QSIURP 2013 Report

Spellchecker to Improve Human-Robot Interaction

Naassih Gopee
Faculty Advisor: Majd F. Sakr, Ph.D.

Introduction

This research focuses on developing a contextual spellchecker to improve the correctness of input queries to the multi-lingual, cross-cultural robot. Queries that have fewer misspellings will improve the robot's ability to answer them and in turn improve the human-robot interaction. We focus on developing a language model based contextual spell-checker to correct misspellings and increase the query-hit rate of the robot.

Our test bed is a bi-lingual, cross-cultural robot receptionist, Hala, deployed at the Carnegie Mellon Qatar reception. Hala can accept typed input queries in Arabic and English and speak responses in Arabic and English as she interacts with a variety of users. All input queries to Hala are logged. These logs allow the study of multi-lingual aspects, the influence of socio-cultural norms and the nature of interactions during human-robot interaction within a multicultural, yet primarily ethnic Arab, setting.

A recent statistical analysis has shown that Hala correctly answers 73.7% of the input queries (hit rate) and hence 26.3% of queries are missed. The queries that are missed are due to either Hala not having the required answer in the knowledge base or due to misspellings in the queries. Based on a sample of missed queries, we have measured that 50% are due to misspellings. Examples of these misspellings are shown below.

Example:

Time Stamp	Query
9/26/2009 11:50:18 AM	what is you work?
9/26/2009 11:50:27 AM	what is your wrk?
9/26/2009 11:50:40 AM	what is your work?

The interlocutor makes multiple attempts in order to correct the misspellings, which might lead to frustration. We conjecture that if we were to implement an effective spellchecker we could increase the hit rate and moreover reduce the need for users to retype the same question several times due to misspellings.

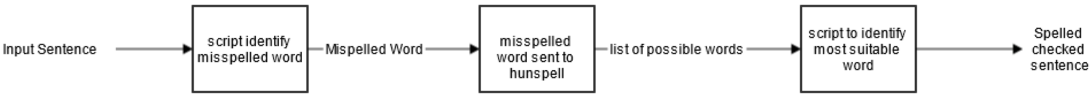
Preparations

Several papers on class-based n-gram model of natural language [1][2] were reviewed to clearly understand how an n-gram model works. This was performed to conclude whether an n-gram model would be a good approach to correct the misspelled words in Hala queries. Upon reviewing Peter Norvig's character based n-gram spellchecker [3], it spurred the idea of using the minimum editing distance before applying the n-gram was better to calculate the list of possible correct words out of a misspelled word.

After evaluating the possibility of a simple character based spellchecker, it was concluded that this approach was not a suitable option. A character based n-gram spellchecker would typically suggest a limited list of words. For instance, the word “cough” when misspelled, as “couf” cannot be correctly spelled with a character base n-gram – it requires more advanced techniques such as morphological analysis. Hence, a more powerful open source spellchecker is needed. The Hunspell spellcheker [5] was identified as a good match. Its engine is based on GNU Aspell. Furthermore, Hunspell’s algorithm is based on n-gram similarity, rule and dictionary based pronunciation data and morphological analysis.

Methodology

Our approach makes use of the Hunspell spellchecker. Misspelled words are passed through the Hunspell spellchecker and its output is a list of possible words. We apply our algorithm, based on the n-gram model to find which word is best suited in a particular context, to the list of possible words. The *n*-gram model calculates the conditional probability *P(w|s)* of a word *w* given the previous sequence of words *s* in the sentence, that is, predicting the next word based on the preceding *n-1* words.


Input Sentence
script identify misspelled word
Mispelled Word
misspelled word sent to hunspell
list of possible words
script to identify most suitable word
Spelled checked sentence

For example, the conditional probability of *P(robot|blue)* consists of calculating the probability of the whole sequence “blue robot.” Put differently, we have to calculate the probability for the word “blue” so that the word following it is “robot”.

It is costly to calculate the probability of a word given all previous sequence of words, therefore we use either a 2-gram model or 3-gram model. It is denoted by $P(w_n|w_{n-1})$ the probability of a word w_n given the previous word w_{n-1}. For instance in a sequence of bigrams, the probability is calculated as follows:

$$P(W_1^n) \approx \prod_{k=1}^{n} P(W_k|W_{k-1})$$

(Basil et al., p.39)[4]

We need to note that the *n*-gram model will be highly dependent on the language corpus it is based on. Given, our context, the n-gram’s model language corpus used in this work is based on Hala’s current knowledge base of answers to potential questions. Since Hala mostly answers questions about CMUQ, we also added the following documents to the corpus:

- Text from CMUQ web pages
- Text from news articles about CMUQ
- Text from CMUQ documentation (annual reports, etc…)

In order to evaluate the accuracy of our spellchecker, we sampled queries for each cases described below from our current 24659 missed queries. Each sample was then run through the spellchecker and then analyzed.

Results

To assess the effectiveness of our system, we evaluate it using 5 different cases of misspelled word location.

