H. Hamdan, S. Albitar, P. Bellot, B. Espinasse, S. Fournier, (2012), « LSIS at TREC 2012 Medical Track – Experiments with Conceptualization, a DFR Model and a Semantic Measure », The Twenty-First Text REtrieval Conference, TREC 2012, Notebook, Vol. Special Publication, pp. 12 p., Gaithersburg, USA, Nov. 2012.

LSIS at TREC 2012 Medical Track – Experiments with conceptualization, the DFR model and a semantic measure

Hussam Hamdan, Shereen Albitar, Patrice Bellot,
Bernard Espinasse, Sébastien Fournier
{firstname.lastname}@lsis.org
LSIS – Aix-Marseille University (AMU)
Domaine Universitaire de St Jérôme
F-13397 Marseille cedex 20 (France)

Abstract
In this paper, we present our participation in the Medical Records Track of TREC2012. We focus on the impact of combining the word space and the concept space in the information retrieval process. For this track, we submitted a baseline run by employing the PL2 weighting model implemented in the Terrier platform, which achieved fair results (0.304 MAP, 0.51P@10). Then, we expanded the documents by performing automatic text conceptualization using UMLS® and the Metamap software on medical records. These textual and conceptual representations, still using the DFR model, led to precision (0.29 MAP, 0.47 P@10). We also automatically extended the topics with UMLS® concepts. This led to a lower precision (0.27 MAP, 0.46 P@10) Lastly, we experimented the usage of semantic IR measures only (0.21 MAP, 041 P@10)..

Keywords: DFR, In_expB2, Automatic Expansion, Medical Record Retrieval, UMLS, Conceptualization, Semantic IR.

1. Introduction

The goal of medical track is to foster research on providing content-based access to the free-text of electronic medical records. To achieve this goal, we propose to combine conceptualization, document and query expansion and the DFR (Divergence from Randomness)[1] matching model. For these purposes, we used the Terrier[1] platform for indexing, retrieval and expansion, and MetaMap®[2] for the conceptualization process.

First of all, we built the free-text index of the medical records and applied a DFR matching model with query expansion. Then, we expended the documents with the concepts extracted from UMLS® and applied a DFR matching model. Finally, we also extended the queries.

The paper is organized as follows: Section 2 describes our system architecture, outlining each component along the three runs. Experimental results will be presented and discussed in section 3. Section 4 gives a conclusion and perspectives.

[1] http://www.terrier.org/
[2] http://metamap.nlm.nih.gov/

2. System architecture

We proposed three strategies to match the user's query and the documents.. We will begin this section by explaining each strategy and by outlining each component.

Each strategy (numbers 1, 2, 3 — see Fig. 1) represents a submitted run. In our first strategy (1) — run LSIS1 —, we indexed the set of documents by employing the Terrier platform and retrieved documents by using the DFR model and performing default query expansion. In the second strategy (2) — run LSIS2 —, we built a second index combining the original documents and their associated concepts after being identified by the Metamap software. As for run LSIS1, we then used the DFR model to retrieve the documents from the topics. Finally, in the third strategy (3) – run LSIS3 –, we added to the second strategy a query conceptualization phase, i.e we matched the extended query (the original tokens and the concepts) with the extended documents. The aim of the second strategy was to measure how much the conceptualization of the documents only affected the weights of the words.

