

Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness

Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger

Abstract—We present a language-independent spell-checker that is based on an enhancement of the n-gram model. The spell checker is proposing correction suggestions by selecting the most promising candidates from a ranked list of correction candidates that is derived based on n-gram statistics and lexical resources. Besides motivating and describing the developed techniques, we briefly discuss the use of the proposed approach in an application for keyword- and semantic-based search support. In addition, the proposed tool was compared with state-of-the-art spelling correction approaches. The evaluation showed that it outperforms the other methods.

Index terms—Spelling correction, n-gram, information retrieval effectiveness.

I. INTRODUCTION

THE problem of devising algorithms and techniques to automatically correct words in texts has become a perennial research challenge. Work began as early as the 1960s on computer techniques for automatic spelling correction and automatic text recognition, and it has continued up to the present. There are good reasons for the continuing research efforts in this area in order to improve quality and performance and to broaden the spectrum of possible applications [1]. For example, even though system programs (language processors, operating systems, etc.) have become increasingly powerful and sophisticated, they do not assist the user (with very few exceptions) in correcting many of the obvious spelling errors in the source input. There are two types of word errors, the real-word error and the non-word error. Real-word errors are misspelled words that have a meaning and can be found in a dictionary. Non-word errors are words that have no meaning and are thus not included in a dictionary. We concentrate on the correction of the non-word error with the proposed algorithm. Damerau (1964) found that 80% of misspelled words that are non-word errors are the result of a single insertion, deletion, substitution or transposition of letters [2]. Therefore, it seems reasonable to base correction algorithms on measures that consider these simple operations. However, approaches based on pure n-

gram statistics (which account for these operations implicitly) have also proven to provide good performance [1, 15].

In this paper, we propose an approach that is based on an enhancement of the n-gram model. Therefore, we first discuss briefly, related work on spelling correction in Section 2. Afterwards, we describe, in detail, in Section 3 our spell checking approach MultiSpell. In Section 4, we present an evaluation based on benchmark data sets in the English and Portuguese language and conclude with a brief discussion.

II. APPROACHES OF SOME SPELL CHECKERS

Algorithmic techniques for detecting and correcting spelling errors in text have a long and robust history in computer science [1]. Many approaches have been applied since people started to deal with this problem. Different techniques like edit distance [4], rule-based techniques [10], n-grams [20], probabilistic techniques [14], neural nets [15], similarity key techniques [16, 17] and noisy channel model [18, 19] have been proposed. All of these are based on the idea of calculating the similarity between the misspelled word and the words contained in a dictionary. In the following, we describe briefly one of the most popular approaches (Aspell) and one recently proposed approach for the Portuguese language (TST) [13] that we used for comparison.

GNU Aspell, usually called just Aspell, is a standard spell-check software for the GNU software system. There are dictionaries for about 70 languages available. GNU Aspell is a Free and Open Source and can be downloaded under <http://aspell.sourceforge.net/>. In contrast to Ispell, which suggests words with small edit-distance, Aspell in addition compares sounds-like equivalents (computed for English words using the metaphone algorithm [21]) up to a given edit distance.

The Ternary Search Trees [13] approach (TST) is a dictionary data structure working with string-keys. It can find, remove and add these keys quickly and also easily search the tree for partial matches. Additionally, near-match functions can be implemented. These give the possibility to suggest alternatives for misspelled words.

For a more conclusive overview of spell-check approaches see [1, 15].

Manuscript received October 23, 2008. Manuscript accepted for publication August 22, 2009.

Farag Ahmed and Andreas Nürnberger are with Data and Knowledge Engineering Group, Institute for Knowledge and Language Engineering, Otto-von-Guericke University of Magdeburg, Germany.

Ernesto William De Luca is with Competence Center Information Retrieval & Machine Learning Distributed Artificial Intelligence Laboratory, Technical University of Berlin, Germany.

III. AN ALGORITHM BASED ON N-GRAM STATISTICS: MULTISPELL

The algorithm we propose, in the following, is a language-independent spell-checker that is based on an enhancement of the n-gram model. It is able to detect the correction suggestions by assigning weights to a list of possible correction candidates, based on n-gram statistics and lexical resources, in order to detect the non-word errors and to derive correction candidates. In the following, we describe first of all the lexical re-source we used (MultiWordNet) and then in detail the proposed MultiSpell algorithm.

A. Lexical Resources

Lexical resources provide linguistic information about words of natural languages. This information can be represented in very diverse data structures, from simple lists to complex resources with many types of linguistic information and relations associated with the entries stored in the resource.

These resources are used for preparing, processing and managing linguistic information and knowledge needed for the computational processing of natural language [3]. An example of such large scale lexical resources is given by linguistic ontologies that cover many words of a language and have a hierarchical structure based on the relationship between concepts.

We propose to use these dictionaries, and especially MultiWordNet [6], the most important lexical resource available. It covers nouns, verbs, adjectives and adverbs. For our purpose, we use the words provided (~80000 entries for the English language) from this resource to correct the misspelled word. Therefore, we extracted all words contained in it with all its linguistic relationships.

B. Computing Similarity Scores Based on N-Grams

The idea of using n-grams in language processing was discussed first by Shannon [8]. After this initial work, the idea of using n-grams has been applied to many problems such as word prediction, spelling correction, speech recognition, translated word correction and string searching. One main advantage of the n-gram method is that it is language independent.

