Query selection methods for automated corpora construction with a use case in food-drug interactions

Georgeta Bordea[1], Tsanta Randriatsitohaina[2], Natalia Grabar[3], Fleur Mougin[1] and Thierry Hamon[2,4]

[1]Univ. Bordeaux, Inserm UMR 1219, Bordeaux Population Health, team ERIAS, Bordeaux, France
Email: name.surname@u-bordeaux.fr
[2]LIMSI, CNRS UPR 3251, Université Paris-Saclay, Orsay, France
[3]CNRS UMR 8163 - STL - Savoirs Textes Langage, Univ. Lille, Lille, France
[4]Université Paris 13, Sorbonne Paris Cité, Villfetaneuse, France

Abstract

In this paper, we address the problem of automatically constructing a relevant corpus of scientific articles about food-drug interactions. There is a growing number of scientific publications that describe food-drug interactions but currently building a high-coverage corpus that can be used for information extraction purposes is not trivial. We investigate several methods for automating the query selection process using an expert-curated corpus of food-drug interactions. Our experiments show that index terms features along with a decision tree classifier are the best approach for this task and that feature selection approaches and in particular gain ratio outperform frequency-based methods for query selection.

1 Introduction

Unexpected Food-Drug Interactions (FDIs) occasionally result in treatment failure, toxicity and an increased risk of side-effects. While drug-drug interactions can be investigated systematically, there is a much larger number of possible FDIs. Therefore, these interactions are generally discovered and reported only after a drug is administered on a wide scale during post-marketing surveillance. A notable example is the discovery that grapefruit contains bioactive furocoumarins and flavonoids that activate or deactivate many drugs in ways that can be life-threatening (Dahan and Altman, 2004). This effect was first noticed accidentally during a test for drug interactions with alcohol that used grapefruit juice to hide the taste of ethanol.

Currently, information about FDIs is available to medical practitioners from online databases such as DrugBank[1] and compendia such as the Stockley's Drug Interactions (Baxter and Preston, 2010), but these resources have to be regularly updated to keep up with a growing body of evidence from biomedical articles. Recent advances in information extraction are a promising direction to partially automate this work by extracting information about drug interactions. This approach has already shown promising results in the context of drug-drug interactions (Segura-Bedmar et al., 2013) but in the case of FDIs, similar progress is currently hindered by a lack of annotated corpora. The work presented in (Jovanovik et al., 2015) for inferring interactions between drugs and world cuisine is based on a largely manual effort of extracting food-drug interactions from descriptions provided in DrugBank.

Although a first corpus of MEDLINE abstracts about FDIs called POMELO was recently made available (Hamon et al., 2017), this corpus has a low coverage of relevant documents for FDIs. The authors made use of PubMed to retrieve all the articles indexed with the *Food-Drug Interactions* term from the MeSH thesaurus[2], but the challenge is that while articles annotated with *Drug Interactions* are abundant, there is a much smaller number of documents indexed with *Food-Drug Interactions*. A bibliographic analysis of the references cited in the Stockley's Drug Interactions in relation to foods shows that only 11% of these articles are indexed with the MeSH term *Food-Drug Interactions*, while almost 70% of the articles are available in MEDLINE (Bordea et al., 2018).

Constructing a high-coverage corpus of FDIs using MeSH terms and PubMed is not trivial because there is a large number of articles that describe food interactions that were published before the introduction of the *Food-Drug Interactions* MeSH term in the early nineties. At the same time, MeSH terms are assigned to scientific articles based on their main topics of interest, miss-

[1]https://www.drugbank.ca

[2]https://www.nlm.nih.gov/mesh/

Proceedings of the BioNLP 2019 workshop, pages 115–124
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics