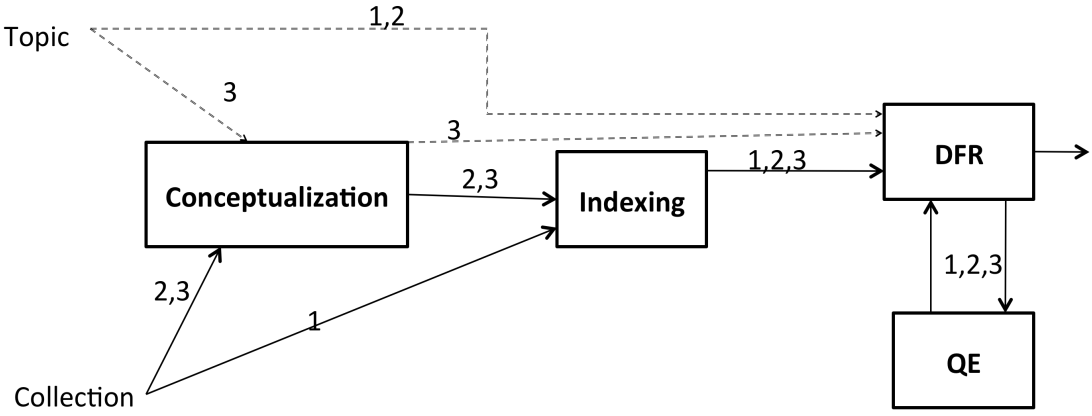

Fig 1.The system's architecture, #1,2 and 3 represent the submitted runs LSIS1, 2 and 3 respectively, DFR designates the model (Divergence From Randomness), QE means (Query Expansion).

2.1. Index Building

We chose the medical report as the indexing unit. We made the indexing for the field TEXT and kept the DOCNO as report identification and VISITID as visit identification (required for distinguishing the reports belonged to the same visit). We used the Terrier IR platform [2] for indexing by applying the Porter stemming algorithm [3] with its standard list of stop words. We applied the same steps for the topics.

2.2. Matching model

We considered that the maximum score between the query-topic q and the visit records d is the relevance score between the query and the visit V.

$$\mathrm{RSV}(V,q) = \mathrm{Max}_{d \in V}\mathrm{score}(d,q) \qquad (1)$$

We submitted runs performed with the DFR model In_expC2 (Inverse Expected Document Frequency model with Bernoulli after-effect and normalization) weighting model [4][5]. Then, we applied query expansion technique based on the default Bose-Einstein 1 (Bo1) expansion model.

According to the In_expC2 model, the relevance score of a document *d* for a query *q* is given by:

$$\text{score}(d,q) = \sum_{t \in q \cap d} qtf \times \text{w}(t,d) \quad (2)$$

where *qtf* is the frequency of term *t* in the query *q*, and w(*t*,*d*) is the relevance score of a document *d* for the query term *t*, given by:

$$\text{w}(t,d) = \left(\frac{F_t+1}{n_t \times (tfn_e)}\right) \times (tfn_e \times \log_2 \frac{\text{N}+1}{n_e+0.5}) \quad (3)$$

where:

–F_t is the term frequency of *t* in the whole collection.

– N is the number of document in the whole collection.

– n_t is the document frequency of *t*.

– n_e is the number of relevant documents containing a term according to the binomial distribution given by:

$$n_e = \text{N} \times (1 - (\frac{1-n_t}{\text{N}})^{F_t}) \quad (4)$$

– $tfn_e$is the normalized within-document frequency of the term *t* in the document *d*. It is given by the second normalization [4][5]:

$$tfn_e = tf \times \log_e \left(1 + c \times \frac{avg_l}{l}\right) (5)$$

where *c* is a parameter for normalization, *tf* is the within-document frequency of the term *t* in the document d, *l* is the document length, and *avg_l* is the average document length in the whole collection.

2.3. Conceptualization using MetaMap

We extended the documents and the queries by the medical concepts extracted from UMLS ontology. For this purpose we used MetaMap, a system developed by the U.S.National Library of Medicine [6]. The comparisons with human subjects have shown that MetaMap is effective in concept identification tasks [7]. MetaMap first analyses the input text and produces a ranked list of possible matching candidate concepts, each candidate concept has a score which will be useful for selecting the appropriate concepts. In fact, MetaMap implements two strategies for selecting concepts: either the 'best concept' strategy which retrieves the best ranked concept or the 'complete' strategy which returns all possible candidates. For the experiments described here, we employed the best concept strategy. Fig 2 shows an example of mapping the original topic number 137 to UMLS concepts.

The mapping to concepts aims to overcome some of the vocabulary mismatch that exists in medical text by mapping different terms to specific concept.

