A Testbed for Indonesian Text Retrieval

Jelita Asian Hugh E. Williams S.M.M. Tahaghoghi
School of Computer Science and Information Technology
RMIT University, GPO Box 2476V, Melbourne 3001, Australia.
{jelita,hugh,saied}@cs.rmit.edu.au

Abstract *Indonesia is the fourth most populous country and a close neighbour of Australia. However, despite media and intelligence interest in Indonesia, little work has been done on evaluating Information Retrieval techniques for Indonesian, and no standard testbed exists for such a purpose. An effective testbed should include a collection of documents, realistic queries, and relevance judgements. The TREC and TDT testbeds have provided such an environment for the evaluation of English, Mandarin, and Arabic text retrieval techniques. The NTCIR testbed provides a similar environment for Chinese, Korean, Japanese, and English. This paper describes an Indonesian TREC-like testbed we have constructed and made available for the evaluation of ad hoc retrieval techniques. To illustrate how the test collection is used, we briefly report the effect of stemming for Indonesian text retrieval, showing — similarly to English — that it has little effect on accuracy.*

Keywords Indonesian, queries, collection, relevance judgements, stemming

1 Introduction

The Text REtrieval Conference, TREC, began in 1992 in response to the need for a common testbed for evaluating Information Retrieval techniques, and the need to meet and discuss those techniques [3]. TREC is divided into *tracks*, and each track into *tasks*. A track investigates a retrieval paradigm, and each task a subelement of that paradigm. For example, the 2003 Web track investigates retrieval techniques for the Web, with two tasks that investigate topic and home (or named-page) finding. The original TREC track was *ad hoc*, proposed to investigate ad hoc searches for new topics in archived data.

The ad hoc paradigm is that used by most users of web search engines. A typical ad hoc query is a phrase or set of keywords that describe an information need, and the correct or *relevant* responses are those documents that meet that information need. The ad hoc paradigm is no longer explicitly investigated as a TREC track, but it is still widely investigated as a task in many tracks such as the genomics, terabyte, and robust tracks. Indeed, it is well-known that the ad hoc queries from earlier TREC conferences are the standard testbeds used by IR researchers developing new techniques.

Indonesia has the world's fourth largest population, with over 230 million people. However, very little research in IR has investigated techniques for Indonesian. It is important for Australian organisations to be able to retrieve, process, and monitor Indonesian for media, political, and intelligence applications. One possible reason that research in Indonesian IR has not flourished as well as other major languages is the absence of a publicly available Indonesian testbed. The Indonesian document collections that do exist [2, 5, 6] either do not have topics and relevance judgements, or are not published.

In this paper, we describe a TREC-like testbed for the evaluation of ad hoc Indonesian text retrieval. We have defined 20 ad hoc queries for which known answers exist in a collection of 3,000 newswire documents. For each query, we have exhaustively examined the documents and identified those that meet the information need as the set of relevant answers. To illustrate the usefulness of the testbed, we report experiments with a stemming scheme we have recently developed [1]. Our results show that stemming has little benefit for the accuracy of Indonesian ad hoc text retrieval.

2 Testbed Construction

A testbed for evaluating ad hoc retrieval requires three parts: a document collection, a list of query topics, and a set of *relevance judgements* [8]. This section explains how we constructed each component. Additional information and the testbed itself are available online[1].

2.1 Collection

A collection for evaluating ad hoc retrieval must be a static set of documents. Motivated by TREC, we obtained a collection of newswire articles by crawling daily news from the popular online Indonesian newspaper Kompas[2]. We retrieved 3,000 articles between January and June 2002 inclusive, leading to a collection of around 0.7 megabytes in size with 38,601 distinct words. Similar to TREC, we followed

Proceedings of the 9th Australasian Document Computing Symposium, Melbourne, Australia, December 13, 2004.

