The Reuters Corpus Volume 1 -

from Yesterday's News to Tomorrow's Language Resources

Tony Rose, Mark Stevenson, Miles Whitehead

Technology Innovation Group
Reuters Limited, 85 Fleet Street, London EC4P 4AJ
{tony.rose, mark.stevenson, miles.whitehead}@reuters.com

Abstract

Reuters, the global information, news and technology group, has for the first time made available free of charge, large quantities of archived Reuters news stories for use by research communities around the world. The Reuters Corpus Volume 1 (RCV1) includes over 800,000 news stories - typical of the annual English language news output of Reuters. This paper describes the origins of RCV1, the motivations behind its creation, and how it differs from previous corpora. In addition we discuss the system of category coding, whereby each story is annotated for topic, region and industry sector. We also discuss the process by which these codes were applied, and examine the issues involved in maintaining quality and consistency of coding in an operational, commercial environment.

1. Introduction

Reuters is a leading global provider of financial information, news and technology to financial institutions, the media, businesses and individuals. It is also the world's largest international text and television news agency, with over 2,000 journalists, photographers and camera operators in 190 bureau, serving 151 countries. On a typical day, Reuters Editorial produces some 11,000 stories in 23 languages, along with approx. 600 pictures, some 23 hours of video and dozens of graphics.

The Reuters Corpus Volume 1 is an archive of 806,791 English language news stories that is freely available to the research community[1]. It includes all English language stories produced by Reuters journalists between 20/8/1996 and 19/8/1997. The stories are formatted using a consistent XML schema that is based on an early version of NewsML[2] - an open standard conceived within Reuters that has since been developed through the International Press Telecommunications Council (IPTC)[3]. This presents clear advantages for researchers, such as easier access to the content and an increased potential for the development of standardized tools for the manipulation and transformation of the data.

A key aspect of this schema is the extensive use of descriptive metadata, whereby all the stories are fully annotated using category codes for topic, region and industry sector. This metadata represents many hours of editorial effort and constitutes a unique and valuable resource, particularly to members of the machine learning community. However, the value of this metadata is inevitably governed by the consistency with which it is applied. Evidently, the coding process may never be perfectly consistent – it is difficult to produce perfectly consistent annotations, particularly when complex coding schemes are involved (Carletta, 1996; Carletta et. al., 1997; Cleverdon, 1991).

One of the major goals of this paper is therefore to document the principles and practices used in applying codes to each of the stories in RCV1. Ideally (for a group of Reuters employees), such a process should be a straightforward exercise: you simply ask the individuals who were directly involved to describe what they did. However, in practice, things are rarely that simple: employees move on, practices change, and purposes change (at the time the corpus data was produced few of those involved could have envisaged that it would be subsequently re-purposed as a text corpus for scientific research). Consequently, at the time of writing (March, 2002), no formal specification of the coding practices exists. However, with appropriate detective work, it is possible to combine related documentation with interviews of key personnel to create a cohesive, reliable account of the process. This paper embodies the major results of that investigative work.

2. Motivations

In the late 1990s, Reuters coding operations were subject to two divergent market forces. Firstly, far greater quantities of information were becoming available through a variety of channels and media. As a result of this 'information overload', additional manual effort was needed to provide richer metadata for more accurate search and filtering. However, competition in the marketplace meant that production costs had to be controlled, and this required a greater degree of automation in the coding process.

At the time, a rule-based categorization system known as 'TIS' (Topic Identification System) was in use. However, the rule-based approach had several drawbacks:

- Creating rules required specialized knowledge, which made it difficult to add new codes or adapt to changes in the input.
- The rules did not provide any indication of the confidence in their output, so there was no way of focusing editorial efforts on the most uncertain cases, nor any way of indicating that new topics were appearing in the stories that would require changes or additions to the code set.

[1] Further details are available from http://about.reuters.com/researchandstandards/corpus/

[2] http://www.newsml.org

[3] http://www.iptc.org

These issues and a number of operational factors mitigated against the further development of TIS, and it was becoming clear that a totally new approach to categorization was required. It was also apparent that any new solution would have to take into account factors such as maintenance overheads, durability, and the ability to accommodate new topics appearing in the data. Nonetheless, the primary concern was still the ability to apply codes accurately, and to measure this the company needed a collection of suitable test data.

