# Comparison of Phrase Indexing for Biomedical and Newswire Documents

**Jose C. Clemente[1]*, Kentaro Torisawa[2] and Kenji Satou[3]**
[1]School of Knowledge Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292, Japan
{clemente,ken}@jaist.ac.jp
[2]DSchool of Information Science, Japan Advanced Institute of Science and Technology
1-1 Asahidai, Nomi, Ishikawa 923-1292
torisawa@jaist.ac.jp

## ABSTRACT

In this paper we compare a simple but widely used approach for multi-word indexing in two large collections of documents belonging to two different genres: newswire articles and biomedical abstracts. While in the first collection indexing results are reasonably accurate, in the second one performance drops noticeably. The special characteristics of the second corpus can explain the difference in results, and questions the validity of a naïve approach to the problem of multi-word indexing, opening an interesting line of research. By comparing the characteristics of both document sets we can bring some light into what aspects would be relevant to develop a domain independent multi-word indexing strategy.

**Keywords:** text mining, information retrieval, phrase indexing, bioNLP.

## 1. INTRODUCTION

### 1. 1. Multi-word indexing

In order to retrieve documents from a collection, we need to describe and index them according to their content, meaning or purpose. We will call *indexing* to the process of assigning certain *keywords* to a document. Because simple words are sometimes not enough to describe a topic, keywords can be in fact composed of multiple words: "Natural Language Processing" describes the field of Computer Science that deals with the study of language. Collocations are also interesting keywords, since their meaning cannot be decomposed into smaller units: "venetian blinds", for instance, makes reference to a certain kind of window cover and not, as could be wrongly infered from the words considered independently, to visually impaired inhabitants of Venetia. When keywords of a document are composed of more than one single word, as in the previous examples, we talk about *phrase indexing* (alternatively, multi-term or multi-word indexing).

Multi-word indexing can make use of more meaningful units to index documents than simple individual words. Names of people, places or organizations, for instance, tend to lose significance when decomposed, so keeping such phrases together would intuitively have benefitial effects for indexing purposes. On the other hand, multi-word indexing is more expensive and time-consuming than single word indexing. Depending on the complexity of the multi-word, we would need to identify Part-of-Speech (POS) tags, syntactic structures, etc. This pre-processing can also introduce undesired errors in the final results.

Nevertheless, multi-word indexing is becoming more relevant with the increasing amount of documents being published nowadays. Specifically, we would like to investigate the use of such technique for biomedical documents. The amount of this type of documents is growing faster than any other kind of textual collection, and this domain is regarded as linguistically more complex than general newswire articles in several aspects: terminology [1], named entity task [2], anaphora resolution [3], etc. Understanding whether previous multi-word approaches are also valid for this domain is therefore of scientific relevance.

### 1. 2. Previous approaches

Although multi-word indexing research has a short history, there are several works in the literature worth reviewing, with the TREC conference publishing some of the most relevant ones (see [4] for an introduction).

The main drawback of previous approaches is that they have been tried in a domain, newswire articles, which has very different characteristics to our domain of interest, biomedical documents. To our knowledge, there is no published work that studies the efficiency of

*Author to whom correspondance should be addressed

multi-word indexing techniques for such collections.

### 1. 3. Objective

The objective of this work is to demonstrate how the domain of interest can greatly affect the performance of multi-word indexing. Through the study of multi-word indexes obtained from two collections of documents belonging to different genres (newswire and biomedical), we can show how the performance of a naïve (but widely used) approach degrades greatly in the biomedical collection. By comparing the characteristics of these document sets, we can obtain some hints on what aspects would be relevant to develop a domain independent multi-word indexing strategy.