In a spelling correction task, an n-gram is a sequence of n letters in a word or a string. The n-gram model can be used to compute the similarity between two strings, by counting the number of similar n-grams they share. The more similar n-grams between two strings exist the more similar they are. Based on this idea the similarity coefficient [9] can be derived. The similarity coefficient δ is defined by the following equation:

$$\delta_n(a, b) = \frac{|\alpha \cap \beta|}{|\alpha \cup \beta|} \quad (1)$$

where α and β are the n-gram sets for two words a and b to be compared. $|\alpha \cap \beta|$ denotes the number of similar n-grams in α and β , and $|\alpha \cup \beta|$ denotes the number of unique n-grams in the union of α and β . Table I shows an example for

the calculation of the similarity coefficient for the misspelled word “secceded” and the correct word “succeeded” using an n-gram with $n=2$ (bigram).

TABLE I
CALCULATING THE BIGRAMS SIMILARITY COEFFICIENT BETWEEN TWO STRINGS.

bi-grams union	<i>succeeded</i>	<i>secceded</i>
<i>su</i>	1	-
<i>uc</i>	1	-
<i>cc</i>	1	1
<i>ce</i>	1	1
<i>ee</i>	1	1
<i>ed</i>	1	1
<i>de</i>	1	1
<i>ed</i>	1	1
<i>se</i>	-	1
<i>ec</i>	-	1
Similarity coefficient	6/10 = 0.6	

C. Revised N-Gram Based Approach

Yannakoudakis and Fawthrop [10] found that in most cases the first letter in the misspelled word is almost always correct and also the misspelled and real word will be either the same length or the length differs just by one. For some examples, we like to refer the reader to the list of commonly misspelled words in English published in [12]. Furthermore, the pure n-gram based approach to compute the similarity coefficient as described above, does not consider the order of the n-grams [22]. This might, however, be important since typing or misspelling errors usually affect only a specific part of the word. Therefore, we revised the computation of a similarity between words to take these two aspects into account.

In the following, we describe our algorithm for $n=2$ (bigrams) for simplicity. However, the approach can be applied for trigrams and n-grams with $n > 3$ as well. We define bigrams of words by their respective position in the word $w_{i,i+(n-1)}$ where i defines the position of the first letter and $i+(n-1)$ the position of the last letter of the considered n-gram. Thus, the last possible position of an n-gram in a word is defined by $j=|w|-n+1$, where $|w|$ defines the length of the word.

In order to consider the findings of Yannakoudakis and Fawthrop as mentioned above, we replace the first and the last n-gram by the first and the last letter of the respective words. Thus, when computing the similarity score these elements are compared directly, independent of the remaining n-grams between them.

In order to deal with the second aspect mentioned above, we define a window of n-grams of the correction candidate words that should be compared, i.e. while in Eq. (1) all n-grams are compared with each other, we only compare n-grams that are in close proximity to the position of the n-gram in the word to be corrected when computing the similarity score. An example is given in Fig. 1, where w' defines the misspelled word and w a correction candidate. Here, the n-gram $w'_{4,5}$ of w' will only be compared to the n-grams $w_{3,4}$,

$w_{4,5}$ and $w_{5,6}$ of the correction candidate w , i.e. even if the n-gram $w'_{4,5}$ is similar to $w_{2,3}$ this would not count towards the similarity score of the words w' and w .

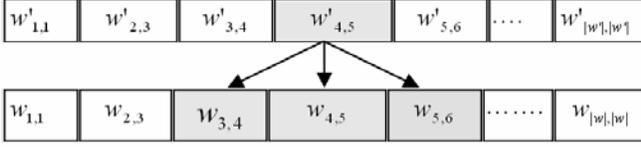


Fig. 1. Bigram comparison for misspelled word w' and a correction candidate w using a comparison window of size 3. Notice that the first and last n-gram represent the first and the last letters only and are therefore always of size one.

Overall, the computation of the similarity score S for a given n-gram size n and a given odd-numbered window size m can be defined as follows, assuming that u is the longer word (if v is longer than u and v can simply be exchanged):

$$S_{n,m}(u,v) =$$

$$\frac{g(u_{1,1}, v_{1,1}) + g(u_{|u|,|u|}, v_{|v|,|v|}) + \sum_{i=2}^{|u|-n+1} \sum_{j=\frac{m-1}{2}}^{\frac{m-1}{2}} g(u_{i,i+(n-1)}, v_{i+j,i+j+(n-1)})}{N} \quad (2)$$

$$\text{where } g(a,b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{otherwise} \end{cases} \quad \text{and}$$

$$u_{i,j} = \begin{cases} \text{substring}(u, i, j) & \text{if } i \leq j \\ "" & \text{otherwise.} \end{cases}$$

Here, $g(u_{1,1}, v_{1,1})$ compares the first and $g(u_{|u|,|u|}, v_{|v|,|v|})$ the last characters of the words u and v and the nested sum counts the number of n-grams in v that are similar to n-grams in a window of size m around the same position in word v . N is computed similarly as in Eq. (1). In Fig. 2 the specific cases that have to be considered when computing the similarity score S are summarized.

D. The MultiSpell Algorithm

The first stage of the MultiSpell algorithm is to compare the keywords given from the user with the correct words contained in the dictionary. First of all, we check based on the used dictionary (here, based on the words extracted from MultiWordNet) if the word is misspelled. If this is the case, the algorithm builds n-grams for the misspelled word. Then we select correction candidates from the dictionary. In order to keep the number of correction candidates as small as possible, we select only words as candidates that are two characters shorter or longer than the misspelled word. This is motivated by the work of Turba [11], who has shown that most misspelled words differ in length only by one character from the correct word.

For the selected words the n-grams are computed and the similarity score is computed according to Eq. (2). The correction candidates can then be simply sorted by the obtained similarity score and the word with the highest score is proposed as the best correction candidate.

