Comparative Study of Indexing and Search Strategies for the Hindi, Marathi, and Bengali Languages

LJILJANA DOLAMIC and JACQUES SAVOY
University of Neuchatel

The main goal of this article is to describe and evaluate various indexing and search strategies for the Hindi, Bengali, and Marathi languages. These three languages are ranked among the world's 20 most spoken languages and they share similar syntax, morphology, and writing systems. In this article we examine these languages from an Information Retrieval (IR) perspective through describing the key elements of their inflectional and derivational morphologies, and suggest a light and more aggressive stemming approach based on them.

In our evaluation of these stemming strategies we make use of the FIRE 2008 test collections, and then to broaden our comparisons we implement and evaluate two language independent indexing methods: the *n*-gram and trunc-*n* (truncation of the first *n* letters). We evaluate these solutions by applying our various IR models, including the Okapi, Divergence from Randomness (DFR) and statistical language models (LM) together with two classical vector-space approaches: *tf idf* and *Lnu-ltc*.

Experiments performed with all three languages demonstrate that the $I(n_e)C2$ model derived from the Divergence from Randomness paradigm tends to provide the best mean average precision (MAP). Our own tests suggest that improved retrieval effectiveness would be obtained by applying more aggressive stemmers, especially those accounting for certain derivational suffixes, compared to those involving a light stemmer or ignoring this type of word normalization procedure. Comparisons between no stemming and stemming indexing schemes shows that performance differences are almost always statistically significant. When, for example, an aggressive stemmer is applied, the relative improvements obtained are ∼28% for the Hindi language, ∼42% for Marathi, and ∼18% for Bengali, as compared to a no-stemming approach. Based on a comparison of word-based and language-independent approaches we find that the trunc-4 indexing scheme tends to result in performance levels statistically similar to those of an aggressive stemmer, yet better than the 4-gram indexing scheme. A query-by-query analysis reveals the reasons for this, and also demonstrates the advantage of applying a stemming or a trunc-4 indexing scheme.

This research was supported by the Swiss National Science Foundation under Grant #200021-113273.

Authors' addresses: L. Dolamic and J. Savoy, University of Neuchatel, Rue Emile Argand 11, 2009, Neuchâtel, Switzerland, email:{Ljiljana.Dolamic, Jacques.Savoy}@unine.ch.

Categories and Subject Descriptors: H.3.1 [**Content Analysis and Indexing**]: Indexing methods; Linguistic processing; H.3.3 [**Information Search and Retrieval**]: Retrieval models; H.3.4 [**Systems and Software**]: Performance evaluation

General Terms: Algorithms, Measurement, Performance

Additional Key Words and Phrases: Indic languages, stemmer, natural language processing with Indo-European languages, search engines for Asian languages, Hindi language, Marathi language, Bengali language

1. INTRODUCTION

Over the last few years there has been increasing interest in Asian languages, especially those spoken in the Far East (e.g., the Chinese, Japanese, and Korean languages) and on the Indian subcontinent. Given the increasing volume of sites, and the number of Internet pages generally available in these languages, not to mention the online users working with them, we clearly need a better understanding of the automated procedures applied when processing them.

As in Europe, the Indian subcontinent can be characterized by the use of many languages. With 23 official languages[1] being spoken in the European Union the situation there would seem to be more complex than in the Republic of India, which has only two official languages (Hindi and English). This general view however hides the fact that approximately 29 languages are spoken by more than one million native speakers there, most of which have official status in the various Indian states.[2] Thus from a linguistic perspective, the situation in India is slightly more complex than in Europe, as evidenced by the four main families to which the various languages belong: the Indo-European (more precisely the Indo-Aryan branch [Masica 1991] including Bengali, Hindi, Marathi, and Pandjabi among others) located mainly in the northern part, the Dravidian family (e.g., Kannada, Malayalam, Tamil, and Telugu) in the southern part, the Sino-Tibetan (e.g., Bodo and Manipur) in the northeastern part, and the Austra-Asiatic group (Santali) in the eastern part of this subcontinent. While Europe is also made up of various language families (e.g., Finnish and Hungarian belong to the Finno-Ugric branch), India's proportion of non-Indo-European languages is much greater than that of Europe. Moreover, compared to the three alphabets used in Europe (Latin, Greek, and Cyrillic), the various Indian languages use at least seven different writing systems.

In this article we focus on three of the most popular Indian languages: Hindi (the native language of ∼180 million speakers), Marathi (∼68 million) and Bengali (∼190 million),[3] as well as the test collections made available

[1]See http://ec.europa.eu/education/languages/languages-of-europe/.
[2]See http://india.gov.in/knowindia/official_language.php.
[3]According to the Web site http://www.ethnologue.com/.

through the first Forum for Information Retrieval Evaluation (FIRE[4]) campaign. This article also describes the main morphological variations and constructions found in these languages, particularly those found most useful from an IR perspective. We thus propose and evaluate various stopword lists and stemmers for these three Indian languages, and then compare them by applying various indexing and search strategies.

The rest of this article is organized as follows. Section 2 presents some related work on stemming. Section 3 describes the main morphological aspects of the three selected languages. Section 4 reveals various stemming approaches while Section 5 depicts the main characteristics of our test-collections. Section 6 briefly describes the different IR models applied during our experiments. Section 7 evaluates and analyzes the performance of the various indexing and search strategies, and to conclude the last section outlines our key findings.

2. RELATED WORK

In the IR domain it is usually assumed that stemming serves as an effective means of enhancing retrieval efficiency through conflating several different word variants into a common form or stem [Manning et al. 2008], and that this efficiency can be achieved through applying morphological rules specific to each language involved. Typical examples for the English language are those described in Lovins [1968] and Porter [1980]. This suffix removal process can also be controlled through the adjunct of quantitative restrictions (e.g., "ing" is removed if the resulting stem has more than three letters, as in "hopping" but not in "ring") or qualitative restrictions (e.g., "-ize" is removed if the resulting stem does not end with "e" as in "seize"). Moreover to improve conflation accuracy, certain ad hoc spelling correction rules can also be employed (e.g., "hopping" gives "hop" and not "hopp"), through applying irregular grammar rules usually designed to facilitate pronunciation.

Simple algorithmic stemming approaches ignore word meanings and tend to make errors, often caused by over-stemming (e.g., "general" becomes "gener", and "organization" is reduced to "organ") or to under-stemming (e.g., with Porter's stemmer, the words "create" and "creation" do not conflate to the same root. This is also the case with the words "European" and "Europe"). For this reason the use of an online dictionary means obtaining better conflation has been suggested as in Krovetz [1993].

Compared to other languages (such as French) with their more complex morphologies [Sproat 1992], English might be considered quite simple and thus using a dictionary to correct stemming procedures would be more helpful for those other languages [Savoy 1993]. For those languages whose morphologies are even more complex, deeper analyses would be required (e.g., for Finnish [Alkula 2001], [Korenius et al. 2004]) and the lexical stemmers [Tomlinson 2004] are not always freely available (e.g., Xelda system at Xerox), thus meaning that their design and elaboration is more complex and labor intensive.

[4]More information available at the FIRE Web site, http://www.isical.ac.in/~clia/.

Moreover, their use involves a large lexicon along with a complete grammar, and thus their application is problematic and requires more processing time, especially with the very large and dynamic document collections (e.g., within the context of a commercial search engine on the Web). Additionally, lexical stemmers must be capable of handling unknown words such as geographical, product, and proper names, as well as acronyms (out-of-vocabulary problem). Lexical stemmers could not therefore be considered as error-free approaches. Moreover, for the English language at least, applying a morphological analysis to obtain the correct lemma (dictionary entry) does not provide better results than a simple algorithmic stemmer [Fautsch and Savoy 2009]. Finally, based on language usage and real corpora, it must be recognized that morphological variations observed are less extreme than those imaginable when the grammar is inspected. For example, in theory, Finnish nouns have approximately 2,000 different forms, yet in actual collections most of these forms occur very rarely [Kettunen and Airo 2006]. As a matter of fact, in Finnish 84% to 88% of inflected noun occurrences are generated by only six of a possible 14 grammatical cases.

While as a rule stemming schemes are designed to work with general texts, some are specifically designed for a given domain (e.g., medicine) or document collection (such as that developed by Xu and Croft [1998] for use in a corpus-based approach). This, in fact, more closely reflects general language usage (including word frequencies and other co-occurrence statistics), rather than a set of morphological rules in which the frequency of each rule (and thus its underlying importance) is not precisely known.

