Biomedical Document Retrieval for Clinical Decision Support System

Jainisha Sankhavara
IRLP lab,
Dhirubhai Ambani Institute of Information and Communication Technology
Gandhinagar, Gujarat, India
jainishasankhavara@gmail.com

Abstract

The availability of huge amount of biomedical literature have opened up new possibilities to apply Information Retrieval and NLP for mining documents from them. In this work, we are focusing on biomedical document retrieval from literature for clinical decision support systems. We compare statistical and NLP based approaches of query reformulation for biomedical document retrieval. Also, we have modeled the biomedical document retrieval as a learning to rank problem. We report initial results for statistical and NLP based query reformulation approaches and learning to rank approach with future direction of research.

1 Introduction and Motivation

Medical and Healthcare related searches are having major focus of internet search now a days. The recent statistics shows that 61% of adults look online for health information (Jones, 2009). This demands proper search and retrieval systems for health related biomedical queries. Biomedical Information Retrieval (BIR) seeks special attention due to the characteristics of biomedical terminologies. Major challenges in biomedical domain are in handling complex, ambiguous, inconsistent medical terms and their ad-hoc abbreviations. Many medical terms are very complex. The average length of biomedical entities is much higher than general entities which makes entity identification task difficult for biomedical domain. Entity identification and normalization helps to better solve the problems of retrieval and ranking of documents for medical search systems, biomedical text summarization, biomedical text data visualization, etc.

As we are focusing here on biomedical document retrieval and ranking system, biomedical literature should be in consideration. Biomedical literature is an important source of study in medical science. Thousands of articles are being added into biomedical literature each year. This large set of biomedical text articles can be used as a collection for Clinical Decision Support System where the related biomedical articles are extracted and suggested to medical practitioners to best care their patients. For this purpose, dataset from Clinical Decision Support (CDS) track is used which contains millions of full text biomedical articles from PMC (PubMed Central)[1]. The statistics of CDS 2014, 2015 and 2016 datasets are given in the table 1. CDS[2] track focuses on retrieval of biomedical articles which are related to patient's medical case reports. These medical case reports which are being used as queries are case narratives of patients medical condition. They describes patients' medical condition i.e. medical history, symptoms, tests performed, treatments etc. For a given query/case report, the main problem is to find relevant documents from the available collection and rank them.

2 Background

'Information Retrieval: A Health and Biomedical Perspective' (Hersh, 2008) provides basic theory, implementation and evaluation of IR systems in health and biomedicine. The tasks of named entity recognition and relation and event extraction, summarization, question answering, and literature based discovery are outlined in Biomedical text mining: a survey of recent progress (Simpson and Demner-Fushman, 2012).

Automatic processing of biomedical text also

[1] http://www.ncbi.nlm.nih.gov/pmc/

[2] http://www.trec-cds.org/

Proceedings of ACL 2018, Student Research Workshop, pages 84–90
Melbourne, Australia, July 15 - 20, 2018. ©2018 Association for Computational Linguistics

Dataset	**CDS 2014**	**CDS 2015**	**CDS 2016**
#Documents	733,138	733,138	1,255,259
Collection size	47.2 GB	47.2 GB	87.8 GB
#Total terms	1,600,536,286	1,600,536,286	2,954,366,841
#Uniq. terms	3,689,317	3,689,317	4,564,612
#Topics	30	30	30
#Rel. docs/Topic	112	150	182
Query forms	Description, Summary	Description, Summary	Note, Description, Summary
Avg. length of Description (in words)	75.8	80.4	119.9
Avg. length of Summary (in words)	24.6	20.4	33.3
Avg. length of Note (in words)	-	-	239.4
Avg. Doc length (in words)	2183	2183	2353