One solution, therefore, was to create a corpus of stories coded to some benchmark standard. This process would have involved pairs of editors independently applying codes to a common set of stories, and then identifying and discussing any inconsistencies that emerged. An iterative process would ensure that the coding practices eventually converged. Evidently, this process happened to a certain degree as part of the editors' everyday interaction, but not in such a way that any "benchmark stories" could be differentiated from the overall operational output.

Moreover, building a substantial benchmark corpus using this approach would have taken considerable time, and ultimately proved too labour-intensive for an operational environment. Nonetheless, it was clear that the company would benefit from having a large corpus of training data for future evaluations, so the construction of the Reuters Corpus began.

3. RCV1 and previous corpora

Interestingly, although RCV1 is the first "official" Reuters corpus, it is not the first time that Reuters news stories have been used for research. An earlier collection of stories, known as the "Reuters-21578" collection, has been available from a public web site for many years[4]. This corpus is an adaptation of the older "Reuters-22173" corpus that consisted of 22,173 Reuters newswire stories dated from 1987. Reuters-21578 has proved an extremely popular resource and has been used in numerous studies. Indeed, it is estimated that this corpus has provided data for over 100 published research papers, particularly in the fields of information retrieval, natural language processing and machine learning (eg. Joachims, 1998; Nigam, 1998; Yang, 1999). This is an indication of the value of such corpora and Reuters has since received many requests to provide further data for research purposes, but up to now we have generally been prevented from doing so due to legal/copyright restrictions. A further motivation behind the new Reuters Corpus was thus to provide a standardized collection, suitable for research purposes, with the minimum possible restrictions.

Despite its popularity, Reuters-21578 does have a number of significant disadvantages, particularly that of overall size (only 21,578 documents). By contrast, RCV1 is some 35 times larger. In addition, Reuters-21578 covered only a fraction of a year, with somewhat inconsistent coverage of that time period. By contrast, RCV1 covers a complete year of editorial output, allowing the investigation of temporal issues such as topic detection & tracking (Wayne, 2000), conflict prediction (Bond et al, 1997) or financial market forecasting (Giles et. al. 1997). In addition, RCV1 was created from a news archive product (i.e. a database) rather than a raw newswire feed, which helps to ensure consistency (since there should be fewer duplicates, corrections, brief alerts, etc.)

However, one advantage that the older collection does possess is that much effort has been applied in identifying suitable training/test splits for various applications, particularly those of text categorization and machine learning (e.g. Lewis, 1992). Work on the new Corpus is only just beginning in this respect, but it is hoped that much of what was learned with Reuters-21578 will be of value in future studies.

At the time of writing, the new Corpus has been supplied to 242 separate organisations, of which 120 are academic institutions and 110 commercial organisations (the remainder being government organisations, individuals, etc.). These are distributed across 40 different countries - with USA having the highest number of applicants (75) followed by the UK (43) and Germany (21). Interestingly, many of these early requests were generated almost entirely by word-of-mouth, before any formal publicity or promotion had taken place. Many of these early applicants were TREC participants, since the Reuters Corpus was the official dataset for the 2001 filtering track (Robertson and Callan, 2001).

4. Coding the Reuters Corpus

4.1. The Coding Scheme

All the stories in RCV1 have been coded for topic, region (geography) and industry sector. This coding scheme was designed to enable effective retrieval from database products such as Reuters Business Briefing (RBB). The stories cover a range of content typical of an international English language newswire, and can vary from a few hundred to several thousand words in length. The data is available on two CD-ROMs and has undergone a significant amount of verification and validation of the content (i.e. removal of duplicates and other spurious entries, normalisation of dateline & byline formats, addition of copyright statements, etc.). An example story is shown in Appendix A (with some truncation of the metadata to conserve space).

4.1.1. Topic codes

The topic codes represent the subject area(s) of the each story. They are organized into four hierarchical groups, with 4 top-level nodes: Corporate/Industrial (CCAT), Economics (ECAT), Government/Social (GCAT) and Markets (MCAT). The code set was designed originally around requirements of business information professionals, although this was broadened to include the needs of end users in large corporates, banks, financial services, consultancy, marketing, advertising and PR firms.

