Two Test Collections for Retrieval Using Named Entity Markup

Jacob Bremerman
jbrem@umd.edu
University of Maryland, College Park

Dawn Lawrie, James Mayfield
(lawrie,mayfield)@jhu.edu
Johns Hopkins University HLTCOE

Douglas W. Oard
oard@umd.edu
University of Maryland, College Park

ABSTRACT

Studying the effects of semantic analysis on retrieval effectiveness can be difficult using standard test collections because both queries and documents typically lack semantic markup. This paper describes extensions to two test collections, CLEF 2003/2004 Russian and TDT-3 Chinese, to support study of the utility of named entity annotation. A new set of topic aspects that were expected to benefit from named entity markup were defined for topics in those test collections, with two queries for each aspect. One of these queries uses named entities as bag-of-words query terms or as semantic constraints on a free-text query term; the other is a bag-of-words baseline query without named entity markup. Exhaustive judgment of the documents annotated by CLEF or TDT as relevant to each corresponding topic was performed, resulting in relevance judgments for 133 Russian and 33 Chinese topic aspects that each have at least one relevant document. Named entity tags were automatically generated for the documents in both collections. Use of the test collections is illustrated with some preliminary experiments.

KEYWORDS

test collection; topic aspects; entity-based search

ACM Reference Format:
Jacob Bremerman, Dawn Lawrie, James Mayfield, and Douglas W. Oard. 2020. Two Test Collections for Retrieval Using Named Entity Markup. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland*. ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3340531.3417452

1 INTRODUCTION

Named Entity Recognition (NER), the identification of mentions of named entities in text, has the potential to improve retrieval in two complementary ways. One use is to improve precision by using an entity type tag as a constraint for word sense disambiguation. For example, "Washington" might refer to either a person or a location, but adding an entity tag as a constraint on that term would allow a query to specify that the "PER" (person) sense of that word was intended. This would allow the system to reward instances of the word bearing the intended sense, when that sense could be correctly detected in a document. An alternative use of NER is to improve recall by allowing a query to specify an entity type rather than a specific entity. For example, question answering systems might automatically rewrite "where did the president visit last year?" to "president visited GPE in 2019," where "GPE" as an entity type would match the name of any geopolitical entity (e.g., "Madrid," "Italy," or "Virginia"). As these examples indicate, the process of adding named entity types to queries might be manual or automated.

As a first step toward exploring how NER can be useful in a diverse range of languages, this paper contributes two test collections that include manual named entity annotations on the queries and automatic named entity annotations on the documents for two information retrieval test collections.[1] We are ultimately interested in developing techniques that will work well in many languages; our initial test collections focus on Russian and Chinese, two languages for which we are not aware of similar collections. Standard test collections for these languages are available from CLEF (for Russian) and from TDT (for Chinese). However, the topics in those test collections were not developed with use of NER in mind. Rather than develop topics *de novo*, we created queries for specific aspects of the existing topics. This allowed us to restrict the documents that needed to be assessed for topic aspect relevance to those that had been annotated by CLEF or TDT as relevant to the original topic on which the topic aspect was based.

In this paper we describe the process by which we built the collections, and we illustrate how the collections can be used to characterize the performance of example systems.

2 RELATED WORK

NER has been shown to be particularly helpful when applied to specific domains. For example, Cowan et al. [3] show that in the travel domain, NER can allow users to search for classes of words such as AMENITY or STAR-RATING. Even when users enter more general queries, they note that extraction based on NER can be used to inform users of particular features of the returned documents. Thus there is evidence that domain-specific NER can be useful in retrieval applications.

Others have found that derivatives of NER can also be useful in certain applications. Notably, Khalid et al. [9] explore the role of named entity normalization (NEN) in question answering (QA). NEN is similar to NER in that it is a way to add metadata to token sequences; it differs in that the tag associated with a token sequence refers to the named entity itself rather than to its class. For example, "the States" and "USA" could both be tagged as UNITED_STATES_OF_AMERICA in NEN. They find that QA performance improves when using IR systems that incorporate NEN. This raises the question—how much of this improvement could NER, an arguably weaker yet more accessible technology, provide on its own?

