CoLe and LYS at BioASQ MESINESP8 Task: similarity based descriptor assignment in Spanish

Francisco J. Ribadas-Pena[1], Shuyuan Cao[1], Elmurod Kuriyozov[2]

[1] Grupo COLE, Departamento de Informática, Universidade de Vigo
E.S. Enxeñaría Informática, Campus As Lagoas, Ourense 32004, Spain
ribadas@uvigo.es, shuyuan.cao@uvigo.es
[2] Grupo LYS, Departamento de Computación y Tecnologías de la Información, Universidade de A Coruña
Facultade de Informatica, Campus de Elviña, A Coruña 15071, Spain
e.kuriyozov@udc.es

Abstract. In this paper, we describe our participation in the MESINESP Task of the BioASQ biomedical semantic indexing challenge. The participating system follows an approach based solely on conventional information retrieval tools. We have evaluated various alternatives for extracting index terms from IBECS/LILACS documents in order to be stored in an Apache Lucene index. Those indexed representations are queried using the contents of the article to be annotated and a ranked list of candidate labels is created from the retrieved documents. We also have evaluated a sort of limited Label Powerset approach which creates meta-labels joining pairs of DeCS labels with high co-occurrence scores, and an alternative method based on label profile matching. Results obtained in official runs seem to confirm the suitability of this approach for languages like Spanish.

1 Introduction

This article describes the joint participation of the CoLe group [3] from the University of Vigo and the LYS group [4] from the University of A Coruña in the Spanish biomedical semantic indexing task of the 2020 BioASQ challenge [6]. Participants in this task are asked to automatically classify abstracts written in Spanish from two medical databases, IBECS and LILACS, labeling those documents with descriptors taken from the DeCS (*Descriptores en Ciencias de la Salud*) structured vocabulary.

In our participation we have followed a similarity based strategy, where the final list of DeCS descriptors assigned to a given article is created from the set

[3] *Compiler and Languages* group, https://www.grupocole.org/

[4] *Language in the Information Society* group, https://www.grupolys.org/

of most similar IBECS/LILACS articles stored in a textual index created from the training dataset. This neighbor based strategy was explored in previous participations in BioASQ challenge [2], where we tested the suitability of this similarity based approach and evaluated several strategies to improve the final ranked list of descriptors.

In the case of text categorization for Spanish written documents in MESINESP Task we have employed this similarity based method using several index term extraction approaches in order to evaluate the effects of document representation in the overal quality of the predicted labels. We have also tried improving the categorization performance using a sort of limited Label Powerset multi-label categorization approach, where meta-labels created by joining pairs of labels with high co-occurrence scores replace the original document labels. Additionally a similarity based method using synthetic documents to represent "label profiles" was evaluated and integrated into our official MESINESP8 runs.

The rest of the paper is organized as follows. Section 2 describes the main ideas behind the proposed similarity based approach for MESINESP8 annotations and also describes the text processing being applied. Section 3 briefly details the use of synthetic meta-labels in our Label Powerset approach and how we use the "label profiles" to annotate MESINESP articles. Finally, section 4 discusses our official runs in the BioASQ challenge and details the most relevant conclusions of our participation.

2 Similarity based descriptor selection

Approaches based on k nearest neighbors (k-NN) have been widely used in the context of large scale multi-label categorization, being employed for MEDLINE documents [1] and for labeling purposes in many other domains. The choosing of k-NN based methods is mainly due to its scalability, minimum parameter tuning requirements and, despite its simplicity, its ability to deliver acceptable results in cases where large amounts of examples are available. The approach we have followed in our BioASQ challenge participation [5] is essentially a large multi-label k-NN classifier backed by an Apache Lucene [6] index. In the case of MESINESP annotation with DeCS descriptors, despite being a complex problem, with more than 33,703 labels in DeCS 2019 arranged in a hierarchical structure, the availability of a fairly large training set (> 318K abstracts) labeled by human experts, a priori supposes a favorable scenario for this k-NN labeling.

