Quality Assessment of the Reuters Vol. 2 Multilingual Corpus

Robin Eriksson
University of Turku
era+rcv2@iki.fi

Abstract

We introduce a framework for quality assurance of corpora, and apply it to the Reuters Multilingual Corpus (RCV2). The results of this quality assessment of this standard newsprint corpus reveal a significant duplication problem and, to a lesser extent, a problem with corrupted articles.

From the raw collection of some 487,000 articles, almost one tenth are trivial duplicates. A smaller fraction of articles appear to be corrupted and should be excluded for that reason.

The detailed results are being made available as on-line appendices to this article.

This effort also demonstrates the beginnings of a constraint-based methodological framework for quality assessment and quality assurance for corpora. As a first implementation of this framework, we have investigated constraints to verify sample integrity, and to diagnose sample duplication, entropy aberrations, and tagging inconsistencies. To help identify near-duplicates in the corpus, we have employed both entropy measurements and a simple byte bigram incidence digest.

Keywords: corpus, quality assurance, quality assessment, methods, metrics

1. Introduction

Corpora – machine-readable collections of language materials – have an important role for the development of language technologies, especially for statistical or machine-learning techniques. Because corpora are created by humans, often under severe budget and timetable constraints, however, we often find that they contain errors. Depending on the goals and tools of the corpus researcher, even a modest amount of errors can be problematic.

In this paper, we study the Reuters Volume 2 Multilingual Corpus collection (NIST, 2005) from a quality assurance perspective. The intent is, on one hand, to attempt to document the quality and error rate of this standard corpus; and on the other, to propose improvements to the corpus itself, and to the collection process for similar corpora in the future. Finally, we provide a sketch for a general-purpose quality assurance method for corpora.

1.1. Motivation

The precision in state of the art text classification is approaching the level where a handful of individual classification errors in the evaluation corpus can threaten to dominate over the error rate of the filter in a technology evaluation. In other words, the result from an experiment with an ideal 100% accurate classifier would be reported as something like 99.5% because of classification errors in the corpus – thus, it is not implausible to think that a less accurate classifier could win an evaluation.

The lack of a published audit of the RCV2 corpus, and our personal observations of errors in the corpus on an anecdotal, nonscientific level, drive us to want to investigate the validity of the corpus, and, where possible, publish corrections for researchers who wish to exclude erroneous samples from the corpus. This should help the text classification and information retrieval community by establishing a reasonably correct, large, real-life corpus from a single domain as a good standard benchmark collection, and eventually, a proper gold-standard test collection.

Furthermore, by outlining and demonstrating our quality assessment method, we hope to be able to contribute to an improved practice for corpus quality assurance within the scientific community of corpus users.

As an aside, a byproduct of the evaluation is a discussion of some biases and peculiarities of the corpus, which could hopefully help future researchers save time on assessing the suitability of the corpus for their specific needs.

1.2. Method Overview

In a nutshell, the quality assurance method amounts to specifying constraints which all samples in the corpus – or, as the case may be, all samples of a particular kind – are required to satisfy. A sample which violates a constraint is suspicious, and needs to be further investigated.

Ideally, we would like to articulate constraints which can be checked automatically. While manual constraint checking is clearly also of value, tools for fully automatic checking never exhibit any subjectivity, and hence are easier to reason about objectively. Furthermore, as corpora grow in size, comprehensive manual review is frequently out of the question entirely, whereas automatic review methods can scale to collections of larger size and complexity.

Having said that, many useful constraints are unsuitable for completely automatic application, but can still be useful for guiding and partially automating manual quality assurance work.

Either way, we can establish a reasonable approximation of the error rate in the corpus by quantifying the constraint violations. This constitutes a quality assessment, and is a valid and often valuable effort in its own right. As we shall see, we can obtain a first approximation of the error rate in a corpus with a fairly modest amount of work.

Section 3 of this article contains a summary of how this general framework was applied in this particular implementation. The on-line Appendices contain the results in individualized form, as an enumeration of proposed corrections to the corpus.