Topic aspects, also sometimes referred to as subtopics, are present in a number of test collections. The TREC Interactive track relied

[1] The test collections are freely available at https://github.com/hltcoe/ner-for-ir-collection/.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3417452

Table 1: CLEF Topic Aspect 179-10 Query Set.

	Aspect	What political parties were involved in the resignation of the NATO secretary general?
Russian	BOW Query	партия генсек НАТО отставить
Russian	NER Query	ORG/партия генсек НАТО отставит
English Translation	BOW Query	party "secretary general" NATO resign
English Translation	NER Query	ORG/party "secretary general" NATO resign

on participants to detect implicit aspects of a topic [12]. The TREC Question Answering track made aspects explicit in the Complex Interactive Question Answering (ciQA) task [4]. The TREC Web track expanded that line of work by identifying two types of aspects (in the context of diversity ranking), one involving topic facets (our focus in this paper), and a second arising from query term ambiguity [2]. Our use of topic aspects in this paper is instrumental, simplifying the relevance judgment process. But our resulting test collections may also be of interest to those whose research calls for collections marked up for topic aspects in Chinese or Russian.

3 AUGMENTING TEST COLLECTIONS

This section describes our process for augmenting test collections to support experimentation with NER in retrieval applications. The topic aspect development process was followed by query generation, relevance judgment and automated NER tagging of the document collections. The Chinese collection was annotated by a single language expert, while the Russian collection was annotated by two language experts. Variations across annotators have been shown to have only small effects on comparisons between systems using aggregate measures [14]. As is common in information retrieval evaluation, we therefore use single-annotation.

3.1 Topic Aspect Generation

Instead of developing new topics from scratch, we built atop two existing test collections, leveraging their relevance judgments. Several of the topics for which relevance had been annotated in these collections implicated named entities in some way, although not always directly. For each such topic, we asked our annotators (fluent speakers of Russian or Chinese who were also fluent in English) to develop several topic aspects, each with a strictly narrower scope than that of the original topic. They then exhaustively judged all documents that had been marked in the original test collection as relevant to the original topic to determine the set of relevant documents for each aspect of that topic.

We selected two collections as starting points: the TDT-3 Chinese collection [7]; and the CLEF 2003/2004 Russian collection [1]. We chose the TDT-3 collection for Chinese because TDT topics are event-oriented, and we expected event-oriented topics to be a rich substrate from which NER-oriented topic aspects could be defined. The TDT-3 collection is also notable for being the world's largest fully-annotated information retrieval test collection, with an explicit human relevance judgment for every topic-document pair. Finally, TDT-3 is a bilingual collection, with topics described as long English narratives, and with relevance judgments for both English documents and Chinese documents. It was thus possible for annotators to examine relevant English documents when designing Chinese topic aspects without ever seeing the vocabulary used to express those concepts in Chinese.

The Russian CLEF 2003/2004 collection, like most test collections, was built using pooling; thus relevance judgments might be missing for some documents. We make the usual assumption that unjudged documents are not relevant, which has been shown to be suitable when system comparisons of averaged measures are used [15]; this limitation should be borne in mind when analyzing specific cases. As with the TDT collections, the CLEF collections have relevance judgments for English documents for the same queries; this allows topic aspect development to be performed without reference to the actual content of specific Russian documents.

The original topics from the TDT and CLEF collections are often broad, so the relevant documents for a given topic may address a wide variety of topic aspects. Annotators were asked to design topic aspects for which they expected to be able to later write queries that might benefit from the ability to refer to some entity type. Each topic aspect was represented as an English question. An example of a generated aspect is: "Which countries sent delegates to the 'women's conference' in Beijing?" This is an aspect generated for CLEF topic 143, "Women's Conference Beijing." This process generated a total of 234 topic aspects, 169 from 17 original CLEF 2003/2004 topics and 65 from 20 original TDT-3 topics.[2]

3.2 Query Generation

While it is common to think of topics (and in our case topic aspects) as a part of the test collection, query generation has often been implicit (e.g., in the standardized use of title and title+description queries in TREC). In our case, however, query generation was central to our goal because neither topics nor topic aspects contain named entity markup—that is the province of queries. The annotators' next task was therefore to generate a pair of queries for each subtopic. Our retrieval task is monolingual, so the annotators wrote their queries in the language of the document collection (Russian for CLEF subtopics and Chinese for TDT subtopics).

