

MALAPROPISMS DETECTION AND CORRECTION USING A PARONYMS DICTIONARY, A SEARCH ENGINE AND WORDNET

Costin-Gabriel Chiru, Valentin Cojocaru, Traian Rebedea, Stefan Trausan-Matu
*“Politehnica” University of Bucharest, Department of Computer Science and Engineering,
313 Splaiul Independentei, Bucharest, Romania*
costin.chiru@cs.pub.ro, amalosoul@yahoo.com, traian.rebedea@cs.pub.ro, trausan@cs.pub.ro

Keywords: malapropism, paronyms dictionary, cohesion, search engine, lexical chains, filters, chunking

Abstract: This paper presents a method for the automatic detection and correction of malapropism errors found in documents using the WordNet lexical database, a search engine (Google) and a paronyms dictionary. The malapropisms detection is based on the evaluation of the cohesion of the local context using the search engine, while the correction is done using the whole text cohesion evaluated in terms of lexical chains built using the linguistic ontology. The correction candidates, which are taken from the paronyms dictionary, are evaluated versus the local and the whole text cohesion in order to find the best candidate that is chosen for replacement. The testing methods of the application are presented, along with the obtained results.

1 INTRODUCTION

During the last years, people have started to write more and more electronic documents using the programs available on everyone’s PC, because they can use some features that a sheet of paper could not offer. One of the most important such feature is the automatic spellchecking. Many people are not paying enough attention to the things they write, knowing that if they make any mistake, the spellchecker will point out the mistake or even correct it. Nevertheless, even the best spellcheckers sometimes fail in correcting a misspelled word, introducing a different word than the original one that has been misspelled, that is close to the initial word from the editing distance point of view, but semantically unrelated. More than that, even people sometimes use other words instead of the ones that they should, due to the lexical or phonetic similarity between these words and to the insufficient knowledge of the language or lack of attention. This unintentional misuse of a word by confusion with another one that sounds similar is called malapropism and cannot be identified by an ordinary spellchecker.

In this paper, we propose a method for the automatic detection and correction of these malapropisms using an ontology (WordNet), a search engine (Google) and a paronyms dictionary. In the next chapter, we present other approaches for malapropisms detection. The paper continues with the architecture of our application and the experiments developed in order to test it, along with a walkthrough example. We wrap up with the results and the conclusions that can be drawn from them.

2 OTHER APPROACHES FOR MALAPROPISMS DETECTION

One of the first approaches about handling malapropisms was proposed by (Hirst and St-Onge, 1998). They presented a method for identifying and correcting malapropisms based on the semantic anomaly discovered through lexical chains. They followed the assumption that malapropisms should be words that do not fit in the context, so they should be found in atomic chains. After detecting the malapropisms, the authors applied the spelling corrections procedures used by a spelling checker in

order to identify some possible corrections for the given malapropism. Then they have tried to see whether these corrections fit better into the lexical chains, and if so, the corrections were made. In order to test their assumption, they have built a corpus of about 500 articles randomly selected from the Wall Street Journal discussing many different topics and they have introduced a malapropism at every 200 words. This action resulted in a 322,645 words corpus having 1,409 malapropisms. After building the corpus, they have used their method for detecting and correcting the malapropisms. The results showed a 28.2% detection rate and a 24.8% correction rate with a false alarm rate of 0.859%. A similar idea is presented in (Hirst and Budanitsky, 2005), where the authors have noticed a 50% recall and 20% precision for malapropisms detection, and between 92 and 97.4% for malapropisms correction.

A different approach (Bolshakov and Gelbukh, 2003) proposes an algorithm for malapropisms detection and correction based on evaluating the text cohesion represented by the number of collocations that can be formed between the words found in the immediate context of each other. The words that do not form any collocation in the context are signalled as being possible malapropisms. The possible corrections are generated and tested if they form at least a collocation with the given context. In (Gelbukh and Bolshakov, 2004) the authors suggest that a paronyms dictionary could be very useful for the generation of possible corrections. Two words are called paronyms if they have only slight differences in spelling or pronunciation, but they have (complete) different meanings. This approach was semi-automatically tested on a set of 16 sentences that were built so that each of them had a malapropism. An accuracy of 68.75% has been achieved. Nevertheless, considering the very small dimension of the test, the accuracy could be affected if the algorithm would be applied on a larger corpus. More than that, since all the sentences had a malapropism, the false alarm rate could not be accurately detected. Variants of this algorithm were also tested against three corpora, one written in Spanish and having 125 malapropisms (Bolshakov et al., 2005) and the other two written in Russian (Bolshakova et al., 2005), having 100 malapropisms each. For Spanish, all the malapropisms have been identified, and around 90% were correctly replaced. For Russian, 99% of the malapropisms have been identified, and around 91% were correctly replaced. Though the obtained results are very good, one should notice that the algorithm uses three constants (P, NEG and Q) that have been empirically

determined to optimize the results and therefore are very text-dependent, as shown by the values chosen for these corpora: P = 3500, NEG \approx -9, Q = -7.5 for the Spanish corpus (Bolshakov et al., 2005), and P = 1200, NEG = -100, Q = -7.5 for the Russian ones (Bolshakova et al., 2005). Again, the corpus contained only phrases that contained a malapropism, so the false alarm rate could not be computed.