Case 1: Single word sentence

Example result 1.1: The word is correctly spellchecked since it one of the word that occurred the most in our corpus.

Incorrect Input	After spellchecked	Correct
Helloo	Hello	Hello

Example result 1.2: The word is incorrectly spellchecked since no word in the list that Hunspell output contains a word that occurred in the corpus.

Incorrect Input	After spellchecked	Correct
byeeeeeeeeeeeeeeee	bitter-sweetness	bye

Case 2: Misspelled word followed by correct word

Example result 2.1: The word is correctly spellchecked since from the list of possible words; “thank” is most likely to follow “you”.

Incorrect Input	After spellchecked	Correct
thsnk you	Thank you	Thank you

Example result 2.2: The bigram “repeat again” does not occur in the corpus. Hence the spellchecker does not know which word from the list is most suitable.

Incorrect Input	After spellchecked	Correct
repat again	repay again	repeat again

Case 3: Correct word followed by misspelled word

Example result 3.1: In the corpus, “morning” is usually followed by “good”. Therefore, the bigram with the highest probability is “good morning”.

Incorrect Input	After spellchecked	Correct
good mornig	good morning	good morning

Example result 3.2: The list of possible words for the misspelled word does not form a bigram that occurs in the corpus when associated with the first word.

Incorrect Input	After spellchecked	Correct

good byeeeeeeee	good bee-keepers	good bye

Case 4: Incorrect word located between two correct words

Example result 4.1: In the corpus, “you” has the highest probability to occur between “are” and “beautiful”.

Incorrect Input	**After spellchecked**	**Correct**
are ypu beautiful	are you beautiful	are you beautiful

Example result 4.2: “you helping” or “helping people” does not occur in the corpus. Therefore the spellchecker could not spell the sentence correctly.

Incorrect Input	**After spellchecked**	**Correct**
you healping people	you heading people	you helping people

Case 5: Long sentence (3 < number of word < 7) having Case 4 and 2 in sequence

Example result 5.1: “whate is” is corrected the same as in case 2 and “is youre job” the same way as in case 4.

Incorrect Input	**After spellchecked**	**Correct**
whate is youre job	what is your job	what is your job

5.2: “langouge” was incorrectly spellchecked due to the same reason as 4.2.

Incorrect Input	**After spellchecked**	**Correct**
my langouge is enghlish	my langouge is enghlish	my language is English

The table below lists our results for the tested samples for the five different cases evaluated. *Correct* indicates when the sentence is correctly spellchecked and *incorrect* when the sentences did not change after passing through the spellchecker. After careful evaluation, we concluded that the output sentences marked incorrect were had due to having transliterated Arabic, or were spellchecked with an incorrect word which resulted in loss of semantics.

Case No.	Definition	Correct (%)	Incorrect (%)
1	Single Word sentence	43	57
2	Misspelled word followed by correct word	53	47
3	Correct word followed by misspelled word	69	31
4	Incorrect word located between two correct words	Bigram: 34 Trigram: 34	Bigram: 66 Trigram: 66

5	Long sentence (3 < number of word < 7) having Case 4 & 2 in sequence	50	50

Conclusion and Future Work

We observed that context enables more accurate spellchecking of a sentence. We observed a 69% improvement (for case 3) for tested sentences. This improvement will lead to a higher hit rate in Hala's knowledge base. For case 5, despite having more context than the previous cases, the hit rate is lower. This is because other sources of errors were introduced such as users making use of abbreviated, borrowed and informal terms or were mixing Romanized form of Arabic expressions within English text.

In our future work we would like to tackle the above-mentioned problems and also work on a Part-of-Speech tagging system that would help in correcting real-word mistakes. We can also study further increasing the n-gram. However, it is highly unlikely that the spellchecker will attain higher accuracy with limited context of a query since the input queries are usually short and hence would not take advantage of a larger n-gram.

References

[1] Brown, P. F., DeSouza, P. V., Mercer, R. L., Della Pietra, V. J., & Lai, J. C. (n.d.). *Class-Based n-gram Models of Natural Language.* Retrieved May 19, 2013, from http://acl.ldc.upenn.edu/J/J92/J92-4003.pdf

[2] Chomsky, N. (1956). Three models for the description of language. IRI Transactions on Information Theory, 2(3), 113-124. http://dx.doi.org/10.1109/TIT.1956.1056813

[3] Norvig, P. (n.d.). *How to Write a Spelling Corrector* [Program documentation]. Retrieved May 19, 2013, from http://norvig.com/spell-correct.html

[4] Bassil, Y., & Alwani , M. (2012). *Context-sensitive Spelling Correction Using Google Web 1T 5-Gram. Information.Computer and Information science*, 5(3), 37-48. http://dx.doi.org/10.5539/cis.v5n3p37

[5] László, N. (n.d.). *Hunspell* (Version 1.3.2) [Computer software]. Retrieved May 19, 2013, from http://hunspell.sourceforge.net