The file *topic_codes.txt* on the RCV1 CD lists 126 codes. However, not all of these were used in the coding of the corpus. For example, there is a set of 11 codes labelled as 'current news', but these and the 2 codes marked as 'temporary' are unused in the corpus data. In addition, a further 10 codes (G11 to G14, plus GEDU and MEUR) also appear to be unused. Therefore, the total

[4] Currently available from: http://www.daviddlewis.com/resources/testcollections/reuters21578/index.html

number of codes actually assigned to the data is 126 - 11 - 2 - 10 = 103.

Under each top-level node there is a hierarchy of codes, with the depth of each implied by the length of the code. For example, the Corporate/Industrial taxonomy is structured as follows:

CCAT (Corporate/Industrial)
➔ C1
➔ C15 ('Performance')
➔ C151 ('Accounts/Earnings')
➔ C1511 ('Annual Results')

However, it should be noted that the nodes at the single-digit level (e.g. C1-C4, E1-E7, G1 and M1) never existed as operational codes, and were therefore never assigned to the data (even though their existence is implied by the hierarchy).

When the stories in RCV1 were coded, the principle was to apply the most specific (i.e. most granular) code that was relevant in each case. However, editors were free to assign any of the codes they felt appropriate (i.e. not just leaf nodes), so in practice they sometimes applied a high level code if none of the more granular codes was appropriate. In addition, there was a principle that at least one topic code and one region code should be assigned to each story. However, in practice there are 2,364 documents that have no topic codes, and a further 13 that have no region codes (a total of 2,377 documents, or 0.29% of the entire corpus). The authors are currently working with the RCV1 user community to provide an appropriate resolution for this (e.g. a script to automatically remove them from the collection).

Once each story had been coded, each individual topic code was automatically 'expanded', such that all its ancestors in the hierarchy would be added as well (e.g. in the example in Appendix A the code E11 requires ECAT to be added, and M11 and M12 require MCAT to be added).

4.1.2. Industry codes

The industry codes were used to indicate the different types of business or industry referred to by each story. There are 870 codes listed in the file *industry_codes.txt*, and they are also arranged as a hierarchy, with the depth of each node implied by the length of the code. However, only the 6- or 8-character variations were intended to be assigned to stories, since the intermediate length codes (i.e. 2-5 characters) are simply the product of a legacy editing system. So for example, of the following codes, only the last (i.e. the 6 character version) is intended for assignment to documents:

I0 AGRICULTURE, FORESTRY AND FISHING
I00 AGRICULTURE, FORESTRY AND FISHING
I000 AGRICULTURE, FORESTRY AND FISHING
I0000 AGRICULTURE, FORESTRY AND FISHING
I00000 AGRICULTURE, FORESTRY AND FISHING

There were therefore only 376 such codes intended for actual use (ignoring the codes labelled TEMPORARY or DUMMY). As with topic codes, once each story had been coded, each individual code was automatically 'expanded', such that all its ancestors in the hierarchy would be added as well. However, the RCV1 data retains only the expansions of 8-character nodes to their 6-character ancestors, and no further. In addition, some 6,771 documents are missing even this degree of expansion (i.e. they possess an 8-character industry code but no 6-character ancestor). The authors are currently working with the RCV1 user community to provide an appropriate resolution for both of these issues (e.g. a script to automatically add the missing expansions).

4.1.3. Region codes

The region codes are used to indicate the geographical regions referred to in a story. They can be thought of as representing three groups:

- Countries (e.g. UK)
- Geographical groups (e.g. BENELUX)
- Economic groupings (e.g. GSEVEN)

However, there is no explicit hierarchical structure to the region codes, and hence (unlike the topic and industry codes) no automatic expansion was performed.

There are 366 region codes listed in the file *region_codes.txt*. However, a further 3 have been found to appear in the corpus: CZ - CANAL ZONE (1 occurrence), CZECH - CZECHOSLOVAKIA (2 occurrences) and GDR - EAST GERMANY (1 occurrence). These codes were erroneously assigned and should be replaced by PANA (PANAMA), CZREP (CZECH REPUBLIC) and GFR (GERMANY) respectively.