Besides these methods that exploit the semantic similarity between words, other approaches are based on statistical methods. Among these, there are methods employing n-grams (Mays et al., 1991; Wilcox-O'Hearn et al., 2008), Bayesian methods (Gale et al., 1993; Golding, 1995), POS tagging (Marshall, 1983), or a combination between the latter two methods (Golding and Schabes, 1996).

3 THE ARCHITECTURE OF THE APPLICATION

Our application has two main modules – one for the malapropisms detection and the other for their correction – and a couple of sub-modules as it can be seen in Fig. 1. In this application, we have used some external technologies (marked by italics in the figure): as a POS tagger we used Q-tag (which can be found online at <http://web.bham.ac.uk/O.Mason/software/tagger/>), but there are many others freely available (<http://www-nlp.stanford.edu/links/statnlp.html#Taggers>); a Web search engine – we have used Google because at the moment it is the most popular search engine; a lexical ontology – we have used WordNet because the APIs provided by the developers are very useful in building the lexical chains; and a paronyms dictionary that has been compiled in our department based on the WordNet ontology – therefore containing only common words. The dictionary has 77,503 words, 22,020 of them (28.4%) having at least one first-level paronym. We called two words to be first-level paronyms if they are at editing distance of 1.

For the malapropisms detection and correction, we have tried to improve the results by combining the methods based on semantic similarity between words with the statistical ones. Therefore, we consider both the lexical chains and words co-appearance as measures for text coherence, while the statistical methods are represented by the way we decide whether the co-appearance of two chunks of text is statistically correct or not.

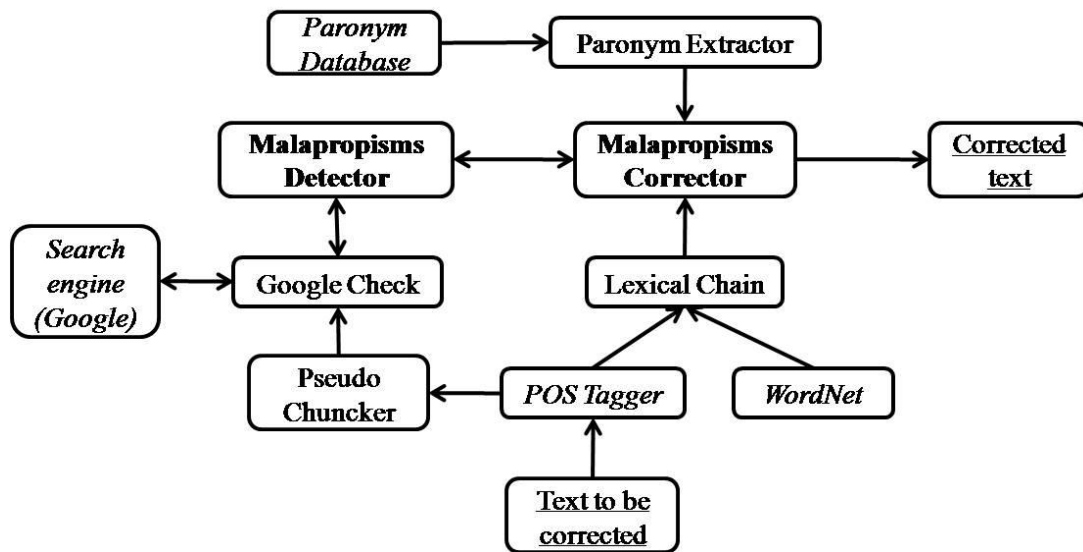


Figure 1: The architecture of the application. The texts written in italics represent third-party technologies that have been used for the application. The underlined texts represent the input given by the user and the output given by the system. The bold texts represent the two main modules of the application: the malapropisms detector and the malapropisms corrector.

We believe that the lexical chains represent the context of the whole text, while the words co-appearance expresses the cohesion of the immediate context of each word. This is why the malapropisms detection and correction is done in two stages: an initial detection that checks for local anomalies is done in the detection module, while in the second stage the results of this phase are revised in the context of the whole text, during the correction.

3.1 Malapropisms detection

This module is responsible for the initial identification of the possible malapropisms by detecting anomalies in the local text coherence. In order to achieve this, we have used the Google search engine. The search engine receives two chunks of text – the means of selecting these chunks is described in the next paragraph – and based on the mutual information inequality it evaluates if their co-appearance is statistically correct in a manner similar to the collocation testing suggested in (Bolshakov and Gelbukh, 2003).