We remark in Fig 2 that *patients* maps the conceptC0030705, *inflammatory disorders* maps C1290884, *receiving* maps C1514756, *TNF-inhibitor treatment* maps C1999216. In fact, these concepts do not represent well the original topic, the concept in SNOMED-CT which represent the *TNF-inhibitor* is C1579324 (*Tumor Necrosis Factor (TNF) inhibitors*), but the concept C1999216 extracted by MetaMap represents

the *inhibitors*, and the concept which represent the *treatment* is *Treating* C1522326. We found several examples that highlight that preprocessing will be needed in the future to improve the conceptualization.

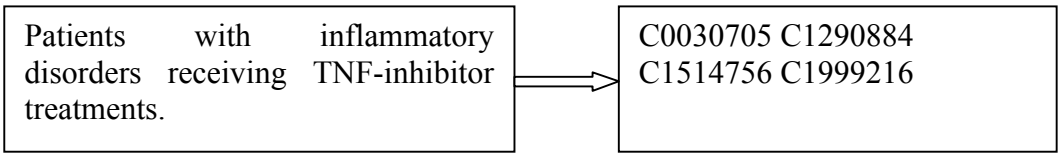

Fig 2. An example of mapping a medical document to UMLS concepts.

2.4. Pseudo-relevance feedback for query expansion

The query expansion (pseudo relevance feedback) mechanism we employed with Terrier, without conceptualization (run LSIS-1) and after conceptualization (runs LSIS-2 and LSIS-3), is a generalization of Rocchio's method[8]. It adds the terms from the top-ranked documents retrieved to the query and reweight the query terms by taking into account the pseudo relevance set. We used the expansion model Bo1 that is based on the Bose–Einstein statistics and on the DFR framework (its efficacy is proven in [2][1][9]). The weight *w* of a term *t* in the top-ranked documents is given by:

$$w(t) = tf_x \times \log_2 \frac{1+P_n}{P_n} + \log_2(1+P_n) \qquad (6)$$

where tf_x is the frequency of the query term in the top-ranked documents, P_n is given by F_t/N, F_t the frequency of the term *t* in the collection, and *N* is the number of documents in the collection. Then, the query term weight *qtw* after merging the top-ranked document terms with the original terms is given by:

$$qtw = \frac{qtf}{qtf_{max}} + \frac{\mathrm{w(t)}}{\lim_{\mathrm{F}\to\mathrm{tfx}} \mathrm{w(t)}} = F_{max} \times \log_2 \frac{1+P_{n,max}}{P_{n,max}} + \log_2(1+P_{n,max}) \qquad (7)$$

where $\lim_{\mathrm{F}\to\mathrm{tfx}} \mathrm{w}(t)$ is the upper bound of w(*t*) (6), $P_{n,max}$ is given by Fmax/N, and F_{max} is the frequency F of the term with the maximum w(t) in the top-ranked documents. If an original query term does not appear in the terms extracted from the top-ranked documents, its query term weight remains equal to the original one.

3. The results

3.1 TREC 2012 results

The results of our system (Table 1) show that the term-based approach LSIS1 gives fair results. It was expected to obtain a little lower precision for LSIS2 (conceptualization of the documents only). But the result for the run LSIS3, where the concepts were added to both collection and topics, shows that our combination (document and query expansions with concepts) did not improve the precision. Indeed, we can remark in Fig. 3 that the behavior of the system has not almost changes within the three strategies for each topic. This led us to say that the term space may well cover the concept space extracted during the conceptualization phase with regard to the retrieved document ranking.

Submitted run	MAP	P@10	R-prec	bpref
LSIS1	**0.3044**	**0.5064**	**0.3340**	**0.3517**
LSIS2	0.2884	0.4681	0.3181	0.3313
LSIS3	0.2690	0.4553	0.3065	0.3094

Table 1. Performance comparison with our three runs (TREC 2012 topics)

3.2 Runs non submitted: concepts only

We experimented two more approaches for testing conceptualization. The first approach (namely "DFR-Concept" hereafter) employs a DFR model for ranking the documents keeping only the mapped concepts (all the original words were removed). The second approach (namely "Semantic IR" hereafter) uses a semantic similarity measure on the concepts in order to rank the documents.