1.3. Towards a Catalog of Corpus Constraints

In this particular case study, we will focus on constraints which are valid for and applicable to the Reuters Vol. 2 Multilingual Corpus. It is envisioned that future articles would explore how this method can be applied to other corpora, and use a similar presentation format to document additional constraints. Thus, we hope to contribute to a future catalog of best current practices for corpus collectors and users.

Appendix A is a sketchy but already somewhat useful outline of a catalog of corpus constraints.

2. Background

This section presents some previous research results, as well as the provenance of the RCV2 corpus.

2.1. Previous Work

While there is an abundance of articles about different specific methods such as duplicate removal, anomaly detection, etc, there is very little on the topic of quality assessment and validation of large data collections in general, or corpora in particular.

Apart from the work on validating the English-language Reuters corpora (see section 2.2 below), the focus of scholars so far seems to have been confined to validation of morphosyntactic annotations.

Eskin (2000), van Halteren (2000), and the publications of the DECCA project at Ohio State University – e.g. Dickinson and Meurers (2003a), Dickinson (2005), Dickinson (2006) – predominantly investigate methods for validating part-of-speech and, in one case (Dickinson and Meurers, 2003b), syntactic relationship tags. In the present framework, these can be summarized as implementations of a single constraint: **consistent tagging.**

2.2. The Reuters-21578 and LYRL Assessments

In 1990, a smaller collection known as Reuters-22173 was released by Reuters Corp. and Carnegie Group Ltd. (Sanderson, 1994) In the following years, the documents were reformatted and additional data files produced, and the collection was distributed between 1993 and 1996. (Sanderson, 1997)

A revision known as the Reuters-21578 corpus was the result of an informal but thorough cooperation among text categorization researchers starting in 1996. The work is detailed in a README file in the Reuters-21578 distribution. (Lewis, 1997)

The original RCV1 corpus was a significantly larger effort. It contains over 800,000 manually categorized newswire stories in English from 1996-1997 (Rose, Stevenson, and Whitehead, 2002).

Khmelev & Teahan noted in their paper (2003) that the RCV1 corpus contained 27,754 duplicates or near-duplicates and some 400 non-English documents. Their assessment merely reports this finding; their article documents their method for making this discovery.

Lewis et al. (2004) contains a detailed account of coding practices for stories at Reuters during the 1996-1997 period, and describes a number of aberrations and anomalies in the corpus. The paper has a substantial number of corrections which can be applied to the corpus; their revision is referred to as RCV1-v2.

2.3. The RCV2 Multilingual Corpus

Slightly later, a multilingual corpus, dubbed Reuters Corpus Volume 2 (RCV2), was released. It contains more than 487,000 newswire articles in thirteen languages from roughly the same time period as the English-language articles in the RCV1. (NIST, 2005)

3. Quality Assessment

The present work merely provides a blueprint for a full-blown methodological framework for corpus quality assessments. It is our hope that the results in this section will persuade the reader that the overall approach is sound and feasible, and produces tangible, actionable results already in this first crude prototype implementation.

3.1. Goals

As Lewis *et al.* point out in their assessment of the RCV1 corpus (2004),

> Existing text categorization test collections suffer from one or more of the following weaknesses: few documents, lack of the full document text, inconsistent or incomplete category assignments, peculiar textual properties, and/or limited availability.

An ideal corpus, then, ought to be large, comprehensive, consistently tagged, with well-documented encoding conventions, and freely available.

In reality, we face opposing forces; well-defined, quality-controlled collections tend to be small and/or very specialized (confined to a restricted domain, for example); large corpora tend to be drawn together from scattered resources with differing conventions both to content and presentation, and thus displaying a wide variation in nonessential features, making it cumbersome to run controlled tests, or challenging to draw general conclusions. Furthermore, availability may be restricted due to e.g. licensing reasons.