Studies of the English language have been more inventive, and various algorithmic stemmers have indeed been suggested for the most popular European languages, especially in conjunction with the CLEF[5] evaluation campaigns [Peters et al. 2008]. The Far East languages such as Chinese, Japanese, or Korean were previously evaluated within NTCIR[6] evaluation campaigns.

Although stemming procedures have been proposed for Hindi [Ramanathan and Rao 2003] and Bengali [Sakar and Bandyopadhyay 2008] as well as morphological analyzers[7] for Hindi and Marathi, there have been no reports of comparative evaluations of these propositions, based on test collections. During the FIRE 2008 evaluation campaign however, Gungaly and Mitra [2008] did propose a rule-based stemmer for the Bengali language (denoted "GM" in our experiments). In this same vein, we should mention the statistical stemmer suggested by Majumder et al. [2007]. Its cluster-based suffix stripping algorithm called "YASS"[8] does rely on a training set (list of words extracted from a document collection), allowing the system to ascertain which pertinent suffixes should be removed. The effectiveness of this statistical stemming could also be studied for other languages, as has been done previously for the Bengali and other European languages.

---

[5]See the Web site http://www.clef-campaign.org/.

[6]For more information, see the Web site at http://research.nii.ac.jp/ntcir/.

[7]Available at http://ltrc.iiit.net/showfile.php?filename=onlineServices/morph/index.htm.

[8]See http://www.isical.ac.in/~fire/Corpus_query_rel/clia-stemmer.tgz.

To date, most evaluation studies have involved IR stemmer performance evaluations for the English language, while those for other languages have been much less frequent. In their retrieval performance evaluations applying two statistical stemmers to five languages, Di Nunzio et al. [2004] demonstrated that across these languages wide variations could be found. Compared to statistical stemmers, Porter's stemmers seem to work slightly better, while for the German language Braschler and Ripplinger [2004] showed that for short queries stemming may enhance mean average precision by 23%, compared to 11% for longer queries. Tomlinson [2004] evaluated the differences between Porter's stemmer and the lexical stemmer, finding that the lexical stemmer tended to produce statistically better results for Finnish and German while for Dutch, Russian, Spanish, French, and English, performance differences were small and insignificant. For Swedish, when compared to a lexical stemming approach, the algorithmic stemmer resulted in statistically superior mean average precision (MAP). Finally, based on a study of eight European languages, Hollink et al. [2004] confirmed some of the above findings. Depending on the language, search systems obtained the best retrieval performances when either ignoring the stemming stage (English, French, and Italian), applying a stemmer (Finnish) or applying a decompounding procedure and then a stemmer (Dutch, Swedish). For the German language, the best performing scheme was based on a morphological analysis.

3. MORPHOLOGY

Given that Sanskrit is their root form, Hindi, Marathi, and Bengali are closely related, and as such their sentence structure follows the same Subject - Object - Verb (or SOV) pattern. Although from an IR perspective this aspect is not of primary importance, the IR models used in this article are based on the bag-of-words assumption wherein the absolute and relative positions of words within a sentence are ignored.

Indeed a closer inspection of the lexicons in these languages reveals that words with similar meanings may have similar spellings. An examples include the word "king", which can be spelled as "राजा" in Hindi and Marathi as "রাজা" in Bengali, while for other terms, spellings may be similar for only two of these languages, and the spelling of other words is completely different in all three languages (e.g., "God" is written as "ईश्वर" in Marathi, "ख़ुदा" in Hindi, and "ঈশ্বর" Bengali). As with other languages, lexicons in these languages are never free from the influence of others and vise versa. English borrows some words from the Indian languages, such as "jungle" (from a Sanskrit stem), "punch" (drink, from Hindi or Marathi), "jute" (vegetable fiber, from Bengali) or "curry" (from the Tamil). Similarly and to a larger extent, many word forms in Indian languages are borrowed from English, especially given its dominant presence in commerce (e.g., taxi, company, bank, budget, ice cream, and gasoline) and in technology (e.g., computer and Internet).

In their written forms, Hindi and Marathi employ the Devanagari script while the Bengali alphabet belongs to the Brahma family. The two scripts are however clearly related and share certain characteristics. All vowels except

the short "a" (written as "अ" in Devanagari and "অ" in Bengali) have two forms: first as an initial or syllable ("-आ", "-আ") and the second as a medial or final vowel (e.g., 'क' + 'आ' = "का" in Devanagari, 'ক' + 'আ' = "কা" in Bengali). Consonants appearing together in special clusters form conjunct letters (ligature) such as ('क' + 'क' = "क्क", 'क' + 'स' = "क्स" (Devanagari) and 'ক' + 'ক' = "ক্ক", 'ক' + 'ষ' = "ক্ষ" (Bengali)).

In the rest of this section we describe the key morphological characteristics of these three languages and analyze their impact on IR design and performance.

3.1 Key Features of Hindi Morphology

Hindi is spoken by approximately 500 million people (non-native included) and ranks second among the world's most spoken languages (Chinese is the first while English and Hindi have the same ranking). The term "Hindi language" however does not refer to a well-defined and clearly standardized language but rather to a relatively large group of dialects wherein interlingual understanding is always possible (just as English is in the UK and the U.S.).

Hindi is written using the Devanagari script, comprising 11 vowels and 33 simple consonants, and also nasal symbols such as anusvar ('ं') and anurasik ('ँ'), plus a symbol for the weak aspiration visgar ('ः') (although very rare in this language). Generally no distinction is made between uppercase and lowercase letters.

In Hindi grammar [Kellogg 1938] there are only two genders, masculine and feminine, while the neuter found in Sanskrit has disappeared. Feminine nouns are usually formed from the masculine, either by replacing the final '-आ'[9] ('ा') by '-ई' ('ी') (e.g., "घोड़ा" (horse), "घोड़ी" (mare)) or by adding '-ई' for nouns ending with a consonant "बंदर" (monkey), "बंदरी" (female monkey). Number is expressed through distinctive singular and plural forms.

This language does not have a definite article (the), and instead of placing prepositions before the noun, it positions them after in the form of postpositions (e.g., "on the table" → "table on"). These are used in certain Western European languages such as German, as in the expression "den Fluss *entlang*" (*along* the river), while the use of this linguistic construction in other Indo-European languages is clearly the exception.

Nouns and adjectives also have two distinct grammatical cases, direct and oblique. The direct case normally indicates the subject of a verb, while the oblique case might be combined with postpositions to form other object or adverbials complements (e.g., "John gives a *bone* to *Fido* in the *garden*").

Number and case are expressed through adding inflectional suffixes and occasionally certain particles to the stem or base form. To obtain the oblique singular form, most masculine nouns ending in '-आ' (written as 'ा' in medial or final form) inflect their final vowel to '-ए' ('े'), and those in '-आं' to '-एं' or into '-ए'. All such nouns inflected in the oblique singular retain the same form in the nominative plural, while for all other masculine nouns the nominative singular and plural have the same form.

---

[9]The drawing of the vowel may change if it is isolated or initial on the one hand, and on the other if it is in a medial or final position.

As an example, the masculine noun "horse" is written as "घोड़ा" in the direct singular while its oblique singular is "घोड़े" and as a rule it is used in conjunction with post-positions to designate other complements, as in "घोड़े को" (dative singular). As for plural forms, the direct case is written as "घोड़े" or the oblique case as "घोड़ों".

Hindi adjectives may be either inflected or uninflected. Uninflected adjectives remain unchanged before all nouns and under all circumstances, the same as with English adjectives (e.g., "सुंदर" (beautiful)). All inflected adjectives usually end in '-आ' (e.g., "काला" (black)) and their inflection depends on the gender and case of the noun they alter (e.g., as for the masculine noun "काला घोड़ा" (black horse), "काले घोड़े" (black horses) or with the feminine noun in "काली बिल्ली" (black cat), "काली बिल्लिज़ां" (black cats)) [Kellogg 1938].

Derivational morphology takes place in Hindi through adding a suffix to the stem, and typically the stem's part-of-speech (POS) changes once the suffix is added (e.g., "-ial" in "commerce" and "commercial"). In most cases the derivation is performed without modifying the stem itself, as in "लघिमा" (lightness) from "लघ" (light), although some changes do occur when forming adjectives, such as "सांसारिक" (worldly) derived from "संसार" (world).

The Hindi vocabulary is borrowed from both the Sanskrit and the Persian languages (with many terms also borrowed from Arabic via Persian), and as such Hindi may thus have two distinct words denoting the same item or a similar object (e.g., "पुस्तक" from Sanskrit or "किताब" from Persian). In these cases one is usually reserved as a technical term and the other for ordinary language. While this phenomenon is not unknown in English (e.g., "car" and "automobile" or "film" and "movie"), it occurs more frequently in Hindi and thus may impact retrieval effectiveness.

