Investigating Multi Word Expressions to Enhance Information Retrieval on Biomedical Text

Topics in Artificial Intelligence CS-598

Sub-Topic - Natural Language Processing in Biomedicine

Shlok Kothari

March 10, 2024

Abstract

This study delves into the realm of multi-word expressions (MWEs) within the domain of biomedical information retrieval, employing Elasticsearch as the tool for investigation. The core objective revolves around evaluating the efficacy of treating MWEs as singular entities versus individual words. The underlying hypothesis posits that in cases of heightened complexity, amalgamating MWEs could potentially enhance retrieval outcomes. Through rigorous experimentation, it was observed that certain categories of MWEs, such as proper names and horizontally dominant MWEs, indeed exhibit improved performance when presented as cohesive units. Conversely, MWEs characterized by substantial variability along the vertical axis did not demonstrate significant improvement when combined. This underscores the importance of considering the nuanced relationships between constituent words within MWEs. However, it was noted that our hypothesis did not universally hold true, as several factors, including vertical and horizontal dominance, influence the optimal treatment of MWEs. Despite promising results, the study underscores the necessity for further research and larger datasets to validate and refine our findings, aiming to optimize information retrieval methodologies within the medical domain.

Introduction

Lexemes are the smallest meaningful units of natural language (e.g., “happy”). Natural language is composed of lexemes that amalgamate through specific syntax and morphology to convey meaning in phrases and sentences. While it may seem trivial, determining what qualifies as a word presents a surprisingly complex challenge. For example, are terms like ’stroll’ and ’strolls’ distinct words or merely variations of a single word? Conventional perspectives in linguistics classify them as distinct inflected word forms of the lexeme ‘stroll’. Secondly, consider the phrases - “spill the beans” and “nursing home”. If an individual is acquainted with the separate words “spill,” “beans,” “nursing,” and “home” but has not encountered these specific combinations, predicting the meanings of “spill the beans” and “nursing home” poses a challenge. To grasp the semantics of these phrases, they must hold a distinct lexical status within their lexicon housing their specific meanings. Expressions like these, characterized by unexpected properties not completely deducible from their constituent words, can be categorized as MWEs.

In this research endeavor, I intend to adopt the definition of MWEs proposed by Sag et al. (2002), characterizing them as “idiosyncratic interpretations that transcend word boundaries (or spaces).” As noted by Jackendoff (1997a, 156), the prevalence of MWEs in a speaker’s lexicon is posited to rival that of single words, potentially even being underestimated, particularly when accounting for lexicalized phrases. Notably, WordNet 1.7 (Fellbaum 1999) indicates that around 41% of its entries constitute MWEs, a figure likely to be conservative for a comprehensive Natural Language Processing (NLP) system[1].

Expanding on this perspective, it is essential to recognize that the significance of MWEs becomes more pronounced in the biomedical domain. The dynamic nature of biomedical lexicons, coupled with the intricate nuances of medical terminology, contributes to a substantial increase in the prevalence and complexity of MWEs. As we delve into the categorization of these expressions, we aim to understand their diverse manifestations and structures, laying the groundwork for a our analysis.

Classifying Multiword Expressions

We can broadly classify MWEs in two categories - **lexicalized phrases** and **institionalized phrases** (terminology adapted from Bauer (1983)). Lexicalized phrases exhibit partially idiosyncratic syntax or semantics and may include 'words' that do not occur independently. They can be further breakdown into **fixed expressions**, **semi-fixed expressions**, and **syntactically-flexible expressions**, arranged roughly in decreasing order of lexical rigidity. In contrast, **institutionalized phrases** are syntactically and semantically compositional but occur with high frequency within a specific context[1].

Fixed Expressions - Fixed expressions are fully lexicalized phrases, resistant to morphosyntactic variation and internal modification. They defy conventional grammar rules and remain immutable in form. For instance, "in vivo" in the biomedical domain is a fixed expression. (cannot be modified morphologically as e.g. "in vivoly" or internally e.g. "in the vivo").

