Duplication in Corpora

Nadjet Bouayad-Agha and Adam Kilgarriff
Information Technology Research Institute
University of Brighton
Lewes Road
Brighton BN2 4GJ, UK
email: {first-name.last-name}@itri.bton.ac.uk

Abstract

We investigate duplication, a pervasive problem in NLP corpora. We present a method for finding it that uses word frequency list comparisons and experiment with this method on different units of duplication.

1 Introduction

Most corpora contain repeated material. In sampled corpora like the Brown Corpus, duplication is not so much of an issue, since the linguistic data is carefully selected proportionally by genre and thus the risk of introducing unwanted duplication is reduced. However, the typical corpus used in NLP is one in which as much data as possible of the desired genre is gathered. The result is a corpus whose nature and content is rather unknown. This issue has not, to our knowledge, been previously discussed in the literature.

While we may expect the repeated occurrence of words or expressions to reflect their use in the language, the repetition of longer stretches of printed material (section-, paragraph- or even sentence-length) most likely do not. Text processing technology allows writers to cut and paste any length of text. Text duplication arises for many reasons: the newspaper that reproduces an article from a weekday edition in the week-end edition, the famous quote that gets cited in every paper of a research community, the warning message that appears at the top of every instruction manual, etc. This is all valid corpus data. However, data duplication can be critical for corpus statistics.

We present a method for finding duplicated material and we evaluate it against a corpus of Patient Information Leaflets (PILS). PILS are those inserts that accompany medicines and which contain information about how to take the medicine, the ingredients, contra-indications, side-effects, etc. The corpus was compiled for a text generation project whose aim was to generate PILS in multiple languages. This means that some of the duplicated material might be used as canned text in our generator.

The PILS corpus is presented in the next section with a brief evaluation of its duplication and the problems it can pose. Section 3 presents a method for finding duplication using word frequency lists and section 4 reports on experiments looking for duplications in the PILS corpus.

2 The Corpus

The source of the corpus is the ABPI[1] (1997) Compendium of PILS. It consists of 546 leaflets (650,000 words) organised by company. There are over 50 companies. The leaflets are characterised by the following:

- Restricted domain. This domain can be divided into further sub-domains according to the type of medicine (such as patches, inhalers, etc.).
- Compliance to some legal requirements about content, order and terminology.
- Use of a house-style particular to a company.

The first two points mean that some conventional ways of saying things have developed, some of which are simply just canned text (e.g. *Keep out of the reach of children*). The last point means that, especially when they are from the same sub-domain, the leaflets from a particular company look as if created from the same template (if not simply copied and modified).

The corpus was made available electronically by OCRing, editing it in Microsoft Word 97 and automatically converting it to HTML format.

The extent of duplication in the corpus was encountered as a problem while we were investigating paragraph initial third person pronouns.[2] We searched for paragraphs containing such pronouns. On inspecting the results, we discovered that a large number —around 40%— were repeated occurrences. In this typical linguistic investigation of a corpus, we were only interested in analysing the different contexts in which the searched

[1] Association of British Pharmaceutical Industry

[2] The aim was to find a constraint on cross-paragraph pronominal reference that we could use in our PILS automatic generator.

(a) Keep your Elixir tablets at room temperature (below 20C) away from sunlight.
(b) Keep your tablets at room temperature (below 20C) away from direct sunlight.

Word Count for (a)		*Word Count for (b)*		*Diff*
20C	1	20C	1	0
at	1	at	1	0
away	1	away	1	0
below	1	below	1	0
		direct	1	1
elixir	1			1
from	1	from	1	0
keep	1	keep	1	0
room	1	room	1	0
sunlight	1	sunlight	1	0
tablets	1	tablets	1	0
tempeature	1	temperature	1	0
your	1	your	1	0
Word Sum	**12**	*Word Sum*	**12**	**2**

Table 1: Using Word Frequency Lists

string occurred. Its multiple occurrence within the same context did not give us a measure of linguistic competence in the text genre we were dealing with.

3 Word Count for Finding Duplication

We used a simple and computationally inexpensive method for finding duplication. We produce word frequency lists for all the units to be compared, and then compare these lists for all pairs.

Table 1 exemplifies the method. Columns 1 and 2 are the word frequency lists for sentences (a) and (b) given above the table. Column 3 is the word count difference between each word, which sums up to 2 whereas the total word count sums up to 24, which makes the ratio (2/24) below 10%. This means that the two units are redundant.

We uses the 10% threshold to allow a very limited variation in uses of words, where for instance, only the name of the medicine changes. However, it is low enough so that, although order information is thrown away, it is likely that these word frequency lists comparisons will reveal duplicates.