If we simply send the content words to Google, we cannot check whether the local text coherence is damaged, because these words are rarely adjacent. This is the reason for also considering the functional words surrounding these content words when evaluating the local text coherence. Therefore, we have built and used a pseudo-chuncker that groups the words in chunks before sending them to the search engine. These chunks contain all the

functional words between any two content words next to each other. After the phrase has been decomposed in chunks, these are sequentially evaluated in order to identify the potential drop in the text coherence. Hence, the first two chunks are sent to the search engine, and the results are parsed in order to find three pieces of information: the number of hits for the first chunk, the number of hits for the second chunk and the number of hits for the co-occurrence of the two chunks (considering the second chunk is right after the first one). From now on, we will address these numbers with the following names: *no_pages1*, *no_pages2* and *no_combined*. These scores are evaluated and then the process continues with the next chunk, evaluating the coherence between the second and the third chunk. The process ends when the coherence between the last two chunks is evaluated.

The coherence evaluation is done based on six progressive filters that depend on these three numbers that are obtained from the search engine. The assumptions behind these six filters are: the fewer hits of the co-occurrences of the two chunks, the greater probability of a malapropism and, the more pages for the individual chunks – having the same number of co-occurrences of the two chunks – the greater probability of a malapropism. In order to model these facts, we used a parameter (beta) that is modified depending on the values of *no_combined*.

The first filter is for the case when *no_combined* has a very small value. The purpose of this filter is to eliminate the noise caused by the indexed pages

that are grammatically incorrect and it does not depend on the number of hits for the individual chunks. If *no_combined* is smaller than the filter's upper threshold (which we considered to be 20) then a possible malapropism is signalled. If *no_combined* is greater than the threshold, one of the next filters is applied and a malapropism is signaled if the following formula is true.

$$\beta * \log\left(\frac{no_combined}{pages}\right) < \log\left(\frac{no_pages1}{pages}\right) + \log\left(\frac{no_pages2}{pages}\right) \quad (1)$$

The *pages* parameter from the formula above represents the number of indexed pages written in the used language. For English, this number could be easily found by sending the word “the” to the search engine and noticing the number of hits. At the moment of writing this paper, more than 11 billion pages written in English were indexed by Google. This value is automatically detected every time the application is launched.

The second filter is applied when a low number of co-occurrences is obtained (less than 500). Here, the value of the parameter *beta* is 1.05 in order to provide a tougher filtering than normal, according to the fact that fewer hits of the co-occurrences of the two chunks imply greater probability of a malapropism. Therefore, the filtering is not dependent on the input text, but on the number of hits of the co-occurrence of the two chunks.

The third filter applies to the co-occurrences that have between 501 and 12,000 hits, being the filter that is the most often used. For this filter, *beta* takes the value 1 instead of 1.05 as in the previous filter, because this is considered the regular filter from the permission point of view. From now on, the permissibility will constantly drop, since the number of hits for the co-occurrence of the two chunks becomes higher and higher, therefore the probability of being a malapropism decreases.

The fourth filter is applied when *no_combined* is between 12,001 and 14,000 and here, *beta*'s value is 0.95. The fifth filter lowers again the probability of having a malapropism by considering *beta* to 0.9 and is applied for chunks that have *no_combined* between 14,001 and 15,000. Finally, the most permissive filter is applied when *no_combined* between 15,001 and 16,000, *beta* having the value of 0.8.

Above this final threshold (16,000), no possible malapropisms are signalled, since having a very large number of hits, one cannot precisely tell if a malapropism occurred or there was just a less often combination of two very popular chunks of text.

The presented thresholds and the coefficient for the co-occurrence of the two chunks that these filters depend on have been empirically determined and they are language and time dependent, but they are text independent. First of all, the values depend on the language, because the number of pages written in different languages is not the same. These values have been detected for English, but if the language is changed, the value of the “*pages*” parameter also changes, and the same happens with the values of *no_pages1*, *no_pages2* and *no_combined*, so the thresholds are not accurate any more. The values are also time dependent, because the Internet is in a continuous expansion and therefore, the number of the written pages available to the search engines continue to increase and in the same time, the probability of finding incorrect text also increases, affecting the thresholds of the presented filters.

Considering the large number of queries that are sent to the search engine, we have also investigated the possibility of using the Google 5-grams corpus “Web 1T 5-gram Version 1 Corpus” (Brants and Franz, 2006) instead of sending our queries to the search engine. Besides his very large size (30 GB of compressed text), which makes it difficult to integrate in any application, we have observed another drawback of this corpus: the document n-grams were not completely covered by the corpus' n-grams – the covering varied from 90% in the case of bigrams to 15% in the case of 5-grams. More than that, the nature of our application made us give up at this corpus, because in the application we do not know a-priori the degree of the n-grams that are going to be used, since this is determined dynamically by the pseudo-chunker.

The purpose of this module is to limit as much as possible the number of misses in the malapropisms detection. The signalled malapropisms generated in this module should cover all the real malapropisms that exist in text. The module also signals a lot of fake malapropisms, but they will be evaluated in the next module and some of them will be ignored.