Then, of course, we have the well-established quality factors listed in any corpus linguistics textbook – selection bias, representativity, balance – as well as mundane processing artefacts like processing errors, encoding errors, tagging errors, etc.

A comprehensive quality assurance methodology should address all of these quality factors.

3.2. Sample Integrity

Our first broad constraint is that the corpus should contain a well-defined set of samples in a reasonable, consistent format.

We note the following problems:

- Odd file naming conventions. We have opted to use each sample's XML file name, which is a running number from 0 through 487394 (with a gap for the aforementioned 44 missing samples) as the primary identifier.
- No DTD or documentation for the XML container format. Some XML elements are empty, some duplicate information in the article text; conversely, the samples themselves contain some metainformation and some boilerplate which properly ought to be separated and explicitly marked up in the XML representation.
- Zip file format error in LATAM43.ZIP. Because of this, 44 samples could not be extracted. This error seems to be present in the corpus master files at NIST as well (Ian Soboroff, personal communication).
- Whitespace markup is inconsistent. Sometimes, <p> tags have been added where no semantical paragraph boundary exists. At the same time, paragraph boundaries and especially tables are often present with no explicit markup.
- The topic, region, and industry tags are nominally machine readable, but the format does not support or enforce any consistency. Indeed, there are many inconsistencies in these tags, both in their representation and in how they are applied. Some of this also deviates from the conventions used within the RCV1, which is doubly unfortunate.
- Language tags haphazardly indicate a different language for some articles, while many articles in an unexpected language are simply bulk tagged with a default tag. On-line Appendix D contains corrections for part of these inconsistencies, but does not attempt to properly correct them; we simply enforce an uninformative but consistent tagging policy.
- There are many articles in English, apparently often as a conscious publishing decision by the newsroom editor. These are generally not identified (wheras articles tagged as being in English generally are not).

Some further integrity problems are identified in subsequent sections.

3.3. Documentation

It appears that the (inferred) documentation for the Reuters Corpus Vol. 1 by Lewis *et al.* (2004) is also applicable to this collection with a few exceptions. The lack of documentation specifying the differences, if any, is a serious impediment for any user of this corpus, as is the absence of any documentation for the collection criteria and possible filtering made at the time the materials were compiled.

The file names in the distribution contain abbreviated directory names which suggest Russia, the Netherlands, Taiwan, Portugal, Spain, Denmark, Norway, Sweden, Latin America (sic; Argentina?), France, Italy, Germany, and Japan as the originating countries. We assume that the Reuters offices in these countries (simply called "locations" below) are the originators of these news stories.

We can speculate that the collection is simply a dump of all articles published by these 13 local Reuters offices during the time frame of August 1996 through July 1997, but additional documentation from Reuters would be most welcome.

3.4. Duplicate Detection

Extracting the bare sample text from the XML documents and calculating a SHA1 hash for each resulting text file revealed a surprising number of duplicates.

On-line Appendix C contains a listing of 48,665 samples which are duplicates of higher-numbered samples.

The choice to keep the highest-numbered sample in any set of duplicates is arbitrary. Article numbering within a set from the same location is roughly chronological, so we speculate that where duplicates differ in metainformation, the final issue will most frequently be a corrected version.

Only a minority of the articles with identical text had been retagged during the process of repeated republication, though. A mere 878 duplicates had topic tags which differed from the tags of the highest-numbered article with the same contents.

The relative frequency of duplicates strongly correlates with the size of the collection from that location. In other words, the bigger a location's collection, the higher the relative amount of published duplicates. Thus, perhaps the number of duplicates is a reflection of the hardships of coordinating efforts within a large, productive office.

3.5. Entropy

By measuring the information entropy of each article, we hope to be able to spot, on one hand, articles which contain atypical text, and on the other, articles which are very similar.

3.5.1. Entropy Measure

The concept of information entropy was defined by Shannon in his seminal text (1948). Informally, the entropy expresses the amount of predictability of a stream of bits. A low entropy corresponds to a predictable stream with a high amount of redundancy, while a high entropy corresponds to an unpredictable stream with a high amount of information.