Semi-Fixed Expressions - Semi-fixed expressions follow strict rules regarding word order and composition but allow some lexical variation, such as inflection, reflexive form changes, and determiner selection. They can be categorized into **non-decomposable idioms**, **compound nominals**, and **proper names**.

- **Non-decompositional idioms** - Some expressions, like "spill the beans," can be interpreted by examining each word individually, such as spilling something and beans representing secrets. However, there are other expressions, like "pull the plug," that defy straightforward analysis when broken down. These are referred to as non-decomposable idioms. Non-decomposable idioms maintain a fixed structure, resisting alterations like adding extra words or changing word order. Their only potential variations involve inflection, as seen in past tense (e.g., "pulled the plug"), or the use of a reflexive form.
- **Compound Nominals** -Compound words like 'car park,' 'attorney general', are structurally fixed, allowing changes for plurality, such as adding an -s. However, certain compound words, like 'attorney general,' deviate from this rule, presenting challenges for pluralization. In biomedical domain, and example of Compound Nominal is 'Gene expression'. 'Gene expression' refers to the process by which information from a gene is used to synthesize a functional gene product, such as a protein. It combines the words 'gene' and 'expression' to represent a specific concept in molecular biology. In this report, we term these constructions as **weak idioms**, acknowledging their fixed structure yet relatively transparent word composition.
- **Proper Names** - Proper names, like the names of sports teams in the U.S., are unique and specific. They usually consist of a place or organization name followed by a special name that sets them apart in their sport. Sometimes the first part of the name can be left out, like in 'the 49ers' instead of 'San Francisco 49ers'. Proper names are ubiquitous in biomedical texts, as they often refer to entities like genes, proteins, diseases, pharmaceuticals, research institutions, and more. These names are crucial for accurately conveying information and maintaining precision in scientific communication.(e.g. 'National Institute of Health Research' (NIHR))

Syntactically-flexible Expressions - Semi-fixed expressions always have the same word order, while syntactically-flexible expressions can have lots of different word orders. We can categorize them as **Verbs-Particles Contruction**, **Decomposible idioms** , and **Light verbs**

- **Verbs-Particle Contruction -** Verb-particle constructions are phrases containing a verb and one or more particles, altering the verb's meaning, as observedin "brush up on" for reviewing. Particles can carry unique meanings for specific phrases or have broader usage.
  These constructions may also position a noun or pronoun in the middle or after the verb and particle, like 'call Kim up' or "fall off a truck." Some require the particle at the beginning, such as 'fall off a truck', while others allow flexibility 'call up Kim'. Adverbs, like "fight bravely on," can be inserted between the verb and particle.
  Due to these variations, establishing rules for all verb-particle constructions is challenging.

- **Decomposable Idioms -** Decomposable idioms are prone to high syntactic variation. Unlike non-decompositional idioms, such as 'kick the bucket' or 'spill the beans,' which stastically maintain a rigid structure, decomposable idioms like 'crack the code' or 'sweep under the rug' exhibit a degree of flexibility in how their components can be rearranged. The extent of this syntactic variation is often unpredictable (Riehemann 2001), making it challenging to establish consistent rules for their structural modifications.

- **Light verbs -** Constructing light verbs involves combining two words in a specific manner, like in "make a mistake" or "give a demo." This can be challenging because predicting which words will go together is not straightforward. While some view these expressions as idoms, in truth, the words maintain their regular meanings, and the way they're used together is just slightly different. These combinations can be reorganized within sentences, for example, turning "give a demo" into "a demo was given" or asking, "How many demos did Kim give?" Consequently, they don't adhere to the typical pattern of words separated by spaces.

Institutionalized Phrases - Institutionalized phrases, exemplified by terms like "traffic light" or "telephone booth," possess both semantic and syntactic compositionality. This means that each word within the phrase retains its individual meaning, contributing to an organized structure that conveys an overall comprehensible meaning. Despite their compositionality, these phrases occasionally exhibit statistical idiosyncrasy, highlighting certain combinations that are conventionally established and used more frequently than potential alternatives.