## 2. EXPERIMENTAL SETUP

For our experiments we used two different document collections: the Reuters corpus [5] and the Ohsumed corpus [6]. The Reuters corpus is a set of newswire articles from the Reuters agency. The Ohsumed corpus is composed of biomedical documents obtained from MEDLINE. Each document is annotated with zero, one or more topics that describe the general content of the text. Topics in Reuters tend to describe more general concepts ("earn", "gold", "corn", etc), while in Ohsumed there is a broad range of topics (from very general, like "human" or "female", to very specific, like "Uveal Neoplasms" , "Blood Pressure Determination" or "X-ray Computed Tomography").

Each corpus was pre-processed by removing stopwords, tokenizing words and sentences, and then running a POS tagger [7]. Although many indexing systems use a stemming algorithm to conflate lexically similar words into a single term, we decided to take a more conservative approach, and used instead a morphological analyzer [8], which only reduces verbs into their infinitive form and plural nouns into their singular. We then run a series of PERL scripts over the resulting documents to obtain all possible multi-word noun phrases (NP). NPs are extracted using the following pattern:

((Adj | Noun)+ | ((Adj | Noun)* (Noun Prep)?) (Adj |Noun)*) Noun

We limited the length of extracted NPs to bigrams, trigrams and fourgrams. Adding longer NPs did not result in a significant improvement of results despite the computational cost of its calculation.

Relevance of the resulting NPs was measured through their tf.idf score [9], and then normalized for each document as follows:

$$tf.idf(w,d) = tf(w,d) * idf(w)$$

$$tf(w,d) = 1 + \log(C(w,d))$$

$$idf(w) = \log\left(\frac{N}{df}\right)$$

$$tf.idf_{norm}(w,d) = \frac{tf.idf(w,d)}{\sqrt{\sum_i tf.idf(w_i,d)}}$$

where $C(w,d)$ is the frequency of word[1] $w$ in document $d$ and $N$ is the total number of documents. The relevance of a word for indexing purposes grows slower than its frequency, therefore we take $tf(w,d)$ to be the logarithm of the number of times a word occurs in a document. The inverted document frequency of a word, $idf(w)$, measures the proportion of documents in which a word occurs, giving higher weight to those words occuring in more documents. Finally, the weight of each word in a document is normalized by the sum of the weights of all words appearing in the same document.

Each topic was annotated with those multi-words of higher normalized *tf.idf* appearing in documents relevant to the topic. We then ran queries to retrieve documents using the best multi-words for each topic (each query would use from 10 to 1 multi-word, those of higher weight for the topic). Finally, we calculated the F-measure [10] for the retrieved set of documents corresponding to every topic as:

$$f = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}}$$

where $P$ stands for precision, $R$ for recall and alpha is a factor that determines the relative weighting of these measures. When alpha = 0.5, as in this work, the formula simplifies to:

$$f = \frac{2PR}{P+R}$$

## 3. RESULTS AND DISCUSSION

We performed the experiment described in the previous

[1] We will use the term "word" to refer to NPs whenever the meaning can be clearly inferred from the context

section on both Reuters and Ohsumed collections. It should be noticed that Reuters documents are annotated on average with less topics than Ohsumed articles. In the following figures, we can observe some examples of the most relevant extracted phrases (bigrams, trigrams and fourgrams) for a topic in both domains:

OHSUMED: Receptors, Insulin/*ME [bigrams]
1. insulin receptor: 0.32645664472183
2. egf receptor: 0.282891574559579
3. sm-c/igf-i insulin: 0.234365071387462
4. growth factor: 0.232797455571258
5. kinase domain: 0.231374713772426

OHSUMED: Receptors, Insulin/*ME [trigrams]
1. receptor sm-c/igf-i insulin: 0.204881576143945
2. epidermal growth factor: 0.187576353295739
3. origin structural similarity: 0.140229555769253
4. outside catalytic domain: 0.140229555769253
5. mechanism signal transduction: 0.140229555769253