4.2. The Coding Process

During the years 1996 and 1997 Reuters produced just over 800,000 English language news stories per year. The process by which these were coded involved a combination of auto-categorization, manual editing, and manual correction. The coding team consisted of around a dozen people working full time (on shifts). The details of the process are as follows:

4.2.1. Auto-coding

First, stories were passed through TIS (the rule based categorizer). TIS contained rules for the majority of the codes in the code set. However, it was believed that the application of certain codes would be difficult to automate completely - in particular, the codes 'GODD' (human interest) and 'GOBIT' (obituaries) were suspected as being beyond the capabilities of most machines. Consequently, no rules existed for these; they had to be applied manually.

In addition to these rules, a number of simple source-processing heuristics were applied that attempted to map existing codes (such as those applied by Reuters Editorial when the story was first filed) to the equivalent codes in the RBB codeset. For example, a story with the Editorial code 'SPO' (Sport) would automatically receive the RBB code 'GSPO'. Similarly, there were heuristics based on other document metadata, e.g. if the slug line of an article contained the string "BC-PRESS DIGEST" then it would automatically receive the highest level General News code (GCAT).

4.2.2. Manual editing

As outlined in Section 4.1.1, it was a principle of the coding process that each story should receive at least one

region code and at least one topic code. Therefore, the next stage after the application of TIS was to check each story to see whether it satisfied this requirement. If so, the story was sent directly to a holding queue (see Section 4.2.3). If not, the story was sent to a human editor. This editor would then assign to the story all codes they felt were appropriate, ensuring that the story received at least one topic code and one region code. They were also free to delete or modify some of the automatically assigned codes. Once this manual editing was complete, the story was sent to the holding queue for final review.

4.2.3. Manual Correction in the Holding Queue

Every 6 hours the contents of holding queue were reviewed by a further editor, whose responsibility was to correct any outstanding coding errors. Finally, once this was complete and the stories had passed through the holding queue, they were batched up and loaded onto the RBB database in blocks.

4.3. Coding Statistics

Since all stories passed through the holding queue, it can be argued that every story in the collection was manually coded, in the sense of having the automated coding checked by at least one editor. Moreover, stories that violated any coding principle (e.g. those lacking at least one topic code and region code) were reviewed by at least two editors, the first of whom always added or changed at least one of the original TIS-assigned codes. This process of manual review represents a significant investment in the maintenance of data quality standards. In addition, further quality control procedures were applied, whereby each month a senior editor would take a sample of stories and assesses them for quality of coding, as well as language, punctuation, spelling etc. The outcome of this process was fed back into the system, and the editors notified of any errors.

Table 1 summarizes, for the year 1997, how many stories were manually edited and how many were manually corrected in the holding queue[5]. The middle line shows that a total of 505,720 stories went straight from TIS to the holding queue (bypassing the manual editing stage), and 334,975 of these (66.2%) were subsequently manually corrected. By contrast, the lower line shows that a total of 312,140 stories went from TIS via manual editing to the holding queue, but only 23,289 (13.4%) of these were subsequently corrected. It is possible that some of this difference could be attributed to the fact that the editors making manual corrections could see which stories had been auto-coded and which had been manually edited, but it nonetheless provides a significant degree of confidence in the degree of consistency between human editors.

	Uncorrected	Corrected
Unedited	170,745	334,975
Edited	288,851	23,289

Table 1: Numbers of stories edited and/or corrected

[5] Note that RCV1 contains stories spanning parts of 1996 and 1997, so the number of stories in the corpus is not the same as the number of stories in Table 1

5. Measuring inter-coder consistency

A fundamental feature of many real-world categorization schemes is that the definition of codes can be inherently quite imprecise, and as such open to interpretation by the various individuals that apply them. Various studies have shown that there can be considerable variation in inter-indexer agreement for different data sets (Bruce and Weibe, 1998; Brants, 2000; Veronis, 1998). In Reuters case, each editor may have a slightly different understanding of the concepts to which each code refers, and this can lead to inconsistencies in their application. Clearly, it would be of great benefit if some quantitative measure of inter-coder consistency could be applied to the RCV1 data. Evidently, the ideal approach would be to compare each story against some benchmark standard, such as that discussed in Section 2. However, even in the absence of such a resource, it was still possible to measure two aspects of coding consistency, using metadata present in the original RBB source files (i.e. a superset of the data that eventually became RCV1).