The example of "traffic light" illustrates the tendency to favor conventional terms over potential alternatives (e.g., "traffic director" or "intersection regulator") due to familiarity and convention. The term **anti-collocations** will be used when describing potential lexical variants that are rarely or never observed.

It's important to understand that the categories outlined for MWEs are just to help you see how diverse and interesting they are. When articles/papers talk about MWEs there is abundant terminology associated with them, some common ones being, *fixed expression, formulaic sequence, idiom, idiomatic expression, lexical/lexicalized phrase, multi-word unit, phraseme, phraseologism, phraseological unit, phrasal lexical item, phrasal lexeme, prefabricated chunk, compound nominal, light verbs.* But often these words overlap or are a bit complicated. Also, MWEs don't fit neatly into one category—they're flexible and can have different features depending on the situation. So, while categories give us a way to think about MWEs, it's good to remember that they're not set in stone. Plus, as researchers keep studying and working in this area, they keep discovering new terms and ideas to add to the mix.

Research specific information

In this research my focus is on investigating potential MWEs that may not fit traditional MWE categories or fall under specific classifications. These MWEs are considered combinational, extracted based on their frequent co-occurrence in various types of text. They can take forms such as Noun + Verb, Noun + Noun, or Verb + Noun, and may vary in strength from strong idioms to weak idioms. The aim is to explore whether there exists a relationship between these words that can be leveraged to enhance retrieval systems.

To delve deeper into understanding the structure of these expressions, we turn to linguistic concepts introduced by Firth, particularly the paradigmatic and syntagmatic axes.

Firth's concept of collocations can be understood through two main axes: the paradigmatic axis and the syntagmatic axis. Let's break down what each axis means and how it applies to language, particularly in the biomedical domain.

1. **Paradigmatic Axis (Vertical):**
   - The paradigmatic axis is like a vertical line connecting words that belong to the same category or class and can be interchanged without changing the overall meaning of the sentence.
   - In biomedical text, for example, consider the words "disease," "condition," and "ailment." These words belong to the same category of health-related issues and can be used interchangeably in certain contexts without altering the meaning significantly.
   - Similarly, in a sentence like "The patient was diagnosed with a disease," you could replace "disease" with "condition" or "ailment" and the sentence would still make sense, demonstrating their paradigmatic relationship.
2. **Syntagmatic Axis (Horizontal):**
   - The syntagmatic axis is like a horizontal line connecting words that are connected or associated with each other in a sentence or phrase.
   - In the biomedical domain, let's consider the phrase "heart attack." Here, "heart" and "attack" are syntagmatically related because they are connected within the phrase to describe a specific medical event.
   - In the sentence "The patient suffered a heart attack," "heart" and "attack" are linked together to convey the occurrence of a specific medical event involving the heart. Additionally, "suffered" is also syntagmatically related to "heart attack" because it describes the patient's experience.

   Overall, the paradigmatic axis helps us understand how words within the same category can be interchanged, while the syntagmatic axis helps us understand how words are connected or associated with each other within a sentence or phrase.

I want to introduce a few terms of my own. I am going to use the term **Vertically dominant** to describe situations where there exists a multitude of variations of a collocation on the paradigmatic axis or the vertical axis. For example, in biomedical text, terms like "fish oil" or "breast cancer" exhibit vertical dominance, as there are numerous variations or synonymous expressions for these concepts.

On the other hand, I will use the term **Horizontally dominant** when there is a strong meaning or connotation associated with the MWE on the syntagmatic or horizontal axis. For instance, terms like "cloud computing" carry a specific and significant meaning within the context of technology and computing.

If the variations or the number of options is less vertically, or if there is a single, more frequently occurring MWE, I will use the term **Vertically light** to describe such instances. These terms aim to provide additional clarity in understanding the complexity and variations of MWEs.

Dataset

I have utilized three different datasets for my exploration.

1. **TREC Genomics (2004 - 2005)**[3][4] -
   The dataset comprises a subset of the MEDLINE database spanning a decade of completed citations from 1994 to 2003.The documents are provided in the MEDLINE XML format, distributed across five files, each ranging in size from 553 MB to 605 MB. This yielded a total of 4,591,008 records, offering a rich resource for research and information retrieval tasks in the biomedical domain.