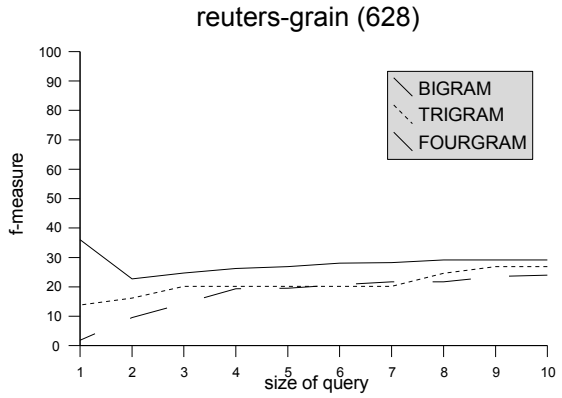
OHSUMED: Receptors, Insulin/*ME [fourgrams]
1. origin structural similarity cysteine-rich: 0.17218403
2. evolutionary origin structural similarity: 0.172184032
3. tyrosine kinase activity epidermal: 0.1721840322676
4. polypeptide chain insulin receptor: 0.1721840322676
5. similarity cysteine-rich region extracellular: 0.172184

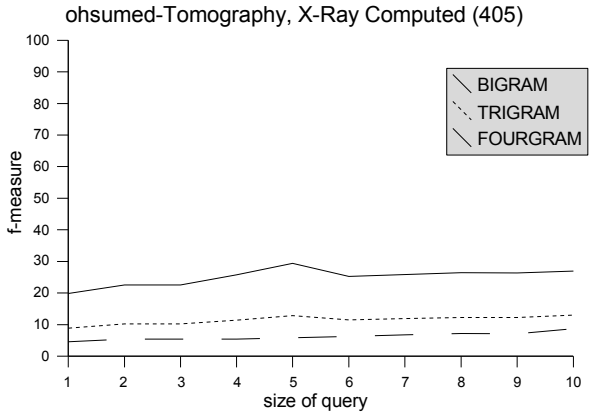
REUTERS: palladium [bigrams]
1. technigen platinum: 0.81127332894067
2. platinum corp: 0.775850862261192
3. platinum group: 0.525506360343429
4. drill section: 0.479151096995816
5. ounce palladium: 0.479151096995816

REUTERS: palladium [trigrams]
1. technigen platinum corp: 1.45747798761892
2. 13-hole drilling program: 0.860809978218734
3. nicel platinum property: 0.860809978218734
4. extensive near-surface zone: 0.437933103297611
5. platinum group base: 0.426579755265928

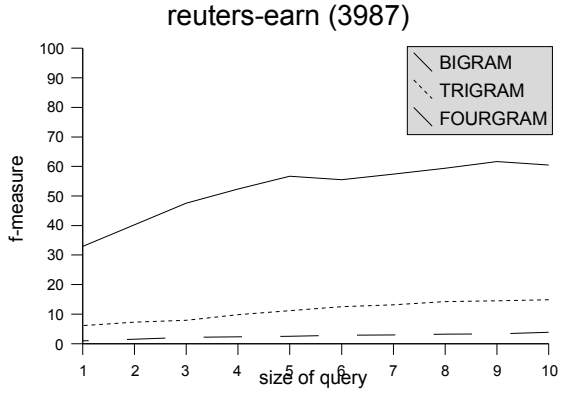
REUTERS: palladium [fourgrams]
1. platinum group base price: 0.747029567862128
2. dlr per troy ounce: 0.664790812767236
...

For illustration purposes, we chose four topics from Reuters and four from Ohsumed, and compared f-value measure on retrieval using high-scoring tf.idf extracted bigrams, trigrams and fourgrams.