Table 1: CDS DATA statistics

suffers from lexical ambiguity (homonymy and polysemy) and synonymy. Automatic query expansion (AQE) (Maron and Kuhns, 1960; Carpineto and Romano, 2012) which has a long history in information retrieval can be useful to deal with such problems. For instance, medical queries were expanded with other related terms from RxNorm, a drug dictionary, to improve the representation of a query for relevance estimation (Demner-Fushman et al., 2011). The emergence of medical domain specific knowledge like UMLS can contribute to the retrieval system to gain more understanding of the biomedical documents and queries. The Unified Medical Language System (UMLS) (Bodenreider, 2004) is a metathesaurus for medical domain. It is maintained by National Library of Medicine (NLM) and it is the most comprehensive resource, unifying over 100 dictionaries, terminologies, and ontologies. Various approaches of information retrieval with the UMLS Metathesaurus have been reported: some with decline in results (Hersh et al., 2000) and some with gain in results (Aronson and Rindflesch, 1997). The next section of this paper includes statistical approaches as well as NLP based approaches.

3 Query Reformulation for Biomedical Document Retrieval

Here, we present statistical and NLP based query reformulation approaches for biomedical document retrieval. Statistical approaches include feedback based query expansion and feedback document discovery based query expansion. An NLP based approach that is UMLS concept based query reformulation is also discussed here.

3.1 Automatic Query Expansion With Pseudo Relevance Feedback & Relevance Feedback

Query Expansion (QE) is the process of reformulating a query to improve retrieval performance and efficiency of IR systems. QE is proved to be efficient in case of document retrieval (Carpineto and Romano, 2012). It helps to overcome vocabulary mismatch issues by expanding the user query with additional relevant terms and by re-weighting all terms. Query Expansion which uses the top retrieved relevant documents is known as Relevance Feedback. It requires human judgment to identify relevant documents from top retrieved documents. While pseudo Relevance Feedback technique assumes the top retrieved documents to be relevant and uses as feedback documents. It does not require human input at all. The Query expansion based approaches for biomedical domain gives better results as compared to retrieval without query expansion (Sankhavara et al., 2014).

Table 2 and table 3 shows the results of standard retrieval (without expansion), Pseudo-Relevance Feedback (PRF) based Query Expansion and Relevance Feedback (RF) based Query Expansion with BM25 and In_expC2 retrieval models (Amati et al., 2003) on CDS 2014, 2015 and 2016 datasets. The retrieval model BM25 is a ranking function based on probabilistic retrieval framework while In_expC2 is also a probabilistic but based on Divergence From Randomness (DFR). These models are available in Terrier IR Plateform[3] (Ounis et al., 2005) which is developed at School of Computing Science, University of Glas-

[3] http://terrier.org

MAP	**CDS 2014**	**CDS 2015**	**CDS 2016**
BM25	0.1071	0.1147	0.062
BM25+PRF$_{10}$	0.1542 (+44%)	0.1805 (+57.4%)	0.0769 (+24%)
BM25+RF$_{10}$	0.205 (+91.4%)	0.1941 (+69.2%)	0.0984 (+58.7%)
BM25+RF$_{50}$	0.2768 (+158.5%)	0.2283 (+99%)	0.1456 (+134.8%)
In_expC2	0.1096	0.1201	0.0632
In_expC2+PRF$_{10}$	0.1623 (+48.1%)	0.1725 (+43.6%)	0.0754 (+19.3%)
In_expC2+RF$_{10}$	0.2117 (+93.2%)	0.1895 (+57.8%)	0.0992 (+57%)
In_expC2+RF$_{50}$	0.2587 (+136%)	0.2191 (+82.4%)	0.1275 (+101.7%)

Table 2: Results (MAP) of Query Expansion with PRF and RF

infNDCG	**CDS 2014**	**CDS 2015**	**CDS 2016**
BM25	0.1836	0.2115	0.171
BM25+PRF$_{10}$	0.2522 (+37.4%)	0.283 (+33.8%)	0.2047 (+19.7%)
BM25+RF$_{10}$	0.3355 (+82.7%)	0.3028 (+43.2%)	0.2428 (+42%)
BM25+RF$_{50}$	0.4186 (+128%)	0.3478 (+64.4%)	0.3094 (+80.9%)
In_expC2	0.2002	0.2132	0.1785
In_expC2+PRF$_{10}$	0.2724 (+36.1%)	0.2734 (+28.2%)	0.2018 (+13.1%)
In_expC2+RF$_{10}$	0.3426 (+71.1%)	0.3015 (+41.4%)	0.245 (+37.3%)
In_expC2+RF$_{50}$	0.4019 (+100.7%)	0.339 (+59%)	0.3219 (+80.3%)

