JOURNAL OF RESEARCH IN SCIENCE, COMPUTING, AND ENGINEERING
Vol. 4 No. 3 December 2007
De La Salle University–Manila

SpellCheF : Spelling Checker and Corrector for Filipino

Cheng, Charibeth, Alberto, Cedric Paul, Chan, Ian Anthony and Querol, Vazir Joshua
College of Computer Studies
De La Salle University - Manila

SpellCheF is a spell checker for Filipino that uses a hybrid approach in detecting and correcting misspelled words in a document. Its approach is composed of dictionary-lookup, n-gram analysis, Soundex and character distance measurements. It is a plug-in to OpenOffice Writer. Two spelling rules and guidelines, namely, the Komisyon sa Wikang Filipino 2001 Revision of the Alphabet and Guidelines in Spelling the Filipino Language (or KWF), and the Gabay sa Editing sa Wikang Filipino (or GABAY) rulebooks, were incorporated into the system. SpellCheF is composed of three modules, namely the lexicon builder, the detector and the corrector. These three modules used both manual-formulated and learned rules to carry out their tasks. Test results showed that the lexicon builder was able to correctly categorize words based on the spelling rules used. It also generated three databases, namely, (1) the KWF-compliant words database, (2) the Gabay-compliant words database, and (3) the database of words common to both KWF and GABAY. The detector module had an overall error rate of 7% in identifying misspellings. Furthermore, it was observed that n-gram analysis performed better than the simple dictionary look-up. The corrector module had a 94% accuracy rate in generating word suggestions. Results also showed that using the soundex code, more suggestions were generated compared to the use of n-gram analysis. However, the first character of the soundex-based suggestions was always the same are the first character of the misspelled word.

Keywords: Filipino language, spelling checker, spelling corrector

INTRODUCTION

A spell checker is a feature found in most document-related applications. It is used to ensure that words in a given document are correctly spelled. Simple spell checkers use a large lexicon containing correctly spelled words. Each word in the document is checked against this lexicon. If a word is not found in the lexicon, it is considered misspelled. A spelling corrector may be included in spell checker, to generate suggestions that may replce the misspelled word.

(Gratilla, 2006) implemented a tri-gram-based Tagalog spell-checker. It eliminated the use of a lexicon to perform spell checking. It learns the acceptable tri-grams in Tagalog given several Tagalog texts. Instead of checking the spelling of word itself, it determines if the tri-grams contained in a word are acceptable.

Filipino is the term used in 1973 and 1987 Philippine Constitutions to designate as the national language. It is sometimes referred to as Tagalog. But the two languages are different. Orthographically, Tagalog has 20 letters while

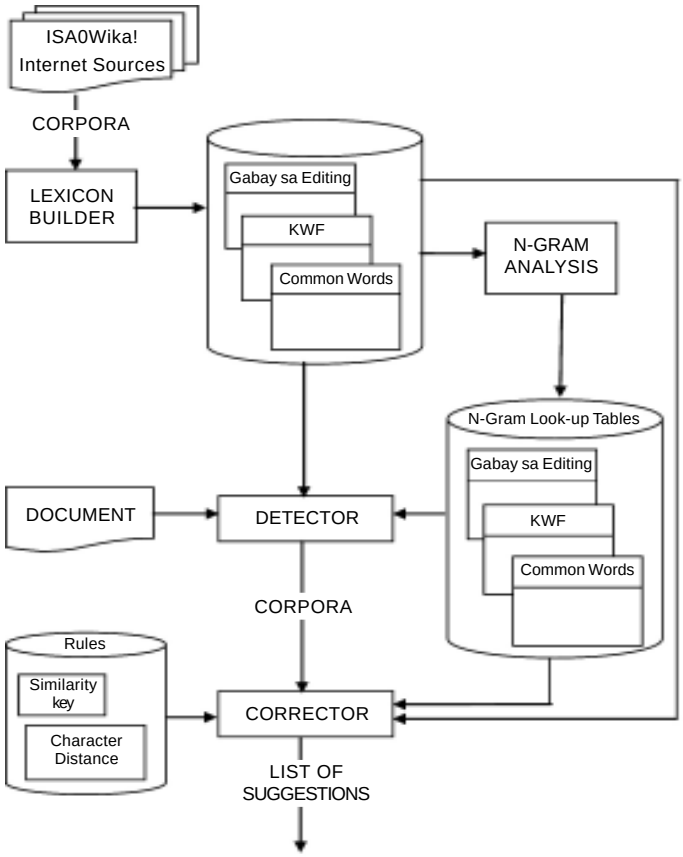

Figure 1. Architectural Design of SpellCheF

Filipino has 28 (which include **C**, **F**, **J**, **Ñ**, **Q**, **V**, **X**, and **Z**).