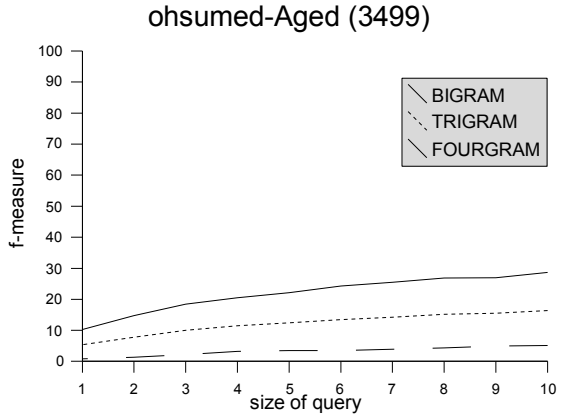


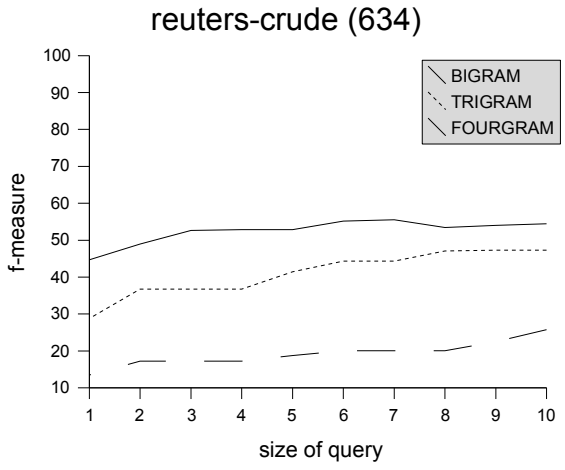

The use of n-grams as multi-word indexes is usually regarded as a simple but extremely effective approach (see [11], p. 533). As it can be seen in the graphs for Reuters topics “grain” (628 annotated documents) and “earn” (3987 documents), results in this corpus prove this claim partially correct. On the other hand, performance in the Ohsumed collection, as seen in topics “Tomography, X-Ray Computed” and “Aged” (405 and 3499 documents, respectively) is far from desirable, specially in the second one. It can also be observed how longer chains (trigrams and fourgrams)



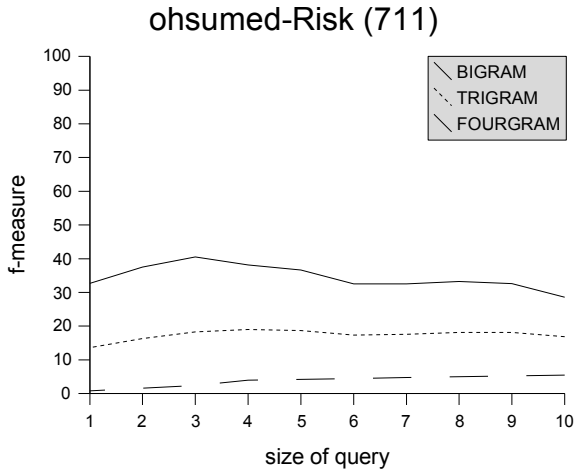
have worse performance than bigrams in both corpus. Longer chains should provide higher precision at the expense of a lower recall, ideally not affecting the f-value. Nevertheless, the low counts of trigrams and fourgrams in these corpora provoked a substantial drop of performance.

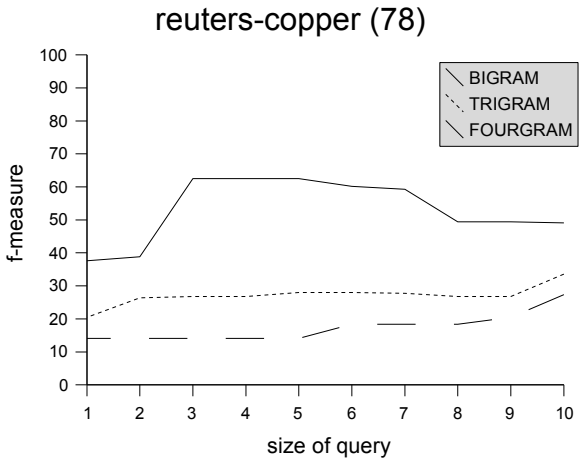



The topic "crude" in Reuters (634 documents) also outperforms the Ohsumed topic "Risk", with similar number of documents (711 documents). Finally, we chose "copper" (Reuters, 78 documents) and "Psychiatric Status Rating Scales" (Ohsumed, 79 documents) as an example of topics to which just a few documents belong. Results are consistent with previous ones, with reasonable performance in Reuters and a significant worsening for Ohsumed.