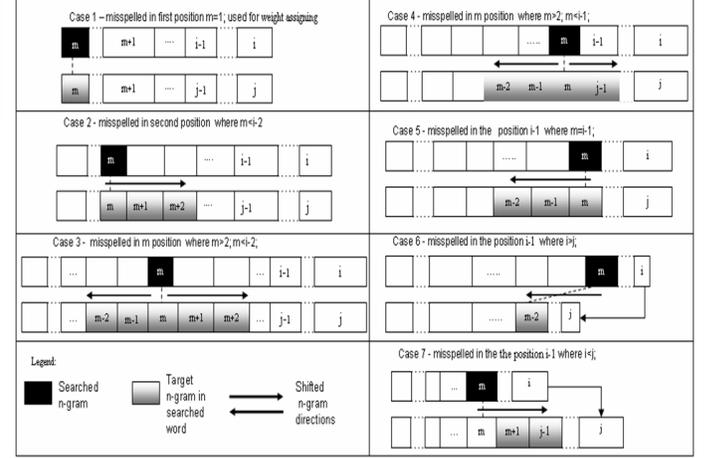


Fig. 2. Comparing n-grams based on the MultiSpell algorithm.

E. Spelling Correction for Keyword- and Semantic-based Search Support

MultiSpell has been also integrated as a pre-processing approach in the Sense Folder Framework [25]. It can be applied to queries and documents, in order to support users during keyword-based and semantic-based search. The first is an important task for retrieving the relevant documents related to the query identifying the misspelled words and correct them for a correct interpretation [23] (see also Fig. 3). The second is specifically trying to improve the semantic search process [24]; therefore several problems have to be addressed, before the semantic classification of documents is started. When users mistype the query in writing, the system has to be able to give correction alternatives to continue the semantic-based search.

The semantic-based search differs from the “normal” search, because users are “redirected” to semantic concepts that could describe their query. This semantic support is provided in the user interface. On the left side of the user interface (see Fig. 4) suggestions are generated by MultiSpell and presented to the user for starting the semantic-based search.

In this case, the use of Multispell is mostly helpful, not only because it performs an efficient correction (as shown in Fig. 3), but also because it can “redirect” the user to a semantic search (see Fig. 4). Thus, if the user types a word that is not contained in the lexical resource used, the system can suggest other “similar” words according to the words found in the resource. Then, a semantic classification is started using the words selected by the user.

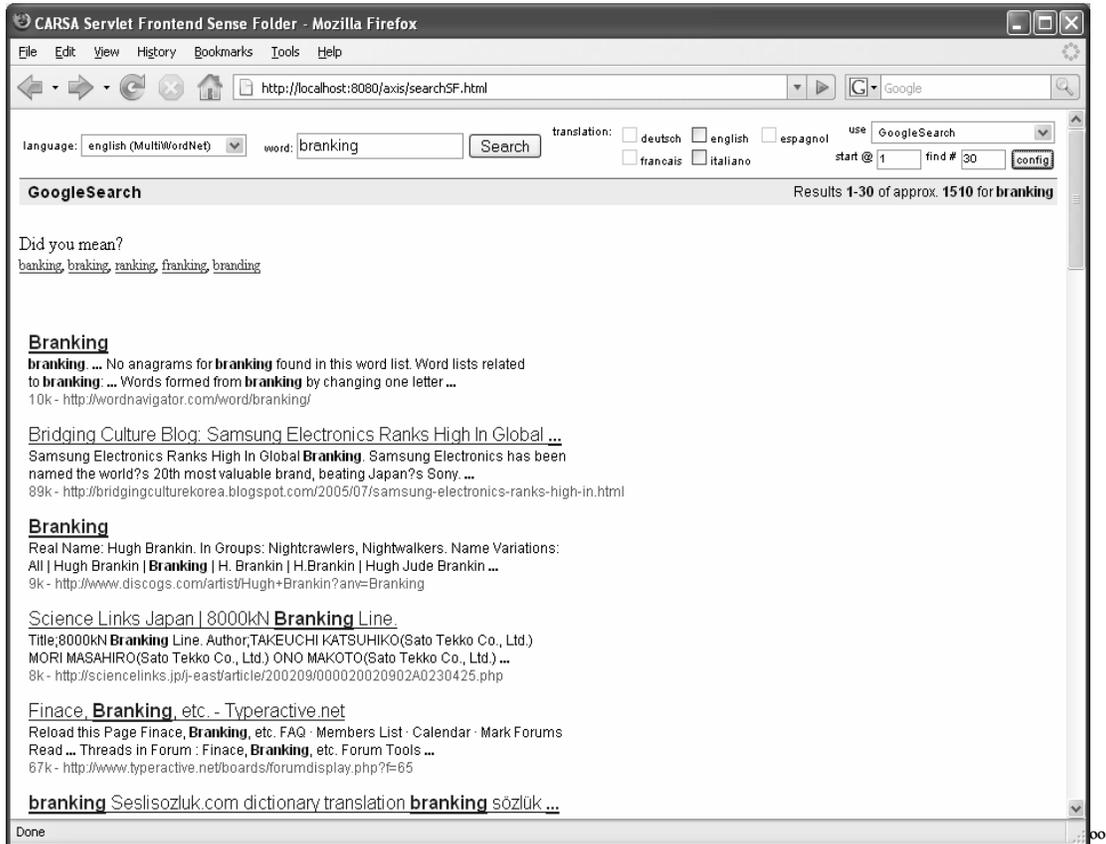


Fig. 3. Corrections for a misspelled word (MultiSpell) in the Sense Folder Framework .