Table 3: Results (infNDCG) of Query Expansion with PRF and RF

gow. Here, we have used terrier plateform for the experiments. Summary part of the query is used for retrieval with top 10 and 50 top documents for feedback in expansion. MAP and infNDCG are used as evaluation metrics (Manning et al., 2008). Higher the value of evaluation measure, better the retrieval result of system. The result improves with PRF and RF based query expansion giving statistically significant results ($p < 0.05$) as compared to no expansion. Here RF is giving 50-60% more improvement than PRF over no expansion. We argue that biomedical retrieval should be done keeping human in the loop. A small human intervention can increase the retrieval accuracy to 60% more.

3.2 Feedback Document Discovery for Query Reformulation

Feedback Document Discovery based query expansion as described in (Sankhavara and Majumder, 2017) learns to identify relevant documents for query expansion from top retrieved documents. The main aim is to use small amount of human judgement and learn pseudo judgement for other documents to reformulate the queries. One approach is based on classification. If we have human judgements available for some of the feedback documents, then it will serve as a training data for classification. The documents were represented as a collection of bag-of-words, the TF-IDF scores of the words represent features and human relevance scores provides the classes. Then the relevance is predicted for other top retrieved feedback documents. The second approach is based on and classification+clustering. It first applies classification in similar way as in first approach and then applies clustering on relevance predicted class by the classification method, thus filtering out more non-relevant documents from relevant ones. Since, the convergence of K-means clustering depends on the initial choice of cluster centroids, the initial cluster centroids are chosen as the average of relevant documents vectors and the average of non-relevant documents vectors from training data.

Here we have used that approach with different features. The TF-IDF features are weighted based on type of words. CliNER tool (Boag et al., 2015) has been used to identify medical entities of type ’problem’, ’test’ and ’treatment’ from documents, which was trained on i2b2 2010 dataset (Uzuner et al., 2011). The i2b2 2010 dataset includes discharge summaries from Partners HealthCare, from Beth Israel Deaconess Medical Center and from University of Pittsburgh Medical Center. These discharge summaries are fully deidentified and manually annotated for concept, assertion, and relation information. Here, we have

	CDS 2014			
	MAP		infNDCG	
	No feature weighting	Feature weighting using CliNER	No feature weighting	Feature weighting using CliNER
Original Queries	0.1071		0.1836	
Queries+RF$_{50}$	0.2768		0.4186	
$\{$Nearest neighbors$\}_{50_200}$	0.2761	0.2747	0.4177	0.4140
$\{$Nearest neighbors + k-means$\}_{50_200}$	0.2794	0.2777	0.4220	0.4195
$\{$Neural net$\}_{50_200}$	0.2790	0.2787	0.4235	0.4240
$\{$Neural net + k-means$\}_{50_200}$	0.2790	0.2807	0.4218	**0.4269**

Table 4: Results of Feedback Document Discovery

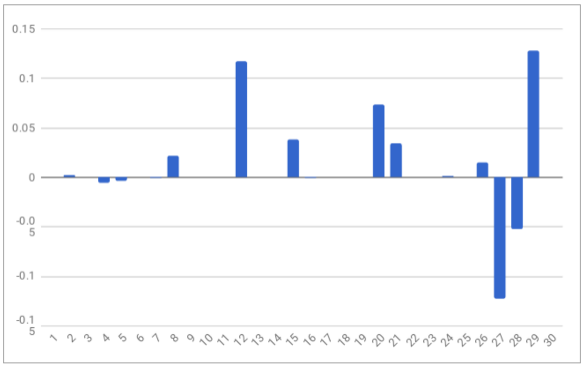

Figure 1: Query wise difference graph of infNDCG for feedback document discovery and relevance feedback

used these discharge summaries along with their concept annotations to train CliNER. This trained model is applied on CDS documents to identify ’problem’, ’test’ and ’treatment’ concept entities. The features related to these entities in CDS documents are weighted thrice, thus giving importance to these entities while learning to identify feedback document. For feedback document discovery with weighted entities, we have used top 50 documents and their corresponding relevance for training, then the relevance was predicted for next top 200 documents and used for expansion of queries. For classification two methods, nearest neighbour classifier and neural net classifier, have been used and k-means is used for clustering with k=2 for relevant and non-relevant documents. In all cases, only relevant identified documents are used for expanding queries. The comparison of results of original queries without expansion, expansion with relevance feedback and expansion with two approaches of feedback document discovery (classification and classification+clustering) for CDS 2014 dataset is given in table 4.