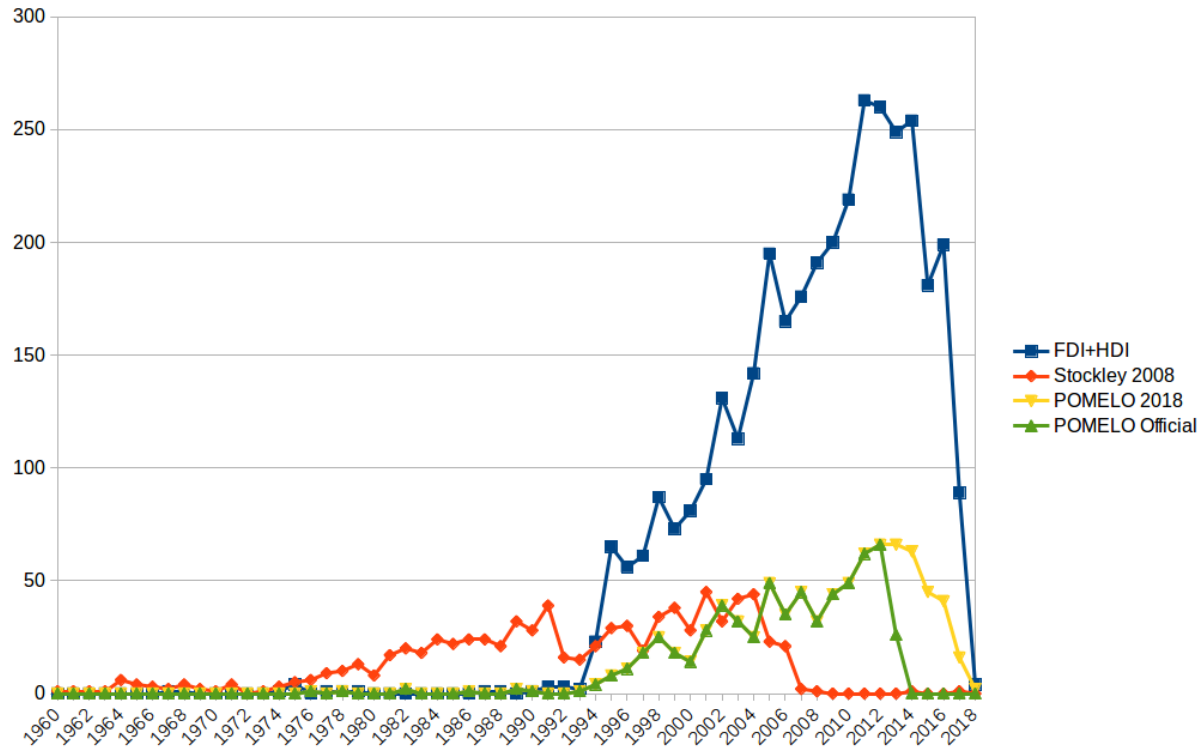


Figure 1: Timeline of MEDLINE articles cited in Stockley 2008 and retrieved using relevant MeSH terms

ing a considerable amount of articles that briefly mention interactions with food. Furthermore, the POMELO corpus has an even more narrow focus on articles related to adverse effects, therefore it covers only 3% of the references provided in the Stockley compendium.

Figure 1 shows a comparison of scientific articles cited in a reference compendium (*Stockley 2008*), with the articles annotated with the *Food-Drug Interactions* MeSH term and the *Herb-Drug Interactions* MeSH term (*FDI+HDI*). It is worth noticing the overall ascending trend of scientific articles that address FDIs, showing an increased interest in this type of interactions. This makes increasingly more costly the effort to manually summarise related information in specialised compendia. The figure also shows the timeline of the articles gathered in the official POMELO corpus (*POMELO Official*) and a more recent retrieval result of the POMELO query (*POMELO 2018*).

We address these limitations by considering several approaches for automatically selecting queries that can be used to retrieve domain-specific documents using an existing search engine. The approach takes as input a sample set of relevant documents that are cited in the Stockley compendium. In this way, the problem of FDI discovery from biomedical literature is limited to the task of interaction candidates search, that is the task of finding documents that describe FDIs from a large bibliographic database. We make use of a large corpus of relevant publications to investigate index terms used to annotate articles about FDIs and we propose an automated method for query selection that increases recall.

The main contributions of this work are:

- a discriminative model for automatically constructing high-coverage and domain-specific corpora for information extraction,
- an approach for automatically selecting queries using index terms as candidates,
- an automated method to evaluate queries based on a sample corpus.

The paper is structured as follows. We begin by discussing several design decisions for the subtask of classifying documents based on relevance, adopting a discriminative model for information retrieval in Section 3. In Section 4, we introduce the subtask of query selection discussing candidate term selection and several methods for scoring queries. Section 5 describes the datasets used

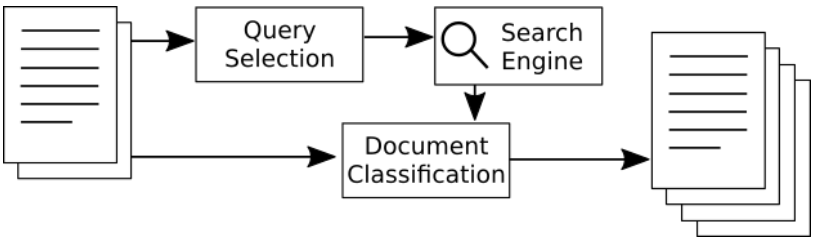


Figure 2: Workflow for automated corpus construction using a collection of sample documents and a search engine

to evaluate our approach for automatically constructing a corpus for FDIs and Section 6 presents the results of an empirical evaluation. Then we provide an overview of related work for this task in Section 7 and we discuss a formal definition for the problem at hand in Section 2. We conclude this work in Section 8.

