Universiteit Leiden
The Netherlands

Opleiding Informatica & Economie

Originality in news media compared to university press

Jim Laros

Supervisors:
Suzan Verberne & Ionica Smeets

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

15/07/2019

Abstract

In this thesis our goal was to see whether news media copy their stories directly from the university press they use as a source. This made our research question as follows: "To what degree do news media directly copy official university press articles?". To get an answer to this question, we used text comparison metrics on our data set, using Python 3. The most important text comparison metric being Rouge-L, with the Dutch data set scoring an average of 54.93%, which is very high. The English data set scoring a 46.52%, which is still an indication of copying. These results imply that indeed a lot of news sources actually directly copy at least large parts of their source texts.

Contents

1 Introduction

1.1 The situation

There is an interesting phenomenon going on in the media, where news websites and newspapers write articles about (medical) discoveries using the press release by the university. It seems very common that these news sources make use of these texts more than just as an inspiration. In this research we aim to get an answer, or at least come close to an answer for the question:
"To what degree do news media directly copy official university press articles?".

To get closer to answering that question we are using text similarity measures to compare a large amount of texts at the same time, making this easier than just individually looking at all the different news texts. How that is done exactly will be explained later on in this thesis. The texts that we are comparing in this thesis are discussing medical scientific results or a combination of the medical and psychological field. This is because the general public is generally very interested in medical advancements since they could apply tips or read about possible cures for diseases they, or friends and family, might have. In the next subsection we explain where the different topics of this thesis can be found.

1.2 Thesis overview

This thesis will start off by explaining some definitions and defining our scope in the next section. In the second chapter we will be discussing some related work of the different theories used in the actual python script. After that we continue with the third chapter in which we will be discussing the data that is used and the preprocessing involved to make the data suitable for this research. In the same chapter we will also be explaining the different methods that have been used to calculate measures to use while analysing the research questions. After this large chapter it is time for the results, these will be shown in chapter 5, Analysis and results. In this chapter the results of the different tests will be shown and quantified. The last chapter provides our conclusion about the research question and also discusses some further research that could be done.

1.3 Purpose of the research

This makes us get to the next point. What do we want to achieve by doing this research. To determine this we have come up with a research question and 2 sub questions.

- Our research question is as follows: To what degree do news media directly copy official university press articles?
- Sub question one: Do certain media outlets copy more than others?
- Sub question two: What are good rating criteria to determine if an article is copied or not?

To answer these questions we will need to use text similarity metrics. We will be using different packages inside of python to use these similarity metrics, and what these metrics do will be explained in the next chapter.

1.4 Deliverables of the thesis

The entire project consists of 2 different deliverables, the thesis itself, and a python script. The thesis contains results that we acquired by using the different functions in the script on the data that we have. The thesis explains which data is used, and how the different functions in the script are built up, and why they are useful for our research questions. Second of all, the script is also an important part of the project, because it has to be quite easy to use, to make sure future research could be done using this script. This means it has to be use-able on new data, also by people who aren't very familiar in the field of study. That way it can actually be useful for studies in the field of linguistics and other comparison/translation text studies. The script will be delivered with a manual to ensure correct usage.

2 Related Work

In this section we will be showing and reviewing some useful work that has already been done in this field of study and we will explain some basics of the metrics we will be using. These metrics are text comparison metrics, BLEU and Rouge to be exact. We will be using Rouge-L specifically for this research. The reviewing of useful work is based on a research by Sumner et al. [1]. We do this because we use the same dataset as the one that is used in that research. We will review some articles they used, and some other articles that also used Sumner's research as an inspiration.

2.1 Text comparison metrics

In this paper we are using multiple types of text comparison measures. Some of these are quite simple, such as literal word overlap comparison. This is a function where we look at each word and look if the word is in both of the texts. There are also some more complicated text comparison metrics that have already been used in different researches. Some of these will be listed below, and are also explained further there.

2.1.1 BLEU

BLEU is one of these text comparison metrics, it is an algorithm that is used to test the quality of translated texts [2]. In this case we classify quality as follows, the closer the machine translated text comes to a text translated by humans, the better it is. BLEU (BiLingual Evaluation Understudy) is one of the most commonly used metrics to test machine translations on accuracy. The formula BLEU uses is as follows, $P = m/w_t$. Where m equals the amounts of words in the candidate sentence that are also in the reference sentence, and w_t equals the total length of the reference sentence. This is the most basic function and the metric does some additions to make sure a accurate score is provided. It also uses a bigram comparison metric where m becomes the amount of bigrams that are in the candidate.

BLEU is reported to often recognise good translations, thus correlating well with human judgement. But BLEU calculates its score mainly by looking at single sentences and is therefore less suited for our problem. This is because we want to know copying (or similarity) on a larger scale than just one sentence at a time. There are some additions that one could do on the basic BLEU metric to make it more suitable for entire texts, but at that point we might as well use Rouge-L. Which is what we decided to do.

2.1.2 Rouge-L

Rouge is a text comparison metric used to measure automatic summarisation and automatic translation [3]. The Rouge software package provides metrics that normally would compare a automatically summarised text versus a text summarised by a real person. A higher score would mean it is closer to the real summarised text, a lower score would mean the opposite. We can use these metrics to get a score of how closely a news media text resembles the official press release. The rouge package uses quite some different methods to calculate such a score. One of these is Rouge-L, this is a method of Rouge where a large part of the score is calculated by checking the

LCS (Longest Common Subsequence). This score is calculated by using the following function.

$$LCS(S_1, S_2)_{MEAD} = \frac{(1+\beta^2)R_{lcs-MEAD}P_{lcs-MEAD}}{R_{lcs-MEAD}+\beta^2 P_{lcs-MEAD}}$$

Figure 1: The formula used for Rouge-L. Source: ROUGE: A Package for Automatic Evaluation of Summaries [3]

The function resembles the original normalised pairwise LCS function quite closely. When the β would be set to one in this function, the only difference would be that Rouge-L uses the union LCS score, where normalised pairwise LCS would take the best LCS score. This makes it very interesting to calculate both Rouge-L and the normalised pairwise LCS, which is also what we have done.