3.2 Malapropisms correction

There are two main purposes of this module. The first one is to determine which of the signalled malapropisms from the previous step are false alarms in order to eliminate them. The second purpose is to detect the most probable candidates for the remaining malapropisms in order to correct the errors. This module uses all three technologies: the paronyms dictionary in order to identify the candidates for the possible malapropisms correction,

lexical chains in order to filter the list of candidates for finding the ones that fit into the context and, finally, the search engine in order to decide which is the best candidate in the case that there are more that fit into the lexical chains.

The module also works sequentially by analyzing every pair of two chunks of words and deciding whether a malapropism or a false alarm has been found, and in the case of a malapropism, what should be the replacement word. If the pair contains no signalled malapropisms, than the process continues with the next chunk, until a signalled malapropism is found. The correction is done in three stages: first of all, the replacement candidates that ensure the local cohesion are identified using the paronyms dictionary; these words are then filtered against the text logic so that the whole text cohesion to be maintained; finally, the replacement word is chosen from the remaining words, based on the information given by the search engine relating to the probability of fitting in the local context.

For the detection of the replacement candidate words, there are three possible situations that are treated separately: a signalled malapropism in the first/last word in a sentence (Fig. 2), an isolated malapropism in the middle of the sentence (Fig. 3), or a malapropisms chain (Fig. 4).

The analysis of a pair of chunks begins with the extraction of all the paronyms of the content word from the chunk signalled as containing a malapropism. Then, every paronym replaces the malapropos word and the local cohesion of the phrase is tested in order to avoid replacing a word with a paronym that is worse than it, from the cohesion point of view. This time the cohesion is tested versus both chunks that surround the problematic one (Fig. 3), except the special case when that chunk is the first or the last one in the phrase (Fig. 2). The cohesion testing between two chunks is done in a similar way as described in the detection module.

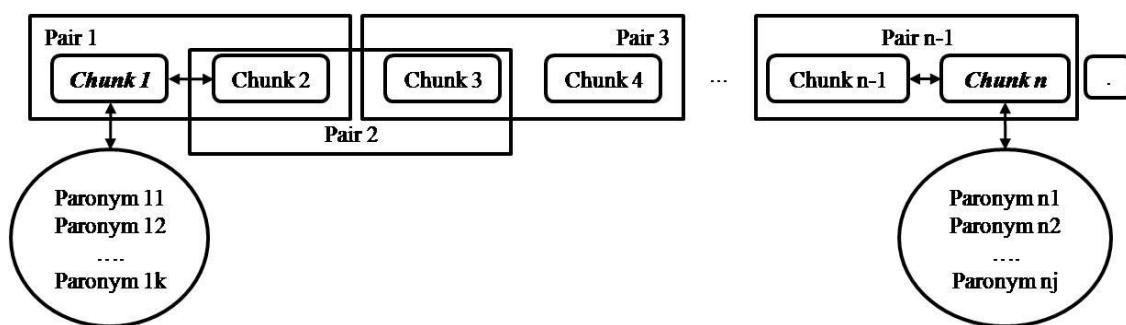


Figure. 2: Isolated malapropism found in the first/last word in the phrase. The malapropism is found in one of the two chunks written in italics. The paronyms of the malapropos words are chosen in conjunction with the only chunk that it relates to.

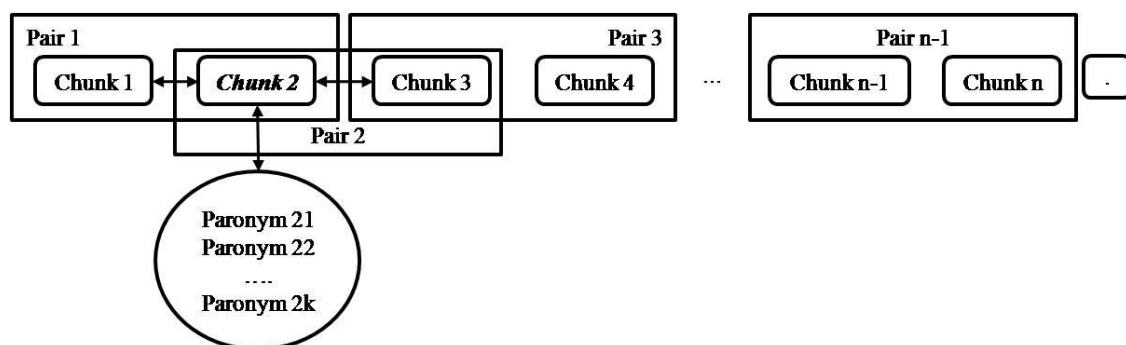


Figure. 3: Isolated malapropism found in the middle of the phrase. The malapropism is found in the chunk written with italics. The paronyms of the malapropos word are chosen in conjunction with both the chunks that surround the one containing the malapropism.

