ISSN 2278-7763

Inflectional Stemming Effect on Evaluation Measures on an Information Retrieval System

[(1)]Dr. Loay Edward George, [(2)]Ibraheem Amer Hameed

[(1)]College of Science, Computer Science Department, Baghadad University, Baghdad, Iraq; [(2)] College of Science, Computer Science Department, Baghadad University, Baghdad, Iraq
[(1)] Loayedward57@yahoo.com
[(2)] Ibrahem_star_2005@yahoo.com

ABSTRACT

With the abundance of textual information available electronically, it is necessary to develop methods that retrieve the most relevant information according to users' needs. These retrieval methods may benefit from natural language constructs, to enhance their results by achieving higher precision and recall rates. In this study, the effect of the inflectional stemming over an English text is experimented. The stemming in the information retrieval system helps in two different places; first; it helps increasing the precision of the retrieved information, second, it helps increasing the elimination of stop words, by storing the stems only in the negative dictionary and then, all derivatives eliminated according to stored stems, this also helps to reduce the storage. So, test results have shown that removal of stop words increased after stemming about 56%. On the other hand, tests have shown also that applying inflectional stemming (with removal of stop words) has increased precision about 9.5% and recall about 10.7%.

Keywords : string, stemming, precision, recall, inflectional , text, natural language, stop words, information

1 INTRODUCTION

In linguistic morphology and information retrieval, stemming is the process for reducing inflected (or sometimes derived) words to their stem, base or root forms—generally a written word form [1]. The stem need not be identical to the morphological root of the word; it is usually sufficient that related words map to the same stem, even if this stem is not in itself a valid root. Stemming of words is an important preprocessing text operation step; it should be handled before the indexing step of input documents in the information retrieval system [1],[2]. The term "stemming" refers to the reduction of words to their roots so that; different grammatical forms or declinations of verbs are identified and indexed (counted) as the same word. The goal of stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For instance: car, cars, car's, cars' will be car [3]. In Information Retrieval (IR), documents contained in document collections are searched by entering a query, which is usually a set of keywords or a sentence in natural language. The frequency of terms both the query and document have in common determines the documents' relevance. Documents are then ordered according to rank in lists. In the simplest case the relevant documents contain exactly the same words as the query. Unfortunately, this method excludes many documents that contain morphological variants of the query term. One solution to this problem is stemming [4]. Stemming alters words by removing affixes and conflates semantically related words. During indexing, terms are associated with the documents they appeared in to make retrieval easier and faster. As stemming conflates terms, the size of the index of a stemmed collection is smaller than the Index-size of the un-stemmed collection. And also the size of the negative lexicon size is reduced, for example, instead of storing the words "reads, reader, reading, reads", only the stem will be stored and this stem will match all of those derivatives. 2 Procedure for Paper Submission

2 MORPHOLOGY

Every word can be classified through a lexical category or part of speech such as article, noun, verb, adjective, adverb, conjunction, preposition, or pronoun. Most of the lexical entities come from four categories: noun, verb, adjective, and adverb. Other categories such as articles, pronouns, or conjunctions have a limited and stable number of elements [4,5]. For example:

The big cat ate the gray mouse
The/article big/adjective cat/noun ate/verb the/article gray/adjective mouse/noun

Morphology is the study of how root words and affixes are composed to form words. Morphology can be divided into inflection and derivation:

- Inflection is the form variation of a word under certain grammatical conditions. In European languages, these conditions consist notably of the number, gender, conjugation, or tense.
- Derivation combines affixes to an existing root or stem to form a new word.

Derivation is more irregular and complex than inflection. It often results in a change in the part of speech for the derived word [5],[6],[7].

Morphological rules enable us to generate all the word forms from a lexicon. Morphological parsers do the reverse

ISSN 2278-7763

operation and retrieve the word root and its Affixes from its inflected or derived form in a text. Morphological parsers use finite state automaton techniques. Part-of-speech taggers disambiguate the possible multiple readings of a word. They also use finite-state automata or statistical techniques [5],[6].