3.2 Key Features of Marathi Morphology

Marathi is spoken in western India by about 70 million people, and thus ranks fourth among the languages spoken there. As in other languages it may include various dialects, along with certain spelling and phonological variations.

Marathi is written in the Devanagari script as well as another variant, the Balbodh script. Marathi contains 52 letters, of which only 50 represent distinct sounds. These sounds are expressed by 14 vowels having different initial-leading forms and also different shapes when following consonants. There are 36 consonants in all, including two compound consonants as well as some nasal symbols.

As in Sanskrit, Marathi nouns may have three possible genders (masculine, feminine, and neutral) and be either singular or plural in number [Navalkar 2001]. Masculine, feminine, or neutral noun forms are derived through applying regular and simple rules (for example, a child "मुलगा" (masculine), "मुलगी" (feminine), "मुलग" (neutral); or for a dog "कुत्रा" (masculine), "कुत्री" (feminine), "कुत्रें" (neutral)). As in other languages there are certain exceptions, such as the noun "camel" which has two distinct forms ("उंट" (masculine), "मांड" (feminine)).

Table I. Examples of Marathi Noun and Adjective Declination

Case		"House"		"wise" (masc.)	
		Singular	Plural	Singular	Plural
Nominative		घर	घरें	शहाणा	शहाणा
Accusative		घर	घरें	शहाणा	शहाणा
Instrumental	by	घरानें	घरांशीं	शहाण्यानें	शहाण्यानें
	with	घराशीं	घरांस	शहाण्यानें	शहाण्यानें
Dative		घराला – स	घरांला – स – ना	शहाण्याला – म	शहाण्याला – म
Ablative		घराहून	घरांहूम	शहाण्याहून	शहाण्याहून
Genitive		घराचा	घरांचा	शहाण्याचा – ची – चें	शहाण्याचे – च्या – चीं
Locative		घरीं	घरीं	शहाण्यांत	शहाण्यांत
Vocative		घरा	घरांनों	शहाण्या	शहाण्या

The plural form of nouns depends on their gender. Masculine nouns ending in '-आ' become plural by changing the final vowel into '-ए', while others normally remain unchanged. The plural form of feminine nouns is usually derived by replacing the tailing '-अ' by '-इ' or by adding '-आ'. Neuter nouns ending in '-ए' usually become '-ईं' in the plural, while the rest become '-एं' (see Table I for a few examples).

Marathi is an inflected language with eight grammatical cases (nominative, accusative, instrumental, dative, ablative, genitive, locative, and vocative). A noun's inflectional termination depends on its case, number, and gender, thus resulting in the complex morpho-syntaxical construction often found in other Indo-European languages, such as Czech [Dolamic and Savoy 2010].

The examples shown in Table I demonstrate how a noun may change its stem to form what is known as a crude (unfinished or imperfect) form and thus accommodate the various case terminations (e.g., the word "घर" (house, nominative singular) becomes "घराचा" (dative and genitive singular). The basic form is usually formed by combining demonstrative pronouns 'या' (e.g., "आंबा" (mango) + 'या' = "आंब्या") or 'ई' ("भिंत" (wall) + 'ई' = "भिंती") with a noun. In certain declinations, these pronouns may also take on their impure forms 'आ' for 'या' (e.g., "घर" + 'आ' = "घरा") and 'ए' for 'ई' (e.g., "कथा" (tale) + 'ए' = "कथे"). Proper names for persons and certain terms used to express respect may reject the 'या' in the basic form, and thus for example the name Ravji "रावजी" becomes "रावजीला" (to Ravji) and not "रावज्याला" [Navalkar 2001].

In Marathi an adjective may be inflected according to the noun to which it is attached. When an adjective ends in '-आ' for example, it is generally inflected, otherwise it remains unaltered before the noun it qualifies. Finally, when an adjective is used as a substantive it is declined as such (see examples in Table II).

In Marathi there are four distinct ways of constructing the derivational morphology. First are the primary derivatives where only the radical vowel and/or consonant are modified (e.g., "डोळा" (an eye) → "डोळू" (an eyelet or a little hole)), and second are those derivatives in which a prefix or a suffix is added to a given stem (e.g., "रवोडी" (mischief) → "रवोडकर" (mischievous)). This method is generally applied in the derivation of new words adapted from the English language (e.g., from "history" we get the adjective "historic" or the related noun

Table II. Examples of Gender-Number Agreement for the Noun "Good"

"Good"	Singular	Plural
Masculine	चांगला	चांगले
Feminine	चांगली	चांगल्या
Neuter	चांगले	चांगली

"prehistory"). A third method of forming new words involves reduplicates (e.g., "लाललाल", literally "red red", meaning "very red"), and finally when two (or more) words are concatenated to form a new compound construction (such as "रण" (battle) + "भूमि" (field) = "रणभूमि" (battlefield)).

3.3 Key Features of Bengali Morphology

Approximately 250 million people (including non-native speakers) speak Bengali (or Bangla) in the eastern part of India and in Bangladesh, and thus it ranks second among the languages spoken in India. Although closely related to that used in Hindi, Bengali has its own script and an alphabet consisting of 35 consonants: 11 vowels plus five modifying symbols.

While the adjectival and nominal morphology in Bengali is very light, its verbs are highly inflected. Nouns are inflected according to seven grammatical cases (nominative, accusative, instrumental, ablative, genitive locative, and vocative), number (singular or plural) and determiners. The vocative is included in this list, yet strictly speaking it is not a case because it is identical in form to the nominative and can be distinguished by various prefixes. Note that adjectives are invariable, and this simplifies the automatic processing of Bengali texts.

Bengali makes no use of gender distinction and thus all nouns are declined using the same terminations. Stems are usually not affected by the application of inflections and case-marking patterns may depend on a noun's degree of animacy (e.g., human beings, living beings other than human or inanimate objects). The noun "সন্তান", for example, appears as such in the singular nominative, instrumental and ablative cases, but varies in other cases "সন্তানকে" (accusative, dative), "সন্তানেক" (genitive) and "সন্তানে" or "সন্তানেতে" (locative), while the plural forms of this noun are "সন্তানেরা", "সন্তান", and "সন্তানদের" [Beames 1891]. To express correct meaning more precisely, Bengali makes use of post-positions rather than the prepositions found in English.

The determiner in Bengali is attached to the noun in the form of a suffix. The definite article for example adds the suffix '-টা' or '-টি' in the singular or in the plural adds the suffix '-দের' (animate) or the suffixes '-গুলা' or '-গুলি' (inanimate). Particles representing determiners must be placed before the case ending (e.g., "ছাত্র" (student) gives "ছাত্রটার" (the student's) and "ছাত্রদের" for the plural (the students')).

Note that additional suffixes may be found, such as those added to indicate measure, words added after the numeral and those that normally precede the noun. This suffix then becomes '-টা' (the same as the definite article) or '-জোণ' (reserved for persons) ("অনেকজন লোক" → "many people").

4. SUGGESTED STEMMING STRATEGIES

For the three Indian languages we created light stemmers, the same strategy we have suggested for other European languages over the past few years [Savoy 2006; Dolamic and Savoy 2010]. In our opinion effective stemming should focus mainly on inflectional suffixes attached to nouns and adjectives (used to sustain most of a document's meaning) and ignore numerous verb forms. Also, attempting to conflate all verb forms under a common stem tends to generate more stemming errors than benefits. Moreover, the stemmed forms obtained removing suffixes related to number, gender, and case variations tend to contain fewer erroneous forms and more often preserve the correct meaning of the word involved. Additionally, most users are more capable of understanding the results of a light stemming procedure returning the dictionary entries ("initiatives" $\rightarrow$ "initiative") than a more aggressive procedure returning obscure terms ("initiatives" $\rightarrow$ "initi").

In our experiments "light" stemmers removed only the inflectional suffixes from nouns and adjectives, and did not account for exceptions present in all natural languages (e.g., "mice" and "mouse"). For the Hindi language we suggested a light stemmer based on 20 rules, while for Marathi we created 51 rules and for Bengali 70 rules.

Suffixes may also be used to derive new words from a stem, usually by changing its part-of-speech (POS) (e.g., "care" and "careful" or "carefulness"). Thus for each language studied we also proposed and evaluated a more aggressive stemmer that not only removed inflectional suffixes from nouns and adjectives, but also removed a limited number of derivational suffixes. To develop this more elaborate stemmer (denoted "aggressive" in our experiments), we designated 49 rules for the Hindi language, 31 rules for Marathi, and 85 for Bengali.