In this way, our annotation scheme starts by indexing the contents of the MESINESP training articles. For each new article to annotate that index is queried using its contents as query terms. The list of similar articles returned by the indexing engine and their corresponding similarity measures are exploited to determine the following results:

- predicted number of descriptors to be assigned

[5] Source code available at `https://github.com/fribadas/mesinesp8`.
[6] `https://lucene.apache.org/`

– ranked list of predicted DeCS descriptors

The first aspect conforms a regression problem, which aims to predict the number of descriptors to be included in the final list, depending on the number of descriptors assigned to the most similar articles identified by the indexing engine and on their respective similarity scores. The other task is a multi-label classification problem, which aims to predict a descriptors list based on the descriptors manually assigned to the most similar MESINESP articles. In both cases, regression and multi-label classification, similarity scores calculated by the indexing engine are exploited. These scores are computed during the query processing phase. Query terms employed to retrieve the similar articles are extracted from the original article contents and linked using a global OR operator to conform the final query sent to the indexing engine.

In our case, the scores provided by the indexing engine are similarity measures resulting from the engine internal computations and the weighting scheme being employed, which do not have an uniform and predictable upper bound. In order for these similarity scores to behave like a real distance metric, we have applied the following normalization procedure:

1. Articles to be annotated are preprocessed in the same way than the training articles indexed by the Lucene engine.
2. In classification time, all of the relevant index terms from the article being annotated are joined by an OR operator to create the search query.
3. In the ranking of similar articles returned by the indexing engine the top result will be the same article used to query the index, this result is discarded but its score value ($score_{\text{MAX}}$) is recorded for future normalization.
4. For each element on the remaining articles set, the number of descriptors is recorded and it is also recorded the list of assigned descriptors, linking to each of them an estimated distance to the article being annotated, equals to $\left(1 - \frac{score}{score_{\text{MAX}}}\right)$, which will be employed in the weighted voting scheme during k-NN classification.

With this information the number of descriptors to be assigned to the article being annotated is predicted using a weighted average scheme, where the weight of each similar article is the inverse of the square of the estimated distance to the article being annotated, that is, $\frac{1}{\left(1-\frac{score}{score_{\text{MAX}}}\right)^2}$.

To create the ranked list of descriptors a distance weighted voting scheme is employed, associating the same weight values (the inverse of squared estimated distances) to the respective similar articles. Since this is actually a multi-label categorization task, there are as many voting tasks as candidate descriptors were extracted from the articles retrieved by the indexing engine. For each candidate label, positive votes come from similar articles annotated with it and negative votes come from articles not including it.

Table 1. Performance comparison of term extraction approaches.

	k	MiF	MiP	MiR	EBF	EBP	EBR	Acc
stems	10	0.3241	0.3342	0.3145	0.3098	0.3391	0.3130	0.1968
	20	0.3473	0.3586	0.3367	0.3319	0.3617	0.3360	0.2131
	30	0.3517	0.3634	0.3407	0.3356	0.3656	0.3391	0.2155
	40	**0.3569**	**0.3679**	0.3465	**0.3404**	**0.3691**	0.3444	**0.2189**
lemmas	10	0.2635	0.2704	0.2569	0.2485	0.2737	0.2520	0.1532
	20	0.2891	0.2974	0.2812	0.2719	0.2994	0.2729	0.1703
	30	0.2988	0.3081	0.2901	0.2806	0.3090	0.2805	0.1765
	40	0.2964	0.3057	0.2876	0.2777	0.3054	0.2774	0.1745
NPs	10	0.2839	0.2899	0.2781	0.2685	0.2915	0.2733	0.1666
	20	0.3079	0.3154	0.3008	0.2911	0.3150	0.2976	0.1823
	30	0.3121	0.3201	0.3044	0.2948	0.3206	0.2992	0.1852
	40	0.3156	0.3237	0.3080	0.2982	0.3237	0.3024	0.1880
DEPs	10	0.1794	0.1917	0.1687	0.1635	0.1892	0.1576	0.0980
	20	0.1982	0.2119	0.1861	0.1801	0.2078	0.1724	0.1092
	30	0.2069	0.2206	0.1948	0.1876	0.2162	0.1794	0.1140
	40	0.2087	0.2224	0.1966	0.1894	0.2174	0.1812	0.1151
all	10	0.3247	0.3352	0.3148	0.3110	0.3389	0.3179	0.1967
	20	0.3499	0.3612	0.3393	0.3351	0.3645	0.3416	0.2151
	30	0.3536	0.3640	0.3437	0.3386	0.3662	0.3455	0.2172
	40	0.3528	0.3634	0.3428	0.3373	0.3648	0.3439	0.2167
UIMA	-	0.2305	0.1677	**0.3687**	0.2475	0.2063	**0.3821**	0.1487

2.1 Evaluation of article representations

In our preliminary experiments we have tested several approaches to extract the set of index terms to represent MESINESP articles in the indexing process.