The results clearly indicates improvement over original queries and relevance feedback. Fig 1 shows query wise difference in infNDCG between $\{$Neural net + k-means$\}_{50_200}$ and Queries+RF$_{50}$. Out of 30 queries of CDS 2014, 2 queries degrade performance but 7 queries improve.

3.3 UMLS Concepts Based Query Reformulation

Medical domain-specific knowledge can be incorporated to the process of query reformulation in Biomedical IR system. There are knowledge based approaches proposed in the literature (Aronson and Rindflesch, 1997; Demner-Fushman et al., 2011; Hersh, 2008). In the biomedical text retrieval, medical concepts and entities are more informative than other common terms. Moreover, medical ontologies, thesaurus and biomedical entity identifiers are available to identify medical related concepts.

Here we have used the resource UMLS. The following three query reformulation experiments are done using it. First: The UMLS concepts are identified from the query text and used with queries. Second: Along with the UMLS concepts, MeSH (Medical Subject Heading) terms are also identified and used in queries. MeSH is a hierarchically organized vocabulary of UMLS. Third: Medical entities are identified manually and used with queries. One example query with all these reformulations is presented in Appendix A.

Table 5 shows the results of these reformulated queries of CDS 2014. PRF and RF based query expansion is also carried out on each form of the queries. The results shows clear improvement when using UMLS concepts in queries as compared to original queries. One more important observation here to make is that, for no-expansion

infNDCG	**CDS 2014**			
	BM25		**In_expC2**	
	MAP	**infNDCF**	**MAP**	**infNDCF**
Original Queries	0.1071	0.1836	0.1096	0.2002
Original Queries + PRF_{10}	0.1542	0.2522	0.1623	0.2724
Original Queries + RF_{10}	0.2050	0.3355	0.2117	0.3426
Original Queries + RF_{50}	0.2768	0.4186	0.2587	0.4019
Queries + UMLS concepts	0.1660	0.1830	0.1597	0.1781
Queries + UMLS concepts + PRF_{10}	**0.1607**	**0.2607**	**0.1486**	**0.2431**
Queries + UMLS concepts + RF_{10}	**0.2164**	**0.3423**	**0.2138**	**0.3459**
Queries + UMLS concepts + RF_{50}	**0.2776**	**0.4232**	**0.2569**	**0.4021**
Queries + UMLS concepts + Mesh terms	0.1039	0.1749	0.1086	0.1792
Queries + UMLS concepts + Mesh terms + PRF_{10}	0.1460	0.2409	0.1411	0.2376
Queries + UMLS concepts + Mesh terms + RF_{10}	0.2052	0.1992	0.3321	0.3291
Queries + Manual Entities	0.1112	0.1860	0.1140	0.2114
Queries + Manual Entities + PRF_{10}	0.1601	0.2634	0.1584	0.2650
Queries + Manual Entities + RF_{10}	0.2112	0.3394	0.2120	0.3414

Table 5: Results of UMLS based query processing

and PRF, the manual entities fail to improve MAP when compared to UMLS entities but certainly give better results in terms of infNDCG.

4 Learning To Rank

Learning to rank (LTR) (Liu et al., 2009) is an application of machine learning in the construction of ranking models for information retrieval systems where retrieval problem is modeled as a ranking problem. LTR framework requires training data of queries and documents matching them together with relevance degree of each match. Training data is used by a learning algorithm to produce a ranking model which computes relevance of documents for actual queries.