Finally, to identify pertinent matches between search keywords and documents we removed very frequently occurring and insignificant terms such as "the," "but," "some," "we," "that," and "have". Based on the guidelines provided by Fox [1990], we proposed a stopword list containing 165 Hindi, 114 Bengali, and 99 Marathi terms. These lists were rather conservative and mainly included only determinants (e.g., "the," "this"), postpositions ("in," "near"), various pronouns ("we," "my") and conjunctions ("and," "while," "because"). They were also rather short compared to those of other Indo-European languages (e.g., for the English language the SMART system [Salton 1971] suggests 571 words).

5. TEST COLLECTIONS

The evaluations reported in this article were based on the test collections built for the Hindi, Marathi, and Bengali languages during the first FIRE 2008 evaluation campaign. These corpora consist of newspaper articles extracted from the *Jagran* newspaper for the Hindi language, from the *Maharashtra Times* and *Sakal* for Marathi (articles spanning the period April 2004 through September 2007), and from the CRI and *Anandabazar Patrika* (a newspaper edited by ABP Ltd.) for Bengali. The encoding system used for both documents and topic formulation is UTF-8.

Table III. FIRE 2008 Test Collection Statistics

	Hindi (HI)	Marathi (MR)	Bengali (BN)
Size (in MB)	718 MB	487 MB	732 MB
# of documents	95,215	99,357	123,047
# of distinct terms	127,658	511,550	249,215
Number of indexing terms per document			
Mean	356.2	264.6	291.88
Standard deviation	400.43	188.96	180.62
Median	256	222	265
Maximum	6,998	5,077	2,928
Minimum	0	28	0
Number of topics	45	73	75
Number rel. items	3,436	1,534	2,610
Mean rel./topic	76.4	21.0	34.8
Median	67	16	28
Maximum	194 (T#60)	123 (T#4)	149 (T#32)
Minimum	1 (T#59,T#66)	1 (T#12, #23) 1 (T#47, #50, #72)	4 (T#23)

Table III lists statistics on the three corpora, showing that the Hindi and Bengali collections are similar in size (in MB) while the Marathi is smaller. The Bengali corpus contains the largest number of documents, while the Hindi or Marathi collections contain a relatively similar numbers. The Hindi corpus has a greater mean document length (based on the mean number of indexing terms per article, following stopword removal), while the Bengali and Marathi corpora have similar mean document lengths (about 275 indexing terms/article), based on the same measuring technique.

The Hindi, Marathi, and Bengali language test collections used in this study contain 45, 73, and 75 topics respectively. The available topics cover various subjects (e.g., Topic #028: "Iran's Nuclear Programme," Topic #034: "Jessica Lall Murder") including cultural issues (Topic #041: "Kolkata Book Fair 2007" or Topic #070: "Remake in Bollywood"), scientific problems (Topic #045: "Global Warming"), or sports (Topic #073: "Zinedine Zidane's headbutting incident at the World Cup"). Certain topics seem to be more national in coverage (Topic #041: "New Labour Laws in France," Topic #058: "Thailand Coup"), while in others the real subject being covered is sometimes difficult to determine, at least based on the title section (Topic #049: "Worldwide natural calamities," Topic #052: "Budget 2006-2007"). Topic descriptions tend to contain many proper names (e.g., geographical with "Singur," "China," "Kolkata," personal names such as "Bush," "Sania Mirza," or products such as "Prince" and "Bofors"), as well as acronyms ("ULFA," "CBI," "HIV," "LOC").

Based on the TREC model, each topic formulation was divided into three logical sections, beginning with a brief title (under the tag <TITLE>, see Figure 1) containing between two and four words, followed by a one-sentence description (tag <DESC>) the user's information need, and finally, a narrative part specifying relevance assessment criteria (tag <NARR>). Full examples written in the Hindi, Marathi, Bengali, and English languages are depicted in Figure 1. In our experiments we used only the title part of topic description, thus more closely reflecting requests sent to commercial search engines. This resulted in

<TOP lang="hi">
<NUM> 30 </NUM>
<TITLE> भारत के रेल मंत्री के रूप में लालू प्रसाद यादव </TITLE>
<DESC> रेलमंत्री के रूप में लालू प्रसाद यादव की भूमिका </DESC>
<NARR> रेलमंत्री के रूप में लालू की भूमिका, आधार संरचना के रूप में रेलवे का उन्नयन, अपने कार्यकाल के दौरान उसके द्वारा प्रस्तुत बजट के गुणदोष, उच्च शक्ति जाँच आयोग बिठा कर गोधरा रेल दुर्घटना की गुत्थी सुलझाने में लालू की भूमिका आदि सूचनाओं को संबद्ध प्रलेख में शामिल किया जाए। इनके अलावा अन्य सूचनाएँ यहाँ संगत नहीं हैं। </NARR> </TOP >

<TOP lang="mar">
<NUM> 30 </NUM>
<TITLE> भारतीय रेल्वे मंत्री म्हणून लालू प्रसाद यादव. </TITLE>
<DESC> भारतीय रेल्वे मंत्री म्हणून लालू प्रसाद यादवांची भूमिका. </DESC>
<NARR> एक रेल्वेमंत्री म्हणून लालूंची भूमिका, पायाभूत सोयींच्या बाबतीत रेल्वेच्या दर्जामध्ये सुधारणा, त्यांच्या कारकिर्दीत त्यांनी तयार केलेल्या रेल्वे अंदाजपत्रकातील सर्व लहान मोठ्या बाबी, गोध्रा येथील रेल्वेगाडी घटनेचा उलगडा करण्या-मध्ये उच्चाधिकार चौकशी आयोग बोलाविण्यामधील लालूंची भूमिका या संबंधीची माहिती संबंधित कागदपत्रात असली पाहिजे. या व्यतिरिक्त इतर माहिती येथे सुसंगत नाही. </NARR>
</TOP >

<TOP lang="bn">
<NUM> 30 </NUM>
<TITLE> রেল মন্ত্রী হিসেবে লালু প্রসাদ যাদব </TITLE>
<DESC> রেল মন্ত্রী লালু প্রসাদ যাদব এবং তাঁর আমলে ভারতীয় রেল সম্বন্ধে নথি খুঁজে বার করো। </DESC>
<NARR> রেল মন্ত্রী লালু প্রসাদ যাদবের সময়কালে ভারতীয় রেলের নিরাপত্তা, পরিকাঠামোগত উন্নতি এবং বিভিন্ন বিতর্ক প্রাসঙ্গিক নথিতে থাকা চাই। </NARR> </TOP >

<TOP lang="en">
<NUM> 30 </NUM>
<TITLE> Laloo Prasad Yadav as the Railway Minister </TITLE>
<DESC> The performance of Laloo Prasad Yadav and the Indian rail in his tenure</DESC>
<NARR> A relevant document should contain information about the safety measures taken by the Indian Railways, or infrastructural improvements planned or undertaken during the tenure of Laloo Prasad Yadav. Information about disputes / controversies surrounding Laloo are only relevant if they pertain to the Railways. </NARR> </TOP >

Fig. 1. Examples of topic description for the Hindi, Marathi, Bengali, and English languages.

a mean query size of 3.8 search terms for Hindi, 3.79 for Marathi, and 3.65 for Bengali (following the removal of stoplist words).

The bottom rows of Table III also compare the number of relevant documents per request, showing that the mean was always greater than the median (e.g., for Marathi, the average number of relevant documents per query was 21.0, and its corresponding median was 16). These findings indicate that for each topic only a comparatively small number of relevant items were found. No relevant records were found in the collection for five Hindi topics (#40, #43, #47, #48, and #50) while for Marathi Topic #70 ("Remake in Bollywood") did not have any relevant items.

Topic #32 ("Relations between Congress and its allies") returned the largest number of relevant articles in the Bengali collection (149), while Topic #59 ("Protests by American citizens against Iraq War") and Topic #66 ("Khadim owner abduction case") returned the smallest number of relevant documents (one in this case and only for the Hindi corpus).

6. INFORMATION RETRIEVAL MODELS

In order to ensure useful conclusions would be obtained when analyzing new test collections, we considered it important to evaluate retrieval performance under varying conditions to develop a broad perspective. We thus evaluated a variety of indexing and search models, ranging from classical *tf idf* indexing schemes to more complex probabilistic models.

To evaluate and analyze different stemming approaches with respect to various IR models, we first used the classical *tf idf* vector-space model wherein the weight attached to each indexing term was the product of its term occurrence frequency (tf_{ij} for indexing term t_j in document d_i) and the logarithm of its inverse document frequency (idf_j). To measure similarities between documents and requests, we computed the inner product after normalizing (cosine) the indexing weights [Manning et al. 2008].

