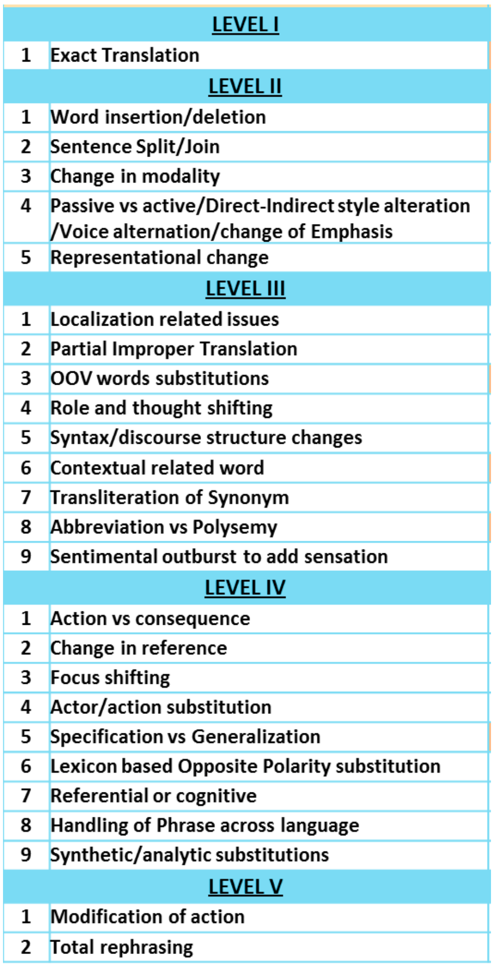
This is mostly in the news related to raids and corruption. One story may emphasize on the property and legal issue and the other equivalent story may give details such as where it happened and total monetary estimate (36 (a) & 36 (b)).

36 a). {..} whose property includes cash, jewellery, agriculture land, flats and plots, has been booked under the disproportionate assets case under the Prevention of Corruption Act 1988

36 b). {..} स्थित निवास के अलावा {..} में ठिकानों पर एक साथ कार्यवाई की गई. उनकी सम्पति की कीमत चार करोड रूपये तक जा सकती है.

(*{..} stith niwas ke alawa {..} mein thikanon par ek saath karyavahi ki gayi. Unki sampati ki keemat char karod rupye tak ja sakti hai*)

Table 5. Complete Typology for English-Hindi Journalistic Text reuse



	LEVEL I
1	Exact Translation
	LEVEL II
1	Word insertion/deletion
2	Sentence Split/Join
3	Change in modality
4	Passive vs active/Direct-Indirect style alteration /Voice alternation/change of Emphasis
5	Representational change
	LEVEL III
1	Localization related issues
2	Partial Improper Translation
3	OOV words substitutions
4	Role and thought shifting
5	Syntax/discourse structure changes
6	Contextual related word
7	Transliteration of Synonym
8	Abbreviation vs Polysemy
9	Sentimental outburst to add sensation
	LEVEL IV
1	Action vs consequence
2	Change in reference
3	Focus shifting
4	Actor/action substitution
5	Specification vs Generalization
6	Lexicon based Opposite Polarity substitution
7	Referential or cognitive
8	Handling of Phrase across language
9	Synthetic/analytic substitutions
	LEVEL V
1	Modification of action
2	Total rephrasing

E.2. Total Rephrasing (High Obfuscation)

When the focal event is the same but entirely diverse sentences supports the fact and it is very difficult to link such news stories because of vagueness of focal event in either new stories(37 (a) & 37 (b) and 38 (a) & 38 (b)).

37 a). Minister had already assured the House that all parties would be taken into confidence by the government on the issue.