Rouge-L is especially useful for our thesis, because if you have sentences, thus many exact same words in a row, that are literally the same, chances are very large that the text has been copied from the other. Rouge-L also helps in detecting if multiple sentences are placed in the same order as the reference text, which can also indicate copying of certain texts. Other forms of Rouge consist of Rouge-1 which is essentially a word comparison function which we already made ourselves. Or Rouge-2 where 2 words get compared with 2 words from the other text, so it is easier to see if texts really overlap. There are even more Rouge measures, but we chose to prioritise Rouge-L.

2.2 The relation between scientific results and press releases

We use the data set of a research done by a large number of authors led by Sumner [1]. This research is not exactly about whether or not news media is original or not, it tries to see where medical exaggerations come from. Very often news articles give direct or indirect advice to readers, even though the official press releases from universities often do not give this advice. Sumner used the English data set we are also using to see if there is any advice given in the news texts, and if there is, if that is also in the university press release it is based upon. For us, that is unfortunately where the comparison between both texts ends, so there is not a lot to use from this research in that particular regard. But in this research they did conclude the following; "For our analysis of advice we found that 40% of the press releases contained more direct or explicit advice than did the journal article (bootstrapped 95% confidence interval 33% to 46%)"[1]. This means for our research, that at least 40% is not literally copied, since otherwise a text can't contain more advice than the university press release.

In the rest of the paper, there are a lot of references to other articles who also try to test the quality of news releases based on medical research. And a lot of these other researches also just looked at the contents of the news posts, and did not really relate them to the original university posts. A different research that also compares news posts versus university press releases, is a research done by Brechman et al [4]. They found out that there are quite some claims that are being made in news texts that aren't actually based on the official university press release. Their conclusion was as follows: "These findings suggest that the intermediary press release may serve as a source of distortion in the dissemination of science to the lay public." This indicates that there is definitely some editing and some addition of information in the news texts, and it does imply that

at least not all texts are directly copied.

When we look at papers that cite the original paper written by Sumner et al. We find a newer research also done by Petroc Sumner and other authors [5]. In this research the goal was to find if exaggerations and caveats are only seen in the news media posts, or also in the official university press releases. What they discovered was that the source of these exaggerations and caveats was most often the original press release. Which means that the writers of the official articles have a bigger influence on the exaggerations and caveats then they first anticipated. A different article that cites our original article by Sumner et al, is a research done by J.S. Taylor [6]. This was a different type of research, which is still very interesting nonetheless. They compared all news articles that are based on a single medical research paper, which were 312 news articles. They then focused on three organisations that provided almost 85% of these articles. The goal was to rate the quality of these articles and what they concluded was that the ratings of these articles went from excellent all the way to weak. This was also seen again in the accuracy of these stories. This research is still interesting for us since it indicates that there are a lot of differences in news articles even though they are based on the same research.

Unfortunately, almost all of these researches don't use text similarity metrics to see if texts are similar or not, most of them were done by hand or by completely other strategies. This does mean that our research has quite some social relevance, as this research has not been done in this exact way before.

3 Data & Methods

3.1 Raw Data

The dataset that is used in this thesis is a dataset that is used in earlier research by Sumner et al. [1]. This set consists of two different parts; English data and Dutch data. Both of these parts have a lot of texts, written by university press or by news media. The data is used to compare texts about the same discovery/subject written by the universities themselves on one hand and the reports of news media on the other.

The dutch data is structured as follows: We have a file called 01-15-001N, the first number means which organisation it belongs to, for instance the LUMC. The second number tells us from which year the research is. And the last number explains which file from the LUMC from 2015 it is. The letter at the end indicates what type of file it is, the N standing for News (Media). For every file like this, there should be a counter file which is called 01-15-001P, which means it is about the same research, but now written by the original researchers. The English set had a different style, that data set just had 2 folders, one for all news files and one for all press files. It used the same style of file naming, without the letter at the end, so we had to manually edit this, to work in the same script. Even though we are using a limited dataset, one could insert any two texts in the script, provided they use this type of file ordering. The exact amount of texts and their average lengths are shown in the tables below.

Dutch publications	Amount of texts	Average word count
University press	135	429
News media	75	863

Table 1: Texts and word count from the Dutch set

When we look at the table above, we can see there is a huge discrepancy in the amount of texts, 60 texts do not have a news article to match them with. It is also easily visible that the news media texts are way longer than the university texts, which also seems really illogical. This is because some university articles have multiple news articles in the same file, making for a very inflated amount of words. This is why we need to preprocess the data which will be discussed in the next section. Unfortunately, the English part of the data had similar issues, as you can see below in table 2. So not only do we need to manually edit the names of the English files, but also do the same preprocessing work as we need to do for the dutch set, which will be discussed in the next section.

English publications	Amount of texts	Average word count
University press	358	544
News media	227	942

Table 2: Texts and word count from the English set

3.2 Preprocessing

3.2.1 Manual preperation

As shown in the last section, the data was not completely ready for use. This meant we had to fix some things before we could use both datasets in different comparison measures. Firstly, all the university press texts that did not have a matching news media article needed to be removed, which made the dataset a lot smaller, with now only 75 texts to work with. The other issue mentioned in the last section is the fact that the news articles sometimes have 2 or 3 articles in the same file. We did not want to manually edit the data, so we wanted to use a script withing Python to split these articles automatically to make them usable for our research. Unfortunately though, there wasn't a recognisable pattern between the articles in the same file to split them on when using docx2text [7]. There were no page ending markings or a set amount of empty lines, this meant we couldn't accurately split the texts in Python.