Better weighting schemes have been suggested for the vector-space model, especially in cases where a term's occurrence in a document might be viewed as a rare event. A good practice may thus be to assign more importance to the first occurrence of a term compared to its successive and repeating occurrences, where the *tf* component is computed as ln(*tf*) + 1 or as ln(ln(*tf*)+1)+1. A term's presence in a shorter document might also provide stronger evidence than in a longer document, and thus to account for document length we could make use of more complex IR models, including the *Lnu-ltc* forms suggested [Buckley et al. 1996]. In this case Equation 1 calculates the indexing weight assigned to a document term (*Lnu*) while Equation 2 provides the indexing weight assigned to query term (*ltc*).

$$w_{ij} = \frac{\left((\ln(tf_{ij}) + 1) \Big/ (\ln(\text{mean tf}) + 1)\right)}{(1 - slope) \cdot pivot + slope \cdot nt_i} \tag{1}$$

$$w_{qj} = ((\ln(tf_{qj}) + 1) - idf_j \Big/ \sqrt{\sum_{k=1}^{t}((\ln(tf_{qk}) + 1) \cdot idf_k)^2} \tag{2}$$

where $nt_i$is the number of distinct indexing terms in document d_i and *pivot* and *slope* are two constants used to adjust term weight normalization values, according to document length. The value of the constant slope was fixed at 0.2 for all languages while pivot represents mean number of distinct indexing terms per document. This formulation prevents the retrieval system from overly favoring short documents compared to those articles longer than the mean, according to the pivot value.

To complement this vector-space model, we implemented three probabilistic models representing three different paradigms. First, we implemented the well-known Okapi (or BM25) approach [Robertson et al. 2000], regularly producing high retrieval effectiveness for various test-collections. Second, we included a model derived from the *Divergence from Randomness* (DFR) paradigm

[Amati and van Rijsbergen 2002], combining two information measures formulated as:

$$w_{ij} = \text{Inf}^1_{ij}(tf) \cdot \text{Inf}^2_{ij}(tf) = -\log_2[\text{Prob}^1_{ij}(tf)] \cdot \left(1 - \text{Prob}^2_{ij}(tf)\right), \tag{3}$$

where for the first information factor, $\text{Prob}^1_{ij}(tf)$ represents the pure chance probability of finding tf_{ij} occurrences of the term t_j in a document. If this probability is high, term t_j would correspond to a non-content bearing word within the context of the entire collection [Harter 1975] while otherwise if $\text{Prob}^1_{ij}(tf)$ is small (or if $-\log_2[\text{Prob}^1_{ij}(tf)]$ is high), the term t_j would provide important information regarding the content of the document d_i. The second information measure depends on $\text{Prob}^2_{ij}(tf)$, the probability of having $tf+1$ occurrences of the term t_j, knowing that tf occurrences of this term have already been found in document d_i. To implement these two underlying probabilities, we selected the $I(n_e)C2$ model based on the following formulae:

$$\begin{aligned} \text{Inf}^1_{ij} &= tfn_{ij} \cdot \log_2[(n+1)/(n_e+0.5)] \\ &\text{with } n_e = n \cdot [1 - [(n-1)/n]^{tc_j}] \\ &\text{and } tfn_{ij} = tf_{ij} \cdot \log_2[1 + ((c \cdot mean\ dl)/l_i)] \end{aligned} \tag{4}$$

$$Prob^2_{ij} = 1 - [(tc_j + 1)/(df_j \cdot (tfn_{ij} + 1))], \tag{5}$$

where tc_j indicates the number of occurrences of term t_j in the collection, n the number of documents in the corpus, *mean dl* the mean length of a document, and l_i the length of document d_i.

Finally we also used an approach based on a language model (LM) [Hiemstra 2000], known as the non-parametric probabilistic model. Within this language model paradigm various implementations and smoothing methods [Zhai and Lafferty 2004] might also be considered, and in this article we adopted the model proposed by Hiemstra [2000] as described in Equation 6, using Jelinek-Mercer smoothing and combining an estimate based on document ($P[t_j \mid d_i]$) and another based on the entire corpus ($P[t_j \mid C]$).

$$\text{Prob}[q_i|q] = \text{Prob}[d_i] \cdot \prod_{t_j \in Q} [\lambda_j \cdot \text{Prob}[t_j|d_i] + (1 - \lambda_j) \cdot \text{Prob}[t_j|C]]$$

$$\text{with Prob}[t_j|d_i] = \left(\frac{tf_{ij}}{I_j}\right) \text{and Prob}[t_j|C] = \left(\frac{df_j}{lc}\right) \text{with } lc = \sum_{k=1}^{t} df_k \tag{6}$$

where λ_j is a smoothing factor (fixed at 0.35 for all indexing terms t_j), df_j the number of documents indexed with the term t_j, and lc a constant related to the underlying corpus C.

7. EVALUATION

To evaluate the various indexing and search strategies, we adopted the mean average precision (MAP) method of measuring retrieval performance (computed by the TREC_EVAL software, based on a maximum of 1,000 retrieved

records). Used by all evaluation campaigns for around 20 years, this performance measure is able to objectively compare various IR models, especially their ability to retrieve relevant items (ad hoc tasks) [Buckley and Voorhees 2005].

Using MAP to measure a system's performance signifies that we attached equal importance to all queries. Comparisons between two IR strategies would therefore not be based on a single query relative to those available in the underlying test collection or specifically created to demonstrate that a given IR approach must be rejected. Thus we believe that it is important to conduct experiments involving the largest possible number of observations (between 45 and 75 queries in our evaluations, depending on the language).

To statistically determine whether or not a given search strategy would be better than another, we applied the bootstrap methodology [Savoy 1997]. This led to a conclusion very similar to that of the *t*-test method but did not require parametric assumptions [Abdou and Savoy 2006]. In our statistical tests, the null hypothesis H_0 stated that both retrieval schemes produced similar MAP performance. This null hypothesis would be accepted if two retrieval schemes returned statistically similar MAP, otherwise it would be rejected. Thus, in the experiments presented in this article, statistically significant differences were detected by a two-sided test (significance level $\alpha = 5\%$).

7.1 IR Model Evaluation

We evaluated the various IR models described in the previous section, applying them to the Hindi (see Table IV), Marathi (Table V) and Bengali test collections (Table VI). These tables report the MAP achieved by the IR models when applying the four different stemming strategies for Hindi and Marathi (e.g., "None," "Light," "Aggress," and "YASS"), and five for Bengali (e.g., "GM" was added to the other four stemming schemes). The last two columns in each table list the retrieval performances produced by two language independent indexing strategies, where the "Trunc-4" column lists results when simply truncating the term into its first four letters (e.g., "goodness" generates "good"), while the last column lists evaluation results obtained by applying the 4-gram indexing approach (e.g., "minister" gives "mini", "inis", . . . , "ster") [McNamee and Mayfield 2004; McNamee et al. 2009]. The fixed length of 4 was selected for both the truncating and *n*-gram methods because it produced the best IR performance for all three languages.

Table IV lists the best performance results for the Hindi language, obtained with either the $I(n_e)C2$ model derived from the Divergence from Randomness paradigm or the vector-space *Lnu-ltc*. As shown by the results listed in bold this latter scheme performed best when we applied an aggressive stemming, YASS stemmer, or when we ignored this word normalization procedure. Tables V and VI show that for both Bengali and Marathi the best performing model was always the $I(n_e)C2$ approach.

A difference in mean performance, particularly when small, did not always indicate differences that might be clearly perceived by the final user. A cross ("†") in these tables indicates which retrieval models resulted in statistically

Table IV. MAP of Various Indexing Strategies and IR Models for the Hindi Language (45 queries)

	Mean Average Precision					
	None	Light	Aggress	YASS	Trunc-4	4-gram
tf idf	0.1548†	0.1756†*	0.1748†	0.1588†	0.1987†	0.1750†
Lnu-ltc	**0.2368**	0.2844*	**0.2981***	**0.2843**	0.2852	0.2516
Okapi	0.2179	0.2601*	0.2811*	0.2598	0.2867*	0.2495†
I(n_e)C2	0.2311	**0.2692***	0.2936*	0.2753	**0.2966***	**0.2629**
LM	0.1872†	0.2369†*	0.2640†*	0.2368†	0.2730†*	0.2199†
Average	0.2056	0.2452	0.2623	0.2430	0.2680	0.2318
% change		+19.3%	+27.6%	+18.2%	+30.4%	+12.8%

Note: Best scores in each column are shown in boldface, significant differences compared to the best performance are indicated by an † while significant differences with no stemming are denoted by an *.