[1] http://www.cs.rmit.edu.au/~jelita/corpus.html
[2] http://www.kompas.com/

```
<DOC>
<DOCNO>news10513-html</DOCNO>
Mayjen Syafrie Samsuddin akan Jadi
Kapuspen TNI JAKARTA (Media): Mantan
Pangdam Jaya Mayjen Syafrie Samsuddin
akan menjadi Kapuspen TNI
menggantikan Marsekal Muda Graito
Husodo.Menurut informasi yang
diperoleh Antara Jakarta Kamis, Syafrie
Samsuddin menjadi Kapuspen TNI dan serah
terima jabatan akan dilakukan pada akhir
Februari 2002. Namun kebenaran
informasi tersebut hingga kini belum
dapat dikonfirmasikan ke Kapuspen TNI.
( M-1 )
</DOC>
```

Figure 1: *An example Kompas newswire document from our test collection, marked up in a TREC-like format.*

```
<top>
<num> Number: 14
<title> nilai tukar rupiah terhadap
dolar AS
<desc> Description: Dokumen harus
menyebutkan nilai tukar rupiah terhadap
dolar AS.
<narr> Narrative: Asalkan dokumen ada
menyebutkan nilai tukar rupiah terhadap
dollar tanpa indikasi menguat atau
melemah sudah dianggap relevan.
Prediksi nilai tukar dianggap
tidak relevan.
</top>
```

Figure 2: *An example topic. Its English translation — not included in the testbed — is shown in Figure 3.*

the principle of keeping the data as close to original as possible, and did not correct any faults such as spelling mistakes or incomplete sentences [8].

The collection of documents are stored in a single file, marked-up using standard TREC tags. The tags <DOC> and </DOC> mark the beginning and end of a document respectively, and each document has a document identifier delimited by the <DOCNO> and </DOCNO> tags. An example document is shown in Figure 1.

2.2 Topic Construction

The next step in building a testbed is to define a set of queries or *topics* that represent user information needs. (The words topic and query are used interchangeably in this paper.) There are different formats of TREC topics from different years of the workshops [7], with recent examples containing fewer fields. We followed the final ad hoc track format from TREC-8 [9].

The ad hoc topics from TREC-8 have three major fields: title, description, and narrative. The title (encapsulated in a <title> element) is a short title that summarises the information need. The description (<desc>) is a longer, one-sentence description of the topic. The narrative (<narr>) gives more detailed explanation that aims to completely describe which documents are relevant to the query. The topics also have the additional <top> and </top> tags to delineate each query in a file and a <num> element that is used to denote the query identifier.

The Kompas newswire is different in topicality and timespan to the newswire collections used at TREC. Therefore, we defined our own topics following the TREC approach. At TREC, candidate topics are brought by participants and the NIST TREC team decide the final topics by approximating the number of relevant documents per topic [7]. In our case, we began by reading all the 3,000 documents to see what topics were available. With only one native speaker involved in the project, we then limited ourselves to defining twenty topics represented in the collection. These topics have two major types: *general* and *specific*. A general topic is one for which many documents meet the information need. For example, many documents answer the query “Who is Megawati Soekarnoputri’s husband?” (topic 9). A specific topic has a small set of document answers and assessors need to read carefully to obtain the answer. An example specific query is “What are the symptoms and causes of asthma?” (topic 10).

An example of a formatted Indonesian topic is shown in Figure 2 and its English translation is shown in Figure 3. To follow the exact format of TREC, the English translations do not form part of the distributed testbed, but are available upon request.

2.3 Relevance Judgements

The final step in constructing a testbed is to make *relevance judgements*, that is, to define which documents are relevant to the information needs expressed by each query. The relevance judgements are then used as the benchmarks to decide whether the documents deemed to be relevant by retrieval systems are indeed relevant according to humans; we report such experiments as an example later.

In TREC, relevance judgements are normally performed though *pooling* [7]. Pooling works as follows: first, for each query, the top 100 documents returned from each *run* (experiment) performed by each participant are returned to NIST; second, a pool is created by grouping together all answers from each run and eliminating duplicates; last, human assessors read each document in the pool and assess its relevance to the topic. The drawback of pooling is that documents that are not retrieved by any system are considered not relevant. Relevance at TREC is binary: a document is either relevant or not, and there is no degree of relevance.

```
<top>
<num> Number:  14
<title> The exchange rate between
rupiah and US dollar
<desc> Description:  Document shall
mention the exchange rate of Indonesian
rupiah against US dollar.
<narr> Narrative:  The document is
relevant as long as it mentions the
exchange rate of rupiah against USA
dollar, even without indication whether
rupiah strengthened or weakened.
Exchange rate prediction is not
relevant.
</top>
```

Figure 3: *English translation of the Indonesian topic shown in Figure 2.*

The number of documents and queries in our testbed is limited, and we did not use pooling. Instead, each of the 3,000 documents was read and judged manually to see whether it was relevant to any of the twenty queries, resulting in an exhaustive tabulation of $20 \times 3,000 = 60,000$ relevance assessments. The relevance judgements are formatted in TREC-like format[3] as follows:

```
14 0  NEWS12738-HTML 0
14 0  NEWS12739-HTML 1
```

The first column indicates the topic number. The second column indicates feedback iteration and this is ignored. The third column is the document identifier appearing between the <DOCNO> and </DOCNO> tags in the documents. The last column indicates whether the document is relevant to the topic. In this example, document NEWS12738-HTML is not relevant to topic 14 and NEWS12739-HTML is relevant. Relevance judgements are ordered by topic, and by document identifier within each topic.