Table 2 shows the MAP, P@10 and R_prec for topics of TREC medical track 2011 and 2012, and Fig. 4 shows a comparison of the two approaches in regard to the MAP of each 2012 topic.

The MAP for each topic was lower in comparison to the term-based approach (Table 1). The main advantage of a 'semantic measure' is to take into account the relationship between concepts in the ontology. This is not accomplished by the DRF-concept approach for which every concept is independent. Unfortunately this theoretical advantage did not produce better results even though the DFR model is not necessarily adapted to conceptual distributions.

A semantic similarity measure exploits an ontology for computing the similarity between two concepts. For computing the similarity between two groups of concepts (the concepts of a topic and the concepts of a document) we have to employ an aggregation measure.

Semantic similarity measures can be generally partitioned in four categories: those based on how close the two concepts in ontology are (structure-based measures), those based on how much information the two concepts share (information content measures), those based on the properties of the concepts (feature-based measures), and those based on combinations of the previous options (hybrid measures) [10].

We experimented a structure-based measure Leacock & Chodorow [11] which exploits the shortest path between the two concepts and the depth of the ontology:

$$Sim_{leacock}(c1, c2) = log\ (\frac{min_i|path_i(c1,c2)|}{2D}) \quad (8)$$

where $\min|\mathrm{path_i}(\mathrm{c1}, \mathrm{c2})|$ is the length of the shortest path between the two concepts c1 and c2, and *D* is the maximum depth of the ontology.

We used an aggregation function [12] for ranking the retrieved documents and computing the similarity between two groups of concepts:

$$Sim(g1, g2) = 0.5\times \left(\frac{\sum_{c\in g1} Maxsim(c,g2)\times idf(c)}{\sum_{c\in g1} idf(c)} + \frac{\sum_{c\in g2} Maxsim(c,g1)\times idf(c)}{\sum_{c\in g2} idf(c)}\right) \quad (9)$$

where Maxsim(*c*,*g*) is the maximum similarity between each concept of the group *g* and the concept *c* given by equation (8).

The results of this approach (Semantic IR in Table 2) for the topics of 2011 and 2012, were not fair, because the measure we used exploits the ontology structure only. These results are weak and we plan to test some other semantic measures that have given good results in other experiments [13].

Run	MAP	P@10	R-prec
DFR-Concept (topics 2011)	0.2160	0.3559	0.2473
DFR-Concept (topics 2012)	0.2103	0.4128	0.2651
Semantic IR (topics 2011)	0.1149	0.2353	0.1715
Semantic IR (topics 2012)	0.1838	0.3362	0.2380

Table 2. Comparison between two concept-based only approaches
(topics 2011 and 2012 — non official results).