The Commission on Filipino Language and the Sentro ng Wikang Filipino of UP-Diliman developed different guidelines in spelling Filipino words, called the *Alfabeto at Patnubay sa Ispelling ng Wikang Filipino* (Komisyon sa Wikang Filipino, 2001), and the *Gabay sa Editing sa Wikang Filipino* (Sentra sa Wikang Filipino, 2004), respectively. Throughout this paper, we will refer to the former as the KWF guideline, while the latter will be referred to as the GABAY guideline. The two guidelines may have similarities, but there are also differences, specifically on the spelling of borrowed words. For instance, the KWF guideline has a rule on when to change the letter **Q** to **K** or to **KW**, while the GABAY guideline did not give any guidelines on how to change **Q**. Another difference is in the spelling of words with letter **X**. **Xerox** (meaning to photocopy) would be spelled as **seroks** if we follow the GABAY guideline, or **xerox** if the KWF guideline is followed.

SPELLCHeF

SpellCheF stands for *Spell Checker in Filipino*. It is able to determine if a word is misspelled based on the KWF and/or GABAY spelling guidelines. When a misspelled word is found, it automatically generates a list of spelling suggestions. The architectural design of SpellCheF is shown in Figure 1. The system is composed of four main parts, namely, the lexicon builder, n-gram analysis, detection, and correction modules. The lexicon builder is responsible for grouping the words from the corpora and creating the lexicon which will then undergo n-gram analysis. The

detector module is responsible for marking possibly misspelled words in a given document, while the corrector module generates a list of suggestions on how to correct the detected misspelled word.

LEXICON BUILDER

The lexicon builder extracts words from source materials and categorizes them according to the spelling guidelines. It generates three databases, the first containing the KWF-compliant words, the second containing GABAY-compliant words, and the third containing the common words of KWF and GABAY. These three databases are used as dictionary by the detector and corrector modules. Table 1 summarizes the differences in the two spelling guidelines.

For example, in KWF digraph **sh**, word with **sh** or **sy** are both acceptable. The rule pattern would be

```
((\w*([s][hy][aeiou]))\w*)
```

where

\w* represents any character of any length;
s may be followed by **h** or **y**, and then followed by the vowels **a**, **e**, **i**, **o**, **u**

Hence, the given pattern will accept words like **shopping, syaping, workshop, worksyap, bisyap, internsyip.**

Based on the differences in Table 1, rules were formulated to filter the words for a specific guideline. These rules are stored in a file and read by the lexicon builder. They are represented in regular expressions, using the **java.util.regex** API[1].

Table 1. Comparison of the Spelling Guidelines

Category	KWF	GABAY
Digraph SH	Work**sy**ap Work**sh**op	Work**sh**op
	Syuting **Sh**ooting	**Sh**ooting
	Sensor**sy**ip	Sensor**sh**ip
Repetition with Clustered Consonants	**Pa**plantsahin **Pa**plantsahin	**Pla**plantsahin
	Piprituhin **Pri**prituhin	**Pri**prituhin
	Kukuwentuhan **Ku**kuwentuhan	**Kwe**kwentuhan
"Diptonggo"		
ia	Krist**ya**no Krist**ya**no	Krist**iya**no
ie	T**ye**mpo T**ye**mpo	T**iye**mpo
io	Rebolus**yo**n Rebolus**yo**n	Rebolus**iyo**n
ua	G**wa**po G**uwa**po	g**uwa**po
ue	S**we**rte S**uwe**rte	S**uwe**rte
ui	B**wi**tre B**uwi**tre	B**uwi**tre
Letter X	Xerox	Seroks

N-GRAM ANALYSIS

An n-gram is an *n* letter subsequence of a string, where *n* usually is 1, 2, or 3. In general, n-gram analysis techniques check each n-gram in an input string against a precompiled table of n-gram statistics to determine whether the n-gram can occur in a word. If it does, its frequency of occurrence in the words of the language is computed. Strings containing n-grams that do not occur in words or occur very infrequently are considered to be misspellings (Kukich, 1992).

There are a variety of n-gram models, including the binary n-grams and the n-gram frequency statistics. In the n-gram table of the binary n-gram, each element in the table has a value set to 0 or 1, depending on whether the n-gram appeared in at least one word in the lexicon (value = 1) or not

[1]http://java.sun.com/docs/books/tutorial/essential/regex/index.html

(value = 0). On the other hand, n-gram frequency statistics represent the probability that a given letter (or letter sequence) will be followed by another given letter. These statistics are generated from a sufficiently large corpus of text covering certain domains.

(Kukich 1992) further classified the binary n-grams into *positional* and *non-positional*. In positional n-gram, given a binary tri-gram array, the *i, j, k*th elements would have a value 1 if and only if there exists one word in the lexicon with letter *l*, *m*, or *n* in positions *i, j, k*. In a non-positional n-gram, the position of the n-gram is not relevant within a word.