While this testbed is not comparable in size to the 475 megabytes of the TREC Disk 5 Los Angeles Times collection, it is a useful resource that can be extended with collaborative input from other researchers.

3 An Example Evaluation

This section shows an example application of our testbed to Indonesian IR research: evaluation of whether stemming — the removal of suffixes, infixes, and prefixes to derive the morphological root — aids retrieval effectiveness. To experiment with our testbed, we used the zettair search engine[4], which has native support for TREC collections, topics, and relevance assessment. We used a modified Indonesian stemmer that we have recently described [1] to preprocess the collection and topics, and compared this to searching with the unmodified data. A detailed discussion of our stemming approach appears elsewhere [1].

Measure	Without Stemming	With Stemming
Average Precision	0.4394	0.4801
Precision at 10	0.3750	0.3550
R-Precision	0.4210	0.4534

Table 1: *Performance before and after stemming.*

For both the stemmed and unstemmed data, we searched with the titles from the 20 queries on our collection of 3,000 documents, returning 100 answers per query. (The normal TREC practice is to return 1,000 answers but with only 3,000 in the collection we recommend a setting of 100.) We found — in unreported experiments — that title-only search is most effective of all possible combinations of fields. These answers were then evaluated against the relevance judgements using the trec_eval program[5], a standard package used by the TREC workshops to evaluate ad hoc tasks [9]. The software reports different interpretations of *recall* and *precision*: recall is the fraction of relevant answers retrieved, and precision is the fraction of answers that are relevant; recall requires that exhaustive relevance judgements be available, and can therefore be calculated using our testbed.

Table 1 shows the results of our experiment. The first row shows average precision, with stemming improving retrieval perfomance by around 4%. Average precision is calculated by determining the precision after each relevant document is retrieved, summing those precision values, and dividing by the number of relevant documents found; the overall average values for each query are then averaged over all queries. The second row shows average precision after processing 10 documents, averaged over all queries, and shows a drop in performance by 2% for stemming. The final row shows the precision for each query where the number of answers processed equals the number of relevant answers for that query, averaged over all queries. The R-precision results favour stemming by around 3%.

We used the Wilcoxon signed ranked test to examine whether the differences in performance are significant. They are not, at both the 95% and 99% confidence intervals. These results are perhaps surprising; Indonesian words have many more variants than those in English, and we expected that the removal of prefixes, infixes, and suffixes should improve retrieval performance. However, these results are consistent with those observed in English text retrieval [4].

To investigate further, Figure 4 shows the per-query performance. For each topic, three bars are shown: to the left, the total number of relevant documents; in the middle, the number of relevant documents found

[3] http://trec.nist.gov/data/qrels_eng/index.html

[4] http://www.seg.rmit.edu.au/zettair/

[5] The trec_eval program can be obtained from ftp://ftp.cs.cornell.edu/pub/smart