The annotators were instructed to generate two queries based on their own notion of how they might express a topic aspect, without detailed instruction in the design of any particular information retrieval system. One query was required to contain an explicit reference to a named entity type, while the other was required to be expressed without the use of any named entity type. Apart from the named entity reference, the queries were expected to be equivalent in intent. The annotators also translated each query into English for reference by researchers who were not fluent in Chinese or Russian. For the NER query in each query pair, we provided our annotators with definitions for the NER tags shown in Table 2, and

[2]The numbers for each language differ due to differences in annotator availability.

we defined for the annotators three ways in which those tags could be used:

> Entity types can stand alone (meaning that an entity of that type should be present) or they may be attached to a query term. Two forms of attachment are possible. | indicates that the entity type is intended to restrict the interpretation of the query term (e.g., PER|Bush refers to a person named Bush, and not to an airport named Bush and not to a shrubbery). / indicates that the query term refers to an entity of that type (e.g., PER/participant indicates that the participant is a participating person and not, for example, a participating organization).

Although we had expected bag-of-words queries (perhaps with quotes to indicate phrases), in practice our annotators freely imagined additional system capabilities, such as a Boolean AND. In all cases the NER tags are rendered entirely in upper case, using ASCII characters, so they are easily distinguished from the Chinese and Russian text, and easily readable in the English translations. Table 1 shows an example topic aspect.

3.3 Relevance Judgments

The annotators made relevance judgments for each of the subtopics using the documents from the original collection that are in the language of the queries (Russian for CLEF and Chinese for TDT). For efficiency, relevance judgments were made for each topic aspect only for documents that had been annotated as relevant to the corresponding topic in the original collection. This decision embeds two assumptions: (1) that the relevance judgment for the original collection were reasonably complete; and (2) that documents judged not relevant to the original topic would not be relevant to the topic aspect. For the first assumption, TDT-3 judgments were exhaustive. CLEF judgments were built using pooling, which is adequate coverage to support system comparisons [15]. To address our second assumption, we instructed annotators to design topic aspects that were clearly subsets of the associated topic so as to avoid the need to judge additional documents.

This resulted in 133 aspects from 17 original topics in Russian and 33 aspects from 9 original topics in Chinese, each of which had at least one relevant document. On average, a topic aspect has 5 relevant documents in the Russian collection and 22 in Chinese.

Table 2: Automatic NER Results. (Query tags show number of NER instances in queries with >0 relevant documents)

	Russian			Chinese		
Tag	F_1	CLEF Tags	Query Tags	F_1	TDT Tags	Query Tags
GPE	0.83	105,990	36	0.77	150,795	15
PER	0.81	31,411	41	0.83	71,699	32
ORG	0.50	8,685	25	0.29	6,047	9
TITLE	0.29	3,304	8	0.62	29,045	0
LOC	0.27	2,476	3	0.56	3,124	0
MONEY	0.00	1,496	2	0.99	12,584	0
DATE	0.05	855	5	0.80	15,783	7
COMM	0.27	790	0	0.58	13,340	0
MIL-G	0.10	571	1	NA[3]	216	0
MIL-N	0.38	476	0	NA	375	0
EVNT	0.53	369	3	0.36	7,169	0
POL	0.08	45	0	0.44	25,582	0
VEH	0.00	27	2	NA	1	0
GOVT	0.25	20	0	NA	644	0
TIME	0.00	19	0	1.00	1,212	0
FAC	0.00	5	3	0.06	2,745	0
COMP	0.10	3	0	NA	0	0
MISC	0.38	12,886	1	0.40	3,347	0
CHEM	NA	0	3	0.62	278	0

Table 3: Collection Statistics.