   Relevant judgments are accessible, enabling precise bpref measurement.

   There are two tracks (2004, 2005) that share the same document corpus. The table lists the features of query dataset in these tracks.

Number of queries	Queries with MWE	Percentage
50	24	46%

   Table 1: 2004 TRACK

Number of queries	Queries with MWE	Percentage
50	30	60%

   Table 2: 2005 TRACK

2. **Clinical Decision Support (2014 - 2017)**[5][6][7] -
   The Clinical Decision Support dataset focuses on the Open Access Subset of PubMed Central (PMC), a digital database housing freely available full-text biomedical literature. Obtained as of January 21, 2014, the dataset comprises 733,138 articles, each represented in NXML format. Additionally, the dataset includes images and supplementary materials.

   Relevance judgements are based on topics, consisting of medical case narratives crafted by medical practitioners. These narratives mimic actual medical records, detailing a patient's medical history, current symptoms, diagnostic tests, eventual diagnosis, and the physician's treatment steps. Given the nature of these topics, the dataset presents a high likelihood of containing MWEs.

There are three tracks (2014, 2015, 2016) that share the document corpus. The table lists the features of query dataset in these tracks.

Number of queries	Queries with MWE	Percentage
30	30	100%

Table 3: 2014 TRACK

Number of queries	Queries with MWE	Percentage
30	29	97%

Table 4: 2015 TRACK

Number of queries	Queries with MWE	Percentage
30	28	93%

Table 5: 2016 TRACK

3. **NFCorpus (2016)**[8] **-**

   The NFCorpus is a comprehensive full-text English retrieval dataset designed for Medical Information Retrieval. It encompasses a total of 3,244 natural language queries, composed in non-technical English, gathered from the NutritionFacts.org site. The dataset includes 169,756 automatically extracted relevance judgments associated with 9,964 medical documents. These documents, written in a language rich in complex terminology, are primarily sourced from PubMed.

Number of queries	Queries with MWE	Percentage
3244	540	17%

Table 6: NFCorpus

Hypothesis and Methodology

In the course of my investigation, I sought to understand the consequences of query modification prior to submission to a retrieval system. Focusing on multi-word expressions within the queries, I experimented with two distinct approaches and compared their results.

Consider the query - “Generating transgenic mice”.
MWE present in this query is - “transgenic mice”

Traditional approach:

generating OR transgenic OR mice

Modified MWE approach:

generating OR transgenic mice

In the first method, the query was presented to Elasticsearch as is, employing the logic ’Generating OR Transgenic OR Mice.’ This traditional approach treats each word independently. However, in the second modification, particular attention was given to multi-word expressions, such as ’transgenic mice.’ In this case, the query was formatted for Elasticsearch using the logic ’Generating OR “Transgenic Mice,”’ treating the multi-word expression as a singular entity.

As previously noted, the diverse nature of multi-word expressions (MWEs), especially within the domain of biomedical text, makes classifying them into specific categories challenging. Given this complexity, formulating solid hypotheses is tricky. However, intuitively, **my only hypotheses is that as the opaqueness of multi-word expressions increases, treating them as singular entities would likely yield improved retrieval outcomes.**

Project Steps (Sequentially)

1. Start the ElasticSearch Docker container (refer to Appendix A for setup details).
2. Connect the Python client to your deployment (see Appendix B).
3. Create the index with settings tailored to the data (refer to Appendix B).
4. Filter the document to extract the article title and abstract.
5. Preprocess the documents through lemmatization, stemming (Krovetz stemmer)[9], and removal of stop words.
6. Index these documents.

Now, let’s move on to the queries.

1. Extract the queries from the file.
2. Identify the multi-word expressions (MWEs) present in the queries.

3. Preprocess the query by stemming, lemmatizing, and removing stop words.
4. Modify the query based on the identified MWEs.
5. Search the index.
6. Compare the results with qrels.
7. Calculate the summation of scores for the top 30 retrieved documents.

