

# USI at BioASQ 2015: a Semantic Similarity-Based Approach for Semantic Indexing

Nicolas Fiorini<sup>1,\*</sup>, Sylvie Ranwez<sup>1</sup>, Sébastien Harispe<sup>1</sup>, Jacky Montmain<sup>1</sup>,  
Vincent Ranwez<sup>2</sup>

<sup>1</sup> LGI2P research center from the École des mines d'Alès, Site de Nîmes, Parc scientifique G. Besse, 30035 Nîmes cedex 1, France

`{name.surname}@mines-ales.fr`

<sup>2</sup> UMR AGAP, Montpellier SupAgro/CIRAD/INRA, 2 place Pierre Viala, Montpellier cedex 1 34060, France

`ranwez@supagro.inra.fr`

**Abstract.** The need of indexing biomedical papers with the MeSH is incessantly growing and automated approaches are constantly evolving. Since 2013, the BioASQ challenge has been promoting those evolutions by proposing datasets and evaluation metrics. In this paper, we present our system, USI, and how we adapted it to participate to this challenge this year. USI is a generic approach, which means it does not directly take into account the content of the document to annotate. The results lead us to the conclusion that methods that solely rely on semantic annotations available in the corpus can already perform well compared to NLP-based approaches as our results always figure in the top ones.

**Keywords:** MeSH indexing, semantic similarity, kNN approach

## 1 Introduction

The task 3a of the BioASQ challenge consists in indexing new biomedical papers with MeSH concepts. The need of document indexed by terms of a thesaurus like the MeSH has already been emphasized several times in Information Retrieval [1, 2]. This task has been historically done by experts, later helped by automated methods [3]. Nowadays, it gets more and more importance with the increasing number of papers to annotate. Indeed, PubMed library has been adding about one million papers yearly since 2008<sup>3</sup>.

BioASQ aims at evaluating the indexing systems according to two criteria: effectiveness and efficiency. The first one is assessed by using common metrics in text classification, the F-measure and the LCA-F [4]. Although the speed of the systems is not directly measured, when a test set is released, the participants have a short amount of time — 21 hours — to send their results.

---

\* Corresponding author

<sup>3</sup> <http://www.nlm.nih.gov/bsd/licensee/baselinestats.html>

Blocking points quickly appear when one wants to properly accomplish this task. For each paper to annotate, the system has to pick  $n$  concepts — where  $n$  varies — among a set of more than 27,000 concepts. The definition of  $n$  as well as the concepts to pick both constitute the first main problem. The second problem is related to the knowledge base. The MeSH is a structured thesaurus, that is, it provides unique concepts that are organized as a graph. Therefore, a simple mapping of the terms contained in the title or abstract of an article is not sufficient to accurately annotate this article.

The literature shows that several investigations have been made to tackle those challenges. Some rely on concept extraction based on the text. This means that the system does some Natural Language Processing (NLP) and aims at finding lemmas in the text that can be mapped on the thesaurus [5, 6]. Those methods can be enriched with other processes such as Machine Learning (ML). It has been extensively used during the last years and many approaches have been tested: learning-to-rank [7], gradient boosting [8], or reflective random indexing [9]. Some authors also stated that ML methods can be more or less appropriate depending on the problem and proposed a meta-learning solution that learns the ML approach to apply for a given problem [10]. So far, hybrid methods — those based on both NLP and ML — produced the best result [11].

We previously presented USI (User-oriented Semantic Indexer), a method that does not do any NLP nor ML [12]. This paper briefly introduces USI and shows how we adapted and tested it for this challenge. We also investigate on how generic methods may or may not contribute to the future of semantic indexing.

## 2 Method

### 2.1 USI Generic Approach

USI is a generic method that aims at indexing entities of any type — be it text files, audio or video samples, genes, etc. — as long as there is an existing collection of entities that are already annotated. It is based on a  $k$  Nearest Neighbors (kNN) approach and it proceeds in two steps:

- identifying the neighbors, i.e documents in the corpus that are close to the one to annotate,
- using this neighborhood for annotating this document.

