THE CONCEPT OF STOPWORDS IN PERSIAN CHEMISTRY ARTICLES: A DISCUSSION IN AUTOMATIC INDEXING

Mohammad Reza Falahati Qadimi Fumani
C. S. Ramachandra
University of Mysore

Abstract

While having a close look at the concept of stopwords, the researchers examined 30 chemistry scientific articles published in high ranking Persian journals to find an answer to the following research questions: (1) Should all punctuation marks, numbers, and English letter combinations (letter-number and letter-punctuation-number combinations) be included in the stoplist and thus be eliminated as candidates for indexing? (2) Is it correct to change all upper case letters into lower case letters, i.e., to downcase letters? And (3) Could Persian be dealt with in automatic indexing without making reference to the characteristics of English? The manual analysis of the sample revealed that the omission of all punctuation marks, numbers, etc. will have a negative effect on recall, since punctuation marks, particularly 'dot' and 'hyphen' appear in the structure of content-bearing elements and even appear abundantly in titles and abstracts of scientific articles. With regard to downcasing, or unifying all upper and lower case letters, the conclusion was that it is much more restricted in Persian compared to English and that some possible places where it may cause problems are the structure of formulas, names of chemical substances, acronyms and proper nouns. Finally, it was found that because of the appearance of English letters, words, numbers, etc. in the body of Persian articles, some characteristics of the English language must also be considered while working on automatic indexing of Persian articles. The overall conclusion was that some sort of compromise is required in labeling numbers, punctuation, acronyms, etc. as either stopwords or content-bearing elements and thus as potential index candidates.

Key terms: computational linguistics, automatic indexing, stopword, stoplist, noise, chemistry articles, punctuation, number, Persian, Farsi, English

Abstracto

Teniendo en cuenta el concepto de las palabras conocidas como "stopwords", unos investigadores examinaron 30 artículos científicos sobre química publicados en diarios persas competitivos para así poder encontrar una respuesta a las siguientes preguntas: (1) ¿Deberían ser incluidos todos los signos de puntuación, números y combinaciones de letras en inglés (letra con número y letra con signo de puntuación y número) en la "stoplist" y por ende ser eliminados de los índices? (2) ¿Es correcto cambiar todas las letras mayúsculas a minúsculas? Y (3) ¿El persa puede utilizarse en índices automáticos sin hacer referencias a las características del inglés? El manual de análisis de la muestra revela que la omisión de los signos de puntuación, números, etc. tendría un efecto negativo ya que los signos como el punto y el guión aparecen en la estructura de los elementos que incluyen contenido y hasta aparecen de manera abundante en títulos y abstractos en artículos científicos. Con relación a la minusculización de las letras o la unión de las letras mayúsculas y minúsculas, la conclusión fue que la variación es más estricta en el persa que en el inglés y que en

algunos casos la estructura de las fórmulas, acrónimos, el uso de nombres propios y los nombres de las sustancias químicas puede resultar problemáticos. Finalmente, se encontró que por la apariencia de las letras en inglés, palabras y números, etc. en el desarrollo de artículos en persa, algunas características del inglés también deben ser consideradas al trabajar con el índice automático de los artículos en persa. La conclusión general fue que algún tipo de compromiso es requerido para categorizar los números, signos de puntuación, acrónimos, etc. como "stopwords" o elementos que contienen contenido y potenciales candidatos de índice.

***Palabras clave**: lingüística computacional, índice automático, "stopword", "stoplist", ruido, artículos de química, puntuación, número, perso, farsi, inglés.*

Mohammad Reza Falahati Qadimi Fumani is a PhD Candidate at Department of Studies in Linguistics in University of Mysore, India, and a Faculty of RICeST's Computational Linguistics Research Department, Iran. His research interests are language processing, Computational Linguistics and its sub-disciplines.

Dr C. S. Ramachandra is a Reader in Linguistics at the Department of Studies in Linguistics in University of Mysore, India. His research interests are Linguistics and its sub-disciplines.