As an alternative, we only used the first 400 words of a text because that is the average word length of a single text. This means that we lose the extra text or two about the same subject. This was done because it wasn't in our scope to fix up the data, but to make a program that given correctly encoded data could provide a score based on how much was copied from the other text. Even though we lost a bit of information by preprocessing the data, we now get way more accurate results when actually performing the experiments on the data. Additionally, to make the English set more easy to use within the script we renamed all files, to make sure all news files contained an N at the end of their filename, while giving all press files a P at the end of the filename. This made it possible to use the English and the Dutch dataset in the same python script without having to edit the entire first function which reads the data and loads it.

3.2.2 Preprocessing in Python

This was the first step of the writing of the script, we needed to make sure all the docx files that we just spoke about in the data section would be stored in the python script, so we could actually compare them. Unfortunately, you can not load docx files into Python by default, so we used the docx2text package [7]. This converts every document to a large string of words which is saved in a list. This means we get a list for all news files and a list for all press files, in the same order, so they can be easily linked to one another. At this point we have a list in which each entry is one giant string that contains the entire text. This is not something we can use to compare the actual contents of the texts. We want to have the text split on individual words, so we can compare these to each other and use algorithms mentioned earlier such as Rouge and Bleu on the texts. This means we need to do one more thing before we can actually gather some results, which is tokenizing the strings.

We did this by using a powerful package in Python, NLTK [8]. NLTK stands for Natural Language Toolkit, and has all types of scripts to help evaluating written natural languages. We are using the tokenize function specifically, which splits the string on every space and on every punctuation mark. This gives us a list of words and punctuation marks, but since we don't want to compare punctuation marks we can remove those. So what we are left with is a list of words, without punctuation marks and also without capital letters. The capital letters are very important to remove because if you want to test if a word is present in both texts and the word is capitalised in

the one text, but not in the other, it will not count as a overlapping word. After this additional preprocessing step in Python, the data can finally be used to gather results.

3.3 Methods

3.3.1 Word length comparison

The easiest test we could perform was comparing the amount of words both texts have. Even though this test does not prove that much, it can be a good indication of copying. The script is very simple, we use the splits provided by the tokenizer mentioned above and count the individual words that remain. Afterwards it is just a simple division to give us a percentage output which tells us how long the news text is compared to the university press text. This test is very helpful to decide if there is not being copied, not so much to see if there is being copied. This is because the texts of the exact same length could very well not have a single overlapping word. On the other hand, if a news text that might have copied is 100-200 words larger than the text it might have copied from we can safely assume the text is not directly copied. This does of course not mean that there is no copying involved, but it does tell us that there is also at least some original content in the news text.

3.3.2 Comparing the individual words

The next function is a bit more complicated, but will also give us a lot more insight in if the text is actually being copied or not. In this function we are testing if a word found in text a1, is also being found in text b1. Then we calculate the percentage of words in a1 that is also in text b1. In this context, a1 is the news text and b1 is the university press text which corresponds to a1. At the end of the function we get all a count of all the words that are in both texts, and we divide this by the length of the text a1, to get a percentage where a score of 100% would imply that every word in text a1 is also in text b1.

This function gives a lot more insight than the previous function, but since the texts have the same focus and are about the same research, you will always get a high percentage. This is because even if you don't copy at all, chances are very high you are using the same words. This means that even if you get a 85% overlap score on two texts, this unfortunately still does not prove the text is copied from the other. This meant we needed more criteria to make sure texts are original or not.

3.3.3 Rouge-L

That is the point where the earlier explained Rouge-L algorithm comes into play. This metric does not only use corresponding words to provide us with a score, but also compares the longest sub sequences in texts. This is a very good method to see if something is copied or not since it is highly unlikely that for instance 2 entire sentences in a row are the exact same. So if that is the case the Rouge-L metric will provide us with a substantially higher score than a text without long matching strings. Even if these two texts both have the exact same percentage in the previous function. Unfortunately, there is not really an easy way to implement Rouge-L in python. There is the Rouge package in python, which when applied to two texts using an example function gives almost 20 different metrics for the texts. However, almost all of the metrics are unusable for our

research, since Rouge is designed for translation purposes instead of comparison purposes. But there is one metric, well actually three metrics that give us a score that is based on the Rouge-L algorithm.

First of all we get scores between 0-100 where just like the other metrics we have had 0 means 0% of the text overlaps with the other and 100 means they are the exact same. Like mentioned, this function provides us with three different scores, the first one being a score that tells us how much text a overlaps text b. The second one being the opposite, thus telling us how much text b overlaps text a. The last and third score is the average of these two first scores. The way our program is build means that the first score is the most important one because that is the score that tells us how much the news media text overlaps the press release. Which is of course what we are trying to find out in this research. The score generated by this function is the most accurate out of the functions the script contains. And we will see in the results and analysis chapter, what kind of percentages indicate clear copying. As mentioned earlier, Rouge-L takes the longest subsequence you can find in both texts into account as well, unfortunately, the python package does not show how long this subsequence is. So we made function number 4.

3.3.4 Finding the longest common subsequence

Like the name suggests, we are going to analyse this problem as a standard LCS (Longest Common Subsequence) problem. A classic problem in computer science, often used on strings with a limited amount of characters, such as just A, B and C in random orders to find out the longest common subsequence between the two. In our thesis, we are trying to use this function on our lists with the tokenised words. Using a commonly used dynamic programming example for LCS type problems, we were able to create a working function that works as follows. When you input two strings, or in our case, lists, this function will give the length of the amount of corresponding words in the same order as the reference text. As an example, the LCS for input sequences ABCDGH and AEDFHR is ADH of length 3. This means that even though the letters ADH are not right beside each other they are still the LCS for this particular case. In our texts, you will often see that the total LCS is based of three or four different copied parts in the same order. This makes it really easy to see when texts are being copied, because when you have 170 words, in a text of around 300 words, that are in the same order in a text, you can be certain that that is not a coïncidence. What would be even more insightful, is a percentage that would tell you how much of a text is directly in the same order. In our previous answer that would be just below 60%, which is a really high percentage for a statistic like this one.