Because the information entropy naturally tends to increase with longer documents, the raw entropy measurement was multiplied by a scaled length coefficient. This coefficient was set to the length of the current article divided by the average article length for this particular location. We call this measure the normalized relative entropy, or simply k.

3.5.2. Observations

Examining the extreme ends of the normalized entropy for the articles from each individual location reveals some problematic articles.

The articles with a low normalized relative entropy tend to be very short, terse telegrams. Some of them are so short as to constitute just a headline. There are also some articles, in particular in the Russian collection, which are incompletely translated from another language. Additionally, these had a much lower k value than most uncorrupted Russian telegrams. A few aberrant articles from some of

the other collections were isolated using this technique as well.
At the other end of the spectrum, the articles with the highest *k* values are often run-in articles, i.e. corrupted by some sort of database glitch or copy/paste error.
There are also articles throughout the collection which contain what we speculate must be vestiges of the original encoding from the underlying system. These manifest as sequences of "computer gobbledygook" resembling ~XD0`?~X98`?~XD0 etc. Based on patterns in the corrupted articles, we hypothesize that the originating system used ISO-2022 or some related encoding for hosting multilingual content, and that the translation from this representation into UTF-8 XML sometimes failed in mysterious ways.
These corrupted articles are listed in on-line Appendix E.

3.5.3. Near-Duplicates

Incidentally, we note that measuring the entropy on different levels measures different distributional patterns. The bit entropy measures the overall bit distribution – a bit stream 01010101 has the same entropy as the unevenly distributed 11110000 at this level of examination, while e.g. the nybble entropy for these two streams is quite different.
Articles with exhibit the same entropy measurements on multiple levels of examination (we have examined the entropy per bit, nybble, byte, and Unicode code point) are likely to be near-duplicates.
However, because it is impossible to draw the line between an accidental duplicate and an intended one with any precision, we cannot determine whether these are problem samples which should be removed from the corpus.
For example, the same measurements often co-occur in daily trading reports which only differ by the date and the rates. The rest of these articles are boilerplate text which constitutes intentional repetition. Without access to actual exchange rates etc, it is impossible to judge whether near-repeats are due to typographical or editing errors, or to the repeating nature of these reports.
Furthermore, it is in the nature of newsroom work to update stories about unfolding events. There are sets of articles in the collection which illustrate how a short telegram about an important event is followed by slightly longer telegrams with more or updated information. It is not only feasible but even likely that the corpus could be useful for investigating this particular aspect of reporting. Hence, no attempt has been made to remove earlier versions of unfolding stories.
This is a departure from the policy of Lewis *et al.,* who systematically pruned articles which are a proper prefix of another article from the RCV1 collection. (Lewis et al., 2004)

3.6. Tagging Consistency

Where there are duplicates or near-duplicates, we can investigate whether they have been tagged consistently. The logic is simple: Articles with substantially similar content ought to have the same tags. (Cf. also Eskin (2000), van Halteren (2000), Dickinson and Meurers (2003a), who however investigate consistency for tagged words, not category tags for articles.)
However, it is not always possible to infer the correct tags even from a substantial collection of duplicate articles.
Be that as it may, 2,607 articles violate the basic tagging policy that each article should have at least one region tag and one topic tag. (One article has neither; 2,283 articles lack region tags, and 323 articles lack topic tags.)
Out of the region tags, we can also observe as a minor aside that some tags are present both with a Z suffix and without one. Out of these, generally, the one without the suffix is incorrect. For example, there are 711 articles with the tag WEUR and 262,672 articles with the correct tag WEURZ.
Like in the RCV1 (Lewis et al., 2004), it is apparent that the official Reuters policy to mark up every leaf tag with the intermediate tags between the root tag and the leaf has not been consistently observed.
Proper review of the tags would require extensive knowledge of the tag set and tagging policy, the languages of the articles, and world and local events in 1996-1997. It is far outside the scope of this article, and outside the expertise available to the author.