Introduction

'Stopwords' is a term that has been present in almost all discussions related to indexing and automatic indexing. Luhn (1958) used the term 'noise', and 'common word list' and described stopwords as the presence in the region of highest frequency of many of the words previously described as too common to have the type of significance being sought would constitute 'noise' in the system. This 'noise' can be materially reduced by an elimination technique in which text words are compared with a stored common-word list.

Although Luhn speaks of a stored common-word list, he himself tries to resolve the problem of 'noise' using the notion of the 'discriminatory power' – the ability to distinguish one article from the rest of articles in a data set – of words. In fact, having established an automatic measure to determine the set of significant words in his document collection, he defines a threshold below and above which terms could be

labeled as ‘noise’ or ‘stopwords’. Tsz-Wai Lo, et al. (2005) defined stopwords as words in a document that are frequently occurring but meaningless in terms of Information Retrieval (IR), i.e., *to, and, for, a, an,* etc. They also state that stopwords do not contribute towards the content or information of the documents and they should be removed during indexing as well as before querying by an IR system (ibid). Many other scientists have also discussed this phenomenon like Fox (1992), Roelleke (2003), Robertson and Sparck-Jones (1976) and readers are requested to refer to these sources for further information.

Literature Review

What can be inferred from the brief introduction given above is that ‘stopwords’, ‘stoplists’, ‘noise’, ‘negative dictionary’ or whatever one may call it, is a concept that could not be easily ignored in discussions related to indexing, whether manual or automatic. The only difference is in the way this concept has been approached by different scholars. With the help of the existing literature, we are able to identify at least four different ways of tackling stopwords, some focusing on finding the least important words, and some focusing on finding the most important ones.

a- *Use of stoplists already generated:* There are a number of such stoplists in the market at present. Francise and Kucera (1982), for instance, worked on the Brown Corpus and consequently extracted 425 stopwords for English. Similarly, van Rijsbergen (1975) produced a stoplist for English comprising 250 words and ‘fluff words’ – words like *below*, *near*, *always*, … that have a low frequency but do not usually have discriminating power. In Persian, Taghva, et al. (2003) produced a stoplist that embodied all variant forms of 12 verbs (saying that each verb had almost as many as 100 variants given its infinitive, imperative and past tense forms. The verbs they listed were as follows: شدن (to become), *بودن (to be), خواستن (to want), داشتن*

(to have), یافتن (to find), توانستن (to be able to), آمدن (to come), گرفتن (to take), آوردن (to bring), کردن (to do), گفتن (to say), دادن (to give) together with their past tense as well as imperative forms, etc. The list they produced also included non-verbal stopwords which mounted to 155 prepositions, conjunctions, etc.

b- ***Automatic extraction of stoplists using a document collection:*** Salton and Buckley (1988), Salton and Yang (1975) and also Spark Jones (1973) are amongst the researchers who discussed term-weighting approaches with the objective of selecting the most important words denoting, or better, representing the whole content of a text. Wilson (2002) reviewed twelve term-weighting models and referred to 'tf' as well as 'tf.idf' and many other methods, calling the 'tf-idf' model one of the most commonly used approaches. The point in 'tf.idf', in contrast to 'tf' is that it considers a text database or a document collection and considers the frequency of not only each term in a single document – this is what 'tf' means – but also its presence in all the documents available in the collection. The methodology to extract stopwords automatically would mark as stopwords all those words that have occurred highly frequently in almost all the documents. Similarly, if the word 'cat' occurs, for instance, thirty times in one document but in no other documents, we can be relatively certain that this text is about cats, and thus the word 'cat' must not be taken as a stopword here. The theoretical justification for such selection procedure is the point stated by Kintsch and Van Dijk (1978). They stated that those propositions that are 'more frequent' are obviously 'more important'.

c- ***automatic production of domain-specific stoplists:*** It is believed that those stoplists produced for a language in general will not be able to function appropriately in specific domains. In fact, it is recommended that stoplists be prepared for each

subject area separately, i.e., for chemistry, geography, physics, etc. Pollock and Zamora (1975), for example, state:

> …the original documents used in our work were all abstracted at CAS for chemical abstracts…We consider that this restriction of the database to a well defined subject area is important in principle, and that it is unrealistic to expect a single algorithm to be able to abstract documents on a wide range of subjects (p. 352).