37 b). महिला आरक्षण बिल पर सहमति कायम करने के लिए मंत्री ने सर्वदलीय बैठक बुलाई है।

(mahila aarakshan bill par sahmati kaayam karne ke liye mantri ne sarva daliya baithak bulayi hai)

38 a). "...Nothing is found against the correctness, legality, propriety or regularity in respect of any of the findings”

38 b). फैसले के खिलाफ कोई ठोस दलील पेश नहीं की गई।

(faisle ke khilaph koi thos dalil pesh nahi ki gayi)

Table 5 shows complete typology for English-Hindi Journalistic Text Reuse

IV. DISCUSSION

Authors participated in the Cross Language Indian News Story Search (CLINSS) task of FIRE 2013 where the task was to link relevant Hindi news stories reporting the same focal event out of a corpus of 50691 stories to their 25 English counterparts. They tried to analyze and identify occurrences of the different paraphrase types as proposed in this typology in the top ten Hindi news stories which were retrieved as the result of the CLINSS task. Out of 250 documents retrieved, 50 were found to be relevant. The relevancy was judged on the basis of whether the document pairs shared the focal event and news event both or only the news event and not the focal event. Those documents which shared the focal events too, showed the tendency of exact translations of group of words occupying major portion in the corpus. Loan words or OOV substitutions was also used frequently in case of equivalent Hindi document. Polysemy hypernymy and hyponymy (specification vs. generalization) and synonymy (contextual related words) also was present in the corpus. Word insertions and deletions and sentence split/join are done to present the same facts with some additional information. The other classification categories were mostly observed in the documents which shared the same news event but different focal event. Different wordings due to change of focal event might have been the reason behind it.

V. CONCLUSION

Defining typologies helps in drawing boundaries to identify across language different equivalent manifestations and helps devising or developing techniques for accurately tracking news stories across language. The analysis of relevant Hindi news stories having the same focal or main event as their English equivalents for different classification types English-Hindi News stories corpora are suggestive of the facts as to which of the paraphrase types are mostly observed in cross-lingual reuse. This analysis may prove beneficial to track and link two such cross-lingual English-Hindi story pairs if proper techniques are devised for dealing with the most prominent paraphrase types.

The work is a novel step towards constructing a paraphrase typology for English Hindi news corpora. The proposed work has tried to bring forth the intricacies involved across the language journalistic text reuse. The paraphrase boundaries are in most of the cases overlapping so the work can be further analyzed to redefine the boundaries to deepest sub-level so that overlapping is reduced to an extent.

ACKNOWLEDGEMENTS

Authors are grateful to FIRE for providing CLINSS corpus of FIRE 2012 and 2013, to English News paper Hindustan Times and to Hindi news paper Dainik Bhaskar.

One of the authors, Aarti Kumar, is thankful to Maulana Azad National Institute of Technology, Bhopal, India for providing her the financial support to pursue her Doctoral work as a full time research scholar.

REFERENCES

[1] I. Androutsopoulos, and P. Malakasiotis, “A Survey of Paraphrasing and Textual Entailment Methods,” Journal of Artificial Intelligence Research, 38(1), 135-187, 2010.

[2] E. Barker and R. Gaizauskas, “Assessing the Comparability of News Texts,” in Proc. Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012.

[3] A. Barreiro, “Make It Simple with Paraphrases. Automated Paraphrasing for Authoring Aids and Machine Translation,” Ph.D. Thesis, Porto: Universidade do Porto, 2008.

[4] Barrón-Cedeño, “On the Mono- and Cross-Language Detection of Text Re-Use and Plagiarism,” Ph.D. Thesis, Spain: Universitat Polit`ecnica de Val`encia, 2012.

[5] R. Barzilay, K. McKeown, and M. Elhadad, “Information Fusion in the Context of Multi-Document Summarization,” in Proc. 27th Annual Meeting of the Association for Computational Linguistics (ACL 1999), College Park (MD), 550-557, 1999.

[6] R. Bhagat, “Learning Paraphrases from Text,” Ph.D. Thesis, Los Angeles: University of Southern California, 2009.

[7] C. Boonthum, “iSTART: Paraphase Recognition,” in Proc. ACL 2004 Student Research Workshop, Barcelona, 31-36, 2004.Available:http://dx.doi.org/10.3115/1219079.1219089.

[8] P. Clough, “Measuring Text Reuse,” Ph.D. Thesis, Sheffield: University of Sheffield, 2003.