3.6.1. Byte Bigram Type Similarity

In addition to the duplicates from the SHA1 investigation and the near-duplicates revealed by the entropy investigation (above), we grouped 26.9% of the articles into clusters based on lexical similarity.
The algorithm is simple: Slide a two-byte window over each sample, noting at every position whether the byte pair in the window has been seen before. We do not perform a count; we simply build a matrix of byte pairs which occur at least once in this document. These matrices can be compared for two documents, producing a measurement of lexical similarity.
(This measurement tends to break down with really long samples, but seems to work well for the samples in this collection. We set the clustering threshold at 65% similarity.)
For example, the biggest cluster in the Russian collection – 252 articles – contains reports of the Ruble exchange rate towards the German Mark and the US Dollar. (It would be nice to have the two currencies in separate clusters, of course.) The next largest cluster, comprising 148 articles, covers the exchange rate for the Ukrainian Hryvnia against the Ruble and the Belarusian Ruble against the US Dollar.
In general, the resulting clusters seem to be useful. Especially the largest clusters are helpful for identifying articles with similar content where you cannot necessarily rely on superficial indicators like headlines and topic tags to be consistent over time. The Reuters corpus contains many recurring articles such as daily stock market summaries which would be challenging to identify and isolate by any other means.
Having said that, some of the smaller clusters are false groupings where basically unrelated pairs of articles have been grouped together because the crude measure of lexi-

cal similarity judges them to be related even though they are not very similar by other measurements.

3.6.2. Observations

Drilling down into these clusters of lexically similar articles, 48,311 articles in 3,135 clusters were selected as somewhat likely to be incorrectly tagged. This is a very rough estimate, but seems plausible as a lower bound for the error rate in tagging.

In order for an automatic assessment to be feasible, the cluster had to be large enough to have a substantial volume of identically tagged samples. Out of samples with more than 20 members, we singled out samples whose tags differed from a majority tag set for the whole cluster assigned to at least 80% of the samples in that cluster.

Manual inspection revealed that there were situations where samples with a different set of tags were in fact apparently correctly tagged (for example, where a routine currency exchange summary had a brief commentary about a particular currency), but if anything, the heuristic we used selected too few, not too many, incorrectly tagged samples with a reasonable precision (albeit an abysmal recall).

Thus we expect this rough number to be too low, but useful as a starting point for further investigations.

4. Conclusions and Future Directions

The experimental results from the previous section allow us to draw conclusions about several aspects of Volume 2 of the Reuters Corpus as well as our experimental methodology for quality assessments.

4.1. Corpus Issues

While the quality assessment unearthed a number of problematic samples in the collection, it remains – by definition – a representative sample of international newswire from 1996-1997.

The majority of the quality problems are authentic, in the sense that they are a necessary, if not unavoidable, consequence of the production setting of a newsroom. You cannot expect "gold standard" proofreading, editing, or category tagging in an environment where the primary success metric is the speed of delivery.

Having said that, a quote from Lewis *et al.* (2004) seems well worth repeating:

> Use of this data for research on text categorization requires a detailed understanding of the real world constraints under which the data was produced.

The documentation by Lewis *et al.* for the RCV1 corpus contains valuable background information about the editorial processes at Reuters at the time the corpus was created, and most of their findings appear to be applicable to RCV2 as well.

However, while the RCV1 documentation is helpful, there are still many undocumented features in the corpus. Reconstructing the semantics of undocumented XML fields this long after the fact is challenging, and finding informants who are able to recollect editorial practices from almost two decades ago is no easier.

The variability in quality between locations is also an indication that editorial policy may have differed between offices, making it harder or impossible to reason about the collection in general terms. Part of the issue may also be the difference between topic subject priorities at different locations. Some offices specialize in reporting financial news, while others have a clear ambition to report sports results or local events as well. This may be a reflection of the competitive landscape in each location. It might simply not make sense for a Reuters office to spend resources on covering local events if all potential customers already subscribe to a well-established high-quality newswire from a local competitor such as a national news agency.