3.3.5 Getting more insight in the texts

Another interesting part about this research, is whether the news media exaggerates when reporting new discoveries. One of the most important steps of discovering this is finding words which are used in news media texts, but not in the official press releases. Which is the opposite function from the word comparison function from section 3.3.2. Let's say the news media commonly uses the word 'extremely' while the university press never uses this word whatsoever, that could be an indication that news media like to exaggerate. We will try to find words across all the different texts that are found in news media texts, but not in the university press. The labelling of these words and checking whether or not they are there to exaggerate is outside of the scope of this project, but we

will definitely try to lay the foundation for similar research.

4 Analysis and Results

In this section all the results will be shown, which all have been gathered using the created python script together with the dataset that is explained in section 3. Firstly we are going to show and analyse all results from the Dutch set. And we do this for every function individually.

4.1 Dutch data set

4.1.1 Function 1: Word length comparison

First off we have the word length comparison, which is a very straight forward metric, with a very simple figure to show for it. We calculated all different percentages for the dutch texts, and it resulted in the following boxplot.

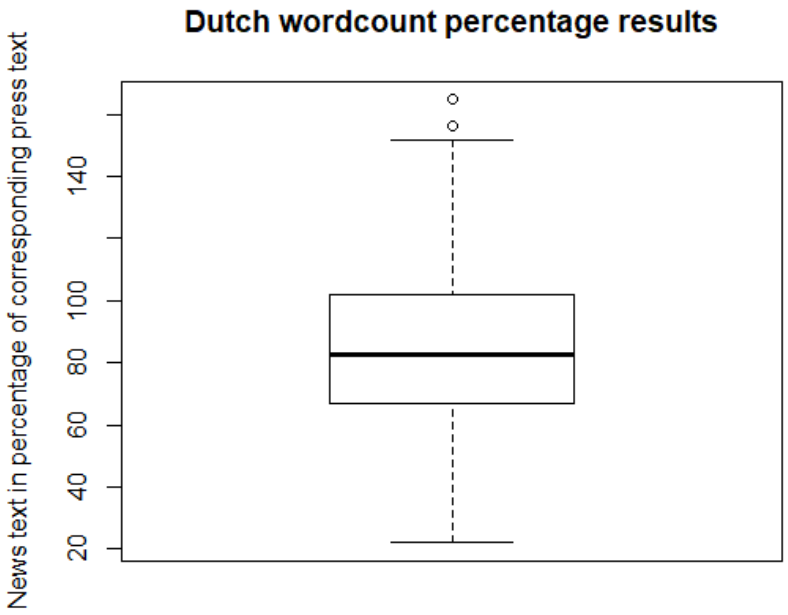

Figure 2: A boxplot that shows the wordcount for the news texts, in a percentage of the corresponding press texts.

In this boxplot you can see that most of the news texts are slightly shorter than the press texts they are based on. Unfortunately, that does not prove or disprove any copying claims we might have on said texts as a whole. What we can see though, is the fact that some of the outliers prove that those texts aren't copied, because you can't fully copy a text when your text is almost 60% longer. For some other individual texts you can also see that the percentage is really low, which means the news article stripped a lot of text from the original press article, which also makes it safe to say they didn't literally copy the entire article. In general this means that it is still very

plausible a lot of news articles are being copied from their respective press articles, but we need to apply more functions to be sure of that.

4.1.2 Function 2: Word comparison between news text a and press text a

This function checks for every word that is in the news text if that exact same word is also in the corresponding press text. We decided to use a histogram with bins every 5% to get a clear view on the results. You can see the histogram below.

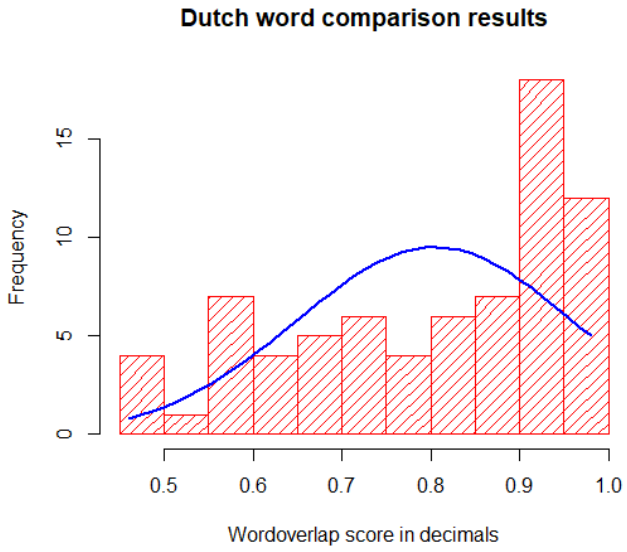

Figure 3: A histogram of the Dutch data set when used on the word comparison function.

The results are quite diverse, but there is one huge trend noticeable, and that is that a lot of texts, almost half of them, are for 85% or more identical to their corresponding press text. The lowest texts still share close to 50% of words with their press texts, but this is very logical. Since both texts are on the exact same subject, even if you write a completely new text, chances are you would still get around 50% or higher when you compare the words. But the further we go to the right in the figure, the more concerning the scores get, because when around 92% or more of the words are the exact same, we can safely assume at least a large part has been copied. Especially when the percentages rise even more. We can see 3 texts with a score of 98%, which means they are almost exactly the same, because there are always some words different, such as the exact time of posting the article and the name of the website or author that posted it. In total we get an average of 80%, which is quite a lot if you think about it. This means that we can see that in the Netherlands, a lot of the press articles are being used as more than just an inspiration. But to be even more sure that this is the case, it is time to apply the renown Rouge-L metric on our data to maybe get a different view.

4.1.3 Function 3: Rouge-L

This is the python implementation of the Rouge-L algorithm described earlier. We have chosen for another histogram, because it is a bit clearer and more detailed than a boxplot in this case.