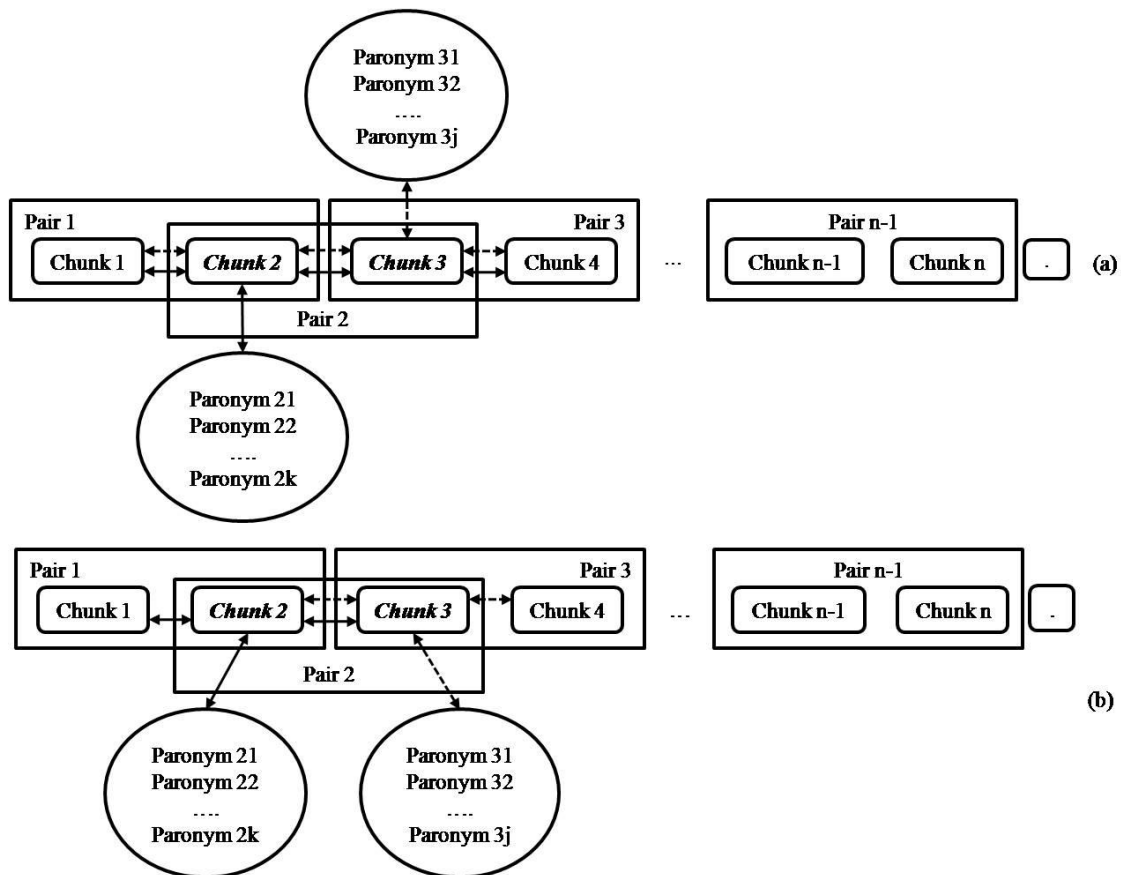


Figure 4: Malapropisms chain. The malapropisms are found in the chunks written in italics. In the first case (a) only one of the malapropos words is corrected so that both malapropisms disappear. The actions described by either the continuous arrows or by the interrupted ones are executed. In the second case (b), the malapropos words are handled independently, so both the actions described by the continuous and interrupted arrows are executed.

Ideally, we obtain a list of paronyms that fit perfectly in the phrase without drops in cohesion between the chunks that the malapropism is part of and the one before and after it. These words become candidates for replacement and the signalled malapropism is marked as a real one that should be corrected. If the paronym fits with only one of the chunks, it is also saved as a possible candidate, but weaker than the regular ones, needing further investigation. The malapropism is not yet marked, but the signal received from the detection is not ignored either.

Sometimes a malapropisms chain is observed in the phrase (Fig. 4). Most of the time, this is caused by a malapropism that makes both its chunk and the next one to be signalled as possible malapropisms. To solve this problem, we start from the premise that only one of the chunks contains a malapropism, and try to find a replacement that corrects both the malapropisms (Fig. 4a) for only one of the two malapropos words. In this case, two pairs of chunks

are corrected together: the one containing both the malapropisms and one of the two pairs containing only one of the signalled malapropisms. If this correction is not possible, then each malapropism is handled independently, trying to correct both of them separately (Fig. 4b). If this is still impossible, then we try to correct at least one of the malapropisms, leaving the other one as it is.

At this point, if none of the paronyms of a signalled malapropos word fits into the local context, without damaging the cohesion, then the signalled malapropism is considered a false alarm.