Table V. MAP of Various Indexing Strategies and IR Models for the Marathi Language (73 queries)

	Mean Average Precision					
	None	Light	Aggress	YASS	Trunc-4	4-gram
tf idf	0.1844†	0.1920†	0.2518†*	0.1886†	0.2299†	0.2394†
Lnu-ltc	0.2152†	0.2542†*	0.3085*	0.2507*	0.3137*	0.2929†*
Okapi	0.2359	0.2759*	0.3438*	0.2626	0.3307*	0.3268*
I(n_e)C2	**0.2416**	**0.2839***	**0.3517***	**0.2770***	**0.3368***	**0.3418***
LM	0.2232	0.2480†	0.3027*	0.2472†	0.3102*	0.2929†*
Average	0.2201	0.2508	0.3117	0.2452	0.3043	0.2988
% change		+13.9%	+41.6%	+11.4%	+38.3%	+35.8%

Note: Best scores in each column are shown in boldface, significant differences compared to the best performance are indicated by an † while significant differences with no stemming are denoted by an *.

significant performance differences, compared to the best performing models. In this case, the classical *tf idf* vector-space model and the language model (LM) typically resulted in significantly lower performance levels. For the other models, the outcome varied according to the language and the indexing scheme involved. It is however evident that performance differences between the *Lnu-ltc* and I(n_e)C2 for the Hindi language were never significant, while for the Bengali corpus performance differences between the I(n_e)C2 and the other approaches always tended to be significant (exceptions can be found only in the "None" column, see Table VI).

7.2 Stemming Evaluation

The Hindi, Marathi, and Bengali morphologies are more complex than that of the English language and thus their MAP values could be improved by applying a stemming procedure that would conflate different surface words with similar meanings under the same stem or indexing unit. If this assumption is true, we could then consider a variety of stemming strategies, be they light or more aggressive. The question then arises as to whether stemming would affect IR performances for these various languages and to what extent?

Tables IV (Hindi), V (Marathi), and VI (Bengali) list the results of our first retrieval performance evaluations in which stemming was omitted, and lists

Table VI. MAP of Various Indexing Strategies and IR Models for the Bengali Language (75 queries)

	Mean Average Precision						
	None	Light	Aggress	YASS	GM	Trunc-4	4-gram
tf idf	0.1876†	0.2015†	0.2144†*	0.2247†*	0.2114†*	0.2102†	0.1987†
Lnu-ltc	0.2539	0.2897†*	0.2979†*	0.3058†*	0.2831†*	0.3242†*	0.2590†
Okapi	0.2662	0.2966†*	0.3066†*	0.3066†*	0.2893†*	0.3310†*	0.2662†
$I(n_e)C2$	**0.2628**	**0.3064***	**0.3132***	**0.3243***	**0.2990***	**0.3390***	**0.2830**
LM	0.2353†	0.2683†*	0.2780†*	0.2747†*	0.2585†*	0.2947†*	0.2418†
Average	0.2412	0.2725	0.2820	0.2878	0.2683	0.2998	0.2497
% change		+13.7%	+17.7%	+20.1%	+11.9%	+25.0%	+4.2%

Note: Best scores in each column are shown in boldface, significant differences compared to the best performance are indicated by an † while significant differences with no stemming are denoted by an *.

the MAP values in the "None" column. The "Light" column lists retrieval performances obtained by applying a light stemmer and the "Aggress" column a more aggressive stemmer**.** In the "YASS" column we reported the MAP values obtained using the statistical stemmer (with a threshold value set at 1.5). As a training set, we generated one word list per language using the respective document collection. Although using the same data for both the training and the test is usually not considered fair [Sebastiani 2002], in our case, we view this as the upper limit of the retrieval performance that might be achieved by this approach. We also evaluated the performance of the GM light, rule-based stemmer for the Bengali language only.

To provide an overview of MAP values for each of these stemming strategies, in each table we listed the average retrieval performance for all five IR models in line just before the last. Finally, the last row (labeled "% change") lists the results obtained by comparing percentage improvement to mean performance obtained when we ignored the stemming stage (listed in the "None" column).

The last two rows in these same tables show that in all approaches that applied stemming, performances were much more effective than those that omitted stemming, a finding that holds for all three languages studied. More precisely for the Hindi language, relative increases ranged from 19.3% with a light stemmer to 27.6% with the more aggressive approach. For Marathi this increase was from 11.4% with the YASS stemmer to 41.6% with the aggressive stemmer. Finally, for Bengali the range of improvement was relatively similar across all four stemmers, ranging from 11.9% with the GM stemmer to 20.1% with the YASS stemmer. Based on this data, we found that a more aggressive stemmer tended to result in better MAP while for some languages (e.g., Bengali) the performance difference between a light and an aggressive stemmer was not significant. Moreover, when compared to MAP found for certain European languages, these relative improvements after stemming for these three Indian languages were quite large (e.g., 4% for English, 4.1% for Dutch, 7% for Spanish, 9% for French, 15% for Italian, 19% for German, 29% for Swedish, 40% for Finnish [Tomlinson 2004], and 45% for Czech [Dolamic and Savoy 2010]).

With the no stemming approach as a baseline and after applying statistical tests, the differences between this approach and the two algorithmic stemmers

were very often statistically significant (marked with "*" in the tables) for all three languages tested. For Marathi only two exceptions were found (see Table V), where performance differences resulting from the classical *tf idf* or the LM models cannot be viewed as significant (*tf idf*: 0.1844 vs. 0.1920, LM: 0.2232 vs. 0.2480).

An analysis of the YASS statistical stemmer shows that the performance differences with a no stemming approach are always statistically significant for the Bengali language (Table VI). For Hindi (Table IV), these differences are never significant, while for Marathi (Table V), the differences fall somewhere between these two extreme cases, with some being significant (*Lnu-ltc* or I(n_e)C2 models) while others are not.

7.3 Query-by-Query Analysis

To determine the outcome of applying stemming, we performed a query-by-query analysis, concentrating on DFR-I(n_e)C2, the best performing retrieval model, and for each query comparing average precision (AP) before and after applying a stemmer.

For Hindi, we found that the primary reason for this improved performance was the application of the stemming approach. The topics and their corresponding relevant documents contained the same words but they were expressed in different grammatical cases, and even though the two strings were not identical before stemming, they were indeed conflated to the same stem, thus resulting in the best performance. As an example, the title of Topic #33 ("President Bush visits India") contained "बुश" and "भारत" ("Bush" and "India" respectively, in the direct case), while a large number of relevant documents contained "बुशा" ("Bush") or "भारतीय" ("India") in the oblique case.

In the Marathi case Topic #41 ("New labor laws in France") is an example that might demonstrate the second advantage of applying a stemming procedure. This topic title contains the term "फ्रान्स" ("France") while the relevant documents contain the terms "फ्रान्सचे", "फ्रान्सचा", "फ्रान्सचया", all of which are conflated to the same stem by the aggressive stemmer and thus result in average precision changes from 0.0111 ("None") to 0.6389 ("Aggress"). The topic title also contains the noun "कायदे" (law) while some relevant documents contain only the derivational term "कायदयावर" (legal). With the light stemmer, the various surface forms were not conflated under the same stem.

For Bengali the largest AP differences between the various indexing strategies were observed with Topic #58 ("Thailand Cup"). The term "থাইল্যাণ্ডের" ("Thailand" in the genitive case) was only found in the topic formulation, while the terms "থাইল্যাণ্ড" and "থাইল্যাণ্ডর" found in relevant documents had been conflated to the same stem for all stemming strategies applied, thus resulting in much better AP for that particular stemming method. Certain relevant documents did however contain the form "তাইল্যাণ্ড", a spelling variation of the country name.

Based on the query-by-query analysis described above we could see that for these languages stemming did have some benefits, while over-stemming or under-stemming could result in decreasing performance. Thus, for the Hindi language when comparing light stemming to no stemming, higher MAP values were obtained for 19 topics, while 18 topics show decreased performance. For

the Marathi and Bengali languages these differences are somewhat greater (e.g., 46 vs. 23 topics and 48 vs. 27 topics language respectively).

With the statistical stemmer YASS, the results showed that stemming strategy seemed to result in retrieval performances comparable to those of the light stemming, for both the Hindi (Table IV) and Marathi (Table V) languages. On the other hand for Bengali (Table VI), the YASS stemmer resulted in better performance than for both the light and aggressive stemming approaches, but these differences were not statistically significant. The restrained set of frequent suffixes used for the Bengali language and the fact that the YASS parameters were better adapted to the Bengali corpus than were the Hindi or the Marathi could possibly explain these results.