2 Problem definition

We address the problem of automatically constructing a domain-specific corpus by making use of a discriminative model for information retrieval that defines the problem of document search as a problem of binary classification of relevance (Nallapati, 2004). This allows us to automatically extract queries making use of a sample of relevant documents and then to use an existing search engine as a black box, as can be seen in Figure 2. Sample documents provided as input are used as positives examples to train a binary classifier that can filter retrieved documents based on their relevance.

The problem of query selection for corpora construction is formally defined following the notation introduced in (Bordea et al., 2018) as follows. Given a test collection C of size n where each document c_i is associated with a vector of index terms v_i of a variable size from a set V of size n defined as follows:

$$v_i = \{t_1, ..., t_k\}$$

where t_j is a term from a controlled vocabulary that describes the contents of document c_i, and k is the number of index terms used to annotate the document. We assume that a subset D of size m of relevant documents known to report FDIs is also given, where $m < n$. The subset of index vectors associated with relevant documents is the set V' of size m and each relevant document d_i is annotated with a vector v' of index terms. We also assume that there is a fixed retrieving function S, where $S(q, d)$ gives the score for document d with respect to query q.

We define query selection as the problem of finding a query scoring function R, that gives the score $R(D, q)$ for query q with respect to the collection of relevant documents D. A desired query scoring function would rank higher the queries that perform best when selecting relevant documents.

3 Document classification

In this section, we give an overview of the features and algorithms used to classify scientific articles based on their relevance for the task of FDI discovery, proposing a supervised method to select relevant documents. Classification models are trained using relevant documents as positive examples and irrelevant documents as negative examples.

Preprocessing. Documents are represented as a bag of words that are normalised by replacing numbers by the '#' character. Additionally, other special characters are removed and each word is lowercased.

Word features. Word features are constructed using 1-grams, 1-grams + 2-grams and 1-grams + 2-grams + 3-grams of words. Take for example a document containing the following expression *Food and drug interactions*. The 1-gram features are *food*, *and*, *drug*, *interactions*; 2-grams features are *food and*, *and drug*, *drug interactions*; 3-grams are *food and drug*, *and drug interactions*. In our task, features are constructed from words contained in all documents.

Feature representation. To train classification models, the dataset is transformed into a matrix of size $N \times M$ where N is the number of documents in the dataset and M is the number of features. For each word feature, three types of feature representation approaches are investigated for representing input data:

- **One-hot encoding.** Raw binary occurrence (RBO) matrices. Each document d is represented as a binary feature-document occur-

rence vector $Rbo = [rbo_0, rbo_1, ...rbo_m]$ of size M where rbo_i = 1 if the feature i is in the document d, 0 otherwise.

- **Term frequency.** Count occurrence matrices. Each document d is represented by a vector of counts of term-document occurrences $Tf = [tf_0, tf_1, ...tf_m]$ of size M where tf_i is the number of occurrences of the feature i in the document d.

- **TF-IDF.** Term frequency-inverse document frequency. Each document d is represented by a vector of products of term frequency (TF) and inverse document frequency (IDF).

Index terms features. There is a large number of infrequent index terms that are used to annotate a small number of training documents. To reduce the feature space, we consider as features only index terms that are used to annotate a minimum number of documents. Additionally, we take into account the IDF of each index term in the full collection, that is the number of documents that are annotated with an index term.

Generalised index terms. Index terms are provided from a vocabulary that is hierarchically structured. We exploit this hierarchy to identify terms related to foods and drugs and we introduce three features called *Foods*, *Drugs*, and *Foods and Drugs* that identify documents annotated with one or both types of concepts of interest for our domain. Table 1 gives several examples of nodes from the MeSH hierarchy that are useful for identifying food and drug related concepts.

Classification algorithms. We compare the performance of five classification algorithms with default parameters provided by Scikit-Learn[3]: (1) a decision tree classifier (DTree), (2) a linear SVM classifier (LSVC), (3) a multinomial Naive Bayes classifier (MNB), (4) a logistic regression classifier (LogReg), and (5) a RandomForest classifier (RFC).

4 Query selection

In this section, we discuss the query selection approach presenting first several methods for selecting candidate terms and then proposing different approaches for scoring candidate terms to select the best queries for automatically constructing a domain-specific corpus.

