A Domain Independent Approach to Clinical Decision Support

Paul McNamee
Johns Hopkins University
Human Language Technology Center of Excellence

Abstract

Continuing our work from the inaugural running of the Clinical Decision Support track in 2014, we submitted runs to the 2015 evaluation. Our approach this year was very similar to that used in 2014 (Xu et al., 2014). Our submitted runs were created using the JHU HAIRCUT retrieval engine, and featured use of character n-gram indexing and use of pseudo-relevance feedback. The main contribution is investigating the retrieval of scientific medical documents using a domain independent approach.

1 Introduction

For this year's Clinical Decision Support (CDS) track we again used the JHU HAIRCUT retrieval engine described by McNamee and Mayfield (2004). One of HAIRCUT's distinctive traits is the use of character n-grams as indexing terms, which have proven to be effective for controlling the effects of morphological variation (McNamee et al., 2009). While morphological variation is not nearly as substantial a problem in English as it is in some other languages, it seems likely that the highly technical terminology prevalent in medical literature could benefit from stemming or n-gram indexing.

Our submissions were produced in less than a person-day of effort. As the document collection remained unchanged from 2014, we used previously created indexes. We had hoped to explore other techniques, including relevance feedback using "collection enrichment" (Kwok and Chan, 1998) from appropriate medical domain side corpora, however we did not have adequate time to conduct those experiments.

Run	Topic Fields	Terms	RF
hltcoewsrf	Summary	words	Yes
hltcoe4srf	Summary	4-grams	Yes
hltcoe5srf	Summary	5-grams	Yes

Table 1: Runs submitted to Task A

Run	Topic Fields	Terms	RF
hltcoewsdrf	Summary, Diagnosis	words	Yes
hltcoe4sdrf	Summary, Diagnosis	4-grams	Yes
hltcoe5sdrf	Summary, Diagnosis	5-grams	Yes

Table 2: Runs submitted to Task B

We did not make any use of domain-specific resources such as ontologies, thesauri, or biomedical IE tools. We sought through our participation to determine the performance that a domain-agnostic, state-of-the-art retrieval engine might obtain.

2 Submissions

We submitted three runs for Task A. Our Task A submissions used the *summary* field and did not use the *description* field; this choice was made based on positive results from the 2014 CDS evaluation, although other participants reported better outcomes using the description field (Roberts et al., 2015). Each run used pseudo relevance feedback. The three runs varied based on the type of tokenization as can be observed in Table 1 above.

For Task B, our submitted runs were just as in Task A, except that if a *diagnosis* field was present, then that text was also used in addition to the *summary* field (see Table 2).

For each task, the principal distinction in our submitted runs is the type of tokenization em-

ployed, either unstemmed words or overlapping word-spanning character n-grams of length 4 or 5. When pseudo relevance feedback was applied, terms were weighted based on comparing term frequencies in documents from the top 20 ranks and bottom 75 (of 1000) ranked documents. When unstemmed words were used, queries were expanded (or limited) to 60 words; when n-grams were used, the number of terms in the revised query was 200, whether 4-grams or 5-grams. These settings were based on values that have yielded favorable results in previous evaluations.

A unigram statistical language model for retrieval was employed (Hiemstra, 2001; Miller et al., 1999) and smoothing was accomplished using linear interpolation:

$$P(D|Q) \propto \prod_{t \in Q} \lambda P(t|D) + (1 - \lambda) P(t|C) \quad (1)$$

Relative document term frequency was used to estimate $P(t|D)$, and $P(t|C)$ was based on the mean relative document term frequency from documents in the collection. The two probabilities were evenly weighted (*i.e.,* a constant of $\lambda = 0.5$ was used) in all conditions. In this model we have generally found retrieval performance to be fairly insensitive to changes in this smoothing parameter, though others have reported differently (Zhai and Lafferty, 2004).

3 Results

NIST provided results for each of our official runs, and we report average results over the 30 topics in Table 3. Sampled metrics include inferred average precision (infAP), inferred normalized discounted cumulative gain (infNDCG), precision at the number of known relevant documents (R-prec), and precision at a fixed cutoff of 10 documents (P@10).

At the time of submission we created runs that used the *description* instead of *summary*, and runs which did not use relevance feedback, though it should be noted that these runs did not contribute to the judgment pools. If the pools created for judging relevance are not biased against post-hoc runs, then we can make direct comparisons using these "post hoc" runs.

Task	Run	infAP	infNDCG	iP10
A	hltcoewsrf	0.0680	**0.2670**	**0.4267**
A	hltcoe4srf	0.0690	0.2419	0.3900
A	hltcoe5srf	**0.0706**	0.2643	0.4000
A	median	0.0414	0.2038	0.3433
A	oracle	0.1258	0.4399	0.6833
B	hltcoewsdrf	0.0841	**0.3275**	0.4933
B	hltcoe4sdrf	0.0863	0.3105	0.4700
B	hltcoe5sdrf	**0.0872**	0.3245	**0.5000**
B	median	0.0633	0.2794	0.4500
B	oracle	0.1670	0.5348	0.7833

Table 3: Averaged results over the 30 topics in 2015, compared to median performance and oracle best results from submitted automatic runs.