Regarding article representation we have evaluated four index term extraction approaches. In these experiments and also in the official MESINESP8 runs we have worked only with the Pre-processed Training set provided by BioASQ organizers. We have discards the dataset of PubMed abstracts translated into Spanish provided by MESINESP organizers due to format issues regarding part of the available translated abstracts. The final training dataset comprised 318,658 records with at least one DeCS code. Index terms which occurred in 5 or less articles were discarded and terms which were present in more than 50 % of training documents were also removed.

Our aim with these experiments was to determine whether linguistic motivated index term extraction could help to improve annotation performance in the k-NN based method we have described. We employed the following methods:

Stemming based representation. This was the simplest approach which employs stop-word removal, using a standard stop-word list for Spanish, and the default Spanish stemmer from the Snowball project[7].

[7] http://snowball.tartarus.org

Morphosyntactic based representation. In order to deal with the effects of morphosyntactic variation in Spanish we have employed a lemmatizer to identify lexical roots instead of using word stems and we also replaced stop-word removal with a content-word selection procedure based on part-of-speech (PoS) tags.

We have delegated the linguistic processing tasks to the tools provided by the spaCy Natural Language Processing (NLP) toolkit [8]. This toolkit offers a set of state-of-the-art components written in the Python programming language, together with a collection of pretrained models, ready to be used in typical natural language processing tasks like dependency parsing, named entity recognition, PoS tagging and morphological analysis.

In our case we have employed the PoS tagging and lemmatization information provided by spaCy to tokenize and assign PoS tags to the MESINESP abstract contents. We employed the standard Spanish models available on spaCy without using any specific data for biomedical related contents.

In order to filter the content-words from the processed MESINESP abstracts, we have applied a simple selection criteria based on the employment of the PoS that are considered to carry the sentence meaning. Only tokens tagged as a noun, verb, adjective, adverb or as unknown words are taken into account to constitute the final article representation.

After PoS filtering, the lemmas (canonical forms of words) corresponding to surviving tokens are employed to normalize the considered word forms in a slightly more consistent way than simple stemming.

Nominal phrases based representation. In order to evaluate the contribution of more powerful NLP techniques, we have employed a surface parsing approach to identify syntactic motivated nominal phrases from which meaningful multi-word index terms could be extracted.

Noun Phrase (NP) chunks identified by spaCy are selected and the lemmas of the constituent tokens are joined together to create a multi-word index term. In the current version of the system no other syntactical units of interests like prepositional phrases or verbal phrases are considered, since nominal phrases use to carry most of the text semantic content.

Dependencies based representation. We have also employed as index terms triples of dependence-head-modifier extracted by the dependency parser provided by spaCy. A dependency parser analyzes the grammatical structure of a sentence, establishing relationships between head words and words which modify those heads. In our case spaCy provides a dependency parsing model for Spanish that identify syntactic dependency labels following the Universal Dependencies(UD) [5] scheme.

Dependence relationships encode information that provides an approximation to high level semantic relationships, giving information regarding the agent of an action (with a *nsubj* relationship between the main verb and the root of the nominal phrase acting as subject) or the object of that action (by means of a `obj` relationship), among others. In our system, the complex

[8] Available at `https://spacy.io/`

index terms were extracted from the following UD relationships [9]: *acl*, *advcl*, *advmod*, *amod*, *ccomp*, *compound*, *conj*, *csuj*, *dep*, *flat*, *iobj*, *nmod* , *nsubj*, *obj*, *xcomp*, *dobj* and *pobj*.