The next step is to filter the candidates for replacing a malapropos word against the text logic. The idea behind this step is that a word that fits in the logical presentation of the text topics should be encountered. To verify if the word fits in the text logic, we extracted the lexical chains of the given text and tried to see if the candidates can be found in one of these chains. The candidates that did not fit in any chain have been discarded. Again, if no

candidates have been kept for a signalled malapropism, the signal is ignored and considered a false alarm. If there is a single candidate for a malapropos word, then that candidate replaces it and a correction is signalled. If there are more candidates to replace a malapropos word, then they are evaluated using the search engine in the same way as in the detection module and the candidate with the best score is chosen as a replacement.

4 WALKTHROUGH EXAMPLE

In order to demonstrate our approach, we shall present an example of the detection and correction mechanisms described in the previous section, considering a simple example:

I am travelling around the *word* [world].

In the first step, we identify the lexical chains of each content word from the phrase, using WordNet. Then, the phrase is split into the following chunks using the pseudo-chunker: *I, am travelling and around the word*. Afterwards, we look for the co-occurrence of any two consecutive chunks and we obtain the following results using Google:

"I am travelling" – 1620000 hits
"am travelling around the word" – 3 hits

Thus, the first combination will be considered correct because of the large number of hits (having over 16000 hits), while the second combination is signalled as a possible malapropism due to the very low number of hits (below 20 hits).

The next step is to look for the paronyms of the content word in the signalled chunk, which are: *cord, ford, lord, sword, ward, wyrd, woad, wold, wood, wordy, work, worm, worn, wort, world*. We replace the content word with each of its paronyms and we look for the number of hits of the newly obtained chunk, along with the previous one. For each such combination, we try to apply one of the filters described in the previous section. The only one that passes through the filters is "am travelling around the world" which has 4120 hits and it is passed through the third filter. Replacing these results into (1) and considering beta equal to 1.00 we obtain a valid relation that shows that *world* is a valid candidate in order to correct the signalled malapropos *word*.

In conclusion, a malapropism is signalled and the corrected form is presented to the user:

I am travelling around the world.

5 EXPERIMENTS

The accuracy of the application depends greatly on the completeness of the paronyms dictionary because the correction method relies on the fact that the candidate words for replacing the malapropos ones are available in the dictionary. And since it contains only first-level paronyms, it means that the application is limited only to the detection of the malapropisms where the correct word and the malapropos one are first-level paronyms. Another limitation of this method rises from the fact that the dictionary has been built based on the concepts from WordNet, without containing declined forms of the words.

Considering these limitations, we had to build some test corpora in order to determine the accuracy of our approach. Therefore, three different types of corpora have been used: the first corpus was built from individual phrases that contained malapropisms in order to evaluate the rate of malapropisms detection and correction; the second contained no malapropisms at all and was used in order to estimate the rate of false alarms; and finally, the third type of corpus consisted of parts of text published on the Internet and modified in the same manner suggested in (Hirst and Budanitsky, 2005) and (Hirst and St-Onge, 1998).

The first corpus was built to evaluate the malapropisms detection and correction rate and contains 31 distinct phrases. The first 11 of them are variants of the examples 1-8, 11, 12 and 14 taken from (Bolshakov and Gelbukh, 2003) that are adapted to suit the limitations of our application by changing the malapropos word that was a second-level paronym of the correct word by a first-level paronym of this word. The phrases 12-15 are the examples 4, 6, 7 and 8 from (Hirst and Budanitsky, 2005), while the phrases 16-18 are the examples 10-12 from (Hirst and St-Onge, 1998). The rest of the corpus has been built by the authors.

1. They are travelling around the *word* [world].
2. The salmons swim upstream to *pawn* [spawn].
3. Take it for *granter* [granted].
4. The *bowel* [vowel] is pronounced distinctly.
5. She has a very loose *vowel* [bowel].
6. He wears a *turfan* [turban] on his head.
7. This is an *ingenuous* [ingenious] machine for peeling bananas.
8. A quite affordable *germ* [term] is proposed.
9. The kinds of Greek columns are Corinthian, Doric, and *Ironic* [Ionic].

10. The desert was activated by *irritation* [irrigation].
11. This is only a scientific *hypothesise* [hypothesis].
12. It is my sincere *hole* [hope] that you will recover swiftly.
13. Maybe the reasons the House Democrats won't let the contras stand and fight for what they believe in is because the Democrats themselves no longer stand and fight for their beliefs. The House's liberals want to pull the plug on the rebels but, lacking the courage to hold a straight up or down vote on that policy and expose its consequences to the U.S. electorate, they have to disguise their *intension* [intention] as a funding "moratorium."
14. American Express says . . . it doesn't know what the *muss* [fuss] is all about.
15. Mr. Russell argues that usury *flaw* [law] depressed rates below market levels year ago . . .
16. Much of that data, he notes, is available *toady* [today] electronically.
17. Among the largest OTC issues, Farmers Group, which expects B.A.T. Industries to launch a hostile *tenter* [tender] offer for it, jumped 2 3/8 to 62 yesterday.
18. But most of yesterday's popular issues were small out-of-the-limelight technology companies that slipped in price a bit last year after the *crush* [crash], although their earnings are on the rise.
19. My *chat* [cat] likes mice.
20. The question is: to eat or not to *beat* [eat].
21. Move your *spawn* [pawn] to attack the queen.
22. A *core* [sore] throat is the thing I want less.
23. *Boss* [Toss] a coin and see whether it is tails or not.
24. He has a beautiful deep *vice* [voice].
25. I want to watch a horror *move* [movie] on TV.
26. We should sharpen our *glade* [blade] and attack the enemy.
27. People said they saw an unidentified flying *abject* [object].
28. *Mild* [Wild] places are hard to find these days.
29. He climbed up the *bill* [hill].
30. The superstition of seeing a *back* [black] cat is one of the most well-known and popular superstitions today.
31. We should pay *deed* [heed] to the words of our elders.