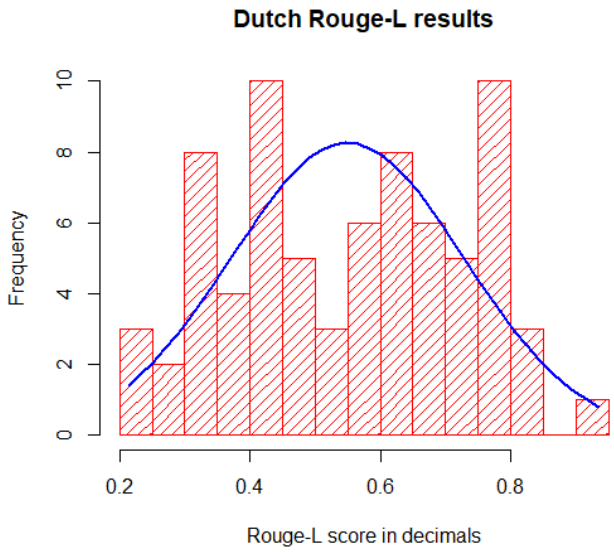

Figure 4: A histogram of the Dutch data set when using Rouge-L.

When we look at the figure above, we can see the Rouge-L metric, and that the scores are quite a bit lower than on the last function. This is because for Rouge-L to give a 100% score, not only do all of the words need to be the exact same, they need be in the same order as well. So if sentences are a little bit rewritten, or shuffled around, this will make a large impact on the Rouge-L score. This was not the case when using the second function. When looking at the original paper in which Rouge was released [3], we can see that our highest score of 93.65% is insanely high, and implies an almost 1 to 1 correlation. When we look on the left side of the figure, we can see some low scores as well, way lower compared to the average of the lower scores from the last function. This is partly because this is of course a completely different metric, but since it still grants a result between 0-100, this is quite relevant. With an average of 54.93%, compared to the average of exact 80% in function 2, there is quite a decrease. This would imply, that even though sometimes almost 90% of the words are the same, the total text might not be as similar as one would expect.

Overall, these Rouge-L scores still give us reason to believe that quite a lot of texts are copied, or at least partially. It is very hard to find what kind of scores are normal for translation. But if we look at the use of BLEU and Rouge-L we can see that researchers often start to doubt results above the 60% because even 2 translated texts by different humans rarely have such high of a score. So if we see scores of around 60% or higher, and we are seeing a lot of those, we can be quite sure that that is no coincidence. So we can safely assume as a preliminary conclusion that the Dutch news press copy quite a bit while writing their articles.

4.1.4 Function 4: LCS

As explained in section 3.3.4, the lcs function produces a number which corresponds to the total amount of words in both texts that are in the exact same order. There can be other words in between, that are not the exact same, but the number that is produced are all the words that are in the same order. To get a result that is easier to read, we decided to divide this number by the amount of words in the news text, to get a percentage. We then plotted all these percentages in a boxplot using R [9], which is shown below.

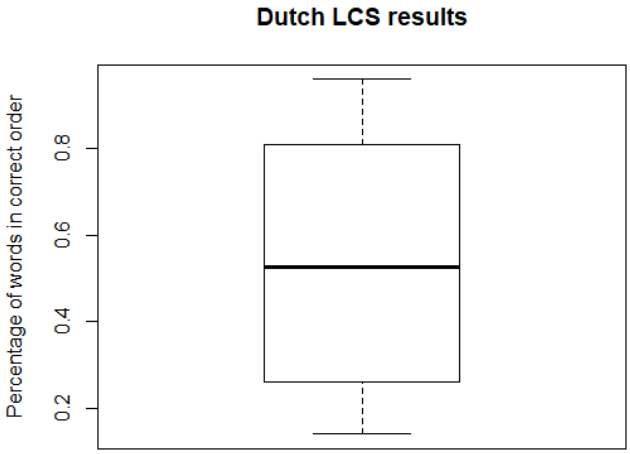

Figure 5: A boxplot of the Dutch data set when applying the LCS function.

In this boxplot, we can see that the scores are quite diverse. The high part of the boxplot is almost at a 100% and the low part is below 10%. This means there is not really a trend noticeable, but we can say something about almost every text above the median. If your text consists of 50% or more of the exact same words in the exact same order, we think it is safe to say, some copying has been involved. Especially when you get around or above the 80% mark, the texts are practically the same except for maybe one paragraph or some quotes that are used from somewhere else. So when we look at these results we can quite definitively say there is quite some copying going on in the Dutch press.

4.1.5 Function 5: Find words in news texts that aren't in the press texts

When we order the words that are in news text, but are not present in the press texts, a lot of useless words come out on top, these are words like: 'telegraaf', 'nu.nl', 'donderdag' (or any other day in the week), 'december' (or any other month), 'universiteit'. These words are really obvious, they are either the name of the news organisation, the date it is published on, or they reference the university they are using as a source. There are some other words that are very interesting to look at:

- conclusie, concluderen (conclusion, concluding)
- wetenschappers, onderzoekers (scientists, researchers)
- altijd, nooit (always, never)
- ze (they)
- blijkt, blijkt uit (turns out, and quite some different translations, since the English don't really have a word for 'blijken'.)

These are some interesting words that we were able to find that are being used very frequently in the 75 news texts, but are very scarcely used in the university press. First off the words around conclusion, which makes quite some sense when you think about it, the news press tries to summarise or explain a university research, which has a conclusion in the end, so it is quite likely news writers want to paraphrase that to make sure the reader can easily what the end result is. The second one is maybe also a bit obvious, but interesting nonetheless, since apparently the writers of news articles tend to reference the people who actually did the research, instead of for instance the entire organisation.