[3] http://scikit-learn.org/stable/

Food concepts	Node	Drug concepts	Node
Plants	B01.650	Pharmacologic actions	D27.505
Food and beverages	J02	Pharmaceutical preparations	D26
Diet, food, and nutrition	G07.203	Heterocyclic compounds	D03
Fungi	B01.300	Polycyclic compounds	D04
Nutrition therapy	E02.642	Inorganic chemicals	D01
Carbohydrates	D09	Organic chemicals	D02
Plant structures	A18	Amino acids, peptides, and proteins	D12

Table 1: Nodes from the MeSH hierarchy used to identify food and drug related index terms

4.1 Candidate terms for query selection

A first step in automatically selecting queries for constructing a domain-specific corpus is to identify candidate terms that are likely to describe and retrieve relevant documents for the given domain. In our experiments, we consider as candidate queries single terms but more complex queries that combine multiple index terms can also be envisaged.

Index terms. Scientific articles are often annotated with high quality index terms from a controlled vocabulary that can be used as queries to retrieve relevant documents. The controlled vocabulary typically provides in addition hierarchical relations between terms that could be further used to identify more general or abstract concepts. One of the limitations of this approach is that index terms summarise the main topics of an article but might miss some of the more fine-grained information.

Document n-grams. All the sequences of words from a document could be considered as candidate terms for query selection but compared to index terms, this approach is more noisy and increases the ambiguity of terms.

Background knowledge. There are several sources of background knowledge that can be considered to identify terms of interest to retrieve documents that describe FDIs. Queries that mention drugs and a food name are likely to retrieve relevant documents for our domain. There are multiple vocabularies and ontologies that partially cover the food domain from different perspectives, but currently the most complete list of foods

can be found by exploiting the DBpedia[4] category structure. DBpedia entities linked to the *Foods* category with the properties *skos:broader* and *dct:subject of* are considered as candidate food terms. Further filtering is required because categories are not necessarily used to identify the type of a DBpedia entity but rather a more loosely defined relatedness relation that often leads to semantic drift when iteratively exploring narrower categories.

Entities are filtered based on their RDF type, based on words but also by excluding categories that are related to foods but are not of interest for FDIs, as can be seen in Table 2. This table is not meant to give an exhaustive list of filters but just a few illustrative examples. We use leaf categories to refer to categories that are taken into consideration as candidate terms but that are not further explored to identify more narrow terms. We identified 15,686 foods from DBpedia and we evaluated the precision of a random sample that is 88%. The recall of this approach was also estimated using a list of 57 foods mentioned in the Stockley 2008 compendium and is 65%.

This is because some of the foods such as *green tea* or *tonic water* can only be found in broader DBpedia categories such as *Food and drink*, *Drinks* or *Diets*, which are more noisy and hence more difficult to filter by hand. The relatively low recall is also due to name variations (e.g., *edible clay* vs. *medicinal clay* in DBpedia), to missing food categories in DBpedia (e.g., *xanthine-containing beverages* and *tyramine-rich foods*), and to errors in the RDF types assigned by DBpedia (e.g., *Brussels sprouts*[5] have the type *Person*).

4.2 Query selection approaches

We consider two types of scoring functions, first based on simple frequency counts of index terms and a second type of scoring functions inspired by existing approaches for feature selection used in supervised classification. The most basic query scoring function is frequency, denoted as the count $c(V', q)$ of query q with respect to the set V' of index vectors associated with relevant documents. The TF-IDF scoring function $tfidf(V', V, q)$ of query q with respect to the set of index vectors associated with relevant documents V' discriminated against the full set of index terms V is defined as:

$$tfidf(V', V, q) = c(V', q)/ln(c(V, q))$$

For the second category of scoring functions, we consider a binary classifier that distinguishes between relevant documents D and an equal number m of randomly selected documents from the test collection C. Assuming that the size of the test collection is much larger than the number of documents known to be relevant, there is a high probability that randomly selected documents are irrelevant. The first scoring function is the information gain that measures the decrease in entropy when the feature is given vs. absent (Forman, 2003) and is defined as follows:

$$InfoGain(Class, t) = H(Class) - H(Class|t)$$

where the entropy H of a class with two possible values (i.e., relevant pos and irrelevant neg) is defined based on their probability p as:

$$H(Class) = -p(pos) * log(p(pos)) - p(neg) * log(p(neg))$$

The gain ratio is further defined as the information gain divided by the entropy of the term t:

$$GainR(Class, t) = InfoGain(Class, t)/H(t)$$

Finally, we also consider the Pearson's correlation as a query scoring function for the same binary classifier.