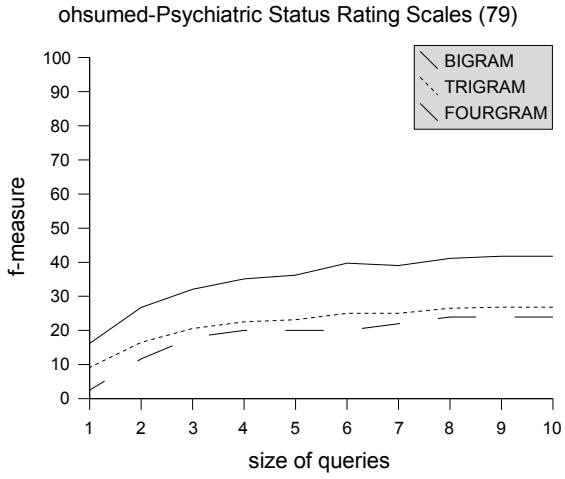

It is clear therefore that the same approach (bigrams, trigrams, fourgrams) using tf.idf as a score for indexing relevance performs considerably worse in biomedical documents. We proceeded therefore to an extensive study of the Ohsumed corpus, comparing it with Reuters, which lead to the discovery of the following several important differences. First, POS tagging for Ohsumed is far less accurate than for Reuters. Because multi-word extraction relies on POS tags, the obtained NPs are sometimes composed of wrong elements, which results in incorrect indexing terms. While this chain-effect problem (incorrect preprocessing provokes noisy input data) is common to other NLP-related tasks, it also shows how domain-specific taggers might be necessary in order to extract relevant knowledge from non-general documents. In general, domains like biomedicine tend to have a higher percentage of unknown words for the tagger, which can explains the lower accuracy of the POS tagger in our experiments.




Second, Ohsumed articles are on average shorter than Reuters documents. Shorter articles with a similar vocabulary size implies there are less indexable terms per article, so documents would be more similar to each other and therefore it would be more difficult to automatically index them. Biomedical corpora are constructed from scientific document abstracts, which are limited in size (usually, from 100 to 200 words).

Full-article corpora are still uncommon, since most published material is not freely available, although this seems to be changing thanks to the efforts of some open access publications. It would be therefore desirable to construct a biomedical corpus using longer articles, which would reduce the risks associated with low counts of index terms [12].




Third, Ohsumed articles contain a higher proportion of semantically empty terms not included in standard stopword lists. Because Ohsumed is constructed exclusively of biomedical documents, stopwords that are common in general texts sometimes do not appear in Ohsumed, and inversely words of low indexing relevance in Ohsumed that are not usually included in standard stopwords lists. The use of stop lists is controversial: although it helps to reduce the index size discarding useless words (those occuring in at least 80% of the documents [13]), stopwords are domain-specific and removing them can reduce the recall of queries.

Forth, average size of NPs is longer in Ohsumed, which results in meaningful units of longer size not being properly indexed. Biomedical documents often include terminology composed of long sequences of names, like chemical compounds or biological products. Determining when a long NP sequence should be kept as an unique index term or split into two or more is still an open problem.