The third item on the list surprised us, since this is a word that could be used anywhere, also in the press texts. Apparently the university tries to refrain themselves from using these terms, to make sure there are no promises being made about a certain solution or result. This could be an interesting phenomenon to look into a bit deeper by a linguistic study. The word 'ze' (they), is quite similar to the second word on this list, it is often used to reference the researchers or the university where the research was done. Lastly we have the verb 'blijken', which is quite a hard word to translate, it can be used in a passive way, when we would best translate it with 'turns out', but is often used in a different way, especially in these texts. In is most often to explain that something is found in the research and then explained, a bit similar how in an English text the term "evidence suggests" is used quite frequently. This is very interesting, because it is another way to reference the article, without using the same words as we described earlier. This seems to be a trend in the words we can find in the news texts and not in the press texts. A lot of words are being added in to reference the original article.

When we look at the different news sources it is also interesting to see that both 'nu.nl' and 'de Telegraaf' have a very low amount of words that are unique to news text, which implicitely also means they copy a lot. But there is something more to the new words some of them use. This is a special type of word usage that is being used mainly by 'de Telegraaf', they tend to accentuate words such as 'wel' and 'maar' by writing 'wél ' and 'máár '.

4.2 English data set

4.2.1 Function 1: Word length comparison

Just as the Dutch set, we also made a boxplot to show to results for the first function in the English set.

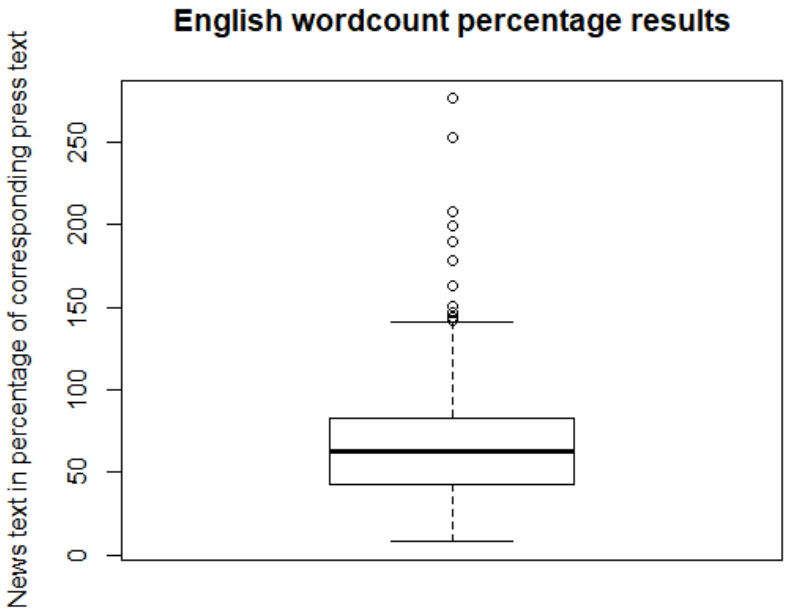

Figure 6: A boxplot that shows the wordcount for the news texts, in a percentage of the corresponding press texts.

In the figure you can see a similar result to the Dutch dataset, except for 2 things. The English set trends a little bit lower than the Dutch set, which would imply that English news sources shorten their press texts a bit more. And that the results are more extreme. We have even larger outliers on both sides, and a filled part of the boxplot that is of a lower percentage than the Dutch texts. A large reason for this is the fact that in general, the English articles are just more varying in length, which also leads to some extremely short news articles of sometimes as little as 40/50 words. This makes it possible to get scores of around 27% which we can see in the boxplot. Unfortunately, the same can't fully be said when we look at the really high percentage outliers. We can explain this phenomenon by looking at the preprocessing part of this thesis in section 3.2. In this section, we explained that some news files, had multiple articles in them. To try to make sure we don't get many false results, we capped the length of a news article to 400 words. For the Dutch set, this works quite well, but because the English data set has articles which such a large variety of different word counts, this becomes tricky. The outlier that is seen at the top of the boxplot exists because there is a really small press text, that has a news text, that consists of 4 articles. Even though there are 4 articles, they almost entirely fall in the 400 word limit, so you get an absurd

percentage of 261% as seen in the figure above. This is a small limitation of our method, but since this boxplot contains over 200 percentages from texts these 2 or 3 outliers do not skew the end results all that much.

4.2.2 Function 2: Word comparison between news text a and press text a

Just like before we use the same figure as we used in the Dutch data set, showing us the results for the second function as clearly as possible for the English data set as well.

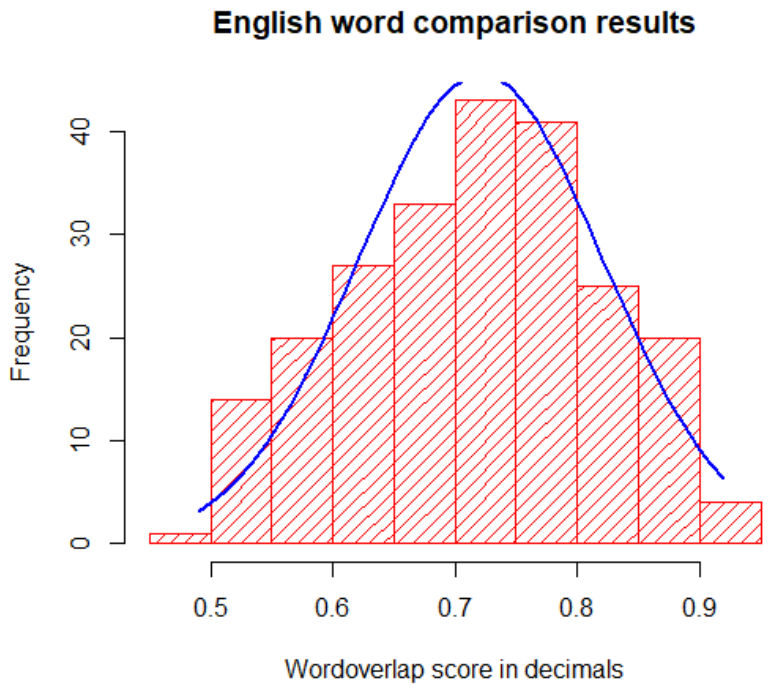

Figure 7: A histogram of the English data set when used on the word comparison function.