Leaflet Pair		*Word Difference*	*Word Sum*
Ast:Pulmicort_Respules_0.5mg	Ast:Pulmicort_Respules_1mg	18	2070
Ben:Amoxil_Capsules	Ben:Amoxil_Sachets	118	1586
Sch:Femodene	Sch:Logynon_ED	399	4571

Table 2: Document Duplication

4 Evaluation of the Method

4.1 Document Duplication

We first applied the method to whole documents. We compared each pair of leaflets (150,000 comparisons).

The results show that 81 of the 546 leaflets compared are redundant. In other words, about 15% of the corpus is redundant.

Leaflet pairs which were duplicates were all from the same company and of the same type of medicine. Table 2 gives three example of these duplicate pairs, each of which are from the same company (*Ast*ra, *Ben*card and *Sch*ering respectively). As the names indicate, the first pair is the same medicine with different strengths, the second pair is the same medicine in different forms and the third pair have different names but are the same type of medicine (i.e. contraceptive pills).

This level of redundancy is still low in comparison with the 40% found in section 2.

4.2 Section Duplication

The input text had been put in electronic form in such a way that the logical division of text was not reflected in the HTML format. Before we could proceed, a program that identified sections had to be written.

4.2.1 The Sectioning Program

The sectioning program proceeds by first identifying the headings. These are typically marked in the HTML representation as paragraphs with often some special typographic features such as bold weight or centered alignment. These features as well as punctuation (such as no ending punctuation, or question mark with a different font) and some other special properties such as the size of the string (less than 15 words) are used to identify the headings.

Next, the levels of nesting are to be decided. This is done simply by starting off with a top node, the document itself, and then maintaining a stack of distinct headers, where each newly encountered header is considered to be on the same level as the previous one if it has the same typographic property; otherwise it gets embedded in the previous division. The result is an SGML marked-up document conforming to a TEI-like DTD.

We formally evaluated the results by computing the recall and precision for the sectioning and nesting of about 10 randomly selected leaflets. Whereas the sectioning precision was about 75%, the nesting precision was below 50%. Indeed, the program is not based on design principles such as size or the higher position of centered headings over non-centered ones. The reason is that the layout of the leaflets is very heterogeneous. Typically, this program only works well when the document has a tree-like highly embedded structure.

4.2.2 The Duplication Experiment

Once the sectioning program has run through all the files, we automatically selected only the top-level sections in the hierarchy.[3] The word frequency script is then run through those section-files, and the sections with less than five words are discarded (those are the sections where the sectioning and the nesting did not work properly).

This produces 4936 section-files which allow for over 12 millions comparisons. The results show that 1207 of the 4936 divisions are redundant, or about 25% of the divisions.

To evaluate whether the results were really duplicates or whether they were simply sections using the same words but in a different manner, we manually looked at 10 section duplicates with a threshold between 8% and 10% and whose word sum is higher than 50, and we found that, in all these cases, the text was essentially duplicated, with minor differences such as difference in names of the medicines or addition/omission of modifiers.

The section duplication performs better than the document duplication. All the duplicated leaflets had duplicated sections, but only 12% of the leaflets had all their sections duplicated. In addition, 9% of the sections duplicates were found across companies.

25% is still below the 40% duplicated patterns found in the case study

[3] Because of the low precision in nesting mentioned in the previous section, if less than four sections were found per document, the section-finding program searches for sections at further levels of nesting.

(section 2). This is probably due to the performance of the sectioning program.

In order to test the validity of the section duplication, we compared it with window-size duplication.

4.3 Window Duplication

We divided every document into a window of 50 words (plus until the end of the orthographic sentence). This gives over 11,000 window-files which were all compared against each other (over 60 million comparisons).

The results show that 1714 of 11218 windows are redundant, which is about 15% of the total number of windows and well below the section duplication results.

5 Summary

Duplication in corpora is a pervasive problem. It would be useful to identify it in order to firstly, understand better the nature of the corpus we are dealing with, secondly, exclude it from analyses where, for example, statistics are computed, and thirdly, have access to the repeated chunks as objects of interest in their own right (for example, to be used as canned text in text generation).

We presented a method for finding duplication based on word frequency lists comparisons. This method allowed us to identify a substantial proportion of our corpus which was redundant.

We found that, for the highly structured documents we are dealing with in our corpus, the *section* was an appropriate unit to look for duplication over the fixed-size window of words.

References

ABPI Compendium (1997), *Compendium of Patient Information Leaflets.* Association of the British Pharmaceutical Industry.

TEI-P3 (1997), *Guidelines for Electronic Text Encoding and Interchange.* Vol. I and II. C.M. Sperberg-McQueen and Lou Burnard (eds). ACH, ACL and ALLC.