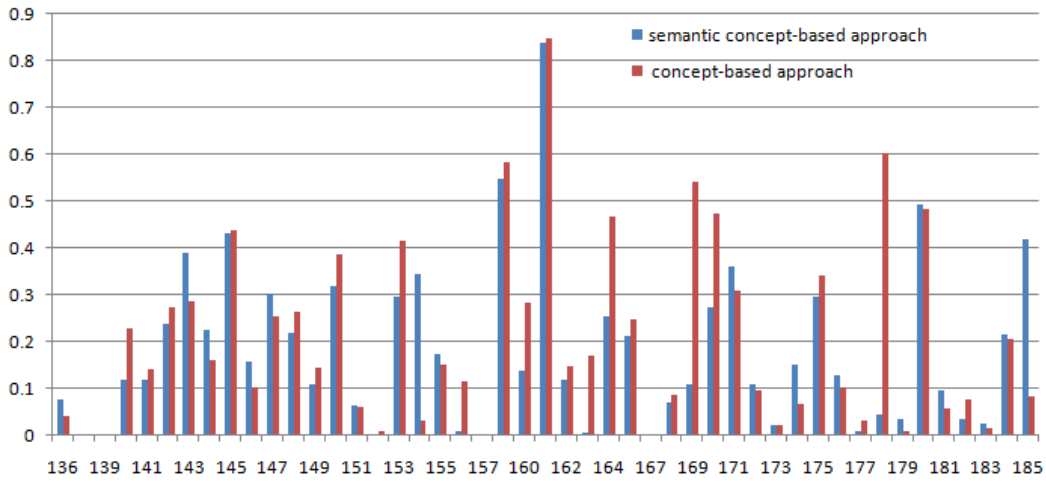

Fig 4. MAP for DFR concept-based and semantic concept-based approaches
(topics 2012 — non official results).

4. Conclusion and perspectives

We have presented our system which uses the Terrier platform for indexing and retrieving, and MetaMap for conceptualization. We focused on the weighting model DFR In_expC2 and measured the impact of expanding documents and topics with concepts. Lastly, we presented some non official runs we experimented by employing a concept only representation of documents and topics. We used a semantic measure that exploits the relationship between concepts. Many measures will be tested in the future and a good integration within the probabilistic model remains to be found.

References

1. Giambattista A, *Probability models for information retrieval based on divergence from randomness*. PhD thesis, University of Glasgow, 2003.
2. Ounis I, Amati G, Plachouras V, He B, Macdonald C et Lioma C. *Terrier: A High Performance and Scalable Information Retrieval Platform*. in *Proceedings of ACM SIGIR'06 Workshop on Open Source Information Retrieval (OSIR 2006)*. 2006. Seattle, Washington, USA.

3. Porter M F, *An algorithm for suffix stripping*, in *Readings in information retrieval*, Karen Sparck J et Peter W, Editors. 1997, Morgan Kaufmann Publishers Inc. p. 313-316.
4. Amati G et Rijsbergen C J V, *Probabilistic models of information retrieval based on measuring the divergence from randomness.* ACM Trans. Inf. Syst., 2002. **20**(4): p. 357-389.
5. Ounis I, Amati G, Plachouras V, He B, Macdonald C et Johnson D, *Terrier Information Retrieval Platform*, in *Advances in Information Retrieval*, Losada D et Fernández-Luna J, Editors. 2005, Springer Berlin Heidelberg. p. 517-519.
6. Aronson A R et Lang F-M, *An overview of MetaMap: historical perspective and recent advances.* JAMIA, 2010. **17**(3): p. 229-236.
7. Pratt W et Yetisgen-Yildiz M, *A Study of Biomedical Concept Identification: MetaMap vs. People.* In Proceedings of American Medical Informatics Association Symposium (AMIA), 2003: p. 529–533.
8. Rocchio J. *Relevance Feedback in Information Retrieval*. in *The SMART Retrieval System*. 1971.
9. Macdonald C, He B, Plachouras V et Ounis I. *University of Glasgow at TREC 2005: Experiments in Terabyte and Enterprise Tracks with Terrier*. in *Proceeddings of the 14th Text REtrieval Conference (TREC 2005)*. 2005. Gaithersburg, MD.
10. Hliaoutakis A, *Semantic Similarity Measures in MeSH Ontology and their application to Information Retrieval on Medline.* Technical Univ. of Crete (TUC), Dept. of Electronic and Computer Engineering, 2005: p. 1-79.
11. Leacock C et Chodorow M, *Filling in a sparse training space for word sense identification.* ms., 1994.
12. Mihalcea R, Corley C et Strapparava C, *Corpus-based and knowledge-based measures of text semantic similarity*, in *Proceedings of the 21st national conference on Artificial intelligence - Volume 1*. 2006, AAAI Press: Boston, Massachusetts. p. 775-780.
13. Sanchez D et Batet M, *Semantic similarity estimation in the biomedical domain: An ontology-based information-theoretic perspective.* Journal of Biomedical Informatics, 2011. **44**(5): p. 749 - 759.