When we look at figure 7, we see a huge difference with figure 3, the histogram of the Dutch data set. Not only are there no really high percentages such as the 98% we have seen in the Dutch dataset, the entire balance is completely different between these two datasets. At first glance, it seems like the English news press, copies a lot less than the Dutch news sites and organisations do. There are still some texts, that are quite likely to be copied, also in the English data set. But we cannot assume that there is a lot of copying in the texts that only have around a 70% word overlap. The average of the English texts is 72% and this is quite a logical amount of overlap when you try to summarise/report about the other text. For now, it seems like the English news press is more original than the dutch news press, but we will need to see the Rouge-L metrics to be more sure.

4.2.3 Function 3: Rouge-L

We are using the same type of figure as we used for the Dutch set, in this case, being a histogram again.

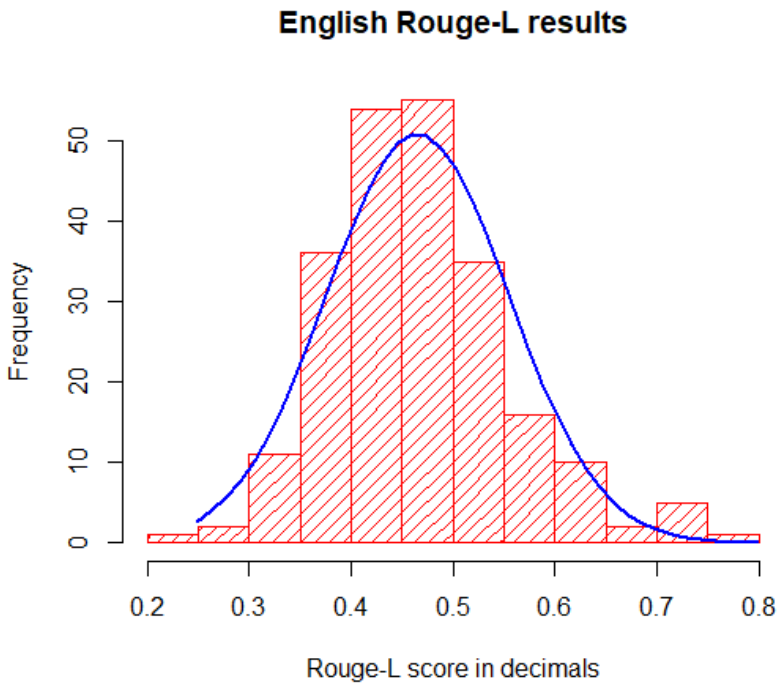

Figure 8: A histogram of the English data set when using Rouge-L.

As explained in the earlier section about Rouge, and also in the Dutch part of the results, a lower Rouge-L score is very logical. But the English Rouge-L scores are substantially lower than the Dutch Rouge-l scores, which further supports our claim from the last section that the English news press copies less than the Dutch does. Furthermore we can see that there are still some cases here, that clearly show not all news articles are 'clean' in that sense. Since cases over 60% are considered suspicious even when looking at translation texts, we can see at least 8 or 9 really suspicious texts. But since these texts aren't supposed to be the same, where a translated text is, we should also expect some copying on even lower percentages. With an average of 46.52%, which is 8.4% lower than the Dutch average, we can conclude that there is not that much copying as maybe expected initially. The other functions will not really give any scores to the texts, but they will give more insight in how the texts are built up.

4.2.4 Function 4: LCS

We are using the same method of plotting as the dutch set, which is a boxplot of the results when using the LCS function on the English data set.

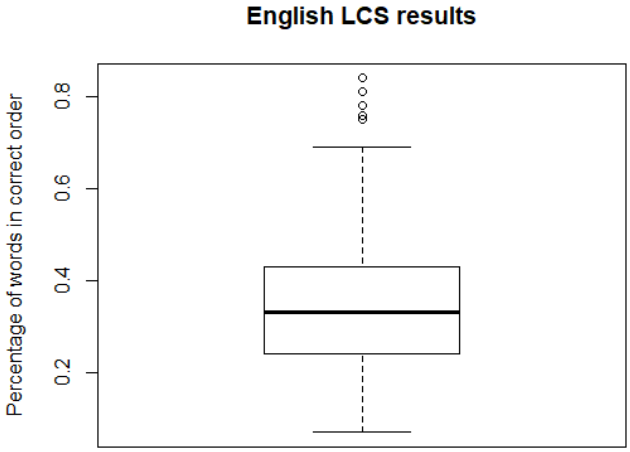

Figure 9: A boxplot of the English set when applying LCS

When we look at this boxplot we can see that the average of the texts isn't nearly as high as the dutch set. The dutch set had a media of 61% where the English set only gets up around to 34%, which is one of the largest differences we have seen so far. There are some outliers that imply some heavy copying, because if almost 85% of your text consists of words in the exact same order you will definitely have taken more than just inspiration of the university text. If we look at the rest of the plot, we can clearly see once again that the English data set contains a lot less copying than the Dutch set. Which is something we can now definitely conclude from the writing this thesis.

4.2.5 Function 5: Find words in news texts that aren't in the press texts

Because we saw a lot less copying in the English texts compared to the Dutch texts, it was really hard to accurately find words that only were used in the news texts and not in the press texts. With sometimes over 125 words per text, a lot of words are found in the news texts, but there are some words that seem to be quite unique to the news texts. For instance the words that are most commonly seen are 'author', 'updated', 'daily mail', 'london' and 'correspondent'. These are all words that score really high because they describe the article or the article that they have used as a source. Even though these are interesting, they don't really show any difference in the contents of the text. Words that are perhaps more interesting, are as follows:

- you
- n't (instead of not)

- so-called
- evidence
- suggests

These words are actually signs that news reporters have a different writing style than the university press members. The university press tends not to write directly aimed towards the reader, and thus avoids the word you. News organisations however sometimes want to make clear that some things impact you, the reader. Something you also see that after the tokenizer splits the words, the university press leaves a lot of 'not's' behind, where the news media often shortens with n't at the end of words. The third one we see is a word that is often used to indicate a word that is not normally used in news articles, and is likely to be explained after. This also implies that writers of the news articles take some time to make the text easier to read for the general public. We would also like to mention that in general, words with the dash in the middle are very often seen in the news articles, which also indicates a different writing style. The last two words on this last are actually often combined to form a statement about the original research. Sentences will start with evidence suggests, and then explain something that is stated in the research, often using similar words in that part of the sentence, but not starting the sentence the same as the original university press text.