4 Discussion

We make several observations from these results.

4.1 Summary vs. Description

We confirm our finding form 2014 that use of summaries results in better performance than use of the description field; performance in all metrics was notably higher using the summary topics. As can be seen in Table 4, MAP, P@10, and the number of retrieved relevant documents are better in each run when the summary field is used instead of the description field.

4.2 Relevance Feedback

While there is no guarantee that performance may not be negatively affected on a particular query, relevance feedback boosts average performance. Table 4 reveals marked improvement on both recall-sensitive and precision-focused metrics. The Task B run using 5-grams and relevance feedback (*i.e.,* hltcoe5sdrf) obtains a relative gain of 26.5% in MAP and finds 13.2% more documents compared to its equivalent run that does not use feedback. Similarly, when words are used as indexing terms, hltcoewsdrf sees a relative gain of 41.4% in MAP and finds 22.2% more documents compared to its non-feedback equivalent.

While the overall score is slightly lower with words, the relative gain from employing relevance feedback is larger than with n-grams, which we attribute to the fact that feedback conveys benefits in morphological normalization which occur naturally with n-gram representations. For example, "epilepsy" and "epileptic" are entirely separate in-

Term	Topic	RF	Dx	MAP	P@10	Recall	Runid
Words	Summary	No	No	0.0946	0.3667	1768	
Words	Description	No	No	0.0810	0.3067	1565	
Words	Summary	RF	No	0.1571	0.4267	2272	hltcoewsrf
Words	Description	RF	No	0.1407	0.3733	2144	
Words	Summary	No	Dx	0.1419	0.4600	2186	
Words	Description	No	Dx	0.1057	0.3667	1799	
Words	Summary	RF	Dx	0.2007	0.4933	2671	hltcoewsdrf
Words	Description	RF	Dx	0.1634	0.4200	2384	
5-grams	Summary	No	No	0.1107	0.3700	1884	
5-grams	Description	No	No	0.1006	0.3400	1832	
5-grams	Summary	RF	No	0.1619	0.4000	2356	hltcoe5srf
5-grams	Description	RF	No	0.1556	0.3733	2341	
5-grams	Summary	No	Dx	0.1611	0.4600	2374	
5-grams	Description	No	Dx	0.1384	0.3967	2226	
5-grams	Summary	RF	Dx	**0.2038**	**0.5000**	**2687**	hltcoe5sdrf
5-grams	Description	RF	Dx	0.1870	0.4333	2651	

Table 4: Mean average precision, P@10, and retrieved relevant for various experimental conditions according to *trec_eval*.

dexing terms, but they share 5-grams such as “epile” and ”pilep’.

4.3 Tokenization

There appear to be gains using character 5-grams over unstemmed words, particularly if feedback is not applied. Table 3 shows improvements in infAP and Table 4 shows marginally higher performance in MAP, P@10, and Recall with 5-grams.[1]

4.4 Diagnosis

Large gains in retrieval performance were observed when the *diagnosis* field is used (*i.e.,* Task B vs. the equivalent Task A) run. This effect is on par with the substantial gain achieved with relevance feedback.

5 Conclusions

Our results in 2015 confirm our observations from the inaugural running of the track in 2014. Our key findings are that relevance feedback or the provision of a diagnosis field conveys a dramatic improvement in performance. We also believe that summaries are superior to the description field, and character n-grams possess advantages over unstemmed words.

[1] Last year we saw relative gains from 3.5% to 12%, depending on the metric).

References

Djoerd Hiemstra. 2001. *Using Language Models for Information Retrieval*. Ph.D. thesis, University of Twente.

K. L. Kwok and M. Chan. 1998. Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’98, pages 250–256, New York, NY, USA. ACM.

Paul McNamee and James Mayfield. 2004. Character n-gram tokenization for European language text retrieval. *Information Retrieval*, 7(1-2).

Paul McNamee, Charles Nicholas, and James Mayfield. 2009. Addressing morphological variation in alphabetic languages. In *Proceedings of the 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 75–82. ACM.

David R. H. Miller, Tim Leek, and Richard M. Schwartz. 1999. A hidden Markov model information retrieval system. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 214–221, New York, NY, USA. ACM.

Kirk Roberts, Matthew Simpson, Dina Demner-Fushman, Ellen Voorhees, and William Hersh. 2015. State-of-the-art in biomedical literature retrieval for clinical cases: a survey of the TREC 2014 CDS track. *Information Retrieval Journal*, pages 1–36.

Tan Xu, Paul McNamee, and Douglas W. Oard. 2014. HLTCOE at TREC 2014: Microblog and clinical deci-

sion support. In *Proceeding of the 2014 Text Retrieval Conference*.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Trans. Inf. Syst.*, 22(2):179–214, April.