SpellCheF implements non-positional tri-gram analysis. The tri-gram analysis module generates three n-gram look-up tables – one for the KWF lexicon, another for the GABAY lexicon, and one for the Common Words lexicon.

THE DETECTOR MODULE

The detector module is responsible for determining if a word is considered misspelled or not. It is invoked through OpenOffice Writer. The user has to indicate whether the detector should follow the KWF or GABAY spelling guidelines in detecting misspelled word.

To illustrate how this module works, consider the unknown word **bayano**. This module will first check if **bayano** is in the lexicon. Since it is not in the lexicon, the system cannot automatically say that it is misspelled, for it may be a valid word but is not yet encountered by the lexicon builder. To determine if this word is possibly acceptable, the n-gram look-up table will be used. This method will check first if the length of **bayano** is greater than three to be able to generate tri-grams. The first extracted tri-gram is **bay**. If **bay** is in the n-gram look-up table, its frequency statistics is compared against the system's threshold of 0.001. If it is less than the threshold, then it means that **bayano** is considered as misspelled. On the other hand, if the frequency statistics is greater than the threshold, the process is repeated for the next tri-gram which is **aya**. Only when all the word's tri-grams' frequency statistics exceed the threshold will the system consider the word as correctly spelled.

THE CORRECTOR MODULE

The corrector module is responsible for generating suggestions that would result to the possible correct spelling of an erroneous word. This module makes use of two different algorithms in generating suggestions: tri-gram analysis and a similarity key algorithm that uses Soundex code.

Suggestions using tri-gram analysis

The first step would be generating suggestions using the tri-gram look-up table. The suggested words' length is ± 1 character difference as the misspelled word. This constraint is based on studies which conclude that most misspellings involve at most one character change from the intended word (Kukich, 1992).

The n-gram look-up table contains entries of the possible "grams" in the Filipino language. Each gram entry also includes its frequency occurrence, possible previous gram and possible next gram. Two methods were used to generate the possible suggestions. The first method assumes that the first tri-gram of the misspelled word is correct. The second method assumes that the last tri-gram of the misspelled word is correct.

Let us consider the misspelled word **panget**. It is assumed that the **pan** gram is correct meaning the actual correct word starts with **pan**. The system will generate suggested words that start with **pan** like **pangat** (fish boiled in sour ingredients), **pangit** (ugly), **pantay** (same or equal), using the tri-gram look-up table.

The second method assumes that the starting characters are wrong, thus the last tri-gram is used to generate suggestions. For example, given the misspelled word **ebogasya**, the generation of the suggestions will begin at the last gram **sya**, resulting to words like **abogasya** (the law profession), **ipinasya** (decided), **nagkasya** (fit), **pantasya** (fantasy).

Suggestions using Soundex

Soundex stands for "Indexing on Sound" (Kukich, 1992). This is a phonetic algorithm for indexing names by their sound when pronounced in English. This research applied Soundex to Filipino. Some Filipino words sound alike but are spelled differently. A Soundex code is created for similar-sounding letters. The idea of the code is to preserve, in a rough way, the salient features of the pronunciation. Vowel letters are discarded and consonant letters are grouped if they are likely to be substituted for each other. The Soundex algorithm is listed below:

1. Keep the first letter (in upper case)
2. Replace remaining characters using the Soundex values as shown in Table 2.

Table 2. Soundex Values

b,f,p,v	1
c,g,j,k,q,s,x,z	2
d,t	3
L	4
m,n	5
R	6
a,e,i,o,u,y,h,w	–

3. Delete adjacent repeats of a number
4. Delete the hyphens.
5. Keep the first three numbers or pad out with zeroes

Consider the misspelled word **imlpuwensya**. Using Soundex, the list of suggestions and their corresponding codes is shown in Table 3.

Table 3. Example of Generated Soundex Codes

Suggestion	Soundex Code
impluwensiya (influence)	I514
i-level (to level)	I414
i-rebuild (to rebuild)	I614
implikasyong (implication)	I514

RANKING THE SUGGESTIONS

After the detector generates a list of correct spelling suggestions, the next step is to rank these suggestions. The character distance method is used to arrange and trim the suggestions. This method uses a Pythagorean-type metric to measure the distance between a misspelled word and a possible correction, based on the QWERTY keyboard layout (Min, et al., 2005). The QWERTY keyboard is represented as a two-dimensional arrays: one for the lower case keys and another for the upper case keys. The suggested word with the shortest distance to the misspelled word is considered as the best suggestion.

i/j	0	1	2	3	4	5	6	7	8	9	10	11
0	1	2	3	4	5	6	7	8	9	0	-	=
1	q	w	e	r	t	Y	u	i	o	p	[	]
2	a	s	d	f	g	H	j	k	l	;	‘	
3	z	x	c	v	b	N	m	,	.	/		

Figure 2. Character Distance Chart

For example, based on the character distance chart in Figure 2, the distance between **w** located at (1,1) and **n** located at (3,5) is

$$||(3,5)-(1,1)|| = \sqrt{2^2+4^2} \approx 4.47$$

Using this character distance, the scores for correction of **imlpuwensya** are shown in Table 4.