Lang	Original Docs	Filtered Docs	Original Topics	Judged Aspects	>0 Rel Aspects
Russian	16,716	9,313	17	169	133
Chinese	12,341	11,922	9	34	33

3.4 Document Tagging

We first tagged all of the documents by tokenizing and sentence splitting the text fields, and then processing each document using a newly-built language-specific NER system for our tag set. Our NER system uses a Bi-LSTM-CRF model. Like many sequence to sequence NER systems [8], this model includes a stacked bi-directional recurrent neural network with LSTM units and a CRF decoder. We combine this system with BERT [5], which is a stack of bi-directional transformer encoders. We keep the BERT frozen during training and testing, feeding the text into BERT and concatenating its final four layers as an input to our Bi-LSTM-CRF.

For the Russian NER system, we sentence split and tokenize the documents using CoreNLP [11]. The model is trained on 433 threads from Russian Reddit that had been labeled with the 19 tags shown in Table 2 inspired by the LDC tags [6]. This training set contains 11,000 named entity instances. Additional manually-annotated data was used to evaluate the model, which scored 0.65 F_1 on the test set. Table 2 includes the F_1 scores on the test set for each tag type. Details on this system's data set are available in Song et al. [13].

For the Chinese NER system, we sentence split using CoreNLP and tokenize at the character level. The model is trained on a collection of four days of the Renmin newspaper, which was labeled with the same tag set as the Russian documents. This training set contains 12,000 named entity instances. Additional manually annotated test data was used to evaluate the model, which scored 0.70 F_1 on the test set. Table 2 includes the F_1 scores on that test split at the type level; zeros are the result of the model finding no instances of that type in the test set. More details on this system's training data are available in Lawrie et al. [10].

3.5 Collection Overview

Statistical overviews for the collections can be found in Table 3. In that table, "Filtered Documents" shows the number of documents

[3] Although there were seven examples in the training partition, there was none in the test partition.

Table 4: Retrieval effectiveness (* for $p < 0.05$)

	MAP	
Queries	BOW	NER
All Russian	0.2667	0.2684
All Chinese	0.3412	0.3579
Chinese, no date constraint	0.3626	0.3886*

after filtering out those which could not be tagged. The relatively large difference between the original collection size and filtered documents for the CLEF collection reflects the fact that a substantial number of CLEF documents did not contain a "text" field, which is the field on which we performed tagging. "Judged Aspects" are aspects for which we have relevance judgments, and ">0 Rel Aspects" are the set for which there is at least one relevant document.

4 PRELIMINARY EXPERIMENTS

To demonstrate some ways in which the collections can be used, we have performed preliminary experiments comparing the effectiveness of Bag of Words (BOW) queries to that of NER queries.

We pre-processed the queries and the documents by tokenizing using Jieba[4] for Chinese and tokenizing and then lemmatizing using pymystem[5] for Russian. We also replaced each instance of named entity markup with an inline type tag immediately before the beginning of the named entity string (e.g., <type="PER">Richard Nixon</type> was replaced with PER Richard Nixon). In Chinese, due to possible mismatch with Jieba tokenization, the tag was inserted at the closest border between tokens. (e.g., Ri<type="PER">ichardNixon</type> is compared to tokenization: "Richard Nixon", and "PER Richard Nixon" is chosen over "Richard PER Nixon"). The resulting documents were indexed using Galago.[6] No stopword removal was performed in either language.

Each query was interpreted as a sequence of terms, treating an NER tag, when present, the same as any other term. For retrieval we used Galago's sequential dependence model, which rewards terms for appearing close to each other and in order. BM25 was used for term weighting. Results are reported as uninterpolated Mean Average Precision (MAP). Table 4 shows the results. Observed differences in MAP were not statistically significant at $p < 0.05$ by a two-tailed paired t-test. Of course, more sophisticated ways of using the NER annotations might show improvements over the baseline; our goal here was simply to illustrate one way in which the test collection might be used.

Aggregate measures like MAP hide many things. We therefore performed a more nuanced analysis by investigating the types of queries and aspects that improve with NER compared to those that perform worse. Several of the queries from the set that had adverse effects when adding NER included a "DATE" tag that was specialized to a specific year (e.g., "DATE|1998"). Instances of dates such as "1998" are likely to refer to a year, so adding the "DATE" tag means that imperfect NER tagging could adversely affect recall without a corresponding improvement in precision. In fact, removing the 7 topic aspects for which a "DATE" tag is used as a constraint from the Chinese query sets results in a statistically significant improvement for NER queries over BOW queries by a two-tailed paired t-test at $p < 0.05$ over the remaining 26 topic aspects. So NER tagging can be helpful, even with our very simple experiment design.