The LTR framework is applied on CDS 2014 dataset where the features for query document pairs are computed similarly as the features used for OHSUMED LETOR dataset (Qin et al., 2010). These features are mainly based on TF, IDF and their normalized versions. Since the whole document pool is too large, document pooling has been done and top K documents (by BM25) for each query are used for feature extraction. SVMRank has been used as a machine learning framework.

Table 6 shows the results of LTR when the features are computed on Title+Abstract part of the documents, on Title+Abstract+Content of the documents (i.e. full documents). With these variations of features, the experiments are carried out on original queries, queries with UMLS concepts and queries with manually identified medical concepts.

infNDCG	**OHSUMED features on T, A and T+A**	**OHSUMED features on T, A and C**
Original Queries	0.097	0.1769
Queries + UMLS	0.0833	0.1556
Queries + Manual	0.1049	0.1785

Table 6: Results of Learning to Rank with different features

	infNDCG
Retrieval (BM25)	0.1836
LTR using human judgements	0.1769
Pseudo LTR K=1000	0.1849
Pseudo LTR K=1500	**0.1872**
Pseudo LTR K=2000	0.1859
Pseudo LTR K=2500	0.1865
Pseudo LTR K=3000	0.1865

Table 7: Results Learning To Rank with pseudo judgements

All these LTR experiments require human judgement for training. To overcome the need of manual judgement, pseudo judgements were considered where out of k training documents, Top k/2 documents are considered to be relevant and other k/2 documents to be non-relevant.

As shown in table 7, the results of LTR trained using pseudo qrels are better than one with actual human judged qrels but the difference is not statistically significant. The results of LTR are comparable to retrieval using BM25.

5 Future Research Directions

Biomedical text processing and information retrieval being a new field of research opens up many research directions. In this article, we have presented a preliminary study of statistical and NLP based biomedical document retrieval techniques for clinical decision support systems. It included query reformulation based information retrieval framework with pseudo relevance feedback, relevance feedback, feedback document discovery and UMLS concept based reformulation for Biomedical domain. Standard IR frameworks PRF and RF works good enough for Clinical Decision Support System. Feedback document discovery based query reformulation which is a statistical approaches can be improvised in future for significant improvement. Another statistical model Learning to Rank is also having future scope for more improvement. The initial framework for NLP based approach UMLS concept based retrieval also shows improvement in the results. Therefore, we plan to combine statistical and NLP based approaches and come up with new better model for biomedical document retrieval for Decision Support Systems. Also, we are planning to do feature weighting using NLP at entity level in feedback document discovery approach.

References

Gianni Amati, Cornelis Joost, and Van Rijsbergen. 2003. Probabilistic models for information retrieval based on divergence from randomness.

Alan R Aronson and Thomas C Rindflesch. 1997. Query expansion using the umls metathesaurus. In *Proceedings of the AMIA Annual Fall Symposium*, page 485. American Medical Informatics Association.

William Boag, Kevin Wacome, Tristan Naumann, and Anna Rumshisky. 2015. Cliner: A lightweight tool for clinical named entity recognition. *AMIA Joint Summits on Clinical Research Informatics (poster)*.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Claudio Carpineto and Giovanni Romano. 2012. A survey of automatic query expansion in information retrieval. *ACM Computing Surveys (CSUR)*, 44(1):1.

Dina Demner-Fushman, Swapna Abhyankar, Antonio Jimeno-Yepes, Russell F Loane, Bastien Rance, François-Michel Lang, Nicholas C Ide, Emilia Apostolova, and Alan R Aronson. 2011. A knowledge-based approach to medical records retrieval. In *TREC*.

William Hersh. 2008. *Information retrieval: a health and biomedical perspective*. Springer Science & Business Media.

William Hersh, Susan Price, and Larry Donohoe. 2000. Assessing thesaurus-based query expansion using the umls metathesaurus. In *Proceedings of the AMIA Symposium*, page 344. American Medical Informatics Association.

S Jones. 2009. The social life of health information. *Pew research center, Washington, DC, Pew Internet & American Life Project*.

Tie-Yan Liu et al. 2009. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331.

Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. 2008. *Introduction to information retrieval*, volume 1. Cambridge university press.