Domain-specific stoplists are prepared through analyzing and examining a collection of texts, i.e., articles, books, etc. related to that subject area and then extracting the stopwords through frequency analysis as well as content analysis as stated in **a** and **b** above. In fact, content analysis is carried out to exclude from the stoplist all those highly frequently occurring terms that may be content-bearing and significant in a given field. The capital letter B, C, D, or K in *Vitamin B*, *Vitamin C*, *Vitamin D* and *Vitamin K* are just some examples where we are dealing with a single meaning-bearing character.

d- ***Indexing all or almost all the words and thus diminishing the role of stoplists:*** One of the major philosophies behind using stoplists is to reduce the 'index storage space' with the hope that although this may reduce 'recall' – the total number of relevant articles retrieved – it may have an increasing effect on 'precision', which is defined simply as the proportion of the retrieved documents that fit the particular needs of the user. So, an increase in the number of stopwords will have a decreasing effect on recall, and that is a point with which the majority of commercial system designers are dissatisfied (Harman, 1992). One reason might be that they want their systems to have the highest recall because quite often they think this is how users may evaluate their system first hand. Another reason they sometimes put forward is that

fluff words and even stopwords may function as meaning-bearing elements, which is a good justification for their inclusion as candidates in the index list, rather than in the stop list.

For example, ORBIT Search Service as stated in Fox (1992) uses only 'and', 'an', 'by', 'from', 'of', 'or', 'the', and 'with' as its 8 stopwords, while MEDLARS System has even fewer stopwords. As stated above, commercial systems are highly sensitive to the size of retrieval and recall of their systems.

Research Questions and Hypotheses

What was mentioned in the last part of 'Literature review' simply illustrates the subject-sensitivity of producing domain-based stoplists. Most important of all, we have already written about words as if they consisted only of letters of the alphabet. But such a presupposition is far from real. A text includes punctuation marks, upper and lower case letters (small and capital letters), as it may include numbers, letter-number and letter-punctuation-number combinations, e.g., *A4 paper, H2O, K-1001* representing 'a specific paper size', 'water' and 'a particular equipment' respectively.

Such cases are quite frequently deleted from the analysis, and so are not taken as index terms. Mayfield and McNamee (1998), for example, in their work with TREC-7 eliminated all punctuation marks, downcased all letters and mapped all numbers to a single digit. For them, a word was any remaining blank-delimited sequence of characters. But, based on examples like Vitamin B, A4, H2O, K-1001, etc. and following the comments of Harman (1992) who believes that before indexing is started, samples of the text to be indexed, and samples of the types of queries to be run, need to be closely examined, the present researchers decided to examine closely and manually a number of Persian scientific articles related to the subject area of chemistry and specifically examine

punctuation marks, numbers, upper and lower case letters, and English letter combinations (letter-number and letter-punctuation-number combinations, etc.) to see if they really should be taken as stopwords, or their occurrence in the stoplist must involve a certain type of compromise. The following research questions were put to the test by the researchers:

1. Should all punctuation marks, numbers, and English letter combinations (letter-number and letter-punctuation-number combinations,…) be included in the stoplist and thus be eliminated as candidates for indexing?
2. Is it correct to change all upper case letters into lower case letters, i.e., to downcase letters?
3. Could Persian be dealt with in automatic indexing without making any reference to the characteristics of English?

The above research questions gave way to the following research hypotheses:

H1: Not all punctuation marks, numbers, and English letter combinations (letter-number and letter-punctuation-number combinations) should be included in the stoplist and thus be eliminated as candidates for indexing.

H2: It is not correct to change all upper case letters into lower case letters, i.e., to downcase letters.

H3: Persian could not be dealt with in automatic indexing without making some reference to the characteristics of English.