RDF types	Words	Categories	Leaf categories
Book	bakeries	Alcoholic drink brands	Beer
Building	books	Carnivory	Ducks
Company	campaigns	Cherry blossom	Geese
Location	disease	Decorative fruits and seeds	Onions
Organisation	history	Forages	Quails
Person	people	Halophiles	Rubus
Place	pizzerias		Swans
Restaurant	science		Whisky
Software	vineyards		Wine

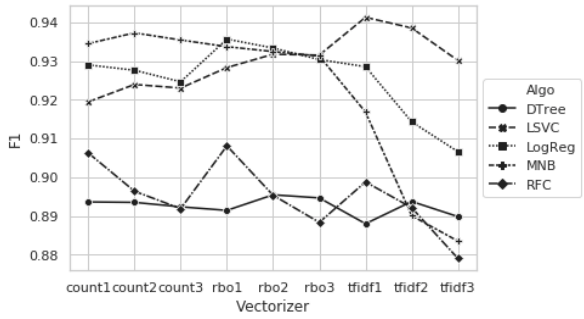
Table 2: Filters used for selecting candidate foods under the DBpedia *Foods* category

5 Experimental setting

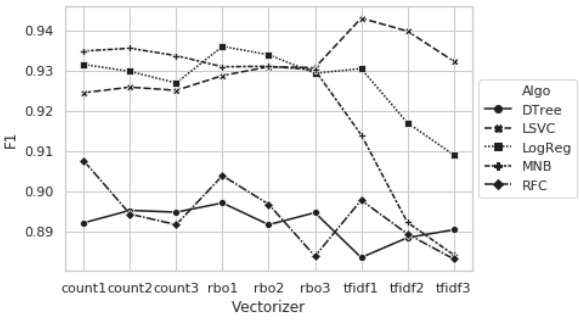
The corpus used in our experiments is manually constructed through a bibliographic analysis of the

[4] https://wiki.dbpedia.org/

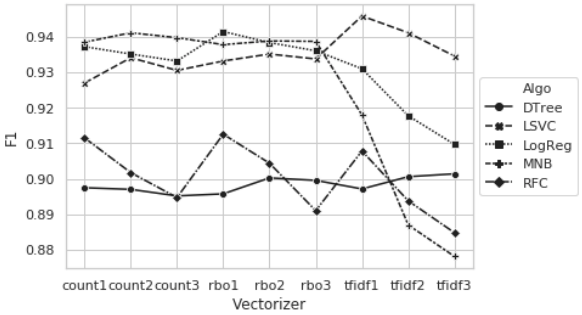
[5] Brussels sprouts: http://dbpedia.org/page/Brussels_sprout



(a) Results on Dataset for Experiment 1



(b) Results on Dataset for Experiment 2



(c) Results on Dataset for Experiment 3

Figure 3: Results of 10-fold cross-validation obtained on each dataset with different classifiers (i.e., decision tree (DTree), linear SVM (LSVC), multinomial Naive Bayes (MNB), logistic regression (LogReg), and RandomForest (RFC)) and vectorizers (i.e., term frequency (count), raw binary occurrence (rbo), and tfidf)

references provided in the Stockley compendium on drug interactions in relation to food. These are considered as positives examples that are used to train a discriminative classifier. The problem of finding negative examples is more challenging because of the problem of unbalanced data and because we aim to train a classifier that is sensitive enough to distinguish between scientific articles that are closely related in topic (i.e., published in the same journals) but that do not describe FDIs.

We manually identify references from pages listed in the index under individual foodstuffs and *Foods*, for a total of 912 references and 460 references, respectively. Using the title and the year of each reference, we retrieve 802 unique PubMed identifiers for references that are available in MEDLINE. In our experiments, we make use of corpora built from MEDLINE abstracts published before 2008 since the version of the Stockley compendium that is available to us was published at this date.

Starting from this collection, several subsets of abstracts are constructed as follows:

(i) references cited in Stockley 2008 (subset *Stockley2008*),

(ii) results of *Food-Drug Interaction* and *Herb-Drug Interaction* MeSH term queries (subset *FDI-HDI*),

(iii) results of the queries *drug and [food name]* where *food name* is one of the 15,686 food names collected from DBpedia as described in Subsection 4.1 (subset *DRUGFOOD*),

(iv) all the MEDLINE abstracts published before 2008 (subset *MEDLINE2008*).

From the first and third subsets, we analyse the list of journals where the articles have been published and all the abstracts published in those journals. In that respect, we have two additional abstract subsets *jrnlAbstracts()* from *Stockley2008* and *jrnlAbstracts()* from *DRUGFOOD* respectively. In our experiments, the set of positive abstracts is the union of Stockley's references with the results of the *FDI-HDI* queries. Table 3 presents the size of the subsets.