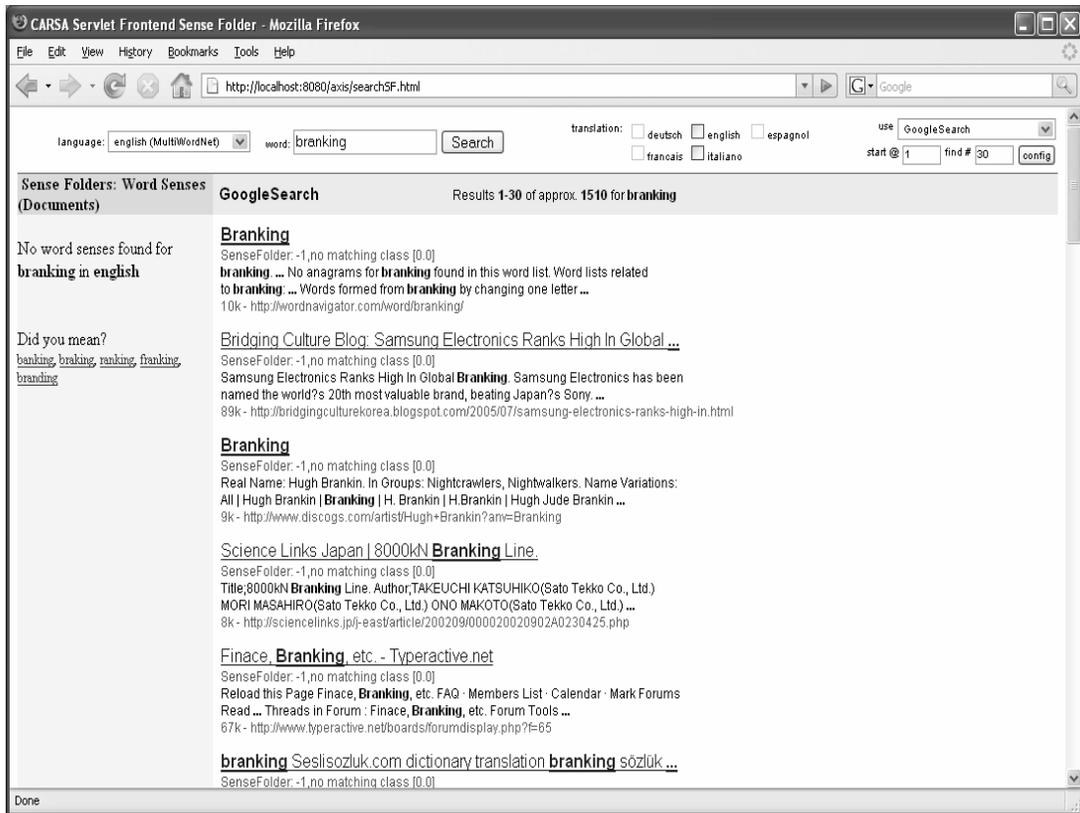


Fig. 4. Using MultiSpell in the Sense Folder Framework for Semantic Search Support.

IV. COMPARISON AND EVALUATION OF RESULTS

In the following, we show results of some experiments done for the English and Portuguese language. The first evaluation was done on the whole English commonly misspelled word list provided in [12]. Afterwards, we compared the results of our spell checker MultiSpell with the results of the TST approach (in one experiment, for the Portuguese language) and of the Aspell approach (in two experiments, for the Portuguese and the English language), showing that the proposed approach always achieved the best results.

For the first evaluation, we used the whole list of commonly misspelled words in English consisting of 3975 words as published in [12]. This list of common spelling mistakes is represented by a table consisting of two columns. The first one shows the misspelled word, the second the correct spelling. For the evaluations, we only considered the correction words that were ranked as best correction word, i.e., even if the second word would have been the correct candidate, this was counted as a wrong correction. We first used all misspelled words of the list, using the bigram case and just the first candidate correction. MultiSpell corrected 3334 misspelled words (84%) and failed for 641 misspelled words (16%) although it provided similar corrections in many cases. For example the word *advice* was suggested instead of *advised* for the misspelled word *advised*. Another example is the provided correction *algebraically* instead of *algebraic* for the misspelled word *algebraical* (see Table V in the Appendix). These suggestions were classified as wrong in our approach, even though they belong to the same word sense. Second, we used trigrams. This showed lower performance and efficiency. MultiSpell corrected 2900 words (73%) and failed for 1075 (27%) as shown in Table II.

A. Evaluation of English Spelling Correction

For the second evaluation, we randomly selected a set of only 120 misspelled words obtained from Wikipedia [12] and not the whole list. All error types and starting letters of the words were taken into account. We compared MultiSpell with Aspell, MicrosoftWord, and Google. Since Aspell provides a list of candidate corrections we took just the first candidate from the list assuming that the first candidate is the most likely one proposed by the algorithm. MicrosoftWord and Google provided only one correction candidate. Table III and Table V (in the Appendix) show that MultiSpell finds the correct spelling for 109 words (90%). In comparison, Google can correct 106 (88%) words, while Aspell and MicrosoftWord 105 words (87.5%). MultiSpell detected 6 of 16 of the multiple correction words (which have more than one possible correction), but it doesn't fail to provide at least one correct suggestion. Aspell detected just two of the multiple corrections and it failed just one time to provide a suggestion for one of the multiple corrections.