When a human editor opens a story, the action is recorded by adding a flag to the 'COMPRO' field of the file. The first letter after the colon is used to indicate a correction (C) or an edit (E). Thus a COMPRO field containing the data 'ED:ETA' indicates a story that had first been through TIS (by default) and was then edited by an editor identified by the letters TA. Likewise, an article with the COMPRO field 'ED:ETA ED:CBY' indicates a story initially coded by TIS, then edited by editor TA and subsequently corrected in the holding queue by editor BY. It is assumed that the last editor shown (i.e. the last to make any changes) is responsible for the final coding of any given story.

One approximate measure of coding consistency is to calculate how frequently an individual editor's coding is corrected. Note that in this context *any* change to the coding of a story counts as a correction (rather than counting corrections on the basis of individual codes). Since editors were equally likely to be first or second to see a given story, the correction rate for a given editor can be calculated thus:

NE = Number of stories coded by a given editor

NF = Number of stories for which a given editor applies the final coding

NC = Number of times an editor is corrected by a second editor, i.e. NE – NF

Correction Rate C = (NC/NE) * 100

The results are shown in Table 2, sorted by C. Note that this data refers to the original RBB source files, i.e. a superset of the data that eventually became RCV1. Editor E101 is TIS, which generally gets corrected around 77% of the time. Since TIS was never solely responsible for an article (i.e. every story was subsequently reviewed by at least one human editor), this was not considered unduly problematic. It should also be noted that editor E3 and was not an active BIP coder, and that editors E71, E73 and E91 were still undergoing training at the time and hence were expected to have higher correction rates.

Based on this data, the average correction rate for manual editors is 5.16%, i.e. slightly more than 1 in 20 stories. However, this figure is likely to be an upper bound

on the true error rates, since some coders were known to open stories when checking them rather than just viewing the stories, which would mean that their initials would be added to the story and flagged as a correction when in fact no changes took place.

EDITOR	NE	NF	NC	C
E101	806804	182782	624022	77.35
E3	153	120	33	21.57
E73	6384	5183	1201	18.81
E91	4602	3744	858	18.64
E71	13653	11995	1658	12.14
E13	36605	33636	2969	8.11
E4	51686	48211	3475	6.72
E9	20608	19241	1367	6.63
E1	47559	44721	2838	5.97
E2	53811	50648	3163	5.88
E7	62179	58840	3339	5.37
E24	46869	44605	2264	4.83
E6	45247	43086	2161	4.78
E0	55030	52473	2557	4.65
E15	53749	51408	2341	4.36
E20	32266	30883	1383	4.29
E8	43290	41440	1850	4.27
E5	42154	40615	1539	3.65
E11	44039	42784	1255	2.85

Table 2: Correction rates for each editor

A second measure of consistency is to compare the distribution of codes applied by each editor, in order to find any evidence of systematic bias. Since all editors were equally likely to code a given story (i.e. they coded all types of story rather than specializing in a particular area), a simple way to measure this is to count the number of stories to which a given editor applied the final coding (call this N), then count the number of times each code appears on those stories, then divide each code count by N.

Once we have calculated the frequency distributions, we can then measure the consistency between the editors by comparing each editor against the mean of all the others in the group. For example, if there were ten editors, then E1's distribution would be compared with the average of distributions E2 to E9, etc. Consistency may then be measured using a simple rank correlation, which produces a value of +1 for perfect consistency and -1 for complete inconsistency (Manning and Schutze, 1999).

The results are shown in Table 3, sorted by correlation. Whilst it is difficult to identify an ideal a-priori value for consistency, the relative degrees of correlation are nonetheless revealing. The mean correlation across all 19 coders is 0.986 and their standard deviation is extremely low at 0.018. It is clear that editor E3, who was not an active BIP coder, has the lowest correlation. The next lowest is TIS (editor 101), although even for this the correlation is still within 0.95 of the group average. The third lowest is editor E73, who was one of those undergoing training at the time.

However, it should be noted that the measures applied here remain somewhat approximate, in that they do not consider the potentially important differences in coding that would appear if we were to compare different editors' choices for an individual story. Nonetheless, it does provide an interesting further insight into the consistency of the RCV1 coding procedures.