Table 4. Corrections for imlpuwensya with their character distance score

Correction	Score
impluwensiya (influence)	2.83
impluwensiya (influence)	11.93
impluwensiya (influence)	16.05

Then the length of the misspelled word is not equal to the length of the suggestion, the score will drastically increase because each additional character will be compared with 0.

RESULTS AND RECOMMENDATIONS

SpellChef was tested using three Filipino documents:

1. An essay written by a student. The document is part of an introduction regarding the government under a known politician;
2. An entertainment article regarding an upcoming TV-series; and
3. A sports article about the Pacquiao – Solis match.

In testing these documents, the dictionary had a total of 45,548 words, where 2,557 words are categorized under GABAY, 4,082 words are under KWF, and 42,797 words are under the common words category. The acceptance threshold used was 0.001. This threshold value was reached after tests were performed on different threshold values. It was observed that a threshold value < 0.001 failed to detect misspelled words, while a higher threshold value caused the system to declare correctly spelled words as misspelled.

The detector module achieved a 7% error rate. This means that 7% of the words the system marked as misspelled are actually correctly spelled. These errors are attributed to the following:

- A small lexicon. The lexicon only contains 45,548 words. Several words in the Filipino language, including the proper nouns and named entities, are not found in the lexicon. Because to this, the probabilities of the grams in the tri-gram look-up table are not that accurate, resulting to the marking of a valid word as misspelled; or the acceptance of a misspelled word.
- Low frequency of the tri-gram of the word in the lexicon, thus the threshold was not met. By increasing the size of the lexicon, the quality of the n-gram frequency statistics would also increase.
- Commonly occurring one- or two-letter Filipino words that are not in the lexicon, such as **`pa`**, **`o`**, **`di`**, **`k`**. Tri-gram analysis is applied only to words with at least three characters.

On the other hand, the corrector module achieved a 94% accuracy rate. All correctly-detected misspelled words were given correct suggestions by the system. The errors came from correctly-spelled words which the detector marked as misspellings.

It is important to note that for a misspelled word, the correct word does not always appear as the first word in the suggestion list. The length of the misspelled word and the suggestion greatly affect the rank of the suggestion. Since the system computes the distances of the two words, there is a higher probability that the suggested word with the same length as the misspelled word will appear at the top of the list.

Results show that the Soundex approach generates suggestions faster than the tri-gram approach. For efficiency and speed concerns, the system only generates suggestions that have less than ten characters using the tri-gram approach. Because of this constraint, the tri-gram approach generates fewer suggestions. However, the quality of the tri-gram suggestions is usually better than the Soundex suggestions. This is because Soundex always assumes that the first letter is correct whereas tri-gram can suggest words whose first character is not the same as the misspelled word. It is, thus, recommended that improvements in the tri-gram based corrector be implemented in the future. Furthermore, instead of using the n-gram model, the syllabication nature of Filipino may be exploited to improve performance of the corrector module. The syllables in Filipino take the form of CV, CVC, CVCC and CCV. These improvements may consider the use of a hash table or better memory management techniques. The best and worst case of the memory management techniques currently used is O(n). When a hash table is implemented, the best case may be improved to O(1).

REFERENCES

Gratilla, J., Junio, M., Sampani, M., Tamayo, J., Bonus, D., Sagum, R. 2006. A Tri-Gram Based Spell Checker for Tagalog. *4th National Natural Language Processing Symposium*. Manila, Philippines. pp. 90-93. February 2006

Komisyon sa Wikang Filipino 2001. Alfabeto at Patnubay sa Ispelling ng Wikang Filipino. Manila.

Kukich, K. Techniques for Automatically Correcting Words in Text. *ACM Computing Surveys*. Volume 24, Number 4. pp. 377-439. December 1992.

Min, K., Wilson, W., Moon, Y. 2005. Syntactic and Semantic Disambiguation of Numeral Strings Using an n-gram Method. *AI 2005: Advances in Artificial Intelligence*. Springer, Berlin. pp. 82-91. 2005.

Sentro ng Wikang Filipino – UP Diliman 2004. Gabay sa Editing sa Wikang Filipino (tuon sa pagbaybay). Bulwagang Tomas Pinpin, IMC Compound A. Ma. Regidor St., Area XI, UP Campus, Quezon City.