TABLE II
COMPARISON BETWEEN BIGRAM AND TRIGRAM IN WHOLE ENGLISH DATA SET (3975 WORDS).

	bigram	trigram
correct	3334 (84%)	2900 (73%)
wrong	641 (16%)	1075 (27%)

TABLE III
COMPARISON OF MULTISPELL, ASPELL, MICROSOFT WORD AND GOOGLE FOR ENGLISH.

	MultiSpell	Aspell	Microsoft Word	Google
correct	109 (90%)	105 (87.5%)	105 (87.5%)	106 (88%)
wrong	11 (10%)	15 (12.5%)	15 (12.5%)	14 (12%)

TABLE IV
COMPARISON OF MULTISPELL, ASPELL AND TST FOR THE PORTUGUESE LANGUAGE.

	MultiSpell	TST	Aspell
correct	97 (80%)	78 (65%)	65 (54%)
wrong	23 (20%)	42 (35%)	55 (46%)

B. Evaluation of Portuguese Spelling Correction

The last evaluation was done for the Portuguese language. Bruno and Mário [13] implemented an algorithm using Ternary Search Trees (TST). The authors show experiments in correcting a list of some Portuguese words and comparing their results with Aspell. Here we compared MultiSpell on the whole list (120 Portuguese words) available from their experiments explained in [13], applying our algorithm and comparing it with the Aspell and TST algorithm. Given that MultiWordNet does not provide any Portuguese word senses, we used the dictionary made available from [13] comparing the approaches. Our algorithm succeeded to correct 97 misspelled words (80%), TST succeeded to correct 78 misspelled words (65%) and Aspell succeeded to correct 65 misspelled words (54%) as shown in Table IV and Table VI (in the Appendix).

IV. CONCLUSIONS

In this paper we proposed a language-independent spell-checker that is based on an enhancement of a pure n-gram based model. Furthermore, we presented evaluations on English and Portuguese benchmark data sets of misspelled words. The obtained results outperformed other state-of-the-art methods. In future work, we plan to further optimize the algorithm and data structure used to compute the similarity scores. Furthermore, the algorithm should be tested on data sets for other languages.

APPENDIX: EVALUATION TABLES FOR ENGLISH AND PORTUGUESE

Table V contains results of word corrections in English, while Table VI contains results of word corrections in Portuguese.

TABLE V
RESULTS OF WORD CORRECTIONS IN ENGLISH.

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
Abberation	aberration	aberration	aberration	aberration	aberration
accomodation	accommodation	accommodation	accommodation	accommodation	accommodation
acheive	achieve	Achieve	achieve	achieve	achieve
abortifcant	abortifacient	aficionados	-	abortifacient	abortifacient
absorbision	absorption	absorbs ion	absorbision	absorption	absorption
ackward	(awkward, backward)	awkward	(awkward, backward)	awkward	(awkward, backward)
additinally	additionally	additionally	additionally	additionally	additionally
adminstration	administration	administration	administration	administration	administration
admissability	admissibility	admissibility	admissibility	admissibility	admissibility
advertisments	advertisements	advertisements	advertisements	advertisements	advertisements
advised	advised	advised	advised	advice	advice
aficionados	aficionados	aficionados	aficionados	aficionados	aficionados
affort	(effort ,afford)	effort	afford	afford	afford
agains	against	agings	agings	against	against
agreement	agreement	agreement	agreement	agreement	agreement
agressively	aggressively	aggressively	aggressively	aggressively	aggressively
agriculturalist	agriculturist	-	-	-	agriculturist
alcoholical	alcoholic	alcoholically	alcoholically	alcoholic	alcoholic
algebraical	algebraic	algebraic	algebraically	algebraic	algebraically
algoritms	algorithms	algorithms	algorithms	algorithms	algorithms
alterior	(ulterior , anterior)	ulterior	(anterior, ulterior)	ulterior	(anterior, ulterior)
anihilation	annihilation	annihilation	annihilation	annihilation	annihilation
anthromorphization	anthropomorphization	anthropomorphizing	-	-	anthropomorphization
bankrupcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy	bankruptcy
baout	(about,bout)	bout	(about,bout)	about	bout
basicly	basically	basically	basically	basically	basically
breakthoug	breakthrough	break though	breakthrough	breakthrough	breakthrough
carachter	character	crocheter	character	character	character
cannotation	connotation	connotation	(connotation ,annotation)	connotation	(connotation ,annotation)
carismatic	charismatic	charismatic	charismatic	charismatic	charismatic
carmel	caramel	Carmel	-	-	caramel
cervial	(cervical, servile)	cervical	cervical	cervical	cervical
clasical	classical	classical	classical	classical	classical
cleareance	clearance	clearance	clearance	clearance	clearance
comissioning	commissioning	commissioning	commissioning	commissioning	commissioning
commemerative	commemorative	commemorative	commemorative	commemorative	commemorative
compatabilities	compatibilities	compatibilities	compatibilities	compatibilities	compatibilities
committment	commitment	commitment	commitment	commitment	commitment
debateable	debatable	debatable	debatable	debatable	debatable
determinining	determining	determinining	determinining	determinining	determining
childbird	childbirth	child bird	child bird	_childbirth	childbirth
definately	definitely	definitely	definitely	definitely	definitely
decribe	describe	describe	describe	describe	describe
elphant	elephant	elephant	elephant	elephant	elephant
emmediately	immediately	immediately	immediately	immediately	immediately
emphysyama	emphysema	emphysema	emphysema	emphysema	emphysema
erally	(orally, really)	orally	really	really	orally
eyasr	(years, eyas)	evesore	years	years	eyas
facist	fascist	fascist	fascist	fascist	fascist
fluorescent	fluorescent	fluorescent	fluorescent	fluorescent	fluorescent
geneology	genealogy	genealogy	genealogy	genealogy	genealogy
gernade	grenade	grenade	grenade	grenade	grenade
girates	gyrates	grates	gyrates	pirates	gyrates