UIMA Concept Mapper representation. In addition to those representations, we also have employed the Concept Mapper [10] module from the UIMA (*Unstructured Information Management Architecture*) framework. This component employs a dictionary with all of the DeCS labels and their corresponding synonyms and searches for exact matches of those DeCS labels into the abstract text. In our case we have added to the document representation as index term each one of those matches in order to maintain its absolute occurrence frequency.

In order to illustrate the index term extraction procedure, figure 1 shows an example of a MESINESP record with the set of representations extracted from the textual contents of its abstract.

Table 1 summarizes the results obtained in our preliminary tests regarding document representation, using as measures MicroPrecision, MicroRecall, MicroF, Example based Precison, Recall and F measure, and Accuracy. The test dataset was created after processing 750 manually indexed records from the *Core-descriptors development set* provided by BioASQ organizers. We have evaluated the performance of the described index term generation methods (`stems`, `lemmas`, `nps`, `deps` and using all of them together) for increasing values of k, the number of similar articles to be used (1) in the estimation of the number descriptors to be assigned and (2) in the voting procedure that will construct the final list of descriptors to attach to a given article.

As can be seen in table 1, the best results were obtained with fairly high values for k (≥ 30). Regarding the index term representations, the runs which employed index term extracted by means of stemming (both stemming alone and stems mixed with the other index terms) provided the best performance. The representations using complex index terms extracted from noun phrase chunks and dependencies triples offered poor performance, maybe because of very infrequent index terms that can have the undesired effect of boosting internal scores in schemes where inverse document frequencies are taken into account.

3 Exploiting DeCS labels

In this section we describe two approaches that try to improve labeling performance taking advantage of the information inherent to DeCS labels. Even DeCS is a fairly large concept hierarchy we have tested the suitability of extending the label space using an approach inspired by the Label Powerset(LP) method employed in multi-label categorization. In this case we create a set of "meta-labels" that replace pairs of DeCS labels which tend to appear together in

[9] Detailed list of UD relationships available at `https://universaldependencies.org/u/dep/`

[10] http://uima.apache.org/d/uima-addons-current/ConceptMapper/RELEASE_NOTES.html

ID	biblio-1000005
DB	LILACS
JOURNAL	Oncol. (Guayaquil)
TITLE	Manejo de Tumores de Mediastino, Serie de Casos
ABSTRACT	Introducción: A pesar del difícil acceso anatómico para los tumores de mediastino, la resección quirúrgica sigue siendo el mejor enfoque diagnóstico y terapéutico. En la presente serie de casos presentamos la experiencia de un centro oncológico en el abordaje de tumores del mediastino y sus resultados. Métodos: En el departamento de Jefatura de Cirugía Oncológica del Instituto Oncológico nacional de Solca-Guayaquil, durante los meses de Enero del 2013 a Enero 2017 se realizó un estudio descriptivo, retrospectivo. Se analizaron todos los casos de pacientes derivado ...
DECS CODES	9562,8650,21034,24375,21044,20174,14341,238,9062,21030,23039
STEMS	manej,tumor,mediastin,seri,cas,introduccion,dificil,acces,anatom,tumor,mediastin, reseccion,quirurg,enfoqu,diagnost,terapeut,objet,present,seri,cas,present,experient, centr,oncolog,abordaj,tumor,mediastin,result,metod,departament,jefatur,cirug,oncolog, ...
LEMMAS	tumores,mediastino,serie,casos,difícil,anatómico,quirúrgica,mejor,diagnóstico,terapéutico, presente,presentamos,oncológico,jefatura,cirugía,oncológica,instituto,oncológico,nacional, solca,guayaquil,enero,enero,realizó,descriptivo,retrospectivo,analizaron,derivados,inicial, ...
NPs	manejo de tumores,mediastino,introducción,acceso,los tumores,mediastino,la resección, el mejor enfoque,el objetivo,la presente serie,la experiencia,un centro,el abordaje,tumores, mediastino,sus resultados,métodos,el departamento,jefatura de cirugía oncológica, instituto oncológico nacional de solca,guayaquil,los meses,enero 2017,un estudio,los casos, ...
DEPs	nmod(tumores,mediastino),conj(diagnóstico,terapéutico),flat(jefatura,cirugía), flat(cirugía,oncológica),flat(instituto,oncológico),flat(instituto,nacional), flat(instituto,solca),flat(instituto,guayaquil),advcl(analizar,previo),obj(previo,marcador), amod(marcador,tumoral),flat(tomografía,tórax),conj(analizar,realizar),nsubj(estudiar,variable), ...
UIMA	14294,14294,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562,9562, 9562,9562,9562,9562,8650,8650,8650,8650,8650,8650,8650,8650,8650,8650,8650,8650,8741,8741, 16771,28219,28219,1731,1731,1731,1731,24373,28632,28632,28599,28599,23274,23274,23484,23484, ...