For the examples 9, 10, 13 from (Bolshakov and Gelbukh, 2003) and 5 from (Hirst and Budanitsky, 2005) we could not find a first-level paronym to replace the original malapropos that was a second-level paronym of the correct word. We have also tested with the original examples from (Bolshakov and Gelbukh, 2003) and (Hirst and Budanitsky, 2005), but in this case we have manually added to the dictionary the second-order paronyms of the correct words.

The second corpus that has been used to test our approach consisted only of phrases that had no malapropisms at all. For this corpus, we have used the examples 1-5, 8 and 14-16 from (Hirst and St-Onge, 1998) and the examples 9 and 10 from (Hirst and Budanitsky, 2005). The rest of the corpus was built from news taken from Yahoo (news.yahoo.com) in mid June 2009:

1. The North's threats were the first public acknowledgment that the reclusive communist nation has been running a secret uranium enrichment program.
2. The resolution also authorized searches of North Korean ships suspected of transporting illicit ballistic missile and nuclear materials.
3. President Barack Obama says he's now found savings that will pay almost all the costs of a massive overhaul of America's health care system.
4. Any honest accounting must prepare for the fact that health care reform will require additional costs in the short term in order to reduce spending in the long term.
5. She has handled only a small number of K-12 education cases during her 17 years on the federal bench, but the trials-- which have focused on such key issues as special education, racial discrimination, and student freedom of expression --could offer clues on future school policy matters if she joins the court.
6. The big goals of the new American general taking charge of the war in Afghanistan start with fixing a problem that bedeviled the man he is replacing: the repeated, inadvertent killing of civilians.
7. Nearly 700,000 calls were received by a federal hot line this week from people confused about the nationwide switch from analog to digital TV broadcasts that occurred Friday.
8. About a third of the calls were about federal coupons to pay for digital converter boxes,

an indication that at least 100,000 people still didn't have the right equipment to receive digital signals.

9. He jokes about all politicians but it's becoming clearer where his sympathies lie — something that Palin and her supporters sensed in their criticisms.
10. The rival candidate said the vote was tainted by widespread fraud and his followers responded with the most serious unrest in the capital in a decade.
11. If everyone at the Tony's were aware that Bret missed his mark, then they should have been aware enough to stop the set piece from hitting him or at least slowed it down until he cleared the stage.

Finally, for the third type of corpus we have considered two distinct corpora: a small one, containing a few paragraphs (199 words) taken from a Fox News, and a larger one, consisting of 2083 words.

In this text, we introduced a malapropism by replacing the word fraud by one of its first-level paronyms, frau. In the bigger corpus – which is too large to be presented here –, we randomly introduced 25 malapropisms in a manner similar to the one used in (Hirst and Budanitsky, 2005) and (Hirst and St-Onge, 1998).

6 INTERPRETATION OF THE RESULTS

For the first corpus, 27 out of the 31 examples were correctly detected and 25 of them were properly corrected, representing an accuracy of 87.05% for the malapropism detection and 80.64% for correction. Only 4 examples were not detected (9, 10, 12, 15) and another two were wrongfully corrected (muss was replaced by mass instead of fuss in example 14, while crush was replaced by rush instead of crash in 18). The tests made on the phrases containing second-level paronyms have shown that these malapropos words could be properly corrected if the paronyms existed in the dictionary. These malapropisms were always detected, and all of them have also been corrected after the corresponding paronyms were manually inserted in the dictionary. This result made us believe that the method could also have good results if applied to the correction of the malapropos words that are second-level paronyms to the correct word,

and suggested us to build another dictionary that contains both first and second-level paronyms.

The second corpus, built in order to evaluate the rate of false alarms introduced by the application, contained 587 words. Only one false alarm was inserted, replacing the word “while” with “white” in the example 16 taken from (Hirst and Budanitsky, 2005) (*And while institutions until the past month or so stayed away from the smallest issues for fear they would get stuck in an illiquid stock,...*). This false alarm was caused by the POS-tagger that we used because it wrongly identified “while” as being a noun and replacing it with the more plausible word “white”. This test has shown that the application has a probability of only 0.17% of introducing a false malapropism in the text. Since the text did not contain any malapropisms, the probability has been computed as the ratio between the number of introduced malapropisms divided by the total number of words from the corpus.