Methodology

Since the present study is an introductory part of a more comprehensive research on Persian automatic indexing and since the analysis at this stage is carried

out manually, but in a way that would fit the later requirements of our future work on the aforementioned topic, the analysis is done on a small number of articles (n=30) extracted from different high ranking Persian journals of chemistry available at the E-Journals database of ***Regional Information Center for Science and Technology (RICeST)***, Shiraz, Iran (www.ricest.ac.ir). To minimize the side-effect of the small number of articles, the following steps were also taken to make the findings more generalizeable:

1. The articles were selected from 20 different high ranking Persian journals related to the subject area of chemistry. The articles were selected completely randomly to avoid subjectivity in data collection.
2. After selection and analysis of the articles, the researchers scanned through some new chemistry journals as well as journals related to some other areas like agriculture, medical sciences, etc. available in the E-Journals database of RICeST to see if the findings in the small sample were also applicable to other articles and domains not inspected yet. The review made the researchers confident that their findings were applicable, more or less, to other articles and other fields as well.

Data Collection Procedures

To collect the data, the researchers surfed RICeST's E-Journals database and randomly selected a number of pdf articles (30 related to chemistry) from 20 journal titles. They also browsed a number of journals related to other disciplines and analyzed the titles of the articles to check for the presence of any English letter combinations (letters, words, phrases, etc.) in them.

Inter-rater scoring was used, that is, the two researchers read each article separately and carefully recorded any application of punctuation marks, numbers, upper and lower case letters as well as English letter combinations. Later, to cope with the two raters' counting differences, the average of the two countings was used in each case. The data collected functioned as the basis for data analysis.

Data Analysis

Data analysis will be given in three parts, each part trying to provide an answer to one of the research questions raised earlier in this study.

Testing Hypothesis One

Before reporting the findings concerning the first hypothesis, the first research question is repeated here for ease of reference:

Question 1: *Should all punctuation marks, numbers, and English letter combinations (letter-number and letter-punctuation-number combinations, …) be included in the stoplist and thus be eliminated as candidates for indexing?*

To find an answer to the above question, the researchers analyzed each sub-part separately.

Punctuation Marks.

Punctuation marks, as already stated, are often taken as stopwords and thus ignored as index term candidates (Note that in this article the term punctuation marks is used in a broader sense covering signs used in mathematical formulas as well.). In this section, punctuation marks will be analyzed, as they appear in the text, to see if the above assumption could be kept or if some compromise must be made. An

analysis of the articles revealed that in each 5-8 pages chemistry article, there were between 150 and 350 punctuation marks excluding those used in the 'reference' section, which comes at the end of each article.

Table 1 shows the whole range of punctuation marks that appeared in the body of the 30 articles examined. In all, 31 different punctuation marks were used by the authors. The findings revealed that *dot* and *comma* comprised more than sixty percent (4100, 62.56%) of the punctuation marks used in the data collection. *Hyphen* and *parentheses* were also used with high frequency (1651, 25.19%). Colon and brackets ranked fifth and sixth and together, with a frequency of 627, comprised 9.57% of the whole punctuation marks. An overview of the items in Table 1 revealed that only the top 6 items comprised 97.32% of the whole punctuation marks.

An important conclusion drawn from the text analysis of the articles is that *hyphen* and *dot* are most likely to appear in words and terms that may be good candidates for indexing. An analysis of 30 key phrase sets related to 30 articles revealed that these two were the only punctuation marks that appeared within the structure of a single key word or key phrase, e.g., *4-Dichlorophenoxiacetic acid*, *Asphalt-polymer*, etc. The analysis showed that such terms could appear in the title of an article as well.