3 Proposed Work

There are nine parts of speech, (i.e., nouns, verbs, adjectives, adverbs, pronouns, conjunctions, prepositions, interjections and articles). As a result of the stop words removal five parts of speech has been totally removed, they are: the pronouns, conjunctions, prepositions, interjections and articles. The other parts of speech are also checked for some of stop words to be also removed before stemming. If the remaining words didn't match with the stop words; they are stemmed out and then rechecked against the stop words. The excluded stop words are put in a predefined exclusion dictionary (or table) to exclude such terms from the text because this kind of words occurred too frequently and their existence is insignificant for mining tasks. So, they are not worthy to be indexed in the system's database. From the experiments it is seen that the removal of the stop words helps in increasing the accuracy of the results and, also, it helps in reducing the size of the database. The main concern of the proposed stemming is to keep up the words relatively meaningful after stemming to the old one, and do not transform the words from one part of speech to another since this may contribute to change the word's meaning.

In general, stemming deals with derivational forms and inflectional forms; the applied stemming algorithm deals with inflectional forms only unlike other algorithms (e.g. porter's algorithm) which deal with both. The inflectional morphology doesn't affect the stem while the derivational affects the stem and may change the meaning. Anyhow, during the conducted tests, it is noticed that stemming the derivational forms is not convenient to be used in the information retrieval system, since it ruins a large number of informative keywords to forms which are semantically correct but become opaque words and autonomous (i.e. have no relation with other words exist in the same article). For example, the words end with " ize", "ise", "en", ify" for verbs and "tion", ary", "ssion" for nouns must remain as they are; but the derivational stemming makes such words further fine-granulated, such that they become unsuitable as keywords, and consequently they negatively affect the quality of the information that expected to be retrieved by information retrieval system.

On the other hand, the inflectional stemming involves the removal of all suffixes added to words as a result of grammatical importance, like the progressive, past tenses for verbs and plural for nouns. Their removal does not change the stemmed word's part-of-speech, this means that the name remains a name, the verb remains a verb. Furthermore, inflectional stemming has nothing to do with the affixes; for example the "un" prefix that's added to the word to negate it remains as it is. Table (1) illustrates an example of the inflectional vs. derivational stemming.

TABLE 1
STEMMING TYPES

(A) INFLECTIONAL STEMMING

inflectional	stem	type1	type2
computers	computer	n.	n.
organizations	organization	n.	n.
purifies	purify	v.	v.
whitened	whiten	v.	v.
serializing	serialize	v.	v.

(B) DERIVATIONAL STEMMING

derivational	stem	type1	type2
compression	compress	n.	v.
organization	organize	n.	v.
friendship	friend	n.	v.
whiten	white	v.	n.
informative	inform	adj.	v.

The stemming alone is not yet enough in processing the words to originate them to their roots; there is another operation which is considered a complementary to the stemming, it is called the lemmatization and used for checking the stem. If the produced stem is not grammatically correct, the lemmatization will add or remove what the stem needs to be an integrated word. If the implemented lemmatization is not correct or not considered at all, many of the stemming results will become false negatives/positives which can effectively cause a defect in the performance of retrieval system because a lot of tokens will be deformed. Not every word needs to be lemmatized; for instance, the suffixes of "reading", "reads" will change to get the stemmed form "read" standing for the infinitive "to read"; in this case both the lemmatized word form and the word stem are equal.

The proposed stemming algorithm is extendable; it can be extended to include other rules. In the reconstructed information retrieval system that used the proposed stemming algorithm the rules involved had led to satisfactory level of word granularity, further stemming rules may defragment the words, and would degrade the information retrieval performance. The proposed stemming algorithm used the regular expressions in matching and searching the texts.

Regular expressions describe regular languages in formal language theory. They have thus the same expressive power as regular grammars. Regular expressions consist of constants and operators that denote sets of strings and operations over these sets, respectively [7]. A regular expression provides a concise and flexible means for "matching" (specifying and recognizing) strings of text, such as particular characters, words, or patterns of characters. Abbreviations for "regular expression" include "regex" and "regexp". The concept of regular expressions was first popularized by utilities provided by UNIX distributions. A regular expression is written in a formal language that can be interpreted by a regular expression processor, which is a program that either serves as a parser generator or examines text and identifies parts that match the provided

ISSN 2278-7763

specification. Many languages provides dedicated classes to deal with regular expression so in the .NET languages the class regex holds all of the methods necessary to deal with text as a regular expression, there are special characters which are called the meta characters which are used as a parser directives to makes it understand what is needed.