Considering the good results that we obtained for these two compiled corpora, we decided to try the application on real texts. The test for the smaller text, containing almost 200 words and one malapropism, has shown that the malapropism has been corrected, but a false alarm has been introduced by replacing the correct word “fighting” with the word “sighting”. This test has shown that we underestimated the rate of false alarms, which for this text was 0.5%.

The test on the larger text (the Yahoo News), shown that 21 malapropisms have been detected out of the 25, and 17 of them were properly corrected. The application has also introduced 10 false alarms, by replacing some correct words. The results of this test have shown an application performance of 84% for the detection rate, 68% for correction and a false alarm rate of 0.48%. Analyzing the false alarms introduced by the application, we have seen that 6 out of the 10 malapropisms introduced by the application were in the vicinity of a proper noun, one of them being exactly a proper noun malapropism (Iran has been replaced by Iraq – this happened because both countries have similar contexts on the Internet: geographically, politically, religiously, etc.). This observation upheld our insight that the application has problems with the proper nouns, the numbers and the metaphors found in the analyzed texts.

7 CONCLUSIONS

In this paper we have presented a fully automatic method for malapropisms detection and correction

for texts written in English, having very good results for this very difficult task: between 84% and 87% for malapropism detection, between 68% and 80% for malapropisms correction and around 0.5% rate of introducing new malapropisms in texts. Moreover, this method could be easily adapted for correcting malapropisms in any language if a lexical database similar to Wordnet and a paronyms dictionary are available for that language.

ACKNOWLEDGEMENTS

The research presented in this paper was partially performed under the FP7 EU STREP project LTfLL.

REFERENCES

- Bolshakov, I.A., Galicia-Haro, S.N., Gelbukh, A., 2005. Detection and Correction of Malapropisms in Spanish by means of Internet Search. *8th International Conference Text, Speech and Dialogue (TSD-2005)*, Karlovy Vary, Czech Rep. In: *Lecture Notes in Artificial Intelligence* (indexed in SCIE), N 3658, ISSN 0302-9743, ISBN 3-540-28789-2, Springer-Verlag, pp. 115–122.
- Bolshakov, I., Gelbukh, A., 2003. On Detection of Malapropisms by Multistage Collocation Testing. NLDB-2003, *8th International Conference on Application of Natural Language to Information Systems*, June 23–25, 2003, Burg, Germany. In: *Lecture Notes in Informatics.*, Bonner Köllen Verlag, ISSN 1617-5468, ISBN 3-88579-358-X, pp. 28–41.
- Bolshakova, E., Bolshakov, I.A., Kotlyarov, A., 2005. Experiments in Detection and Correction of Russian Malapropisms by Means of the Web. In: *International Journal on Information Theories & Applications*. V.12, N 2, p 141-149.
- Gale, W. A., Church, K. W. and Yarowsky, D., 1993. A method for disambiguating word senses in a large corpus. In *Computers and the Humanities*, 26:415–439.
- Gelbukh, A., Bolshakov, I.A., 2004. On Correction of Semantic Errors in Natural Language Texts with a Dictionary of Literal Paronyms. Jesus Favela, Ernestina Menasalvas, Edgar Chávez (Eds.) *Advances in Web Intelligence (AWIC-2004, 2nd International Atlantic Web Intelligence Conference, May 16–19, 2004, Cancun, Mexico)*. In: *Lecture Notes in Artificial Intelligence* (indexed by SCIE), N 3034, Springer-Verlag, ISSN 0302-9743, ISBN 3-540-22009-7, pp. 105–114.
- Golding, A., 1995. A bayesian hybrid method for context-sensitive spelling correction. In *The Third Workshop on Very Large Corpora*, pages 39–53.
- Golding, A. and Schabes, Y., 1996. Combining trigram-based and feature-based methods for context sensitive spelling correction. In *34th Annual Meeting of the Association for Computational Linguistics*.
- Hirst, G., Budanitsky, A., 2005. Correcting Real-Word Spelling Errors by Restoring Lexical Cohesion. In: *Computational Linguistics. Natural Language Engineering*, 11:87–111.
- Hirst, G., St-Onge, D., 1998. Lexical Chains as Representation of Context for Detection and Corrections of Malapropisms. In: C. Fellbaum (ed.) *WordNet: An Electronic Lexical Database*. The MIT Press, p. 305-332.
- Marshall, I., 1983. Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB corpus. In *Computers and the Humanities*, 17:139–150.
- Mays, E., Damerau, F. J. and Mercer, R. L., 1991. Context based spelling correction. In *Information Processing and Management*, 27(5):517–522.
- Wilcox-O’Hearn, A., Hirst, G. and Budanitsky, A., 2006. Real-word spelling correction with trigrams: A reconsideration of the Mays, Damerau, and Mercer model. In *CICLing-2008, 9th International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 605–616, Haifa, Israel.