This kind of approach has already been used in the literature [7], [11], [13] as part of a bigger process mixing NLP and ML. The aim of USI is to get rid of the content processing so that it can be applied on any document. Besides, several studies highlighted the fact that the full text is rarely accessible for biomedical papers for instance [7]. There are several downsides of such a generic approach. First, the quality of annotations is not guaranteed as less properties are used to annotate the document than with classical systems. Second, the approach still

has to be effective enough not to have an exploding computation time. Finally, it requires to be able to find neighboring documents in the corpus.

USI tries to deal with the first problem by using semantic similarities. In fact, when the neighborhood has been set, USI selects the set of concepts in the annotations of the neighbors that is the most semantically similar with every other neighbor. Formally, for a given document to annotate, let us define its neighbors as  $K$  with  $k = |K|$ .  $K$  contains documents  $d$  that are respectively annotated by a set of concepts  $A_d$  each. The list of all those annotations for the neighborhood forms a family of set  $\mathfrak{A}_K$ . For example, consider two neighbors  $d_1, d_2$  respectively annotated by  $A_1 = \{c_a, c_b\}, A_2 = \{c_a, c_c\}$  where  $c_a, c_b, c_c$  are concepts — here, MeSH headings; then  $\mathfrak{A}_k = \{\{c_a, c_b\}, \{c_a, c_c\}\}$ . In order not to look for the solution in the whole thesaurus, we define a search space  $A_0$  such that

$$A_0 = \bigcup_{A_d \in \mathfrak{A}_k} A_d. \quad (1)$$

An optimal solution  $A^* \subseteq A_0$  has to be found. To do so, USI follows an objective function defined as:

$$A^* = \operatorname{argmax}_{A \subseteq A_0} \{f(A)\}, f(A) = \frac{1}{k} \sum_{A_d \in \mathfrak{A}_K} \operatorname{sim}(A, A_d) - \mu|A|, \quad (2)$$

with  $\mu \in [0; 1]$ .  $\operatorname{sim}(A, A_d)$  represents the semantic similarity of  $A$  with  $A_d$ . This semantic similarity can be any one of the numerous semantic similarity measures of the litterature. Here, we need to assess the similarity of two sets of concepts so a groupwise measure seems appropriate. However, the calculation of direct groupwise measures — such as the Dice or Jaccard indices — is time-consuming. We thus chose to use an indirect groupwise measure called Best Match Average (BMA) [14], a composite average based on pairwise similarity values that can be efficiently updated when removing a single concept from the testing solution  $A$ . With  $\operatorname{sim}_p(a, b)$  being a pairwise semantic similarity measure for two concepts  $a, b$ , BMA is defined as follows:

$$\operatorname{sim}(A, B) = \frac{1}{2|A|} \sum_{c \in A} \operatorname{sim}_m(c, B) + \frac{1}{2|B|} \sum_{c \in B} \operatorname{sim}_m(c, A), \quad (3)$$

where  $\operatorname{sim}_m(c, A) = \max_{c' \in A} (\operatorname{sim}_p(c, c'))$ . This means finding for each concept of  $A$  the most similar one in  $B$  and vice versa. Hence BMA needs the selection of a pairwise measure and an associated Information Content (IC) measure. The choice of these metrics obviously impacts USI performance and is further detailed in §2.2.  $\mu$  is a concision constraint and represents the decrease of the similarity with the neighbors that is allowed for the removal of one concept of  $A$ . For example, say  $\mu = 0.02$ . Let  $A_1, A_2$  be sets of concepts such that  $A_1 = \{c_a, c_b, c_c\}, A_2 = \{c_a, c_b\}$ . Assume that the similarity with the neighbors — the first part of  $f(A)$  in equation 3 — is 0.70 for  $A_1$  and 0.69 for  $A_2$ . Then,  $f(A_1) = 0.64$  and  $f(A_2) = 0.65$ . Therefore in that case  $A_2$ , a more concise annotation, is favored. For the challenge the value of  $\mu$  has been empirically optimized

on the training sets ( $\mu = 0.01$ ).

The second problem encountered when designing a non-content-based method — i.e., keeping the computation time low — is even bigger as we need to find a subset of  $A_0$  maximizing an objective function. The computation of semantic similarities is also time-consuming for any non-trivial measure. We thus implemented a heuristic algorithm that starts from  $A_0$  and removes concepts until the value of  $f(A)$  stops increasing. The concept that is removed at each iteration is the one that leads to the best increase of  $f(A)$ . This algorithm computes a solution in polynomial time and we optimized it to reduce its time complexity [12].