Melvin Earl Maron and John L Kuhns. 1960. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM (JACM)*, 7(3):216–244.

Iadh Ounis, Gianni Amati, Vassilis Plachouras, Ben He, Craig Macdonald, and Douglas Johnson. 2005. Terrier information retrieval platform. In *European Conference on Information Retrieval*, pages 517–519.

Tao Qin, Tie-Yan Liu, Jun Xu, and Hang Li. 2010. Letor: A benchmark collection for research on learning to rank for information retrieval. *Information Retrieval*, 13(4):346–374.

Jainisha Sankhavara and Prasenjit Majumder. 2017. Biomedical information retrieval. In *Working notes of FIRE 2017 - Forum for Information Retrieval Evaluation*, pages 154–157.

Jainisha Sankhavara, Fenny Thakrar, Prasenjit Majumder, and Shamayeeta Sarkar. 2014. Fusing manual and machine feedback in biomedical domain. In *Proceedings of The Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014*.

Matthew S Simpson and Dina Demner-Fushman. 2012. Biomedical text mining: a survey of recent progress. In *Mining text data*, pages 465–517. Springer.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

A Example Query

```
<topic number="1" type="diagnosis">
<summary>
  58-year-old woman with hypertension and obesity presents with
exercise-related episodic chest pain radiating to the back.
</summary>
<UMLS_entities>
  hypertension obesity exercise related chest pain radiating back nos
</UMLS_entities>
<MeSH_entities>
  Vascular Diseases Overnutrition Overweight Motor Activity Human
Activities Torso Bone and Bones Neurologic Manifestations Sensation
</MeSH_entities>
<manual_entities>
  woman hypertension obesity exercise-related episodic chest pain
radiating back </manual_entities>
</topic>
```

B Example Document

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC 3990010/pdf/1745-6215-15-124.pdf

C OHSUMED features

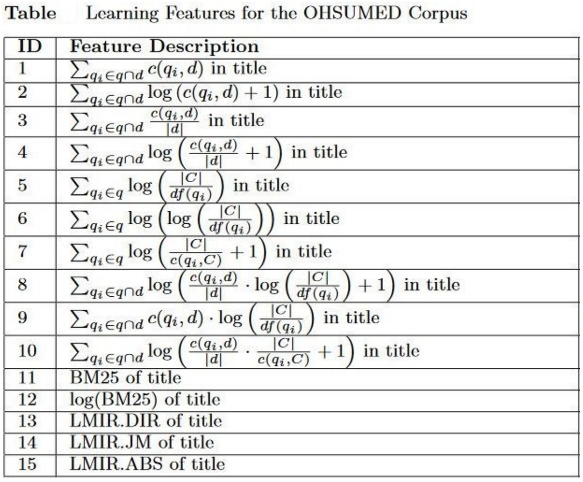

Table Learning Features for the OHSUMED Corpus

ID	Feature Description
1	$\sum_{q_i \in q \cap d} c(q_i, d)$ in title
2	$\sum_{q_i \in q \cap d} \log(c(q_i, d) + 1)$ in title
3	$\sum_{q_i \in q \cap d} \frac{c(q_i,d)}{\lvert d \rvert}$ in title
4	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} + 1\right)$ in title
5	$\sum_{q_i \in q} \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in title
6	$\sum_{q_i \in q} \log\left(\log\left(\frac{\lvert C \rvert}{df(q_i)}\right)\right)$ in title
7	$\sum_{q_i \in q} \log\left(\frac{\lvert C \rvert}{c(q_i,C)} + 1\right)$ in title
8	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right) + 1\right)$ in title
9	$\sum_{q_i \in q \cap d} c(q_i, d) \cdot \log\left(\frac{\lvert C \rvert}{df(q_i)}\right)$ in title
10	$\sum_{q_i \in q \cap d} \log\left(\frac{c(q_i,d)}{\lvert d \rvert} \cdot \frac{\lvert C \rvert}{c(q_i,C)} + 1\right)$ in title
11	BM25 of title
12	log(BM25) of title
13	LMIR.DIR of title
14	LMIR.JM of title
15	LMIR.ABS of title