Fig. 1. Example of the index term extraction methods.

training dataset. The other aspect regarding DeCS labels that we have explored is to exploit the idea of "label profiles". These profiles represent the concepts behind each DeCS label by means of a synthetic document that aggregates the contents of all of the abstracts annotated with a given label.

3.1 Limited Label Powerset

Label Powerset(LP) [3] [4] is a problem transformation approach to multi-label classification that seeks to convert a multi-label classification problem into a multi-class classification problem. The LP transformation creates one multi-class classifier trained on all unique label combinations found in the training data. This approach is unfeasible in the case of DeCS labeling due the large amount of different labels in the hierarchy: 33,703 labels in DeCS 2019, of which 23,197 are actually present in the training dataset.

Our approach limits classical LP multi-label categorization to the cases where only combinations of highly correlated pairs of labels are taken into account. To select the pairs of labels to be joined we have computed the Normalized Pointwise Mutual Information (NPMI) between each pair of DeCS labels, l_i and l_j, across

Table 2. Most frequent codes in original dataset and when "meta-labels" are applied.

Original MESINESP Dataset (number of codes: 23199)

rank	freq	code	label
1	225535	21034	"seres humanos" (*"humans"*)
2	120433	21030	"femenino" (*"female"*)
3	105182	21044	"masculino" (*"male"*)
4	56161	331	"adulto" (*"adult"*)
5	41571	9062	"persona de mediana edad" (*"middle aged"*)
6	33115	29315	"adolescente" (*"adolescent"*)
7	27275	2694	"niño" (*"child"*)
8	15795	20174	"anciano" (*"aged"*)
9	14514	28612	"factores de riesgo" (*"risk factors"*)
10	14325	2715	"niño preescolar" (*"preschool child"*)
11	12408	7399	"lactante" (*"infant"*)
12	11121	22226	"recién nacido" (*"newborn"*)
13	11038	28611	"estudios retrospectivos" (*"retrospective studies"*)
14	10654	28596	"estudios transversales" (*"cross-sectional studies"*)
15	10567	841	"animales" (*"animals"*)

Meta-labels with NPMI threshold at 0.25 (number of codes: 50771)

rank	freq	code	label
1	116576	21030.21034	"femenino"∧"seres humanos" (*"female"∧"humans"*)
2	102586	21034.21044	"seres humanos"∧"masculino" (*"humans"∧"male"*)
3	89717	21034	"seres humanos" (*"humans"*)
4	84759	21030.21044	"femenino"∧"masculino" (*"female"∧"male"*)
5	44104	331.21030	"adulto"∧"femenino" (*"adult"∧"female"*)
6	37698	331.21044	"adulto"∧"masculino" (*"adult"∧"male"*)
7	34076	9062.21030	"persona de mediana edad"∧"femenino" (*"middle aged"∧"female"*)
8	31766	9062.21044	"persona de mediana edad"∧"masculino" (*"middle aged"∧"male"*)
9	27602	331.9062	"adulto"∧"persona de mediana edad" (*"adult"∧"middle aged"*)
10	25381	21030.29315	"femenino"∧"adolescente" (*"female"∧"adolescent"*)
11	22792	21044.29315	"masculino"∧"adolescente" (*"male"∧"adolescent"*)
12	16576	331.29315	"adulto"∧"adolescente" (*"adult"∧"adolescent"*)
13	11438	28612	"factores de riesgo" (*"risk factors"*)
14	11344	9062.29315	"persona de mediana edad"∧"adolesente" (*"middle aged"∧"adolescent"*)
15	11044	2694.29315	"niño"∧"adolescente" (*"child"∧"adolescent"*)

the training dataset employing the following formula:

$$NPMI(l_i, l_j) = \frac{PMI(l_i, l_j)}{-log(P(l_i, l_j))}$$

Where PMI is the Pointwise Mutual Information computed by:

$$PMI(l_j, l_j) = log(\frac{P(l_i, l_j)}{P(i_i) \cdot P(l_j)})$$

And where $P(l_i, l_i)$ is computed as $\frac{|\text{docs. labeled with } l_i \text{ and } l_j|}{|\text{docs. in training collection}|}$ and $P(l)$ is computed as $\frac{|\text{docs. labeled with } l|}{|\text{docs. in training collection}|}$.