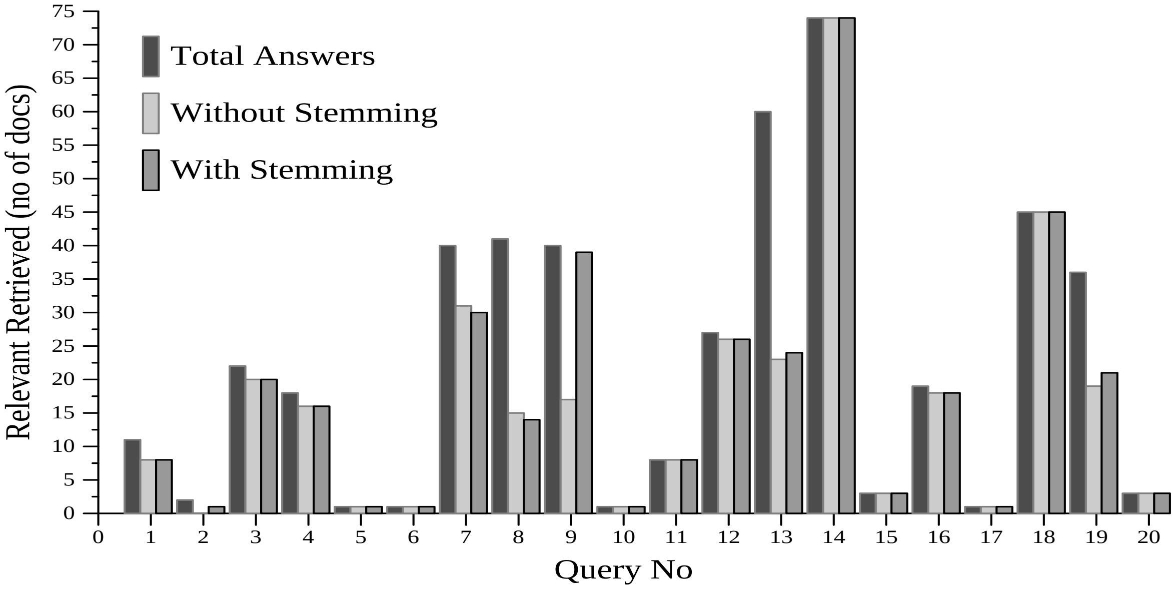

Figure 4: *Topic-by-topic performance with and without stemming. For each topic, the left column shows the number of relevant documents. The middle column shows the number retrieved without stemming, and right column the number retrieved with stemming. The queries used are only the titles.*

without stemming; and, on the right, the number of relevant documents found with stemming. The results show that — with the exception of topics 2, 9, and 19 — there is little difference between performance with and without stemming.

We suspect that this is because some relevant documents answer the query implicitly and do not contain the query terms. For instance the query for “nama bos Manchester United”(the name of the boss of Manchester United) does not retrieve one document that discusses “the manager of MU”. A human assessor understands that “manager” is a synonym of “boss” and “MU” is the acronym of “Manchester United”; automated retrieval systems generally use words directly from the query, and stemming is ineffective here.

4 Conclusion

In this paper, we have described the first testbed for Indonesian Information Retrieval. It includes 3,000 documents from newswire texts, 20 topics, and exhaustive relevance judgements. The testbed is stored in the TREC format, and can be used in TREC-like ad hoc evaluations with standard TREC retrieval and evaluation tools.

We have presented a brief experiment with the collection. The experiment shows that an accurate stemmer does not significantly aid retrieval performance on ad hoc queries, despite our expectations to the contrary. We intend to investigate stemming further as a result of this first evaluation. We also plan further fundamental work on techniques for Indonesian Information Retrieval using our testbed.

References

[1] J. Asian, H.E. Williams and S.M.M Tahaghoghi. Stemming Indonesian. In V. Estivill-Castro (editor), *Proceedings of the Twenty-Eighth Australasian Computer Science Conference (ACSC2005)*, 2005. To Appear.

[2] I. Fahmi, March 2004. Personal Communication.

[3] D.K. Harman. Overview of the First TREC conference (TREC-1). In *Proceedings of the First Text REtrieval Conference (TREC-1)*, pages 1–20. NIST Special Publication 500-207, 1992.

[4] D. A. Hull. Stemming Algorithms: A Case Study for Detailed Evaluation. *Journal of the American Society of Information Science*, Volume 47, Number 1, pages 70–84, 1996.

[5] F. Tala. A Study of Stemming Effects on Information Retrieval in Bahasa Indonesia. Master’s thesis, University of Amsterdam, July 2003.

[6] V. B. Vega. Information Retrieval for the Indonesian Language. Master’s thesis, National University of Singapore, July 2001.

[7] E. M. Voorhees and D.K. Harman. Overview of the Sixth TREC conference (TREC-6). In E.M. Voorhees and D.K. Harman (editors), *Proceedings of the 6th Text REtrieval Conference (TREC-6)*, pages 1–24. NIST Special Publication 500-240, 1997.

[8] E. M. Voorhees and D.K. Harman. Overview of the Eighth TREC conference (TREC-8). In E.M. Voorhees and D.K. Harman (editors), *Proceedings of the 8th Text REtrieval Conference (TREC-8)*, pages 1–24. NIST Special Publication 500-246, 1999.

[9] E. M. Voorhees and D.K. Harman. Overview of the Ninth TREC conference (TREC-9). In E.M. Voorhees and D.K. Harman (editors), *Proceedings of the 9th Text REtrieval Conference (TREC-6)*, pages 1–14. NIST Special Publication 500-249, 2000.