Finally, there is the definition of neighbor documents. This task is actually an information retrieval task. For a given document (or its title, abstract, etc. for an article), finding the ones in the corpus that are the most similar. There is one main tool for this purpose in PubMed called PMRA. It is even usable through a web service [15], so it perfectly meets the needs of USI for this challenge.

## 2.2 USI Adaptations for BioASQ

The BioASQ challenge is a fantastic opportunity for us to test USI with several variations in a real case context. Indeed, USI has parameters for which we have no or poor idea of the value they should take. They are:

- the size of the neighborhood,
- the semantic similarity measure and an associated IC metric,
- the impact of integrating baseline results.

We thus propose several systems presented in table 1. Although we already studied the impact of the size of the neighborhood, we wanted to confirm those results on a larger scale through this challenge. We previously concluded that considering a too large neighborhood ( $> 20$ ) may significantly negatively impact the results. The reason is that when  $k$  increases, it introduces noise in  $A_0$  that is difficult to get rid of. In our previous tests, best results were obtained with  $k = 10$ . We reproduced this experiment by proposing two systems implementing USI as described in §2.1, with  $k = 10$  and  $k = 20$ . The choice of the semantic similarity measure for these systems was Lin [16] using Seco’s IC [17] as those were the default values in USI and quite neutral metrics.

**Table 1.** Description of the systems submitted to BioASQ 2015.

System name	Description
USI 10 neighbors	Default version of USI where 10 neighbors are selected
USI 20 neighbors	Default version of USI where 20 neighbors are selected
USI abstract	“USI 10 neighbors” where semantic similarity is chosen using an abstract framework
USI baseline	“USI abstract” integrating the provided baselines

Recent work introduced an abstract framework that encompasses many semantic similarities [18] and provides a Java library (SML) to use it [19]. We used

this library for testing the broad spectrum of semantic similarities deriving from the ratio model below.

$$sim_{RM}(a, b) = \frac{f(A \cap B)}{\alpha f(A \setminus B) + \beta f(B \setminus A) + f(A \cap B)} \quad (4)$$

$a, b$  are concepts,  $A, B$  are their respective sets of features and  $f$  a function defined on the sets of features. Many well-known similarity measures are special cases of this ratio model. For instance for the Lin measure,  $\alpha = 0.5$ ,  $\beta = 0.5$  and  $f$  is an *IC* function. We tested many values of  $\alpha$  and  $\beta$  along with five IC metrics: Resnik [20], Sanchez [21], Sanchez adapted<sup>4</sup>, Seco [17] and Zhou [22]. We selected the triplet  $\alpha, \beta$ , and IC for USI that led to the best results on the training sets. The resulting approach is called **USI abstract**.

Finally, we created a system that takes the baselines into account. We noticed with the training set that our method is pretty bad at predicting the check tags. Check tags are over-represented headings, which is a specificity of the MeSH<sup>5</sup>. USI does not consider this kind of specific aspect when annotating because of its genericity and the lack of NLP. Besides, we thought that we could increase the quality of the results by (i) enriching  $A_0$  with concepts from the baseline and (ii) compare the results of the baseline and USI to select the most accurate concepts. MeSH Now [11] won the challenge last year and is now a baseline for this challenge. As a result, the main objective of participating to this challenge is to be better than baseline systems, that is **BioASQ baseline**, **Default MTI**, **MTIFL** and **MeSH Now BF** — knowing that **MeSH Now BF** should be the most effective. The authors of **MeSH Now BF** even provide their results on a dataset of 5,000 document called BioASQ5000<sup>6</sup>. Those results can be used to simulate a participation to the challenge as a baseline and check whether or not our results are better than **MeSH Now BF** ones. Several modifications of the abstract framework system have thus been implemented for creating the **USI baseline** system. First of all, the check tags of **MeSH Now BF** results — an ordered list later referred to as  $A_{baseline}$  — are automatically added to the output. Second of all, for each document,  $A_0$  is enriched with  $A_{baseline}$  before being processed. Finally, the concepts of  $A^*$  are submitted to a post-process that keeps them if:

- they are present in both  $A^*$  and  $A_{baseline}$ ,
- or they are present in  $A^*$  only but removing it from  $A$  would decrease the objective function score by more than  $\theta_1 \in [0; 1]$ ,
- or they are present in  $A_{baseline}$  only and it is in the  $\theta_2 \in \mathbb{N}$  top concepts of  $A_{baseline}$ ,

otherwise they are deleted.  $\theta_1$  and  $\theta_2$  are optimized by using the BioASQ5000 dataset.