Table 1: Punctuation marks used in 30 chemistry articles.

column	punctuation mark	frequency	usage
1	Comma	2200 (33.57%)	separating words and phrases
2	Dot	1900 (28.99%)	closing sentences, within acronyms and formulas
3	Hyphon	947 (14.45%)	distance, chemical substances, titles and sub titles
4	Parentheses	704 (10.74%)	further explanation, synonyms, year of

			publication, formula or algorithm number, reference number in some articles
5	Colon	342 (5.22%)	before quotations, in formulas, further explanation, introducing sub parts
6	Brackets	285 (4.35%)	reference number in some articles, formulas
7	!, #, …., ?, «, », ±, ××, ×, ‘‘ ”, …, (( )), {}, ~, >, <, ≥, +, -, /, *, %, =, ", ;	176 (2.68%)	exclamation mark, question mark, ellipsis, mathematical formulas
Total		6554 (100%)	

Note: Punctuation marks in item 7 were all reported together due to their low frequency of occurrence. The first 6 items were concentrated on because they were present in all the 30 articles analyzed and also because of their high frequency of use. Further, all punctuation marks in item 7 are used in Persian as per se the only exception being question mark and semi-colon for which their English parallel symbols were used to avoid font problem and possible confusion by non-Persian readers.

Numbers.

Numbers appeared in title as well as the body of the articles in a number of ways some of the most important of which are as follows: title of an article, e.g., *2,4-Dichlorophenoxiacetic acid*; sections and sub-sections, e.g., *1- Introduction, 3-1 Materials of the study*; references used by the author, e.g., *[1], [1,2,8*; names of chemical substances and formulas, e.g., *Cacl2, H2O*; measurements (weight, temperature, speed, …), e.g., *21g/l (grams in liter)*, *35 C*˚; tables and figures, e.g., *Table 1, Figure 2,* etc.

Three points found concerning numbers are as follows. First, sometimes numbers occur separately in the text like, *2, 25* and 10^6. Second, in some cases numbers are combined with punctuation marks, e.g., *98%, 85-90* and *[1,2-5]*. Thirdly, numbers may be used with single English letters. In such cases, usually punctuation marks are also used in between. So, they may even be called *letter-punctuation-number* combinations, e.g., *c-*

18, k-1001, RP-18e, etc. compared to simple *number-letter* or *letter-number* combinations like, *35 C˚* and *A4.*

In all, four ***inconsistencies*** were found concerning numbers. First, in Persian articles, numbers may be written using both Persian and English numerical scripts. In tables, figures and formulas more English numbers were found, but within the text Persian numbers were more frequent. Second, even within the text itself inconsistencies were found. One interesting example, though used by few authors, was the use of Persian word صفر meaning *zero* rather than its equivalent mathematical numerical in reporting a list of numbers like *0, 2, 4, 9, 11.* Third, numbers requiring a punctuation were not used so consistently, i.e., *2,4-Dichlorophenoxiacetic acid* and *2,4 Dichlorophenoxiacetic acid* both appeared in the text, one with hyphen another without it. Finally, measurements were not used consistently. Sometimes they were reported using Persian words and sometimes using English symbols, i.e., *m/l* and *میلی گرم بر لیتر (*equivalent to English *milligram per liter)* both appeared in the sample studied and even in the same article.

English Letter Combinations.

One source of difficulty with Persian texts is that they may embody letters from the English alphabet. These letters may form acronyms, words and phrases or be combined with numbers and punctuation marks. This may be due to Persian having borrowed a large number of scientific words and terms from English. In fact, a number of these borrowed scientific terms have already been recognized as universal terms used by scientists throughout the world. Such concepts may be included in Persian texts exactly in the form they appear in English. Some authors, of course, are apt to write English words with Persian letters, e.g., موند for *Monod*. In the sample studied, several types of English words appeared as follows:

1- Whole phrases, i.e., *Chromolith Performance*, *Ralstonia eutropha*, etc.

2- Acronyms like, *HPLC*, *KNAUER*, etc.

3- Single words like, *Monod*, etc.

4- Single English letters, i.e., t, k, p, C°, etc.

Though not all these cases may appear together in a single article, as the review of the sample articles showed, each article may embody one or more of the items mentioned above. For example, a single article embodied 6 English phrases (*type 1*), 13 acronyms (*type 2*), 3 single words (*type 3*) and 44 single English letters (*type 4*). Three indexing problems that arise as a result of English words are as detailed below.