The problem of constructing a domain-specific corpus for FDIs is characterised by unbalanced training sets with the non-relevant class representing a large portion of all the examples, while the relevant class has only a small percent of the examples. Dealing with unbalanced class distributions is inherently challenging for discriminative

	Abstracts	Jrnls	jrnlAbstracts()
Stockley2008	895	339	3,344,842
FDI-HDI	3593		
DRUGFOOD	309,327	7421	23,383,538
MEDLINE2008	16,733,485		

Table 3: Overview of different corpora used in our experiments and their size in number of documents

algorithms resulting in trivial classifiers that completely ignore the minority class. We deal with the problem of unbalanced data by under-sampling the majority class such that the training examples in both classes are equal. We define three sets of 4,500 randomly sampled abstracts as negative training examples that successively contain an increasing number of restrictions based on document relevance, publication venue and year of publication:

Experiment 1: abstracts in *jrnlAbstracts()* from *DRUGFOOD* subset that are not cited in Stockley, *FDI-HDI* and *DRUGFOOD* abstracts;

Experiment 2: abstracts in *jrnlAbstracts()* from *DRUGFOOD* subset that are not cited in Stockley, *FDI-HDI* and *jrnlAbstracts()* from *Stockley2008* abstracts;

Experiment 3: MEDLINE abstracts published before 2008 in *jrnlAbstracts()* from *DRUGFOOD* subset which are not cited in Stockley, *FDI-HDI*, *jrnlAbstracts()* from *Stockley2008* and *jrnlAbstracts()* from *DRUGFOOD* abstracts.

6 Results

In this section, we give an overview of the results obtained under different settings. We begin by discussing the results obtained for document classification and we continue with a discussion of the results obtained for the subtask of query selection. In both cases, the classical measures of precision, recall and F-score are used, but in the case of query selection, we adapt these measures to reflect our interest in discovering unseen documents.

6.1 Document classification evaluation

For the purpose of selecting relevant documents regarding food-drug interactions, we evaluate several configurations to construct an efficient classification model. Three sets of experiments are designed around the three training datasets described in the previous section. For each case, we evaluate the models using average of Precision (P), Recall (R) and F1-score (F1) using 10-fold cross-validation. Figure 3 shows the cross-validation results for different word-based features described in Section 3. The best results in terms of F1-score are obtained across all datasets for TF-IDF features with an SVM classifier. TF-IDF of unigram features combined with SVM classifier produce the best F1-score on all datasets. Focusing on these configurations, results are detailed in Table 4 where we can notice that the recall is higher for the third dataset. The best F1-score presents a low standard deviation, which shows that the obtained model is relatively stable. We conclude that results are better on datasets that use a more restrictive filter for selecting the negative examples (Experiment 3). This demonstrates that the random sampling approach for the majority class can benefit from using a more informed strategy than selecting documents from the full collection.

Exp.	Precision	Recall	F1-score + Std
1	0.962	0.921	0.941 ± 0.010
2	0.965	0.922	0.943 ± 0.007
3	0.964	**0.928**	**0.946*** ± 0.004

Table 4: Results of 10-fold cross-validation on the three datasets using an SVM classifier and 1-gram TF-IDF features. The best result is marked with a star

The next set of experiments is focused on evaluating the performance of features based on index terms as can be seen in Table 5. All the index terms that are used to annotate at least 10 documents from our collection are considered as features, ignoring the less frequent index terms. In general, the results are comparable or better than the best results using word features in terms of F1-score. In the case of index terms features, the best results are obtained for the decision tree classifier that outperforms the linear SVM classifier on all three datasets. The same conclusion can be drawn from these experiments in relation to the random sampling approach as the best results are obtained again for the third experiment.

6.2 Query selection evaluation

The challenge for evaluating queries is that it is preferable to rely on the training examples alone for evaluation. But each selected query will retrieve documents that might be relevant but that are not contained in the provided dataset. To address this issue, we use the best performing classification approach described in the previous section to predict the relevance of retrieved documents instead of computing precision based on the docu-

Exp.	Algorithm	Precision	Recall	F1-score
1	DTree	**0.963**	**0.961**	**0.962**
	LSVC	0.947	0.942	0.944
	LogReg	0.960	0.954	0.957
	MNB	0.941	0.941	0.941
	RFC	0.959	0.955	0.957
2	DTree	0.962	0.958	0.960
	LSVC	0.954	0.950	0.952
	LogReg	**0.964**	0.959	**0.961**
	MNB	0.944	0.943	0.943
	RFC	0.963	**0.960**	0.961
3	DTree	**0.967***	**0.965***	**0.966***
	LSVC	0.959	0.956	0.957
	LogReg	0.965	0.961	0.963
	MNB	0.946	0.946	0.946
	RFC	0.963	0.961	0.962

Table 5: Results of 10-fold cross-validation using different classifiers: decision tree (DTree), linear SVM (LSVC), multinomial Naive Bayes (MNB), logistic regression (LogReg), and RandomForest (RFC) with index terms features. The overall best results are marked with a star

ments known to be relevant alone. Our assumption is that the high performance achieved by the classifier allows us to compute a reliable estimate of precision. Although not perfect, this evaluation strategy allows us to avoid the need for further manual annotation or relevant documents. Recall is calculated for a limited number of retrieved documents as some of the MeSH index terms such as *Humans* and *Animals* are broad enough to be used for annotating most of the documents in the test collection.