EDITOR	CORRELATION
E3	0.922
E101	0.955
E73	0.973
E9	0.985
E11	0.986
E15	0.989
E71	0.989
E91	0.989
E13	0.990
E1	0.992
E5	0.992
E8	0.992
E0	0.993
E20	0.993
E2	0.995
E7	0.996
E24	0.996
E4	0.997
E6	0.997

Table 3: Correlation rates for each editor

6. Conclusion

This paper has described the Reuters Corpus (RCV1), and has attempted to outline some of the ways in which it represents an improvement over previous corpora such as Reuters-21578. We have described the RCV1 category codes in considerable detail, and have outlined the process by which these codes were applied to the corpus data. In addition, we have described the background to the creation of RCV1 and the business motivations behind its

release. Moreover, we have attempted to measure the degree of inter-coder agreement of the RCV1 data, and have presented two approximate measures that suggest a high degree of coding consistency.

However, the approach to coding embodied in RCV1, based on TIS and manual correction, has since been superseded. Reuters has since moved on to adopt statistical categorization techniques, in which the rules are induced from large amounts of training data, and an inbuilt feedback loop is used to initiate the involvement of human editors and analysis tools to decide when new training data or topic codes are required.

Reuters is currently considering the possibility of releasing other volumes of data. In particular, we hope to be able to compile a multi-lingual corpus, containing non-English language stories from the same period as RCV1. This would constitute a comparable corpus that would hopefully be of use in the development of multi-lingual applications such as machine translation systems and cross-language information retrieval systems. In the longer term, we plan to investigate the possibility of providing corpora based on non-text media, such as an image corpus or a news corpus composed of composite stories (i.e. text and associated images). Evidently, photographic images do not decrease in value in the same way as textual news stories, so the commercial implications of such an initiative are likely to be somewhat more involved. In this respect, we actively encourage suggestions from the research community regarding the type of corpora that would most effectively serve their current and future needs.

7. References

Bond, D., Jenkins, J., Taylor, C. and Schock, K 1997. Mapping Mass Political Conflict and Civil Society: Issues and Prospects for the Automated Development of Event Data. *Journal of Conflict Resolution*, 41,4:553-579.

Brants, T., 2000. Inter-annotator Agreement for a German Newspaper Corpus. *Proceedings of Language Resources and Evaluation Conference 2000*, Athens, Greece.

Bruce, R. and Weibe, J., 1998. Word Sense Distinguishability and Inter-coder Agreement. *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-98)*, Granada, Spain.

Carletta, J., 1996. "Assessing Agreement on classification tasks: the kappa statistic" *Computational Linguistics* 22,2:249-254.

Carletta, J., Isard, A., Isard, S., Kowtko, J., Doherty-Sneddon, G., and Anderson, A., 1997. The Reliability of a Dialogue Structure Coding Scheme. *Computational Linguistics* 23(1), 13-31.

Cleverdon, C. 1991. The Significance of the Cranfield Tests of Index Languages. *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-12.

Giles, C., Lawrence, S. and Tsoi, A. 1997. Rule Inference for Financial Prediction using Recurrent Neural Networks. *Proceedings of the IEEE/IAFE Conference on Computational Intelligence for Financial Engineering*, Piscataway, NJ.

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Proceedings of the ECML-98, 10th European Conference on Machine Learning*, Chemintz, Germany.

Lewis, D., 1992. An evaluation of phrasal and clustered representations on a text categorization task. *Proceedings of the Fifteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37-50, 1992.

Manning, C. and Schutze, H, 1999. *Foundations of Statistical Language Processing*. MIT Press, Cambridge, MA.

Nigam, K., 1998. Learning to Classify Text form Labelled and Unlabelled Documents. *Proceedings of AAAI-98, 15th Conference of the American Association for Artificial Intelligence,* Menlo Park, CA.

Roberson, S. and Callan, J., 2001. Guidelines for the TREC-2001 Filtering Track – Version 1.0 http://trec.nist.gov/data/t10_filtering/T10filter_guide.htm

Veronis, J., 1998. A Study of Polysemy Judgements and inter-annotator agreement. *Programme and Advanced Papers of the Senseval workshop,* Herstmonceux Castle, England.