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
gouvener	governor	governor	<u>souvenir</u>	<u>gouverneur</u>	<u>convener</u>
gurantees	guarantee	guarantee	guarantee	guarantee	guarantee
guerrila	(guerilla, guerrilla)	guerrilla	guerrilla	guerrilla	(guerilla, guerrilla)
guerrillas	(guerillas, guerrillas)	guerrillas	guerrillas	guerrillas	(guerillas, guerrillas)
Giuseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe	Giuseppe
habaeus	(habeas, sabaeus)	habeas	<u>habitués</u>	habeas	<u>sabaeus</u>
hierarcical	hierarchical	hierarchical	hierarchical	hierarchical	hierarchical
heros	heroes	heroes	heroes	heroes	<u>herbs</u>
hypocracy	hypocrisy	hypocrisy	hypocrisy	hypocrisy	hypocrisy
independance	Independence	Independence	-	Independence	Independence
intergration	integration	integration	integration	integration	integration
intrest	interest	interest	interest	interest	interest
Johanine	Johannine	Johannes	Johannes	Johannes	Johannine
judisuary	judiciary	judiciary	judiciary	-	judiciary
kindergarden	kindergarten	kindergarten	kindergarten	kindergarten	kindergarten
knowlegeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable	knowledgeable
labatory	(lavatory, laboratory)	(lavatory, laboratory)	(lavatory, laboratory)	laboratory	(lavatory, laboratory)
lonelyness	loneliness	loneliness	loneliness	loneliness	loneliness
legitamate	legitimate	legitimate	legitimate	legitimate	legitimate
libguistics	linguistics	linguistics	linguistics	linguistics	linguistics
lisence	(license, licence)	licence	<u>silence</u>	licence	licence
mathmatician	mathematician	mathematician	mathematician	mathematician	mathematician
ministry	ministry	ministry	ministry	ministry	ministry
mysogynist	misogynist	misogynist	misogynist	misogynist	misogynist
naturally	naturally	naturally	naturally	naturally	naturally
ocuntries	countries	countries	countries	countries	countries
paraphernalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia	paraphernalia
Palistian	Palestinian	<u>Alsatain</u>	<u>politian</u>	Palestinian	Palestinian
pamflet	pamphlet	pamphlet	pamphlet	pamphlet	pamphlet
psychic	psychic	psychic	psychic	psychic	psychic
Peloponnes	Peloponnesus	Peloponnes	Peloponnes	Peloponnes	Peloponnesus
personell	personnel	personnel	personnel	personnel	personnel
posseses	possesses	possesses	possesses	possesses	possess
prairy	prairie	<u>priory</u>	prairie	prairie	<u>airy</u>
qutie	(quite, quiet)	quite	quite	<u>cutie</u>	<u>queue</u>
radify	(ratify, ramify)	ratify	ratify	ratify	ramify
recommended	recommended	recommended	recommended	recommended	recommended
reciever	receiver	receiver	receiver	receiver	<u>reliever</u>
reconaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance	reconnaissance
restauration	restoration	restoration	restoration	restoration	<u>instauration</u>
rigeur	(rigueur, rigour, rigor)	<u>rigger</u>	rigueur	-	(rigueur, rigour)
Saterdag	Saturday	Saturday	Saturday	Saturday	Saturday
scandinavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia	Scandinavia
scaleable	scalable	scalable	-	scalable	scalable
secceeded	(seceded, succeeded)	succeeded	succeeded	seceded	succeeded
sepulchure	(sepulchre, sepulcher)	sepulcher	<u>sepulchered</u>	sepulcher	sepulchre
themselves	themselves	themselves	themselves	themselves	themselves
throught	(thought, through, throughout)	(thought, through)	(thought ,through)	<u>throat</u>	(thought ,through, throughout)
troups	(troupes, troops)	(troupes, troops)	troupes	troops	troops
simultaneous	smultaneous	simultaneous	simultaneous	simultaneous	simultaneous
sincerley	sincerely	sincerely	sincerely	sincerely	sincerely
sophicated	sophisticated	<u>suffocated</u>	<u>supplicated</u>	-	sophisticate
surrended	(surrounded, surrendered)	surrounded	surrender	surrender	surrounded
unforetunately	unfortunately	unfortunately	unfortunately	-	unfortunately
unnecesarily	unnecessarily	unnecessarily	unnecessarily	-	unnecessarily
usally	usually	usually	usually	usually	usually
useing	using	using	using	using	<u>seeing</u>
vaccum	vacuum	vacuum	vacuum	vacuum	vacuum

Misspelling	Correct Spelling	Aspell	Microsoft word	Google	MultiSpell
vegetables	vegetables	vegetables	vegetables	vegetables	vegetables
vetween	between	between	between	between	between
volcanoe	volcano	volcano	volcano	volcano	volcano
weaponary	weaponry	weaponry	weaponry	weaponry	weaponry
worstened	worsened	worsened	worsened	-	worsened
wupport	support	support	support	support	support
yeasr	years	years	years	years	yeast
Yementite	(Yemenite, Yemeni)	Yemenite	Yemenite	Yemenite	Yemenite
yuonger	younger	Younger	younger	younger	sponger

TABLE VI
RESULTS OF WORD CORRECTIONS IN PORTUGUESE.