[First,] English terms may appear in the title of an article and in the key terms assigned by the author(s). This necessitates that the Persian indexing system be able to consider such items as well - either using their well accepted Persian equivalent terms, if any, or using them as key terms if during querying users are apt to use such English terms as well and if such terms have found their way well into the terminology of Persian scientific articles and information resources. Fifteen out of eighty (18.75 %) articles in 'Iran's Agriculture Journal', Volume 33; seven out of sixty-five (10.76%) articles in 'Iran's Journal of Chemical Engineering', Numbers 2-5, and eleven out of twenty eight (39.28%) articles in 'Iran's Journal of Medical Sciences', Volume 5, embodied English words, acronyms and even phrases in their titles.

[Secondly,] Sometimes authors make inconsistent use of English words. The word *Monod*, for example, first appeared as the English word *Monod* in the abstract of an article, which was repeated many times throughout the text. This word also appeared amongst the keywords assigned by the author, but this time it was written using the Persian alphabet and as موند. Since this was a core term and one of the major content - bearing terms in that article - to find this use was made of the expert views of specialists

in chemistry and Library and Information Science (LIS) – the weighting system in AI must be able to handle this sort of inconsistency by either merging them into one form or by using an English to Persian chemistry dictionary to match such terms and count them as belonging to the same concept. That is, a link may be established between this dictionary and the Automatic Indexing system so that the English word used in the article will be checked in that dictionary and consequently Persian translations of that word will also be checked in the text and all will be included in the frequency count of that single lemma. Of course, the efficiency of this procedure depends, to a great extent, on the quality of the dictionary used. Another point is that this solution may not be able to remove the problem completely, as it will also increase the whole storage capacity the ultimate AI system will require.

[Thirdly,] Single letters. As stated before, single English letters like *t*, *g*, etc. are also used in Persian texts. The problem is that again authors make use both of these English terms and their Persian translation. For example, in formulas, tables and even the body of the articles, an author may use *t* to refer to 'time', but s/he may also use the word زمان (Persian word equivalent to English word 'time') quite frequently.

All the acronyms found in the articles appeared in English script and were not translated into Persian by any of the authors. So, it seems that acronyms are used uniformly in Persian articles, and therefore may be tackled by making reference to an English-Persian lexicon or an acronym finder.

Testing Hypothesis Two

The second research question of the paper aimed at finding an answer to the following question: *Is it correct to change all upper case letters into lower case letters, i.e., to downcase letters?*

The analysis of the articles revealed that Persian does not have downcasing - upper and lower case letters - in the sense that English does. In fact, Persian sentences, unlike English, do not begin with capital letters. Similarly, all Persian acronyms are written with lower case letters (because of the absence of dichotomy between small and capital letters) and often have a word shape, so it will be meaningless to speak about upper and lower case letters regarding Persian words. To show importance in Persian, words are bolded, underlined or written with a larger font size.

If this is the case, why discuss it here? The reason is that as mentioned earlier, English words, acronyms and even letters are frequently used in Persian scientific articles, so the system must be able to tackle them. The philosophy behind downcasing in English is that since the initial word of a sentence starts with a capital letter, which will be different from the way it appears elsewhere in a sentence (unless it is a proper noun), it must be downcased so that the AI system will be able to include them as a single term while running the frequency count, i.e., *Information & information* will be counted as tokens of the same lemma. In corpus linguistics studies, this is one way that lemmatizing comes into play. That is, having the program recognize when words are different instantiations of the same lemma (Smith, 1991, pp. 70-71). Therefore, discussing downcasing is relevant given the English words that are embedded in Persian scientific texts.

The manual testing of English words that appeared in the small sample of Persian articles revealed the following points concerning downcasing: (1) Since we are dealing with Persian texts, the issue of downcasing is much more restricted in Persian than in English. (2) One possible problem that may arise is in the formulas because in formulas capital and small versions of a single letter may refer to two

distinct concepts like, Cc. We should, however, take into account the fact that in none of the articles inspected, were formulas taken as key terms; formulas were always used to demonstrate some point previously stated in the paper or that were expressed using some terms. (3) Downcasing certain proper nouns (i.e., White House, etc.) may cause problems in English but in Persian because of the specific and restricted use of English terms (only technical terms appear amongst Persian texts) it seems not to play a significant role.