Wayne, C. 2000. Multilingual Topic Detection and Tracking: Successful Research Enabled by Corpora and Evaluation. *Proceedings of Language Resources and Evaluation Conference (LREC) 2000*, 1487-1494.

Yang, Y., 1999. An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval,* 1(1/2):67-88.

8. Acknowledgements

This paper owes much to the work of Chris Harris, who performed the original analysis of the inter-coder consistency described in Section 5.

This paper has also benefited greatly from discussions with Dave Lewis, whose unfeasibly large appetite for detail on RCV1 never ceases to amaze us. We are also grateful for the input and efforts of other Reuters personnel (past and present), notably Trevor Bartlett, Dave Beck, Chris Porter, Jo Rabin, Richard Willis and Andrew Young.

9. Appendix A


```
<?xml version="1.0" encoding="iso-8859-1" ?>
- <newsitem itemid="2286" id="root" date="1996-08-20" xml:lang="en">
<title>MEXICO: Recovery excitement brings Mexican markets to life.</title>
<headline>Recovery excitement brings Mexican markets to life.</headline>
<byline>Henry Tricks</byline>
<dateline>MEXICO CITY</dateline>
- <text>
<p>Emerging evidence that Mexico's economy was back on the recovery track sent Mexican markets into a buzz of excitement Tuesday, with stocks closing at record highs and interest rates at 19-month lows.</p>
<p>"Mexico has been trying to stage a recovery since the beginning of this year and it's always been getting ahead of itself in terms of fundamentals," said Matthew Hickman of Lehman Brothers in New York.</p>
<p>"Now we're at the point where the fundamentals are with us. The history is now falling out of view."</p>
<p>That history is one etched into the minds of all investors in Mexico: an economy in crisis since December 1994, a free-falling peso and stubbornly high interest rates.</p>
<p>This week, however, second-quarter gross domestic product was reported up 7.2 percent, much stronger than most analysts had expected. Interest rates on governent Treasury bills, or Cetes, in the secondary market fell on Tuesday to 23.90 percent, their lowest level since Jan. 25, 1995.</p>
<p>The stock market's main price index rallied 77.12 points, or 2.32 percent, to a record 3,401.79 points, with volume at a frenzied 159.89 million shares.</p>
<p>Confounding all expectations has been the strength of the peso, which ended higher in its longer-term contracts on Tuesday despite the secondary Cetes drop and expectations of lower benchmark rates in Tuesday's weekly auction.</p>
<p>With U.S. long-term interest rates expected to remain steady after the Federal Reserve refrained from raising short-term rates on Tuesday, the attraction of Mexico, analysts say, is that it offers robust returns for foreigners and growing confidence that they will not fall victim to a crumbling peso.</p>
<p>"The focus is back on Mexican fundamentals," said Lars Schonander, head of researcher at Santander in Mexico City. "You have a continuing decline in inflation, a stronger-than-expected GDP growth figure and the lack of any upward move in U.S. rates."</p>
<p>Other factors were also at play, said Felix Boni, head of research at James Capel in Mexico City, such as positive technicals and economic uncertainty in Argentina, which has put it and neighbouring Brazil's markets at risk.</p>
<p>"There's a movement out of South American markets into Mexico," he said. But Boni was also wary of what he said could be "a lot of hype."</p>
<p>The economic recovery was still export-led, and evidence was patchy that the domestic consumer was back with a vengeance. Also, corporate earnings need to grow strongly to justify the run-up in the stock market, he said.</p>
</text>
<copyright>(c) Reuters Limited 1996</copyright>
- <metadata>
- <codes class="bip:countries:1.0">
+ <code code="MEX">
- <codes class="bip:topics:1.0">
+ <code code="E11">
+ <code code="ECAT">
+ <code code="M11">
+ <code code="M12">
+ <code code="MCAT">
</codes>
<dc element="dc.publisher" value="Reuters Holdings Plc" />
<dc element="dc.date.published" value="1996-08-20" />
<dc element="dc.source" value="Reuters" />
<dc element="dc.creator.location" value="MEXICO CITY" />
<dc element="dc.creator.location.country.name" value="MEXICO" />
<dc element="dc.source" value="Reuters" />
</metadata>
</newsitem>
```