Correct Form	Spelling Error	TST	Aspell	MultiSpell
acerca	àcerca	acerca	acerca	acerca
açoriano	açoreano	açoriano	coreano	açoriano
alcoolemia	alcoolemia	alcooolÚmia	-	alcoolemia
ameixial	ameixeal	ameixial	ameixial	ameixial
antártico	antártico	catártico	antártico	antártico
antepor	antepôr	-	antepor	antepor
ártico	artico	artigo	aórtico	aórtico
artifice	artifece	artífice	artífice	artífice
bainha	baínha	bainha	bainha	bainha
bebé	bébé	bebé	bebe	bebé
bege	beje	bege	beije	bejense
bênção	benção	bencao	-	bênção
beneficência	beneficiência	beneficência	beneficência	beneficência
biopsia	biópsia	biópsiu	-	biopsia
burburinho	borborinho	burburinho	burburinho	burburinho
caiem	caem	-	-	cabem
calvício	calvíce	calvício	calvício	calvício
camoniano	camoneano	camoniano	camoniano	camoniano
campeão	campião	campeão	campeão	campeão
chiita	xiita	chiita	xiitas	xiitas
comboio	combóio	comboio	comboio	comboio
compor	compôr	-	compor	compor
comummente	comumente	comovente	comumente	comumente
constituia	constituía	-	-	constituia
constituiu	constituíu	constituiu	constituiu	constituiu
cor	côr	-	cor	cor
crânio	crâneo	crânio	cárneo	crânio
definição	defenição	definição	definição	definição
definido	defenido	definido	-	defendido
definir	defenir	definir	definir	definir
desequilíbrio	desequilíbrio	desequilíbrio	desequilíbrio	desequilíbrio
despretensioso	despretensioso	despretensioso	despretensioso	despretensioso
dignatários	dignitários	dignatários	digitarias	dignatários
dispender	despender	dispender	-	despendes
dispêndio	dispendio	dispundio	dispundio	dispendioso
écran	ecran	-	écran	écran
emirados	emiratos	estratos	méritos	emirados
esotérico	isotérico	-	-	esotérico
esquisito	esquesito	esquisito	esquisito	esquisito
estratego	estratega	estratego	-	estratego
feminino	femenino	feminino	feminino	feminino
feminismo	femininismo	-	feminismo	feminismo
fôr	for	-	-	forcar
gineceu	geneceu	gineceu	gineceu	gineceu
gorjeta	gorjeta	gorjeta	gorjeta	gorjeta
granjeiar	grangear	granjeiar	granjeiar	granjeiar
guisar	guizar	guisar	gizar	guinar
halariedade	hilaridade	hilariedade	-	polaridade
hectare	hectar	hectare	-	hectare
hiroshima	hiroxima	aproxima	próxima	hiroshima
ilacção	elação	ilação	ilação	delação
indispensável	indispensável	indispensável	indispensável	indispensável
inflação	inflaçãõ	-	-	inalacção
interveio	interview	intervir	Inter viu	intervim
intervindo	intervido	intervindo	-	intervindo
invocar	evocar	invocar	-	evocai

Correct Form	Spelling Error	TST	Aspell	MultiSpell
ípsilon	ipslon	ípsilon	ípsilon	ípsilon
irisar	irizar	irisar	razar	irisar
irupção	irrupção	-	-	irupção
jeropiga	geropiga	jeropiga	Georgia	jeropiga
juiz	juíz	-	juiz	Juiz
lampião	lampeão	lampião	sarjeta	campeão
lêem	lêm	lés	lema	lêem
linguista	linguísta	-	linguista	linguista
lisonjear	lisongear	lisonjear	lisonjear	lisonjear
logótipo	logotipo	logo tipo	logo tipo	logótipo
maciço	massiço	mássico	mássico	massudo
majestade	magestade	majestade	majestade	majestade
manjerico	mangerico	manjerico	manjerico	manjerico
manjerona	mangerona	tangerina	tangerina	manjerona
meteorologia	metereologia	meteorologia	meteorologia	meteorologia
miscigenação	miscegenação	miscigenação	miscigenação	miscigenação
nonagésimo	nonagessimo	nonagésimo	nonagésimo	nonagésimo
oceânia	oceania	oceânia	Oceania	oceânia
oficina	ofecina	oficina	oficina	oficina
opróbrio	opróbio	aeróbio	próbio	opróbrio
organograma	organigrama	organograma	-	organograma
paralisar	paralizar	paralisar	paralisar	paralisar
perseverança	preseverança	perseverança	perseverança	perseverance
persuasão	persuação	persuasão	persuasão	persuasão
pirinés	pirenéus	-	pirinés	pirinés
pretensioso	pretencioso	pretensioso	pretensioso	pretensioso
privilégio	previlégio	privilégios	privilégios	privilegios
quadricromia	quadricomia	quadricromia	quadriculai	quadricromia
quadruplicado	quadriplicado	quadruplicado	quadruplicado	quadruplicado
quasimodo	quasimodo	-	quisido	quasimodo
quilo	kilo	quilo	Nilo	dilo
quilograma	kilograma	holograma	holograma	holograma
quilómetro	kilómetro	milímetro	milímetro	quilómetro
quis	quiz	quis	qui	juiz
rainha	raínha	rainha	rainha	rainha
raiz	raíz	-	raiz	raiz
raul	raúl	raul	Raul	raul
rectaguarda	retaguarda	rectaguarda	-	rectaguarda
rédea	rédia	rêdea	radia	radia
regurgitar	regurjitar	regurgitar	regurgitar	regurgitar
rejeitar	regeitar	rejeitar	regatar	receitar
requero	requero	requere	requero	requer
réstia	rêstea	réstia	resta	réstia
rubrica	rúbrica	rúbreca	rubrica	rubrica
saem	saiem	saíam	saem	caiem
saloíce	saloice	baloice	saloíce	saloíce
sarjeta	sargeta	sarjeta	sarjeta	Sarjeta
semear	semiar	semear	semear	Semear
suiça	suiça	suiça	suiça	Suíça
supor	supôr	-	supor	Supôs
trânsfuga	transfuga	transfira	transfira	trânsfuga
transpôr	transpor	-	-	transportar
urano	úrano	-	-	grano
ventoinha	ventoínha	ventoinha	ventoinha	ventoinha
verosimil	verosímel	-	-	verosímil
vigilante	vegilante	vigilante	vigilante	vigilante
vôo	voo	-	-	ovo
vultuoso	vultoso	vultuoso	-	vultosos
xadrez	xadrês	xadrez	ladres	xadrez
xamã	chamã	chama	chama	chamã
xelindró	xilindró	cilindro	cilindro	xelindró
zângão	zangão	zangai	-	mangão
zepelin	zeppelin	zepelim	zepelim	zepelin
zoo	zoô	zoo	coo	zoo