Analysis and Discussion

For the analysis and discussion, a total of 141 queries were examined. The MWEs were grouped into specific categories based on their characteristics. Here are the results -

- **Proper Names and Fixed Expressions** - This comprises strong or proper names that represent specific entities within the biomedical domain. While these terms may not exhibit significant vertical or horizontal variability, they often carry specific connotations or denote well-defined concepts.
  - multiple sclerosis
  - bronchoalveolar lavage
  - irritable bowel syndrome
  - new york
  - ex vivo
- **Vertically Dominant and Synonymous** - MWEs in this category exhibit high variability along the vertical axis, with multiple variations or synonymous expressions. From the data the examples are -
  - nuclease assay
  - microarray data
  - insulin receptor
  - casein kinase
  - viral capsid
  - old male
  - mouse kidney
  - dna repair
  - peptidoglycan recognition
  - shortness of breath
  - male fertility
- **Horizontally Dominant** - They represent horizontally dominant MWEs where the combination of words carries a strong connotation or denotes a specific concept.

  - white blood cells
  - plant based diet
  - mediterranean diet
  - food poisoning
  - green tea
  - brown fat
  - red dye

My initial analysis reveals two important findings -

1. Proper names and horizontally dominant words, such as ”Multiple sclerosis” or ”Plant based diet,” often exhibit improved performance when treated as a single entity in retrieval systems, suggesting a strong association or specific meaning linked to their combination.
2. Vertically dominant words, such as ”fish oil” or ”breast cancer,” represent situations where there exists a multitude of variations or synonymous expressions for the same concept, leading to high variability along the vertical axis. However, it is observed that due to this variability, they do not show improvement in retrieval results.

However, further deep analysis reveals some additional findings that warrant consideration.

- **Vertically Dominant MWEs with Equal Word Significance** - They represent MWEs where there is vertical dominance, meaning variations or alternatives may exist for either word, yet both words hold equal importance in conveying the intended meaning.
  - ovarian cancer
  - abdominal pain
  - hip arthroplasty
  - coronary artery
- **Internally modifiable Horizontally Dominant** - MWEs in this category exhibit a strong connotation or specific meaning associated with the combination of words, resulting in horizontal dominance. However, these horizontally dominant MWEs contain terms that can be internally modified or rearranged while still conveying the same intended meaning. Variations in the composition of such MWEs may arise due to differences in writing styles, preferences, or terminological conventions among authors or users. Examples of such cases include:
  - mad cow disease | mad cow | disease of mad cow
  - adenomatous polyposis | adenomatous polyposi coli
  - aminobutyric acid | gamma-aminobutyric acid
  - respiratory distress | respiratory distress syndrome
  - smoking history | history of smoking
- **Substitutional Horizontally Dominant** - In this category, degradation occurs when one or more terms within the horizontally dominant MWE can be substituted with alternative terms without significantly altering the intended meaning.
  - productive cough | wet cough

  - heart rhythm | cardiac rhythm
  - oxidative stress | oxide stress
  - bipolar disorder | bipolar I

- **Ambiguous Horizontal Dominance** - This category encompasses horizontally dominant MWEs that possess multiple interpretations or meanings. Instances are -
  - nucleoside diphosphate kinase (gene and an enzyme)
- **Balanced Horizontal Dominance -** This category comprises horizontally dominant MWEs where both constituent terms hold significant importance individually, contributing to the overall meaning of the expression. Examples -
  - hemiplegic migraine

Deeper analysis reveal further key findings -

3. Combining vertically dominant words of equal importance, such as "ovarian cancer," does not improve retrieval results.
4. Internally modifiable horizontally dominant MWEs degrade performance, suggesting a need for analyzing different modifications of these MWEs in the corpus.
5. Horizontally dominant words, such as 'productive cough' and 'wet cough,' are substitutable. Understanding the vertical variation associated with them provides an opportunity to enhance retrieval. Given the light vertical variation, treating them as a single entity can be advantageous.
6. When there is horizontal dominance and both entities are of equal importance, it is preferable not to combine these terms as a single entity for the retrieval system.