[9] P. Culicover, “Paraphrase Generation and Information Retrieval from Stored Text,” Mechanical Translation and Computational Linguistics, 11(1-2), 78-88, 1968.

[10] I. Dagan, O. Glickman, “Probabilistic Textual Entailment: generic Applied Modeling of Language Variability”. Available:http://u.cs.biu.ac.il/~dagan/publications/ProbabilisticTE_fv07.pdf.

[11] B. Dolan, C. Quirk, and C. Brockett, “Unsupervised

Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources," in Proc. 20th International Conference on Computational Linguistics (COLING 2004), Geneva, 350-356, 2004.Available: http://dx.doi.org/10.3115/1220355.1220406

[12] B. J. Dorr, et. al., "Semantic Annotation and Lexico-Syntactic Paraphrase," in Proc. LREC 2004 Workshop on Building Lexical Resources from Semantically Annotated Corpora, Lisbon, 47-52, 2004.

[13] M. Dras, "Tree Adjoining Grammar and the Reluctant Paraphrasing of Text," Ph.D. Thesis, Sydney: Macquarie University, 1999.

[14] C. Dutrey, D. Bernhard, H. Bouamor, and A. Max, "Local Modifications and Paraphrases in Wikipedia's Revision History," Procesamiento del Lenguaje Natural, 46, 51-58, 2011.

[15] L. Faigley, and S. Witte, "Analyzing Revision. College Composition and Communication," 32(4), 400-414, 1981. Available: http://dx.doi.org/10.2307/356602.

[16] A. Fujita, "Automatic Generation of Syntactically Well-Formed and Semantically Appropriate Paraphrases," Ph.D Thesis, Nara: Nara Institute of Science and Technology, 2005.

[17] R. P. Honeck, "A Study of Paraphrases," Journal of Verbal Learning and Verbal Behavior, 10, 367-381, 1971. Available: http://dx.doi.org/10.1016/S0022-5371 (71)80035-X.

[18] R. Kozlowski, K. F. McCoy, and V. K. Shanker, "Generation of Single-Sentence Paraphrases from Predicate/Argument Structure Using Lexico-Grammatical Resources," in Proc. International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003), Sapporo, 1-8, 2003.

[19] D. Munteanu, and D. Marcu, "Improving Machine Translation Performance by Exploiting Comparable Corpora" Computational Linguistics, 31 (4), pp. 477-504, December 2005.

[20] M. Recasens, and M. Vila, "On Paraphrase and Coreference," Computational Linguistics, 36(4), 639-647, 2010. Available: http://dx.doi.org/10.1162/coli_a_00014.

[21] F. Rinaldi, J. Dowdall, K. Kaljurand, M. Hess, and D. Mollá, " Exploiting Paraphrases in a Question Answering System," in Proc. 2nd International Workshop on Paraphrasing: Paraphrase Acquisition and Applications (IWP 2003), Sapporo, 25-32, 11 July 2003.

[22] M. Shimohata, "Acquiring Paraphrases from Corpora and Its Application to Machine Translation," Ph.D. Thesis, Nara: Nara Institute of Science and Technology, 2004.

[23] M. Vila, M. Antonia Marti, and H. Rodrguez, "Paraphrase Concept and Typology-A Linguistically Based and Computationally Oriented Approach," Procesamiento del Lenguaje Natural, pp 83-90, 2011.

[24] M. Vila, M. A. Martí, and H. Rodríguez, "Is This a Paraphrase? What Kind? Paraphrase Boundaries and Typology," Open Journal of Modern Linguistics, 4, 205-218., 2014. Available: http://dx.doi.org/10.4236/ojml.2014.41016.

[25] P. Clough, and R. Gaizauskas, "Corpora and Text Re-Use," In A. L ̈udeling, M. Kyt ̈o, and T. McEnery, editors, Handbook of Corpus Linguistics, Handbooks of Linguistics and Communication Science, Mouton de Gruyter, 2009, pages 1249—1271.

[26] J. Robin, "Revision-Based Generation of Natural Language Summaries Providing Historical Background: Corpus-Based Analysis, Design, Implementation, and Evaluation," Ph.D. thesis, Department of Computer Science, Columbia University, NY, 1994.