The measure $NPMI(l_i, l_j)$ normalizes the values of PMI in $[-1, 1]$, resulting in -1 for a pair of labels never occurring together, 0 for independence, and +1 for complete co-occurrence of labels l_i and l_j.

In our experiments we have evaluated three thresholds (0.25, 0.50 and 0.75) to create new "meta-labels" joining pairs of labels whose NPMI scores are over them. Table 2 compares the most frequent codes in the original dataset and when "meta-labels" with NPMI scores above 0.25 replace the original codes.

Once these "meta-labels" are identified we create new training documents replacing in the set of DeCS labels associated to each record the two original labels with the corresponding new "meta-label". To annotate the test articles we apply the k-NN procedure described in previous sections over the new training documents where "meta-labels" were placed.

3.2 Label profiles

Another approach that we have tested in order to capture the semantics of the DeCS labels is the use of "label profiles" that try to represent the contents associated to each DeCS label and match incoming test documents to those profiles.

To create those DeCS label profiles we have followed a very simple approach that is easily integrated into our Lucene backed k-NN multi-label categorization scheme.

1. For each DeCS label we collect the index terms extracted from the abstracts of documents annotated with that label.
2. With those lists of index terms we create a synthetic Lucene document concatenating the terms to create a big document that holds the representation of the "label profile" for the corresponding label.
3. All those synthetic documents representing "label profiles" for every DeCS label are indexed into a Lucene index.

To annotate an incoming article abstract text is processed as described in precedent section to extract its index terms. With those index terms the Lucene index of "label profiles" is queried and the top most similar synthetic documents are recorded to annotate that article with their corresponding labels. The idea behind this approach is to improve the main k-NN annotation procedure, which follows a content-based method, with a complementary method focused on the labels and its semantic aspects.

Table 3. Official results for BioASQ MESINESP8 Task.

system	rank	MiF	EBP	EBR	EBF	MaP	MaR	MaF	MiP	MiR	Acc.
best	1/25	0.4254	0.4382	0.4343	0.4240	0.3989	0.3380	0.3194	0.4374	0.4140	0.2786
iria-mix	8/25	0.3892	0.5375	0.3207	0.3906	0.5539	0.2263	0.2318	0.5353	0.3057	0.2530
iria-1	10/25	0.3630	0.5055	0.2980	0.3643	0.5257	0.1908	0.1957	0.5024	0.2842	0.2326
iria-3	11/25	0.3460	0.5432	0.2674	0.3467	0.5789	0.1617	0.1690	0.5375	0.2551	0.2193
iria-2	12/24	0.3423	0.4699	0.2837	0.3408	0.4996	0.1715	0.1719	0.4590	0.2729	0.2145
iria-4	14/25	0.2743	0.3070	0.2635	0.2760	0.2655	0.2925	0.2619	0.3068	0.2481	0.1662
BioASQ_Baseline	15/25	0.2695	0.2681	0.3239	0.2754	0.3733	0.3220	0.2816	0.2337	0.3182	0.1659

4 Official MESINEPS8 runs and discussion

Although we have tested several alternatives to try to improve the results obtained by the Lucene based k-NN method, only the most simple ones have been submitted to the official batches of BioASQ challenge.