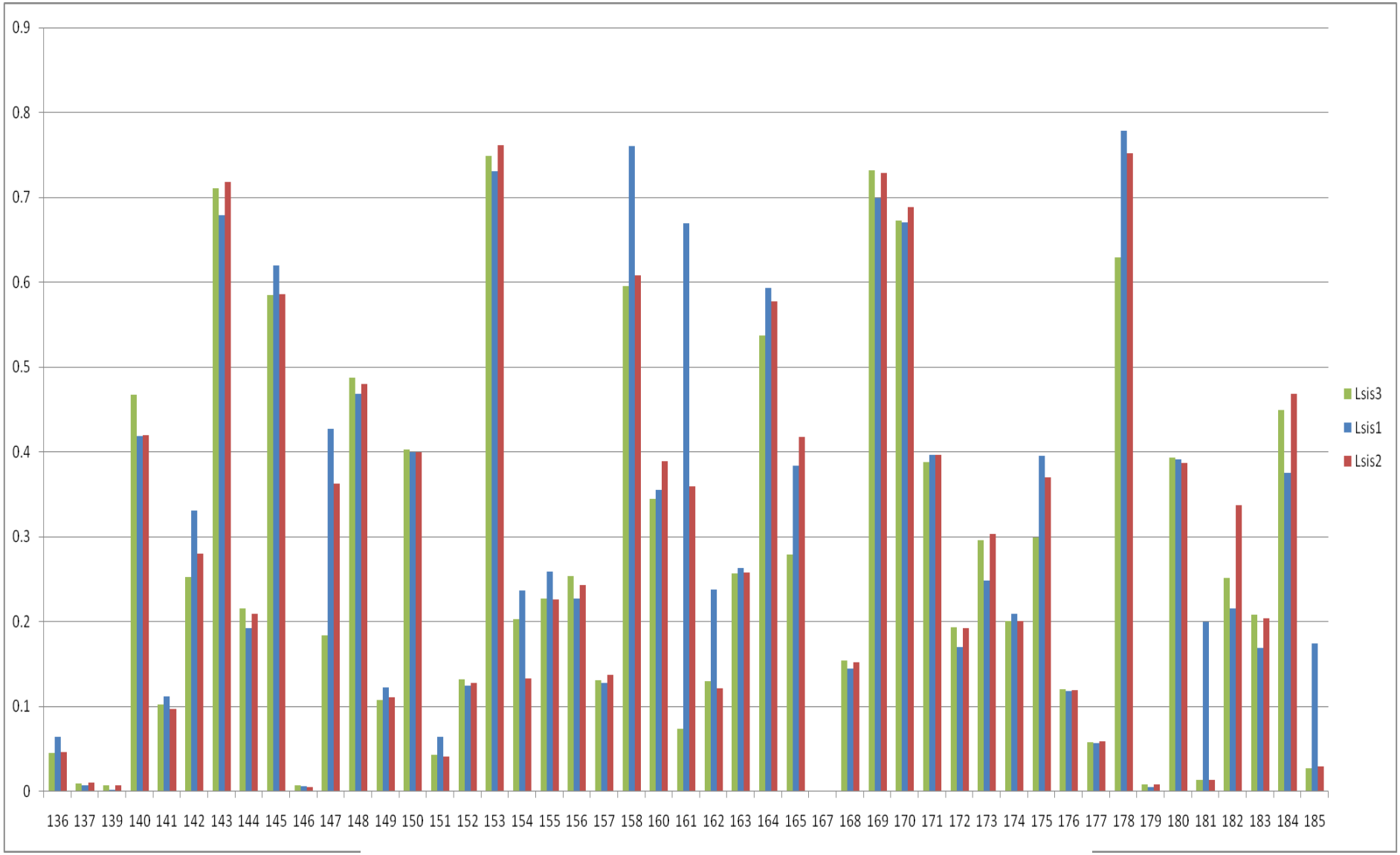

Fig 3. MAP for each topic for 3 submitted runs LSIS1,2,3 (official results)