<sup>4</sup> This is from another formula in [21]

<sup>5</sup> [http://www.nlm.nih.gov/bsd/indexing/training/CHK\\_010.htm](http://www.nlm.nih.gov/bsd/indexing/training/CHK_010.htm)

<sup>6</sup> <http://www.ncbi.nlm.nih.gov/CBBresearch/Lu/Demo/MeSHNow/>

### 3 Results

The challenge proposes to evaluate systems on three batches. Each batch is a set of five test sets of different sizes (the biggest contains 21,014 tests, the smallest contains 2,153). At a rate of one test set per week, the challenge lasted 15 weeks.

Once a test set has been published, each paper of this set is manually annotated by experts to constitute the gold standard annotation. This takes a long time and results are not final as of June 7th, 2015.

The table 2 shows the — provisional — best results of all our systems compared to the baseline on batch 1 week 2 test set. The main evaluation metrics are a flat measure (micro F-measure, or MiF) and a hierarchical one (LCA-F) [23]. It appears that the most elaborated version of USI ranks first among the others. It also performs better than all the baselines and it ranks third with other teams systems — this is not displayed in the table. However, other versions output mixed results.

**Table 2.** Results obtained (as of June 7th, 2015) on batch 1 week 2 with all USI systems compared to the baselines. Bold values represent the best scores among these systems.

System name	MiF	MiR	MiP	LCA-F	LCA-R	LCA-P
USI <b>baseline</b>	<b>0.5624</b>	0.5417	<b>0.5847</b>	0.4677	<b>0.4765</b>	0.4895
Default MTI	0.5618	<b>0.5460</b>	0.5786	<b>0.4788</b>	0.4754	0.5186
MTIFL	0.5552	0.5008	0.5230	0.4672	0.4391	<b>0.5377</b>
USI <b>abstract</b>	0.5159	0.4951	0.5384	0.4416	0.4364	0.4882
USI 10 <b>neighbors</b>	0.5114	0.4883	0.5368	0.4430	0.4375	0.4864
USI 20 <b>neighbors</b>	0.4846	0.4725	0.4973	0.4450	0.4457	0.4820
MeSH Now BF	0.4361	0.4332	0.4391	0.3817	0.3987	0.3992
BioASQ <b>baseline</b>	0.1083	0.0865	0.1451	0.1541	0.1398	0.2850

It must be noted that MeSH Now BF performed quite poorly on this test set as compared to all other ones for which the scores are comparable to USI **baseline** ones. The results confirm our previous tests that 10 neighbors bring good enough results with little noise. USI **baseline** also has the best precision score among these systems. Finally, we note that the scores are very close (sometimes the difference is less than 1%), which shows that it becomes more and more difficult to create better methods than the existing ones. USI appears to be a fast and generic method that provides one of the best results comparable with some that need to parse the abstract/title with sometimes an extensive learning phase.

## 4 Conclusions and Future Work

Every year, the BioASQ challenge reveals novel and powerful methods. The most elaborated version of USI, `USI baseline`, ranked in the top results of all test sets. However, this approach does not directly take document contents into account nor does any NLP. Although some modifications require NLP (PMRA, `MeSH Now BF` baseline that uses MetaMap), the results show that our generic method can be applied to specific use cases with a bit of adaptations. Given an information retrieval system — which exists in any field such as video retrieval or non-biomedical text retrieval —, USI can be an easy-to-set-up solution for annotating document with terms from a thesaurus. It does not need big training sets for optimizing its few parameters and the computation time has been shown to be better than for some ML methods [12].

On the other hand, USI does not perform the best compared to other systems, it seems to only be able to provide good enough results. Unfortunately, we did not have time to integrate an NLP module to `USI baseline`. USI has been designed for being easily included in a pipeline with other modules. A proper NLP module would certainly greatly improve the results and the creation of a pipeline for annotating biomedical papers is a direction of future work.