In table 3 the official performance measures obtained by our runs in the MESINESP8 Task are shown. The official runs sent during our participation were created using the following configurations.

iria1. This run created the representation of MESINESP articles using all of the index term extraction methods described in section 2.1. During indexing and querying, terms appearing in 5 or less abstracts and terms used in more than 50% of total documents were discarded. The number of neighbors used by the k-NN classifier is 30 and the predicted number of descriptors to be returned was increased a 10% in order to ensure slightly better values in recall related measures.

iria2. For this run the same setup as `iria1` was employed, but instead of using the original train dataset this runs employed the limited Label Powerset approach and indexed a new training dataset annotated with "metalabels" created by joining pairs of DeCS labels with a NPMI scores above 0.25.

iria3. This run was simply the intersection of the labels predicted by `iria1` and `iria2`.

iria4. This run created a set of "label profiles" over the train dataset employed in `iria2`, that is, documents annotated with "metalabels" created by joining pairs of labels with NPMI scores over 0.25. In this case the number of labels to predict was fixed to 10 and the number of neighbors used by the k-NN classifier was 15.

iria-mix. This run was based on run `iria1`, adding the predictions of `iria4` and the exact matches provided by UIMA Concept Mapper.
Labels predicted by `iria4` but discarded by `iria1` were added to the final list of candidate labels. The same procedure was applied to add the exact matches identified by means of Concept Mapper but not predicted by neither of `iria1` and `iria4`.

The results of our participation in the MESINESP8 task of the BioASQ biomedical semantic indexing challenge were not far from the results of the most competitive teams, showing that similarity based methods can still be considered for large scale indexing tasks. As positive aspects of our participation, we have confirmed that k-NN methods backed by conventional textual indexers like Lucene are a viable alternative for this kind of large scale problems, with minimal computational requirements and fairly good results in the case of Spanish biomedical abstracts. We have also conducted a comprehensive evaluation of the performance of several alternatives to index term extraction, ranging from simple ones, based on stemming rules, to more complex ones were natural language processing was required.

The future lines of work are related with the improvement of natural language processing. In this participation we have employed general domain NLP models. Biomedical documents have many specific characteristics that suggest that custom NLP models trained with text from this domain will help to improve the performance of our classifier. Likewise, the use of ”meta-tags” in this work opens a future line of research on the exploitation of the semantics inherent to the co-occurrence of tags.

Acknowledgements

F.J. Ribadas-Pena and S. Cao have been supported by the Spanish Ministry of Economy, Industry and Competitiveness (MINECO) through the ANSWER-ASAP project (TIN2017-85160-C2-2-R), and by the Galician Regional Government (Xunta de Galicia) under project ED431D 2017/12.
E. Kuriyozov received funding from the ANSWER-ASAP project (TIN2017-85160-C2-1-R) from MINECO, and from Xunta de Galicia (ED431B 2017/01, ED431G 2019/01). He is also funded for his PhD by El-Yurt-Umidi Foundation under the Cabinet of Ministers of the Republic of Uzbekistan.

References

1. D Trieschnigg, P Pezik, V Lee, F De Jong, W Kraaij, D Rebholz-Schuhmann: MeSH Up: effective MeSH text classification for improved document retrieval. Bioinformatics 25 (11), 1412-1418, 2009.
2. Francisco J. Ribadas-Pena, Luis M. de Campos, Victor M: Darriba-Bilbao, Alfonso E. Romero. CoLe and UTAI at BioASQ 2015: Experiments with Similarity Based Descriptor Assignment. Working Notes of CLEF 2015 - Conference and Labs of the Evaluation forum. Toulouse, France, September 8-11, 2015. CEUR Workshop Proceedings, vol. 1391.
3. Tsoumakas, G., Katakis, I: Multi-label classification: An overview. International Journal of Data Warehousing and Mining 3, pp. 1–13 (2007)
4. Boutell, M., Luo, J., Shen, X., Brown, C: Learning multi-label scene classification. Pattern Recognition, Vol. 37(9), pp. 1757–1771 ()2004)
5. Nivre, J., M.-C. de Marneffe, F. Ginter, Y. Goldberg, J. Haji˘c, C. Manning, R. McDonald, S. Petrov, S. Pyysalo, N. Silveira, R. Tsarfaty, y D. Zeman: 2016. Universal dependencies v1: A multilingual treebank collection.LREC.
6. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M.R., Weissenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Arti´eres, T., Ngomo, A.C.N., Heino, N., Gaussier, E., Barrio-Alvers, L., Schroeder, M., Androutsopoulos, I., Paliouras, G.: An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics 16, 138 (Apr 2015)