Results

Given these findings, I am now going to apply this newly acquired knowledge to a new dataset-nfcorpus. The procedure I followed involved classifying the multi-word expressions found in queries of nfcorpus and predicting whether they would improve or degrade retrieval performance based on my findings. It's important to note that for the subsequent analysis, I have focused only on significant improvements and decrements in query performance, where the difference $>\sim 10$.

Here are my results

Table 7: Accuracy of Predicting Query Degradation

Total Degraded	**Predicted Degradation**	**Correct Prediction (%)**
156	118	75.64%

Table 8: Accuracy of Predicting Query Improvement

Total Improved	**Predicted Improvement**	**Correct Prediction (%)**
24	13	54.16%

Using my findings, I achieved a degradation prediction accuracy of 75.64%, indicating that out of the total number of queries that experienced degradation, I correctly predicted degradation in nearly three-quarters of the cases. This suggests a robust ability to anticipate instances where combining multi-word expressions might lead to a decrease in retrieval system performance. Furthermore, the accuracy of predicting query improvement was 54.16%. While this accuracy is slightly lower than that for degradation prediction, it still demonstrates a considerable capability to identify scenarios where combining multi-word expressions enhances retrieval outcomes. This finding underscores the importance of considering the nuanced relationships between terms within multi-word expressions when optimizing information retrieval strategies.

Conclusion

The aim of this project was to investigate the impact of treating multi-word expressions (MWEs) as singular entities in the context of query modification for information retrieval using Elasticsearch. By experimenting with different approaches, the goal was to understand the consequences of presenting MWEs in a unified manner and assess whether this modification enhances or diminishes retrieval results.

While my hypothesis suggested that as the opaqueness of multi-word expressions increases, treating them as singular entities would likely yield improved retrieval outcomes, my findings revealed a more nuanced picture.

My analysis revealed several key findings:

1. **Proper names and horizontally dominant words**, such as "Multiple sclerosis" or "Plant-based diet," often exhibited improved performance when treated as a single entity in retrieval systems. This suggests a strong association or specific meaning linked to their combination.
2. **Vertically dominant words**, such as "fish oil" or "breast cancer," represent situations where there exists a multitude of variations or synonymous expressions for the same concept, leading to high variability along the vertical axis. However, due to this variability, they did not show improvement in retrieval results.
3. **Internally modifiable horizontally dominant MWEs** degraded performance, suggesting a need for analyzing different modifications of these MWEs in the corpus.
4. **Horizontally dominant words**, such as 'productive cough' and 'wet cough,' are substitutable. Understanding the vertical variation associated with them provides an opportunity to enhance retrieval. Given the light vertical variation, treating them as a single entity can be advantageous.

My test results demonstrated a degradation prediction accuracy of 74.88%, indicating a robust ability to anticipate instances where combining multi-word expressions might lead to a decrease in retrieval system performance. Furthermore, the accuracy of predicting query improvement was

58.24%, underscoring the importance of considering the nuanced relationships between terms within multi-word expressions when optimizing information retrieval strategies.

Evaluation

1. **Data Size and Diversity**: The evaluation of the project could be influenced by the size and diversity of the dataset used. The nfcorpus dataset, while significant, might be considered relatively small for comprehensive analysis. It's also important to note that the nfcorpus dataset consisted of strings of medical words as documents rather than regular medical document text, which may have impacted the analysis and generalizability of the findings.
2. **Experimental Methodology:** The project employed an experimental methodology to analyze the performance of treating MWEs as singular entities. This involved predicting the impact of combining MWEs on retrieval outcomes and assessing the accuracy of these predictions. While the approach provides valuable insights, the effectiveness of the methodology could be further explored through rigorous testing and validation.
3. **Findings and Implications:** Despite the limitations, the findings of the project reveal promising insights into the potential benefits of treating certain types of MWEs as single entities in retrieval systems. Proper names and horizontally dominant expressions demonstrated improved performance when combined, suggesting a strong association or specific meaning. This highlights the importance of considering the contextual significance of MWEs in retrieval tasks.
4. **Need for Further Research:** The evaluation underscores the need for further research to validate and extend the findings of the project. More extensive datasets encompassing diverse domains and larger sample sizes could provide a more comprehensive understanding of the impact of MWE treatment on retrieval outcomes. Additionally, hypothesis testing and statistical analysis could offer robust validation of the observed trends.