4 Information Retrieval Measures

As the proposed stemming is operated on an information retrieval system so its efficiency and effectiveness will be calculated implicitly within the information retrieval system. The measures usually employed to evaluate the performance of an information retrieval system are precision and recall. Precision (Pr) is "How many of the retrieved results are relevant" and the recall (Re) is "How many of the relevant results are retrieved"; mathematically, they defined as [10].

5 Test Results

The usage of the stemming in an information retrieval system in two different places:

1. Stemming/Stop Word Removal Dependency: As stated in table (2) removal of the stop words will be increased after stemming about 56%. The first column in table (2) is the query number and the second is the percentage of character reduction in case of no stemming operation, while the third column is the reduction percentage when stemming operation is conducted. Figure (1) illustrates the test results.

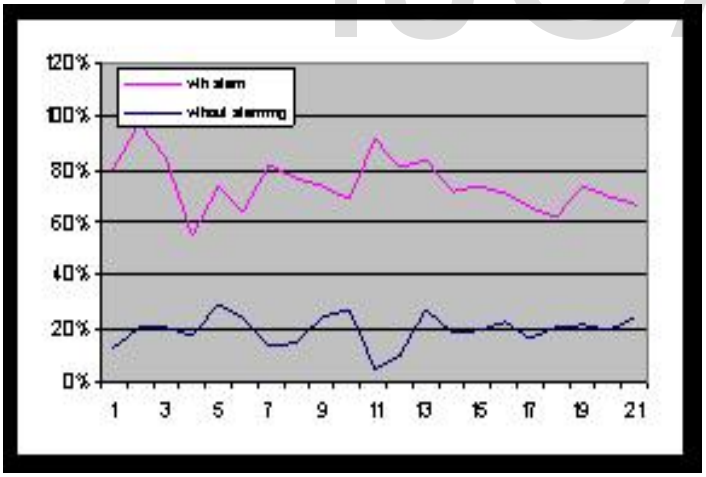

Fig. 1 Stop words removal and stemming relationship.

TABLE 2
STOP WORDS REMOVAL AND STEMMING RELATIONSHIP

Q no.	without stemming	with stem
1	13.23%	66.65%
2	20.69%	76.69%
3	20.34%	64.72%
4	17.67%	37.13%
5	29.08%	44.43%
6	24.87%	38.59%
7	14.11%	67.26%
8	14.47%	62.43%
9	24.29%	49.72%
10	27.17%	42.06%
11	4.58%	86.63%
12	9.93%	70.58%
13	27.56%	55.97%
14	18.35%	53.49%
15	19.24%	54.66%
16	22.28%	48.39%
17	17.06%	48.78%
18	20.35%	41.92%
19	21.27%	52.04%
20	19.79%	49.63%

2. Text operations effect on retrieval: the main text operations are stemming and stop word detection and removal; and in order to evaluate their effectiveness on retrieval then, system performance measures (the precision and recall;) and the time is calculated with and without applying these text operations. The tests show that applying these operations has increased the precision about 9.5% and the recall about 10.7%, but, the execution time has increased about 45%, as shown in table (3) and Figure(2) illustrates the test results.
3.

ISSN 2278-7763

TABLE 3
THE TEST RESULTS OF THE RETRIEVAL USING THE TEXT OPERATIONS

Q no.	without stemming	with stem	Q no.	without stemming	with stem
1	13%	67%	22	11%	61%
2	21%	77%	23	9%	74%
3	20%	65%	24	21%	51%
4	18%	37%	25	26%	41%
5	29%	44%	26	26%	41%
6	25%	39%	27	20%	57%
7	14%	67%	28	18%	59%
8	14%	62%	29	17%	60%
9	24%	50%	30	17%	60%
10	27%	42%	31	18%	76%
11	5%	87%	32	27%	70%
12	10%	71%	33	19%	75%
13	28%	56%	34	40%	64%
14	18%	53%	35	27%	48%
15	19%	55%	36	19%	44%
16	22%	48%	37	16%	51%
17	17%	49%	38	19%	45%
18	20%	42%	39	17%	65%
19	21%	52%	40	21%	40%
20	20%	50%	41	16%	63%
21	25%	43%		20%	56%