The high error rate in some of the subcorpora is slightly troubling. For example, an experimental result for classifying the German collection could have an under- or over-reported result of fifteen percentage points just because of the high amount of trivial duplicates in that collection.

For comparison, Lewis *et al.* found "between 2,500 and 30,000" duplicates in RCV1, but conceded that the amount of duplicates was insignificant. Based on a ceiling of 30,000 out of approximately 800,000, the duplication ratio is at most 3.375% in RCV1. Similarly, the amount of (discovered and reported) non-English documents in RCV1 is a measly 400 (0.05%).

As detailed above, the overall duplicate rate in the RCV2 corpus is 9.98% (48,665 articles). The English articles in Section 3.2 total 2,171 (0.45%) but this number may be too low, and should perhaps also include at least a significant fraction of the outright corrupted articles (351 articles in on-line Appendix E; again, probably under-reported) for a total upper bound of 0.52%. Altogether, the problematic samples exceed 10% of the entire RCV2.

4.2. Summary of Techniques

Ultimately, it would be useful to have a detailed catalog of constraints, where each constraint can be straightforwardly checked using a simple dedicated tool. Meanwhile, devising checks for each constraint involves multiple ideas and tools.

Iterative application of these techniques helped identify some error sources, allowing for the successive development of a library of search expressions for finding additional problematic samples to exclude or at least investigate, even when the techniques above did not directly unearth these samples.

In the broadest possible terms, we can define the following – perhaps obvious – high-level techniques.

- Duplicate detection. Identical samples – perhaps only identifiable after some preprocessing – should have identical metadata. In some collections, identical duplicates are undesirable; this collection belongs to this set.
- Similarity clustering. Similar samples should often have identical metadata. This is much broader and more complex than constraints based on identity, but can still be quite fruitful.
- Metadata clustering. Conversely, samples with similar metadata can be expected to be similar in content by

some metric. Where this is not the case, outliers can be selected for investigation.

While the goal of this methodology is to relieve us from manual inspection as far as possible, it is arguably unavoidable in the end. In fact, these experiments were originally devised based on observations during informal browsing of the corpus.

In practice, all of the fragmented and corrupt samples in on-line Appendix E were found by manual search, however directed by samples which violated a constraint. So, for example, many articles with low entropy turned out to contain corruption; many of them had markers which could then also be found in other articles with higher entropy which were also corrupt.

While methodologically, we would like to relegate manual inspection as a technique to only use as a last resort, it should in practice also be the very first technique a researcher deploys. Many a flawed experiment could have been avoided or corrected in time if the scholar had been more familiar with his or her material at the outset.

4.3. Final Conclusions

This paper demonstrates the feasibility of performing a quick quality assessment on a corpus using fairly simple tools. The entire effort, including writing this paper, was completed in about three months; with proper tools and preparations, the budget could probably be squeezed down to a few person-weeks or even person-days for a rapid assessment.

While we have obtained a rough estimate of the error rate and types of errors in the RCV2 corpus, significant work remains before it can be properly useful as a standard test collection.

Simultaneously, we have demonstrated that the proposed framework for constraint-based quality assessments offers a general, potentially lightweight foundation for corpus quality work. The catalog of constraints is still simple and crude, but already suggests additional experiments and new constraints to explore.

4.4. Future Directions

This publication is the first in a planned series of corpus quality case studies. Our plan is to extend and expand the catalog of corpus constraints by attempting to apply the experimental methodology to diverse and varied corpora.

The Reuters corpus is specifically a text categorization / information retrieval corpus. It is easy to see how additional constraints could be articulated for other corpus types – treebanks, dialogue corpora, recorded speech, etc. At the same time, some of the constraints here are so general that they can readily be applied and potentially already add value to these other corpus types.

4.5. On-Line Results

Detailed on-line appendices are published simultaneously with this article at http://github.com/rcv2/. Appendix A is a first draft of a general constraint catalog, whereas Appendices C through E enumerate the problematic samples, with a summary in Appendix B.