5 Conclusions and Further Research

5.1 Conclusion

To conclude this research we are going to answer our research question as well as we can with the data we have acquired. Firstly we will give a small summary of what the aim of this thesis was and what we have done. We started with our problem, that being the fact that there seemed to be a whole lot of plagiarism by news websites and newspapers alike. We went to test this by using two different data sets, one from the Netherlands, with texts from the year 2015. And the other data set being from England, with articles written in the year 2011. To answer our research question, we wanted to use text comparison metrics on 2 corresponding texts, to give scores to the texts, so we can easily see if there is a lot of copying or not. We used Python 3 for this, and we made several functions in a python-script to analyse the texts. We also used 2 renown measures, BLEU and Rouge-L, unfortunately, BLEU gave us no additional information over Rouge-L, this is why we only kept Rouge-L in the script. When the data was preprocessed we could give these scores to all the texts in our data sets.

The results were shown in the last chapter in detail, but what we saw was that for the Dutch texts in both Rouge-L and in the word comparison function, the scores were really high, sometimes almost 98% similarity between the news text and the university press text. Even though we saw some high scores for the English data set as well, it was nowhere as high as the Dutch set. Which would lead us to the following answers on our main research question: "To what degree do news media directly copy official university press articles?" First off we can specify that for the most Dutch news sites/newspapers this is a lot, with an average of 80% similarity when directly comparing the words and scoring a really high 54.93% score on the Rouge-L metric. It is hard to quantify how many texts directly copied, but with scores this high you are bound to find at least multiple copied sentences every other text.

Secondly we have our English data set, which scored quite a bit lower. With an average of 72% in the word comparison function, it scores a solid 8% lower than the Dutch set. And when we look at the Rouge-L metric, this score is even lower, 46.52% to be exact, this is 8.4% lower than the Dutch set again. This means it is pretty safe to say that the English media copies a bit less than the Dutch media does, but you can also expect to still see quite some copied sentences and even sometimes almost entire paragraphs in the English news media.

When we look at the sub questions, one of the two is very easily answered by saying that either Rouge-L, direct word on word comparison, or even LCS is a good measure to score similarity between texts. The other question, which is as follows: "Do certain organisations copy more than others?", is a lot harder to answer. In the Dutch texts there is a clear pattern that is seen, 'nu.nl' and 'de Telegraaf' tend to copy a lot, both of these organisations often had scores of over 90% in the word comparison function, and scored above average on the Rouge-L metric almost every time. Unfortunately the data set was not properly labelled to automatically calculate the scores grouped for each organisation. This would be great for further research, which brings us to the next section.

5.2 Further Research

Building on the last sentence from the previous section, it would be really nice to do more research in certain groups of media. Do website/television based news suppliers, such as "nu.nl", "nos" or even "rtl" more often copy texts than news sources who originate from newspaper such as 'The Daily Mail', 'de Telegraaf', 'the Guardian', 'de Volkskrant' and others. Another important part of this study that could be built upon is the factual correctness and other linguistic choices that the news article writers make. We are unfortunately not capable of doing this type of research, since it is not in our field of study. Also as mentioned in section 3.3.5, it would be nice to research more about the meaning of the texts and if news writer sometimes alter the semantic implications of the texts, instead of just reporting what has been discovered.

References

[1] Petroc Sumner et al. "The association between exaggeration in health related science news and academic press releases: retrospective observational study". In: *BMJ* 349 (2014). DOI: 10.1136/bmj.g7015. eprint: https://www.bmj.com/content/349/bmj.g7015.full.pdf. URL: https://www.bmj.com/content/349/bmj.g7015.

[2] Todd Ward Kishore Papineni Salim Roukos and Wei-Jing Zhu. "BLEU: a Method for Automatic Evaluation of Machine Translation". In: *ACL* (2002). URL: https://www.aclweb.org/anthology/P02-1040.

[3] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: (2004). URL: anthology.aclweb.org/W/W04/W04-1013.pdf.

[4] Jean Brechman, Chul-joo Lee, and Joseph N. Cappella. "Lost in Translation?: A Comparison of Cancer-Genetics Reporting in the Press Release and Its Subsequent Coverage in the Press". In: *Science Communication* 30.4 (2009). PMID: 25568611, pp. 453–474. DOI: 10.1177/1075547009332649. eprint: https://doi.org/10.1177/1075547009332649. URL: https://doi.org/10.1177/1075547009332649.

[5] Petroc Sumner et al. "Exaggerations and Caveats in Press Releases and Health-Related Science News". In: *PLOS ONE* 11.12 (Dec. 2016), pp. 1–15. DOI: 10.1371/journal.pone.0168217. URL: https://doi.org/10.1371/journal.pone.0168217.

[6] Joseph W. Taylor et al. "When Medical News Comes from Press Releases-A Case Study of Pancreatic Cancer and Processed Meat". In: *PLOS ONE* 10.6 (June 2015), pp. 1–13. DOI: 10.1371/journal.pone.0127848. URL: https://doi.org/10.1371/journal.pone.0127848.

[7] Ankush Shah. *docx2text*. Version 0.7. Nov. 25, 2017. URL: https://pypi.org/project/docx2txt/.

[8] NLTK Project. *NLTK*. Version 3.4.3. June 6, 2019. URL: https://www.nltk.org/.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: http://www.R-project.org/.