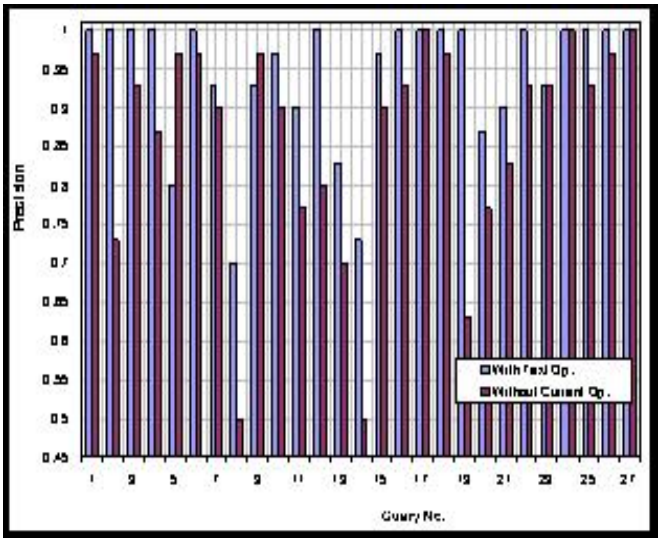

Fig. 1. The effect of text operation on the retrieval process

6 CONCLUSION

With the abundance of textual information available electronically, it is necessary to develop methods that retrieve the most relevant information according to the needs. These retrieval methods may benefit from natural language constructs to enhance their results by achieving higher precision and recall rates. The study showed that the inflectional stemming helps the information retrieval systems to increase the accuracy of the retrieval and also helps to remove the stop words efficiently. The test results showed that the inflectional stemming helped to increase precision by 9.5% and recall by 10.7%. After stemming, the reduction percentage of characters of stop words is reduced to 56%.

REFERENCES

[1] Bauer, L., "Introducing Linguistic Morphology", Georgetown University Press, 2nd edition, ISBN 9780878403431 (0878403434), 2003.

[2] Klavans, J.L., Kan, M.Y.: "The Role of Verbs in Document Analysis". In Proc. Coling-ACL, Vol. 36, pp. 680-686. Association for Computational Linguistics (1998)

[3] BijanKhan, M.: The Role of the Corpus in Writing a Grammar: An Introduction to a Software. Iranian Journal of Linguistics, 19(2) (2004)

[4] Agirre, E., Nunzio, G.M.D., Ferro, N., Mandl, T., Peters, C.: Multilingual Textual Document Retrieval (Ad Hoc), in Evaluating Systems for Multilingual and Multimodal Information Access. In Proc. 9th Workshop of the Cross-Language Evaluation Forum, CLEF 2008, Aarhus (2008)

[5] Shah, C., Bombay, I.I.T., Mumbai, P., Maharashtra, I., Bhattacharyya, P.: A Study for Evaluating the Importance of Various Parts of Speech (POS) for Information Retrieval (IR). In Proc. International Conference on Universal Knowledge and Languages (ICUKL) (2002)

[6] Carlberger, J., Kann, V.: Implementing an Efficient Part-Of-Speech Tagger. Software Practice and Experience, 29(9), 815-832 (1999)

[7] Kenneth R. Beesley and Lauri Karttunen. Finite-state non-concatenative morphotactics. In SIGPHON-200. Fifth Workshop of the ACL Special Interest Group in Computational Phonology., pages 1–12, Luxembourg, August 5-6 2000. Association for Computational Linguistics.

[8] Michael Meeuwis. Lingala. Lincom Europa, Munchen, Germany, 1998.

[9] Gregory T. Stump. Inflectional Morphology. A Theory of Paradigm Structure. Cambridge University Press, Cambridge, England, 2001.

[10]Christopher, D., Manning, "An Introduction to Information Retrieval", Springer, ISBN-13 978-3-540-31588-9, 2009.