The following example may well illustrate the limited scope of downcasing in Persian. If in English, IT (Information Technology) is downcased, it will look like the pronoun *it* which itself may refer to different things in different contexts, but if IT appears in a Persian text such a misunderstanding will not arise because the text of the article is Persian. The only point is that we must have lexicons to identify IT or *it* as Information Technology. A more complicated linguistic elaboration may be required for cases like AI, which may refer to *Automatic Indexing*, *Artificial Intelligence*, etc.

Testing Hypothesis Three

The third research question of the study read as follows: *Could Persian be dealt with in automatic indexing without making any reference to the characteristics of English?*

The answer to this question has already been indirectly given in our earlier discussions in Sections 5-1 and 5-2. As stated earlier, Persian articles, at least those related to specific subject areas, embody English terms as well and these terms could not

be treated as marginal because they have already found their ways into titles, author-given key terms, abstracts, etc. of Persian articles. These words have been borrowed from English and because they have preserved their English forms, it seems logical to incorporate some characteristics of English into Persian AI systems as well.

Conclusions

Based on what was already discussed throughout this article, it may be concluded that as Harman (1992) mentioned, some sort of compromise is required in labeling numbers, punctuations, acronyms, etc. as either stopwords or content-bearing elements and thus potential index candidates. With regard to punctuation, special attention must be paid to 'dot' and 'hyphen'. With regard to numbers, attention must be paid to them if they appear in the structure of terms appearing in title, abstract, or the key terms given by the author(s). Finally, downcasing should not be done with blind eyes – acronyms, formulas, names of chemical substances and proper nouns require special attention.

References

Fox, C. (1992). *Lexical analysis and stoplists.* Upper Saddle River, NJ: Prentice-Hall.

Francis W., & Kucera H. (1982). *Frequency analysis of English usage*. New York, NY: Houghton Miffin.

Harman, D. (1992). Automatic indexing. *NIST interagency report 4873*, Retrieved June 25, 2007, from http://www.itl.nist.gov/iad/894-02/works/pubs/ir4873.html

Luhn, H. P. (1958). The Automatic creation of literature abstracts. *IBM Journal*, April,

Presented also at *IRE National Convention*, New York, March 24, 1958.

Mayfield, J., & McNamee, P. (1998). Indexing using both N-grams and words. Retrieved on June 5, 2006, from http://www.citeseer.ist.psu.edu/mayfield98indexing.html

Pollock, J. J., & Zamora, A. (1975). Automatic abstracting research at chemical abstracts service. In M. & M.T. Maybury (Eds.). A*dvances in automatic text summarization*. Massachusetts Institute of Technology.

Robertson, S. E., & Sprack-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, *27(3),* 129-146.

Roelleke, T. (2003). A frequency-based and a poison-based definition of the probability of being informative. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* Toronto, Canada, 227-234.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, *24(5),* 513-523.

Salton, G., Yang, C. S., & Yu, C. T. (1975). A theory of term importance in automatic text analysis". *Journal of the American Society for Information Science*, *21(1),* 33-44.

Smith, G. W. (1991). *Computers and human language.* New York: Oxford University Press.

Sprack Jones, K. (1973). Index terms weighting. *Information Storage and Retrieval, 9,* 619-633.

Taghva, K., Beckley, R. & Sadeh, M. (2003). A list of Farsi stopwords. Retrieved September 7, 2006, from http://www.isri.unlv.edu/publications/isripub/taghva2003-01.ps

Tsz-Wai Lo, et al. (2005). Automatically building a stopword list for an information retrieval system. Retrieved February 10, 2007, from http://www.ir.dcs.gla.ac.uk/terrier/publications/rtlo_DIRpaper.pdf

Van Rijsbergen, C. J. (1975). *Information retrieval.* London: Butterworths.