4.6. Acknowledgements

This work was made possible by a generous grant from the Kone Foundation.

Thanks to Filip Ginter for pointers to relevant research, and to Ian Soboroff and David D. Lewis for patiently responding to my emails.

Bibliographical References

Dickinson, Markus (2005). Error Detection and Correction in Annotated Corpora. PhD thesis, The Ohio State University. http://linguistics.osu.edu/sites/linguistics.osu.edu/files/dissertations/dickinson05.pdf.gz.

——— (2006). From Detecting Errors to Automatically Correcting Them. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL-06)*, 265–72. Trento, Italy. http://www9.georgetown.edu/faculty/mad87/papers/dickinson-06.html.

Dickinson, Markus, and W. Detmar Meurers (2003a). Detecting Errors in Part-of-Speech Annotation. In *EACL '03: Proceedings of the Tenth Conference of the European Chapter of the Association for Computational Linguistics*, 107–14. http://www.aclweb.org/anthology/E/E03/E03-1068.pdf.

——— (2003b). Detecting Inconsistencies in Treebanks. In *Proceedings of the Second Workshop on Treebanks and Linguistic Theories (TLT 2003)*, 45–56. Växjö, Sweden. http://ling.osu.edu/~dm/papers/dickinson-meurers-tlt03.html.

Eskin, Eleazar (2000). Detecting Errors Within a Corpus Using Anomaly Detection. In *Proceedings of the 1st North American Chapter of the Association for Computational Linguistics Conference*, 148–53. Association for Computational Linguistics. http://acl.ldc.upenn.edu/A/A00/A00-2020.pdf.

Khmelev, Dmitry V., and William J. Teahan (2003). A Repetition Based Measure for Verification of Text Collections and for Text Categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 104–10. SIGIR '03. New York, NY, USA: ACM. doi:10.1145/860435.860456.

Lewis, David D. (1997). Reuters-21578 Text Categorization Test Collection. http://www.daviddlewis.com/resources/testcollections/reuters21578/readme.txt.

Lewis, David D., Yiming Yang, Tony G. Rose, and Fan Li (2004). RCV1: A New Benchmark Collection for Text Categorization Research. *Journal of Machine Learning Research* 5: 361–97. http://www.jmlr.org/papers/volume5/lewis04a/lewis04a.pdf.

Rose, Tony, Mark Stevenson, and Miles Whitehead (2002). The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In

Proceedings of the Third International Conference on Language Resources and Evaluation, 827–32. http://www.lrec-conf.org/proceedings/lrec2002/pdf/80.pdf.

Sanderson, Mark (1994). Reuters Test Collection. In *Proceedings of the 16th BCS IRSG Colloquium*. http://www.seg.rmit.edu.au/mark/publications/my_papers/BCS_IRSG_94.pdf.

——— (1997). *Duplicate Detection in the Reuters Collection*. TR-1997-5. Glasgow G12 8QQ, UK: Department of Computer Science at the University of Glasgow. http://eprints.whiterose.ac.uk/4571/1/Duplicates.pdf.

Shannon, Claude E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal* 27 (3). Institute of Electrical & Electronics Engineers (IEEE): 379–423. doi:10.1002/j.1538-7305.1948.tb01338.x.

van Halteren, Hans (2000). The Detection of Inconsistency in Manually Tagged Text. In *Proceedings of the COLING-2000 Workshop on Linguistically Interpreted Corpora*, 48–55. Centre Universitaire, Luxembourg: International Committee on Computational Linguistics. http://www.aclweb.org/anthology/W00-1907.

Language Resource References

NIST (2005). “Reuters Corpus, Volume 2, Multilingual Corpus, 1996-08-20 to 1997-08-19 (Release Date 2005-05-31, Format Version 1, Correction Level 0).”

Appendices

Appendices are being made available on-line on GitHub; please see https://github.com/rcv2/rcv2r1/