## 5 Acknowledgements

We thank Dr. Zhiyong Lu and Dr. Yuqing Mao for providing us with extended results obtained with `MeSH Now BF` on their BioASQ5000 dataset and for their valuable remarks. This work was partly supported by a grant from AVieSan national program (French Alliance nationale pour les sciences de la Vie et de la Santé) and by the French Agence Nationale de la Recherche (ANR-10-BINF-01 Ancestrome).

## References

1. Lin, J., Wilbur, W.J.: PubMed related articles: a probabilistic topic-based model for content similarity. *BMC bioinformatics* 2007, 8, 423 (2007)
2. Lu, Z., Kim, W., Wilbur, W.J.: Evaluation of query expansion using MeSH in PubMed. *Information retrieval* 12, 69-80 (2009) volume
3. Aronson, A.R., Mork, J.G., Gay, C.W., Humphrey, S.M., Rogers, W.J.: The NLM indexing initiatives medical text indexer. *Medinfo*, 11(Pt 1), 268272 (2004)
4. Kosmopoulos, A., Partalas, I., Gaussier, E., Paliouras, G., Androutsopoulos, I.: Evaluation measures for hierarchical classification: a unified view and novel approaches. *CoRR* abs/1306.6802, (2013)
5. Jonquet, C., Shah, N.H., Musen, M. A.: The open biomedical annotator. *Summit on Translational Bioinformatics*, 2009, 56 (2009)
6. Aronson, A.R., Lang, F.-M.: An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17, 229-236 (2010)

7. Huang, M., Névéol, A., Lu, Z.: Recommending MeSH terms for annotating biomedical articles. *Journal of the American Medical Informatics Association* 18, 660-667 (2011)
8. Delbecque, T., Zweigenbaum, P.: Using co-authoring and cross-referencing information for MEDLINE indexing. *AMIA Annual Symposium Proceedings*, 147 (2010)
9. Vasuki, V., Cohen, T.: Reflective random indexing for semi-automatic indexing of the biomedical literature. *Journal of Biomedical Informatics*, 43(5), 694700 (2010)
10. Jimeno-Yepes, A., Mork, J.G., Demner-Fushman, D., Aronson, A.R.: A One-Size-Fits-All Indexing Method Does Not Exist: Automatic Selection Based on Meta-Learning. *Journal of Computing Science and Engineering*, 6(2), 151160 (2012)
11. Mao, Y., Wei, C.H., Lu, Z.: NCBI at the 2014 BioASQ challenge task: large-scale biomedical semantic indexing and question answering. *Working Notes for CLEF 2014 Conference*, 1180, 1319-1327 (2014)
12. Fiorini, N., Ranwez, S., Montmain, J., Ranwez, V.: USI: a fast and accurate approach for conceptual document annotation. *BMC Bioinformatics*, 16, 83 (2015)
13. Mao, Y., Lu, Z.: NCBI at the 2013 BioASQ challenge task: Learning to rank for automatic MeSH indexing. Technical report (2013)
14. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics*, 7, 302 (2006)
15. Sayers, E.: E-utilities Quick Start. *Entrez Programming Utilities Help*. (2010)
16. Lin, D.: An information-theoretic definition of similarity. *ICML*, 98, pp. 296-304 (1998)
17. Seco, N., Veale, T., Hayes, J.: An intrinsic information content metric for semantic similarity in WordNet. *ECAI*, 10891090 (2004)
18. Harispe, S., Sanchez, D., Ranwez, S., Janaqi, S., Montmain, J.: A framework for unifying ontology-based semantic similarity measures: A study in the biomedical domain. *Journal of biomedical informatics*, 48, 38-53 (2014)
19. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: The Semantic Measures Library and Toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. *Bioinformatics*, 30(5), 740-742 (2014)
20. Resnik, P.: Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *J. Artif. Intelligence Res.* 1999, 11, 95-130 (1999)
21. Sanchez, D., Batet, M., Isern, D.: Ontology-based information content computation. *Knowledge-Based Systems 2011*, 24, 297-303 (2011)
22. Zhou, Z., Wang, Y., Gu, J.: A New Model of Information Content for Semantic Similarity in WordNet. *Proceedings of the 2008 Second International Conference on Future Generation Communication and Networking Symposia*, IEEE Computer Society, 85-89 (2008)
23. 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., Artiéres, T., Ngonga Ngomo, A.-C., 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 (2015)