Word-based query candidates are not further considered at this stage because the best classification performance is achieved for 1-gram features which are deemed to be too ambiguous for our purposes. Table 6 gives an overview of the top 30 1-gram features selected using the SVM classifier. Several names of drugs such as *aminophylline*, *cyclosporine*, and *ephedrine* that are known to have interactions with foods are among the highest ranked features. Foods such as *caffeine*, *coffee*, *cola* and *grapefruit* are also known for their high potential of interactions with drugs. Among these features, names of plants with drug interactions are present including *biloba* and *kava*. Although interesting on their own, we conclude that these features are too generic to be used as queries to extract articles about FDIs without further combining them with other features or index terms.

On the other hand, index term candidates are much more precise, including many terms that refer to food-drug interaction mechanisms such as *Biological Availability* and *Cytochrome P-450 CYP3A*. Also included in this list are chemical compounds such as *Flavanones* and *Furocoumarins* that are contained in certain foods such as *grapefruit* and that interact with many drugs.

absorption	cyclosporine	interaction
alcohol	diet	kava
aminophylline	drug	lithium
anticoagulation	effects	medication
biloba	ephedrine	milk
bioavailability	ergotism	monograph
caffeine	food	nutrition
cheese	grapefruit	oral
coffee	herb	pharmacokinetic
cola	ingestion	phytotherapy

Table 6: Top 30 1-gram features selected using the SVM classifier

Table 7 gives an overview of the results obtained by each scoring function discussed in the previous section. Performance is computed for the top 20 ranked queries for each method. All the methods score high the *Food-Drug interactions* MeSH term but we remove this term from the results because it was used to construct the FDIs corpus. Overall, the best performance is obtained by the Gain ratio scoring function. Selected queries using this approach include: *Biological Availability*, *Drug Interactions*, and *Intestinal Absorption*. Gain ratio outperforms other approaches because it penalizes high frequency terms that are too broad, such as *Adult*, *Aged*, and *Female*.

Scoring function	Predicted P@100	Recall @16k	Predicted F1-score
Frequency	0.2020	0.0032	0.0584
TF-IDF	0.2590	0.0084	0.0784
Info gain	0.2755	0.0084	0.0812
Gain ratio	**0.3755**	**0.0557**	**0.0970**
Correlation	0.2590	0.0081	0.0770

Table 7: Scoring functions evaluated for the top 20 MeSH terms using predicted precision at top 100, recall at top 16k and the combined predicted F1-score

7 Related work

Hand-crafted queries based on MeSH terms are often used for retrieving documents related to adverse drug effects (Gurulingappa et al., 2012), but there is a much smaller number of documents available for specific types of adverse effects such as FDIs and herb-drug interactions. The prob-

lem of building queries for finding documents related to drug interactions has been recently tackled for herb-drug interactions (Lin et al., 2016). This work addresses a less challenging usage scenario where users have in mind a pair of herbs and drugs and are interested in finding evidences of interaction. Queries are manually constructed by a domain expert using MeSH synonyms for herbs and drugs together with the following MeSH qualifiers: *adverse effects*, *pharmacokinetics*, and *chemistry*. Two additional heuristics rank higher retrieved articles that are annotated with the MeSH terms *Drug Interactions* and *Plant Extracts/pharmacology*. Another limitation of this work is the size of the evaluation dataset that is based on a single review paper (Izzo and Ernst, 2009) that provides about 100 references. In contrast, we propose an automated approach for query selection and we make use of a considerably larger dataset of relevant publications for training and evaluation.

The food-drug interaction discovery task proposed here is similar in setting with the subtask on prior art candidates search from the intellectual property domain (Piroi et al., 2011). In the CLEF-IP datasets, topics are constructed using a patent application and the task is to identify previously published patents that potentially invalidate this application. Keyphrase extraction approaches were successfully applied to generate queries from patent applications (Lopez and Romary, 2010; Verma and Varma, 2011). The input is much larger for our task, that is a corpus of scientific articles describing FDIs manually annotated with index terms from the MeSH thesaurus. A main difference between our work and the CLEF-IP task is that we mainly focus on evaluating different methods for query selection by relying on the PubMed search engine. This makes our task more similar to the term extraction task (Aubin and Hamon, 2006), as we aim to identify relevant terms for a broad domain rather than for a specific document, as done in keyphrase extraction.