Interestingly, some dependence on the original topic is also evident. For example, the NER queries for the two topic aspects derived from Chinese TDT topic 30006 yielded Average Precision (AP) below that of the corresponding BOW query, whereas NER queries for 4 of the 5 topic aspects derived from topic 30009 yielded AP above the BOW query (the 5th yielded no difference). All four of these topic 30009 aspects sought documents about entities of a certain class related to other terms in the query, such as "What were the reactions of various nations to the policy?" The ability to use the more general "GPE" to refer to 'nations' instead of the more specific word "country" in the BOW query yielded better results.

5 CONCLUSION AND FUTURE WORK

We have augmented two existing test collections to support evaluation of the impact of named entity recognition on retrieval effectiveness. The test collections also include mappings between topics and topic aspects that may be useful for diversity ranking experiments and for other cases in which aspect analysis is called for. Our preliminary experiments illustrate a potential use case for the collections. We are making the augmented collections freely available for research use, including standoff annotations for named entities. Much remains to be done. Perhaps most obviously, our simple mechanism for adding NER tags to queries only begins to hint at the range of techniques that might be explored given the rich structure of the queries in this collection, including comparison of automatically generated NER queries to those that were added by our annotators. This paper takes the first step—describing how the collections were created and illustrating ways to use them.

[4]https://github.com/fxsjy/jieba
[5]https://github.com/nlpub/pymystem3
[6]https://www.lemurproject.org/galago.php

REFERENCES

[1] F Borri, C Peters, and N Ferro (Eds.). 2014. *Working Notes for CLEF 2004 Workshop.* CEUR Workshop Proceedings, Vol. 1170. CEUR-WS.org.
[2] C Clarke, N Craswell, and I Soboroff. 2007. Overview of the TREC 2009 Web Track. In *TREC*.
[3] B Cowan, S Zethelius, B Luk, T Baras, P Ukarde, and D Zhang. 2015. Named entity recognition in travel-related search queries. In *Twenty-Ninth IAAI Conference*.
[4] HT Dang, D Kelly, and J Lin. 2007. Overview of the TREC 2007 Question Answering Track. In *TREC*, Vol. 7. 63.
[5] J Devlin et al. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *CoRR* abs/1810.04805 (2018).
[6] J. Getman, J. Ellis, Z. Song, J. Tracey, and S. Strassel. 2019. Overview of linguistic resources for the TAC KBP 2019 evaluations. In *Text Analysis Conference*.
[7] D Graff, C Cieri, S Strassel, and N Martey. 1999. The TDT-3 Text And Speech Corpus. In *DARPA Broadcast News Workshop*.
[8] Z Huang, W Xu, and K Yu. 2015. Bidirectional LSTM-CRF Models for Sequence Tagging. *CoRR* abs/1508.01991 (2015). https://arxiv.org/pdf/1508.01991.pdf
[9] M Khalid et al. 2008. The Impact of Named Entity Normalization on Information Retrieval for Question Answering. In *ECIR*.
[10] D. Lawrie, J. Mayfield, and D. Etter. 2020. Building OCR/NER Test Collections. In *Proceedings of The 12th Language Resources and Evaluation Conference*. 4639–4646.
[11] C Manning et al. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *ACL System Demonstrations*. 55–60.
[12] P Over. 1998. TREC-6 interactive track report. In *TREC*.
[13] C. Song, D. Lawrie, T. Finin, and J. Mayfield. 2020. Gazetteer Generation for Neural Named Entity Recognition. In *The 33rd International FLAIRS Conference*.
[14] E Voorhees. 1998. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. In *SIGIR (SIGIR '98)*.
[15] J Zobel. 1998. How Reliable Are the Results of Large-Scale Information Retrieval Experiments?. In *SIGIR*.