[27] J. Milićević, "Semantic Equivalence Rules in Meaning-Text Paraphrasing," In L. Wanner (Ed.), Selected Lexical and Grammatical Issues in the Meaning-Text Theory, Amsterdam: John Benjamins, 2007, pp. 267-296.

[28] C.D. Manning, P. Raghavan, and H. Schutze, "Introduction to Information Retrieval," Vol. 1, Cambridge: Cambridge University, Press; 2008.

[29] D. Guan and W. Yuan, "A Survey of Mislabeled Training Data Detection Techniques for Pattern Classification", IETE Technical Review, vol. 30, issue-6, pp. 524-530, Nov-Dec 2013.

[30] M. B. Bashir, M. S. A. Latiff, A. A. Ahmed, A. Yousif, and M. E. Eltayeeb, "Content - based Information Retrieval Techniques Based on Grid Computing: A Review," IETE Technical Review, vol. 30, issue-3, pp. 223-232, May-Jun 2013.

Authors' Profiles

Aarti Kumar was born in Patna, India in 1963. She has done her Masters in Botany in 1983 from Patna University, India and in Computer Applications in 2005 from Indira Gandhi National Open University, India and is a university topper of Bachelors in Education (1999) from Barkatullah University, India.

She is currently pursuing her Ph. D. in Computer Applications from Maulana Azad National Institute of Technology (MANIT), Bhopal, India. Her area of research is Cross-Language Information Retrieval, more specifically English-Hindi Journalistic Text Reuse. She has a teaching Experience of 18 years and a research experience of more than three years. Her published works include:

- "Query Formulation for Heuristic Retrieval in Obfuscated and Translated Partially Derived Text", Journal of Information Science Theory and Practice(JISTaP), Korea Institute of Science and Technology Information, Vol. 3, No. 1 March 2013 issue, pp24-39, pISSN2287-9099, eISSN 2287-4577, DOI Prefix10.1633.
- "An evolutionary survey from Monolingual Text Reuse to Cross Lingual Text Reuse in context to English - Hindi", International Journal of Scientific & Engineering Research, Volume 6, Issue 2, February-2015 ISSN 2229-5518 pp 996-1003
- Pre-Retrieval based Strategies for Cross Language News Story Search" accepted in ACM Journal as Post-Proceedings of the 2013 Forum for Information Retrieval Evaluation (FIRE) December 04 - 06 2013, New Delhi, India.

Mrs. Kumar is an Associate Member, Information Retrieval Society of India (IRSI) and a Professional Member, Association for Computing Machinery (ACM).

Sujoy Das was born in Patna, India in 1969. He has done his Masters in Computer Applications in 1991 and Ph. D. in 2009 from MANIT, Bhopal.

He is working as Associate Professor in Department of Mathematics & Computer Applications, Maulana Azad National Institute of Technology (MANIT), Bhopal, India .He has a teaching experience of 20 years and his research interests include areas such as Cross Language Information Retrieval, Query Expansion and Cross Language Text Reuse. His published works include:

- Query Expansion Strategy based on Pseudo Relevance Feedback and Term Weight Scheme for Monolingual Retrieval. CoRR abs/1502.05168 (2015)
- Improving Performance of English-Hindi CLIR System using Linguistic Tools and Techniques. IHCI 2009: 261-271
- Disambiguation Strategies for English-Hindi Cross Language Information Retrieval System. IHCI 2009: 306-315
- Evaluation of the English-Hindi Cross Language Information Retrieval System Based on Dictionary Based Query Translation Method. ICIT 2007: 56-61

Dr. Das is a Member of Information Retrieval Society of India (IRSI) and a Professional Member, Association for Computing Machinery (ACM).

How to cite this paper: Aarti Kumar, Sujoy Das, "Typology for Linguistic Pattern in English-Hindi Journalistic Text Reuse", International Journal of Information Technology and Computer Science (IJITCS), Vol.8, No.8, pp.75-86, 2016. DOI: 10.5815/ijitcs.2016.08.09