Biomedical documents need in consequence a different approach to multi-word indexing. If we are to rely on POS tagging, taggers should be re-trained specifically for such kind of documents. We suggest the use of already annotated corpus such as GENIA [14] as a golden standard for retraining the POS tagger. The use of hand-made stopword lists is controversial, with different studies showing advantages and disadvantages of their utilization. It is generally agreed though that different domains require different sets of stopwords, therefore developing a method that does not rely in such artificial lists would be of great interest. Finally, being able to deal with the longer NP chains present in biomedical texts represents a more challenging problem, and we are currently working to obtain a multi-word extraction technique based on the concept of *chained NPs* [15] that would partially avoid the problems described above.

## 4. CONCLUSIONS

In this paper we have presented a comparison of a performance of a simple but widely used measure for phrase indexing on two different corpora: newswire and biomedical documents. Preliminary results show how f-value of queries constructed using bigrams, trigrams and fourgrams selected by their tf.idf score decrease significantly on biomedical documents. We found four possible causes for this worsening of performance. First, erroneous pre-processing in POS tagging leads to noisy input data. Second, shorter average size of biomedical documents makes them more difficult to index. Third, standard stopword lists are not well suited for the non-generic documents. Fourth, long NP phrases present in biomedical documents are more difficult to detect and index correctly. This clearly shows that domain-specific documents need a different approach for phrase indexing. We are already working in a new measure for phrase indexing that can avoid the mentioned problems.

## ACKNOWLEDGEMENTS

We would like to specially thank Dr. Tho Hoan Pham and Prof. Jun'ichi Kazama for their valuable comments and suggestions on this paper. This work was supported by Grant-in-Aid for Scientific Research on Priority Areas © “Genome Information Science” from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

## REFERENCES

[1]. Satou, K., and Yamamoto, K. 2004. Utilizing weakly controlled vocabulary for sentence segmentation in biomedical literature. *In Silico Biology 5, 0008*.
[2]. Kazama, J,. et al. 2002. Tuning Support Vector Machines for Biomedical Named Entity Recognition. *Proc. of the Workshop on Natural Language Processing in Biomedical Domain*.

[3]. Clemente, J.C., and Torisawa, K., and Satou, K. 2004. Improving the Identification of Non-Anaphoric it using Support Vector Machines. *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*.
[4]. Strzalkowski, T., et al. 1998. Natural Language Information Retrieval: TREC-7 Report. *Proc. of the Seventh Text Retrieval Conference.*
[5]. Rose, T.G., and Stevenson, M. And Whitehed, M. 2002. The Reuters Corpus Volume 1: From yesterday's news to tomorrow's language Resources. *Proc. of the Third Intl. Conf. on Language Resrouces and Evaluation*
[6]. Hersh, W., et al. 1994. OHSUMED: An interactive retrieval evaluation and new large test collection for research. *Proceedings of SIGIR'94.*
[7]. Brill, E. 1994. Some Advances In Rule-Based Part of Speech Tagging. *Proceedings of AAAI'94.*
[8]. Minnen, G., and Carroll, J. and Pearce, D. 2001. Applied morphological processing of English. *Natural Language Engineering, vol. 7, Issue 3.*
[9]. Salton, G., and Buckley, C. 1988 Term weighting approaches in automatic text retrieval. *Information Processing and Management, 24:513-523.*
[10]. van Rijsbergen, C.J. 1979. *Information Retrieval*. London: Butterworths. Second Edition.
[11]. Manning, C. D., and Schutze, H. Foundations of Statistical Natural Language Processing. The MIT Press. Fourth Edition.
[12]. Dunning, T. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61-74.
[13]. Baeza-Yates, R., and Ribeiro-Neto, B. 1999. *Information Retrieval*. New York: Addison-Wesley.
[14]. Tomoko, O., and Tateisi, Y., and Mima H., and Tsujii, J. 2002. GENIA Corpus: an Annotated Research Abstract Corpus in Molecular Biology Domain. *Proceedings of the Human Language Technology Conference, HLT'2002*.
[15]. Frantzi, K.T., and Ananiadou, S. 1996. Extracting nested collocations. *Proceedings of the 16th International Conference on Computational Linguistics, COLING'96*.