REFERENCES

- [1] K. Kukich, "Techniques for automatically correcting words in text," *ACM Computing Surveys*, 24(4), 377-439, 1992.
- [2] F. J. Damerau, "A technique for computer detection and correction of spelling errors," *Communications of ACM*, 7(3):171-176.7, 1964.
- [3] W. Peters, "Lexical Resources," NLP group, Dept. of Comp. Sc., Uni. of Sheffield, 2001.
- [4] R. A. Wagner and M. J. Fisher, "The string to string correction problem," *Journal of Assoc. Comp. Mach.*, 21(1):168-173, 1974.
- [5] A. Stanier, "How accurate is Soundex matching?" *Comp. in Genealogy*, vol. 3:7, 1990.
- [6] C. Fellbaum, "WordNet, an electronical lexical database," Cambridge, MIT Press, 1998.

- [7] E. Pianta, L. Bentivogli, and C. Girardi, "MultiWordNet: developing an aligned multilingual database," in *Proc. of 1st Int. Conf. on Global WordNet*, 2002.
- [8] C. E. Shannon, "Prediction and entropy of printed English," *Bell Sys. Tec. J.* (30):50–64, 1951.
- [9] U. Pfeifer, "Retrieval Effectiveness of Proper Name Search Methods," *Information Processing and Management*, 32(6):667–679, 1996.
- [10] E. J. Yannakoudakis and D. Fawthrop, "An intelligent spelling error corrector," *Information Processing and Management*, 19:1, 101-108, 1983.
- [11] T. N. Turba, "Length-segmented lists," *Comm. of the ACM*, 25:8, pp 522-526, 1982.
- [12] Wikipedia, list of Common Misspelling Word List, [http://en.wikipedia.org/wiki/Wikipedia: List_of_common_misspellings](http://en.wikipedia.org/wiki/Wikipedia:List_of_common_misspellings), 05.10.2006.
- [13] B. Martins, M. J. Silva, "Spelling Correction for Search Engine Queries," in *EsTAL - España for Natural Language Processing*, Alicante, Spain, 2004.
- [14] K. Church and W. A. Gale, "Probability scoring for spelling correction," *Statistics and Computing*, Vol. 1, No. 1, pp. 93–103, 1991.
- [15] V. J. Hodge and J. Austin, "A comparison of standard spell checking algorithms and novel binary neural approach," *IEEE Trans. Know. Dat. Eng.*, Vol. 15:5, pp. 1073-1081, 2003.
- [16] J. J. Pollock and A. Zamora, "Collection and characterization of spelling errors in scientific and scholarly text," *Journal Amer. Soc. Inf. Sci.*, Vol. 34, No. 1, pp. 51–58, 1983.
- [17] ———, "Automatic spelling correction in scientific and scholarly text," *Comm. ACM*, Vol. 27, No. 4, pp. 358–368, 1984.
- [18] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in *Proc. 38th Annual Meet. of the Assoc. for Comp. Ling.*, Hong Kong, 2000, pp. 286–293.
- [19] K. Toutanova and R. C. Moore, "Pronunciation modeling for improved spelling correction," in *Proc. 40th Annual Meeting of the Assoc. for Comp. Ling.*, Hong Kong, 2002, pp. 144–151.
- [20] Jin-ming Zhan, Xiaolong Mou, Shuqing Li, Ditang Fang, "A Language Model in a Large-Vocabulary Speech Recognition System," in *Proc. of Int. Conf. ICSLP98*, Sydney, Australia, 1998.
- [21] S. Deorowicz and M. G. Ciura, "Correcting Spelling Errors by Modelling Their Causes," *Int. Journal of Applied Mathematics and Computer Science*, 15(2):275–285, 2005.
- [22] B. Khaltar, A. Fujii, and T. Ishikawa, "Extracting loanwords from Mongolian corpora and producing a Japanese-Mongolian bilingual dictionary," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the ACL*, Sydney, Australia: ACL. Pages: 657 – 664, 2006.
- [23] E. W. De Luca and A. Nürnberger, "Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation," *International Journal of Intelligent Systems*, 21:693–709, 2006.
- [24] E. W. De Luca and A. Nürnberger, "Rebuilding Lexical Resources for Information Retrieval using Sense Folder Detection and Merging Methods," in: *Proc. of the 5th Int. Conf. on Language Resources and Evaluation (LREC 2006)*, 2006.
- [25] E. W. De Luca, "Semantic Support in Multilingual Text Retrieval," Shaker Verlag, Aachen, Germany, 2008.