7.4 Word-Based vs. n-gram Indexing Strategies

The last two columns of the previous tables report our test results for the Indian languages: Hindi, Marathi, and Bengali, showing the retrieval performances obtained by "Trunc-4" and "4-gram", the two language independent approaches we applied. For these indexing strategies we ignored each language's underlying morphology and syntax, assuming that the first part of the word (trunc-*n*) or character sequence (*n*-gram) would provide the information needed to obtain a pertinent match between search keywords and document surrogates.

Based on the retrieval performances obtained for the Hindi (Table IV), Marathi (Table V) or Bengali (Table VI) languages, the simple trunc-*n* (or trunc-4 in our evaluation) resulted in higher performance levels. On average for both Hindi and Bengali, this indexing approach led to the best retrieval effectiveness. Moreover with the *n*-gram approach the mean performance differences were relatively large (Hindi: 0.2680 vs. 0.2318, -13.6%; Bengali: 0.2998 vs. 0.2497, -16.7%) while for Marathi the average retrieval performances were similar (0.3043 vs. 0.2988, -1.8%).

To verify whether these differences could be considered significant we performed a statistical test using the "Trunc-4" indexing approach as a baseline with both the Hindi and Bengali languages. For Hindi (see Table IV), performance differences were always statistically significant, compared to the "None" and "4-gram" approaches (except for *tf idf* model). When compared to the light, aggressive or YASS stemming approaches, the trunc-4 differences were never statistically significant. For Bengali (see Table VI) the performance differences were never statistically significant when compared to both the aggressive or YASS stemming approaches, yet for other approaches they were mostly significant (except for *tf idf* model).

For Marathi (Table V) the best overall performance was achieved using the aggressive stemming approach, and with this best performance as baseline, the performance differences were always statistically significant when compared to the "None", light or YASS stemming approaches, but they were not significant when compared to "Trunc-4" or "4-gram" strategies.

To provide a more complete picture we should mention that for Bengali and Hindi, the MAP differences between "None" and 4-gram indexing strategies were never statistically significant.

To obtain a better understanding of when word-based or language independent indexing strategies could lead to the best performances, we analyzed a few specific queries. As a first explanation for performance differences, we noted certain spelling variations in topic formulations and in relevant documents. With the Hindi corpus for example in Topic #44 ("Terrorist attacks in Britain"), the term terrorist was spelled "अआतकवादी" while it was spelled "आतकवादी" in the relevant documents. In the Marathi corpus, we found a similar situation with Topic #39 ("Attacks on American soldiers in Iraq") where the corresponding script was "बाँबस्फोट" while in the relevant documents it was "बॉबस्फोट" or "बॉमबस्फोट". In these cases the 4-gram was the best performing strategy, producing at least one match for both terms. In Marathi with Topic #9 we also encountered a spelling problem, where "Israel" was spelled "ईस्राएल" in the topic formulation and "इस्राईल" in the relevant documents. In this case also the 4-gram approach performed better than a word-based indexing scheme.

For the Marathi case, Topic #41 ("New labor laws in France") can be cited as an example of improved retrieval effectiveness after applying the word-based indexing approach. The topic title contains the term "फ्रान्स" ("France") while the relevant documents contain the following terms "फ्रान्सचे", "फ्रान्सचा", "फ्रान्सचया", all of which were conflated to the same stem by the aggressive stemmer, and thus causing the average precision to change from 0.1946 with trunc-4 to 0.6389 with the aggressive stemmer due to over-stemming by trunc-4 strategy. For this query the fixed limit of 4 was clearly too small.

7.5 Stopword List Evaluation

During the indexing of documents or queries, it is assumed that very frequent word forms having no precise meaning (e.g., "the," "you," "of," "is") may be removed. In fact, each match between a query and a document should be based on pertinent terms, rather than retrieving documents simply because they contain words such as "an," "ours," or "but".

In our final experiments we compared the retrieval effectiveness of various IR models, with and without the suggested stopword list, for each of the three languages studied. For Marathi (the stopword list contained 99 words) the mean difference across the five retrieval models tested was around 1%, while for Bengali (114 stopwords) this difference was around 2%. For both these languages the differences were never statistically significant.

For Hindi (165 forms in the stopword list) however the mean differences between various search models with or without applying a stopword list were larger. Table VII shows the results obtained when ignoring the stemming stage (columns labeled "No stemmer") and after applying a light stemmer ("Light stemmer"). As shown in the last two rows, the average performance differences were around 20% for both stemming strategies. Using the retrieval performance with stopword removal as a baseline, any significant MAP differences detected were also listed in Table VII and marked with the symbol "*". This shows that the differences were always statistically significant, except for those obtained using the *Lnu-ltc* model with no stemmer applied (0.2368 vs. 0.2182).

Table VII. MAP with and without Stopword Removal for the Hindi Corpus (45 queries)

	Mean Average Precision			
	No stemmer		Light stemmer	
	With	Without	With	Without
tf idf	0.1548	0.1024*	0.1756	0.1187*
Lnu-ltc	0.2368	0.2182	0.2844	0.2547*
Okapi	0.2179	0.1593*	0.2601	0.1969*
I(n_e)C2	**0.2311**	**0.2020***	**0.2692**	**0.2374***
LM	0.1872	0.1664*	0.2369	0.2138*
Average	0.2056	0.1697	0.2452	0.2043
% change	+21.2%		+20.0%	

Note: Best scores in each column are shown in boldface, significant differences compared to the no stemming performance are indicated by an *.

Based on our analysis of mean query length across the three languages, we found that removing the stopwords only slightly changed the averages for the Marathi (from 4.04 to 3.78 search terms) or Bengali languages (from 3.80 to 3.64 terms). For Hindi, however, this difference was somewhat greater, decreasing from 4.80 indexing terms without stopword removal to 3.8 terms if this step was performed.

For Hindi this important improvement resulting from stopword removal can be partially explained by an analysis of Topic #27 ("Relations between India and China"), showing an AP of 0.2690 after removal vs. 0.0532 before stopword removal. The situation was similar with Topic #38 ("Uneasy truce between Greg Chappell and Sourav Ganguly"), with an AP of 0.6055 (after) vs. 0.2221 (before) or Topic #54 ("HIV and AIDS epidemic"), providing an AP of 0.7271 (after) vs. 0.2929 (before). In these three topics, it was the term "और" meaning "and" that made the difference. This word did not have high document frequency because in Hindi other words can be used to express "and" (in fact its frequency in the underlying document is similar to a word like "China"). This resulted in the term being incorrectly viewed as pertinent for the given queries and this decreased the resulting retrieval effectiveness.

8. CONCLUSION

This article presents the main morphological characteristics of the Hindi, Marathi, and Bengali languages. To facilitate IR operations in each of these Indo-Aryan languages we suggest two algorithmic stemmers, one to remove only inflectional suffixes (denoted "Light") and a second to remove certain frequently occurring derivational suffixes (listed in our tables under the heading "Aggress"). As an alternative stemming approach, we apply and evaluate the YASS statistical stemmer. To compare these word-based indexing models with the language independent approaches, we then include the *n*-gram and trunc-*n* indexing schemes. We also propose a stoplist for each of these languages which for contains the Hindi language 165 words, 99 for Marathi, and 114 for Bengali.

To evaluate and compare these various indexing approaches, we use five different IR models corresponding to different probabilistic approaches (Okapi,

one model derived from the Divergence from Randomness (DFR) paradigm, and another from a language model) as well at two vector-space approaches, namely the classical *tf idf* and the *Lnu-ltc* weighting schemes.

For all three languages and independently of the indexing approaches, we find that the I(n_e)C2 approach derived from the Divergence from Randomness (DFR) paradigm tends to produce the best retrieval performance. Only for the Hindi corpus (see Table IV) does the *Lnu-ltc* vector-space model result in better performances for two indexing strategies (applying an aggressive stemmer or ignoring this word normalization procedure). Based on the application of a statistical test for each of these three languages, we conclude that performance differences are statistically significant when comparing the best performing model with both the classical *tf idf* approach and the language model (LM). For the Bengali corpus when comparing the best IR model (I(n_e)C2) and the others (see Table VI) however the differences are always significant (denoted by a "†").

Upon an analysis of performance differences resulting from the application of the various stemming strategies, for all three languages we find that a stemming approach tends to perform significantly better than an indexing scheme without a stemmer. Moreover, an aggressive stemmer usually results in better MAP than a light stemmer, and these performance differences are even statistically significant, although this holds for the Marathi language only (see Table V).

When applying the YASS statistical stemmer, the performance differences obtained when a no stemming approach was applied are always significant with the Bengali language (see Table VI), but never with Hindi (Table IV), while for we obtained mixed results with Marathi language (Table V).