In conclusion, while the project offers valuable insights into the impact of treating MWEs as singular entities in retrieval systems, it also highlights the need for further research and validation. By addressing the limitations and building upon the findings, future studies can contribute to advancing the effectiveness and efficiency of information retrieval techniques.

Appendices

Appendix A:

Setting up elasticsearch on ilabs and importing the index

1. Set up docker on ilabs using the guide - Docker Installation Rutgers. Refer to non-root section

   Elasticsearch container requires extra resources than available to non-root users on Rutgers ilab machines. However, we can request resources by emailing - Rutgers CS Help

2. Email Rutgers CS Help requesting following resources
   (a) Increase processes to 7000 processes
   (b) Increase open files to 100K

3. Create a new docker network and pull the elasticsearch image using the commands

```
$ docker network create elastic
$ docker pull docker.elastic.co/elasticsearch/elasticsearch:7.6.0
```

4. Download the indices from the Google Drive using your Rutgers email by accessing the google drive link Indexes. Each index directory includes a subdirectory. Please make a note of the location where you download the indices.

5. Execute the container using the provided command, replacing '/path/to/index' with the actual location where you downloaded the index.

```
$ docker run
    --name elasticsearch
    --net elastic
    -p 9200:9200
    -v /path/to/index/subdirectory:/usr/share/elasticsearch/data
    -e ”discovery.type=single-node”
  docker.elastic.co/elasticsearch/elasticsearch:7.6.0
```

6. Pull and run Kibana using the commands

```
$ docker pull docker.elastic.co/kibana/kibana:7.6.0
$ docker run
    --name kib01
    --net elastic
    -p 5601:5601
  docker.elastic.co/kibana/kibana:7.6.0
```

7. Verify whether the containers are running by using the following commands:

```
$ docker ps # tells docker container status
$ curl localhost:9200 # if successful it will return a json
```

Appendix B:

Setting up elasticsearch python client and index

1. Install elasticsearch python client using pip.

   ```
   $ pip install elasticsearch=7.6.0
   ```

2. Connect your python client to your deployment

   ```
   import elasticsearch
   from elasticsearch import Elasticsearch

   es = Elasticsearch(hosts =
           [{”host”:”localhost”,
           ”port”:9200,
           ”scheme” : ”http”}],
           timeout=40)

   # Test your connection
   # If successful it should print True
   print(es.ping())
   ```

3. Configure the settings for your index according to the data. This configuration is an example which I used for Genomics data -

```
Settings = {
        "settings": {
            "number_of_shards": 1,
            "number_of_replicas": 0,
            "analysis": {
                "analyzer": {
                    "custom_analyzer": {
                        "type":"custom",
                        "tokenizer": "whitespace",
                        "filter": ["lowercase","kstem"]
                    }
                },
                "filter":{
                    "kstem": {
                        "type":"kstem"
                    }
                }
            }
        },
      "mappings": {
        "properties": {
          "ArticleTitle": {
            "type": "text",
            "analyzer": "custom_analyzer"
          },
          "AbstractText":{
              "type":"text",
              "analyzer":"custom_analyzer"
          },
          "PMID":{
              "type": "keyword",
              "index": "false"
          }
        }
      }
    }
  }
}
```

Listing 1: Elasticsearch Configuration Settings (Genomics Data)

4. Create the index using the Setting variable

```
es.indices.create(index='genomicsindex',
                  ignore= [400,404],
                  body=Settings)
```

5. Check if the index has been created using the code

```
# Prints all the indices. Check for 'genomicsindex'
indices = es.indices.get_alias("*")
for index in indices:
    print(index)
```

Appendix C:

Data idiosyncrasies During the process of identifying multi-word expressions (MWEs) within the queries, it was observed that certain words lacking semantic sense were identified. These words were subsequently removed from the dataset to ensure the integrity of the analysis. The following words were identified and removed:

1. s scarcoma
2. s disease
3. s milk
4. t forget
5. s wort
6. s health
7. s esophagus

Table 9: Combined Statistics for All Datasets

Dataset	**Number of Queries**	**Queries with MWE**	**Percentage with MWE**	**Queries Improved**	**Percentage of Improvement**
nfcorpus	3244	540	16.90%	91	16.85%
Genomics Dataset 2004	50	24	46%	7	29.17%
Genomics Dataset 2005	50	30	60%	8	26.67%
Clinical Dataset 2014	30	30	100%	3	10.00%
Clinical Dataset 2015	30	29	97%	14	48.28%
Clinical Dataset 2016	30	28	93%	13	46.43%

References

1. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. Proceedings of Computational Linguistics and
2. Jackendoff, Ray. The Architecture of the Language Faculty. Cambridge, Mass. ; London, Mit Press, 1997.
3. Hersh, William R., et al. 'TREC 2004 Genomics Track Overview.' 1 Oregon Health & Science University, 2004
4. Hersh, William R., et al. 'TREC 2005 Genomics Track Overview.' 1 Oregon Health & Science University, 2005
5. Simpson, Matthew S., Ellen M. Voorhees, and William Hersh. "Overview of the TREC 2014 Clinical Decision Support Track."12 National Institutes of Healthe, 2014.
6. Roberts, Kirk, et al. "Overview of the trec 2015 clinical decision support track." TREC. 2015.
7. Roberts, Kirk, et al. "Overview of the TREC 2016 Clinical Decision Support Track."
8. Vera Boteva, Demian Gholipour, Artem Sokolov and Stefan Riezler: A Full-Text Learning to Rank Dataset for Medical Information Retrieval Proceedings of the 38th European Conference on Information Retrieval (ECIR), Padova, Italy, 2016
9. Chen, Ruey-Cheng, and Reza Gharibi. "KrovetzStemmer." GitHub, GitHub Inc., 25 Commits, github.com/rmit-ir/KrovetzStemmer.
10. Wikipedia contributors. "Multiword expression." Wikipedia, The Free Encyclopedia. Wikipedia, The Free Encyclopedia, 5 Feb. 2024. Web. 13 Mar. 2024.
11. Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, Amalia Todirascu; Multiword Expression Processing: A Survey. Computational Linguistics 2017; 43 (4): 837–892.
12. De Marneffe, Marie-Catherine, et al. Multi-Word Expressions in Textual Inference: Much Ado about Nothing?
13. Remy, François, et al. Detecting Idiomatic Multiword Expressions in Clinical Terminology Using Definition-Based Representation Learning. 2023.
14. Chakraborty, Tanmoy, and Sivaji Bandyopadhyay. Multiword Expression Multiword Expression Multiword Expression Multiword Expressions S S S under the Esteemed Guidance Of. 2009.
15. Agata Savary, Carlos Ramisch, Silvio Cordeiro, Federico Sangati, Veronika Vincze, Behrang QasemiZadeh, Marie Candito, Fabienne Cap, Voula Giouli, Ivelina Stoyanova, and Antoine Doucet. 2017. The PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017), pages 31–47, Valencia, Spain. Association for Computational Linguistics.
16. Baldwin, Timothy, and Su Nam Kim. "Multiword Expressions." Handbook of Natural Language Processing, edited by Nitin Indurkhya and Fred J. Damerau, 2nd ed., CRC Press, Boca Raton, 2010, pp. 267-292
17. Villavicencio, Aline, et al. "Introduction to the Special Issue on Multiword Expressions: Having a Crack at a Hard Nut." Computer Speech & Language, vol. 19, no. 4, 1 Oct. 2005, pp. 365–377, https://doi.org/10.1016/j.csl.2005.05.001. Accessed 21 Apr. 2023.

18. Butler, Keith. Literature Survey: Multi-Word Expressions (MWEs) for CISC889 Fall 2013.