The dataset used in (Jovanovik et al., 2015) to infer interactions between drugs and world cuisine is based on textual information from DrugBank about food-drug interactions and optimum drug intake time with respect to food. But this information was manually extracted and structured. The most closely related work to ours is (Bordea et al., 2018) where the authors propose an approach for query selection based on index terms. We extend this work by considering multiple types of classification algorithms and by analysing different query candidates beyond index terms.

8 Conclusion and future work

In this paper, we introduced a large dataset of articles that describe food-drug interactions annotated with index terms to investigate an approach for query selection that allows us to discover other food-drug interactions using an existing search engine. We investigated different strategies for addressing the problem of unbalanced data and we showed that a more informed approach that takes into consideration publication venue and year gives better results than a naive approach for random sampling. We proposed an automatic evaluation of retrieved results using a high-performance classifier and we showed that feature selection approaches outperform frequency-based approaches for this task, with an approach based on gain ratio achieving the best results in terms of predicted F1-score.

In our experiments mainly focused on queries constructed using a single index term, therefore a first direction for future work is to investigate more complex queries that combine multiple terms. The number of queries that have to be evaluated would increase considerably especially for combinations with word-based features. Another improvement would be to compare our results with keyphrase extraction approaches instead of analysing all the n-grams and to generate queries using background knowledge about drugs and foods. Finally, the datasets proposed here are based on an older version of the Stockley compendium from 2008. The results presented in this work could be more relevant if a more recent version is considered as this is a highly dynamic field of research.

9 Acknowledgments

This work was supported by the MIAM project and Agence Nationale de la Recherche through the grant ANR-16-CE23-0012 France and by the KaNNa project and the European Commission through grant H2020 MSCA-IF-217 number 800578.

References

Sophie Aubin and Thierry Hamon. 2006. Improving term extraction with terminological resources. In *Advances in Natural Language Processing*, pages 380–387. Springer.

Karen Baxter and CL Preston. 2010. *Stockley's drug interactions*, volume 495. Pharmaceutical Press London.

Georgeta Bordea, Frantz Thiessard, Thierry Hamon, and Fleur Mougin. 2018. Automatic query selection for acquisition and discovery of food-drug interactions. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 115–120. Springer.

Arik Dahan and Hamutal Altman. 2004. Food–drug interaction: grapefruit juice augments drug bioavailability-mechanism, extent and relevance. *European journal of clinical nutrition*, 58(1):1.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. *Journal of machine learning research*, 3(Mar):1289–1305.

Harsha Gurulingappa, Abdul Mateen Rajput, Angus Roberts, Juliane Fluck, Martin Hofmann-Apitius, and Luca Toldo. 2012. Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *Journal of biomedical informatics*, 45(5):885–892.

Thierry Hamon, Vincent Tabanou, Fleur Mougin, Natalia Grabar, and Frantz Thiessard. 2017. Pomelo: Medline corpus with manually annotated food-drug interactions. In *Recent Advances in Natural Language Processing (RANLP)*, pages 73–80.

Angelo A Izzo and Edzard Ernst. 2009. Interactions between herbal medicines and prescribed drugs. *Drugs*, 69(13):1777–1798.

Milos Jovanovik, Aleksandra Bogojeska, Dimitar Trajanov, and Ljupco Kocarev. 2015. Inferring cuisine-drug interactions using the linked data approach. *Scientific reports*, 5:9346.

Kuo Lin, Carol Friedman, and Joseph Finkelstein. 2016. An automated system for retrieving herb-drug interaction related articles from medline. *AMIA Summits on Translational Science Proceedings*, 2016:140–149.

Patrice Lopez and Laurent Romary. 2010. Experiments with citation mining and key-term extraction for prior art search. In *CLEF 2010-Conference on Multilingual and Multimodal Information Access Evaluation*.

Ramesh Nallapati. 2004. Discriminative models for information retrieval. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 64–71. ACM.

Florina Piroi, Mihai Lupu, Allan Hanbury, and Veronika Zenz. 2011. Clef-ip 2011: Retrieval in the intellectual property domain. In *CLEF (notebook papers/labs/workshop)*.

Isabel Segura-Bedmar, Paloma Martínez, and María Herrero Zazo. 2013. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 341–350.

Manisha Verma and Vasudeva Varma. 2011. Exploring keyphrase extraction and ipc classification vectors for prior art search. In *CLEF (Notebook Papers/Labs/Workshop)*.