Language independent indexing strategies such as *n*-gram and trunc-*n* are valid alternatives, especially for unfamiliar languages. For the three languages studied, truncation after the first *n* characters tends to produce better MAP than the *n*-gram scheme. When comparing word-based indexing strategies using an aggressive stemmer, the mean differences tend to be relatively small, +2.2% for the Hindi test-collection (0.2680 "Trunc-4" vs. 0.2623 "Aggress") or -2.4% for Marathi (0.3043 "Trunc-4" vs. 0.3117 "Aggress"), yet for Bengali mean performance differences are larger (0.2998 "Trunc-4" vs. 0.2820 "Aggress", -5.9%).

When comparing all indexing schemes, we find that for the Hindi language the best approach is either trunc-4 or word-based when applying either a light or an aggressive stemmer, even though performance differences between these three schemes are usually not statistically significant. For Marathi our statistical tests detect no significant differences between an aggressive stemmer, the trunc-4 or the 4-gram schemes. According to the statistical tests we applied for Bengali, the trunc-4, aggressive stemmer and YASS approaches all resulted in similar performance levels.

When comparing retrieval performances with or without the removal of stopwords, it appears there are no real and significant differences for the Marathi and Bengali languages. For Hindi, however, the use of a stopword list significantly improves retrieval performances, with average differences being around 20% (see Table VII).

REFERENCES

ABDOU S. AND SAVOY, J. 2006. Statistical and comparative evaluation of various indexing and search models. In *Proceedings of the Conference on Alliance of Information and Referral Systems (AIRS'06)*. Lecture Notes in Computer Science, 362–373.

ALKULA, R. 2001. From plain character strings to meaningful words: Producing better full text databases for inflectional and compounding languages with morphological analysis software. *Inform. Retriev. 4*, 3–4, 195–208.

AMATI, G. AND VAN RIJSBERGEN, C. J. 2002. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inform. Syst. 20*, 4, 357–389.

BEAMES, J. 1891. *Grammar of the Bengali Language, Literary, and Colloquial*. Clarendon Press, Oxford, UK.

BRASCHLER, M. AND PETERS, C. 2004. Cross-language evaluation forum: Objective, results, achievements? *IR J. 7*, 1–2, 7–31.

BRASCHLER, M. AND RIPPLINGER, B. 2004. How effective is stemming and decompounding for German text retrieval? *IR J. 7*, 3–4, 291–316.

BUCKLEY, C., SINGHAL, A., MITRA, M., AND SALTON, G. 1996. New retrieval approaches using SMART. In *Overview of the 3rd Text Retrieval Conference (TREC'96)*. D. K. Harman Eds., 25–48.

BUCKLEY, C. AND VOORHEES, E. M. 2005. Retrieval system evaluation. In E. M. Voorhees, D. K. Harman Eds., *TREC Experiment and evaluation in information retrieval*. The MIT Press, Cambridge, MA, 53–75.

DI NUNZIO, G. M., FERRO, N., MELUCCI, M., AND ORIO, N. 2004. Experiments to evaluate probabilistic models for automatic stemmer generation and query word translation. In *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science, Springer, Berlin, 220–235.

DOLAMIC, L. AND SAVOY, J. 2010. Indexing and stemming approaches for the Czech language. *Inform. Proc. Manage.* To appear.

FAUTSCH, C. AND SAVOY, J. 2009. Algorithmic stemmers or morphological analysis: An evaluation. *J. Am. Soc. Inform. Sci. Technol. 60*, 1616–1624.

FOX, C. 1990. A stop list for general text. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'90). 24*, 19–35.

GUNGALY, D. AND MITRA, M. 2008. Using language modeling at FIRE 2008 Bengali monolingual track. In *Working Notes of the Forum for Information Retrieval Evaluation (FIRE'08)*. http://www.isical.ac.in/~fire/paper/lm_at_fire.pdf.

HARTER, S. P. 1975. A probabilistic approach to automatic keyword indexing. Part I: On the distribution of specialty words in a technical literature. *J. Am. Assoc. Inform. Sci. 26*, 197–216.

HIEMSTRA, D. 2000. *Using language models for information retrieval*. CTIT Ph.D. Thesis.

HOLLINK, V., KAMPS, J., MONZ, C., AND DE RIJKE, M. 2004. Monolingual document retrieval for European languages. *Inform. Retriev. 7*, 1–2, 33–52.

KELLOGG, S. H. 1938. *A Grammar of the Hindi Language*. Kegan Paul, Trench, Trubner & Co. Ltd., London, UK.

KETTUNEN, K. AND AIRO, E. 2006. Is a morphologically complex language really that complex in full-text retrieval? In *Advances in Natural Language Processing*, 411–422. Lecture Notes in Computer Science. Springer, Berlin.

KORENIUS, T., LAURIKKALA, J., JÄRVELIN, K., AND JUHOLA, M. 2004. Stemming and lemmatization in the clustering of Finnish text documents. In *Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM'04)*. The ACM Press, 625–633.

KROVETZ, R. 1993. Viewing morphology as an inference process. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'93)*. 191–202.

LOVINS, J. B. 1968. Development of a stemming algorithm. *Mechan. Trans. Comput. Linguist. 11*, 1, 22–31.

MAJUMDER, P., MITRA, M., PARUI, S. K., KOLE, G., MITRA, P., AND DATTA, K. 2007. YASS: Yet another suffix stripper. *ACM Trans. Inform. Syst. 25*, 4, 18.

MANNING, C., RAGHAVAN, P., AND SCHÜTZE, H. 2008. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge, UK.

MASICA, C. P. 1991. *The Indo-Aryan Languages*. Cambridge University Press, Cambridge, UK.

MCNAMEE, P. AND MAYFIELD, J. 2004. Character *n*-gram tokenization for European language text retrieval. *IR J. 7*, 1–2, 73–97.

MCNAMEE, P., NICHOLAS, C., AND MAYFIELD, J. 2009. Addressing morphological variation in alphabetic languages. In *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'09)*. 75–82.

NAVALKAR, G. R. 2001. *The Student's Marathi Grammar*. Asian Education Services, New Dehli.

PETERS, C., JIJKOUN, V., MANDL, T., MÜLLER, H., OARD, D.W., PEÑAS, A. AND SANTOS, D. EDS. 2008. Advances in multilingual and multimodal information retrieval. Lecture Notes in Comuter Science. Springer-Verlag, Berlin.

PORTER, M. F. 1980. An algorithm for suffix stripping. *Program 14*, 3, 130–137.

RAMANATHAN, A. AND RAO, D. 2003. A lightweight stemmer for Hindi. In *Proceedings Workshop of Computational Linguistics for the South Asian Languages (EACL'03)*. 42–48.

ROBERTSON, S. E., WALKER, S., AND BEAULIEU, M. 2000. Experimentation as a way of life: Okapi at TREC. *Inform. Proc. Manage. 36*, 1, 95–108.

SAKAR, S. AND BANDYOPADHYAY, S. 2008. Design of a rule-based stemmer for natural language text in Bengal. In *Proceedings of the International Joint Conference on Natural Language Processing for Less Privileged Languages (IJCNLP'08)*. 65–72.

SALTON, G. ED. 1971. *The SMART Retrieval System: Experiments in Automatic Document Processing*. Prentice-Hall, Englewood Cliffs, N.J.

SAVOY, J. 1993. Stemming of French words based on grammatical category. *J. Am. Soc. Inform. Sci. 44*, 1, 1–9.

SAVOY, J. 1997. Statistical inference in retrieval effectiveness evaluation. *Inform. Proc. Manage. 33*, 4, 495–512.

SAVOY, J. 2006. Light stemming approaches for the French, Portuguese, German, and Hungarian languages. In *Proceedings of the ACM Symposium on Applied Computing (SAC'06)*. 1031–1035.

SEBASTIANI, F. 2002. Machine learning in automated text categorization. *ACM Comput. Surv. 34*, 1, 1–47.

SPROAT, R. 1992. *Morphology and Computation*. The MIT Press, Cambridge, MA.

TOMLINSON, S. 2004. Lexical and algorithmic stemming compared for 9 European languages with Hummingbird SearchServer™ at CLEF 2003 (2004). In *Comparative Evaluation of Multilingual Information Access Systems*, Lecture Notes in Computer Science. Springer-Verlag, Berlin, 286–300.

XU, J. AND CROFT, B. 1998. Corpus-based stemming using cooccurrence of word variants. *ACM Trans. Inform. Syst. 16*, 1, 61–81.

ZHAI, C. AND LAFFERTY, J. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inform. Syst. 22*, 2, 179–214.