# Spell Checker - NLP Assignment

Aparna Garimella

Surya Teja

Sai Krishna Badrinayana

Department of Computer Science and Engineering
Indian Institute of Technology, Madras
Chennai, 600036
India

19th September, 2013

# 1 Problem Statement

There are three stages of spell correction in our assignment -

- Word-level
- phrase-level
- sentence-level

# 2 Data Set

## 2.1 Word-level Spell Correction

- the set of valid words is taken from the `UNIX dictionary` - to check the validity of a given word
- list of words for prior calculations is taken from `Norvig` files
- `Norvig` files for constructing the confusion matrices

## 2.2 Phrase and sentence level Spell Correction

The `Reuter's subset` is used for both the tasks.

# 3 Technical Details

## 3.1 Confusion matrices

Confusion matrices for the operations `insertion`, `deletion`, `transposition` and `substitution` are generated based on the unigram and bigram counts from the `Norvig` files, using the formulae provided in the class.

## 3.2 Prior Calculations

Prior probabilities are calculated using the `Norvig` files and stored for prediction.

## 3.3 Smoothing

Smoothing is done using `add-`$\delta$ `smoothing`. The value of $\delta$ is chosen based on the total size of the corpus available for the corresponding task. $\delta$ is set to ..... after trying out several values for it. Smoothing is done for the following tasks.

- confusion matrices
- priors
- counts/probabilities of context words

## 3.4 BKTree

BKTree is constructed to give all those words which are at a given edit distance from the given word. It also gives the scores for those words, using the confusion matrices.

### 3.5 Homophones

Homophones are used for all the three stages of spell correction using `Metaphone` algorithm.

#### Word-level Spell Check

For this task, homophones, along with the *candidate* words at edit distance of 1 and 2 are taken. The edit distance of the homophones with the given wrongly spelt word are also computed. All these words are now sorted based on their edit distance scores and the top few words are suggested along with their scores.

#### Phrase and Sentence-level Spell Check

For this task, the context is also to be taken into account. The confusion set for the wrongly spelt word is constructed from the *candidate* words at edit distance 1 and 2 and the homophones of the wrongly spelt word. Now using context words (or collocations), these words are ranked according to their posterior probabilities. The top few words are suggested.

### 3.6 Context Words - Context Window

Several values for the context window size are tried out ($k = 2, 3, 4, 5$). It concluded that for phrase-level spell correction, $k = 2$ works better, and for sentence-level spell correction, $k = 3$ works better. But, it is to be noted that these might not be strictly correct in all cases -

- the number of test cases we tried is limited in number, so for words which we checked, these values of $k$ worked better
- the data set could be such that it gives good values for some values of $k$ while some other (not anticipated) values for other values

For same reasons, similar values for the window $k$ are chosen for collocations as well.

### 3.7 POS tagging

`Stanford Parts Of Speech Tagger` is used for tagging the sentences in the corpus. This is used for finding the `syntactic features` (collocations) for the wrongly spelt word.

### 3.8 Pruning of the candidate context words

The number of context words we get could be sometimes really huge, which would make the process of suggestion computationally expensive. Hence, there is a need for pruning the candidate context words.

Consider the following two scenarios.

- A particular context word occurs very rarely in the corpus, that is, its count itself is very low - it might occur just once or twice in the whole corpus, and those occurances are with the given word. In this case, it is very unlikely that it could help discriminating among the confused words.
- A particular contex word might be a context word for all the words in the confusion set equally, in which case it does not help discriminating among the confused set.

Hence, such words are pruned. In addition to these words, `stop words` are also removed from the candidate context words.

# 4 Assumptions

- Spelling error is assumed to occur only due to mis-spelt words, illustrated as follows. '*pese* of cake' will be corrected, but not '*peace* of cake'

# 5 Procedure

## 5.1 Words Spell Check

Standalone words are given here and corrections are suggested.

**Training Phase**

- Norvig files are parsed and a list of words is constructed along with their counts.
- The confusion matrices are then constructed for the operations `insertion`, `deletion`, `transposition` and `substitution`.
- `BKTree` is constructed for the given words - using this, the words at a given edit distance from a given words can be efficiently found in $\mathrm{O}(logn)$ time.
- All the priors and confusion matrices are smoothed to account for the unknown elements' values.
- Also, for each word in the training corpus, list of homophones are found and stored.

**Testing**

- Given a misspelt word $w$, top few words at edit distance 1 and 2, along with the set of homophones to the given word are extracted to form the candidate set of suggestions.
- Now, edit distance is computed for the homophones as well and the candidate set is sorted based on the edit distance measure.
- Top few words are suggested as possible corrections.

## 5.2 Phrase-Level and Sentence-level Spell Correction

For both phrase and sentence levels of spell correction, the context is to be taken into account, in addition to the edit distance. We followed the same procedure for both phrase-level and sentence-level spell correction, with the only difference being the window size (*more context* in the case of sentences). For phrase-level spelling correction, given a phrase, the following things are done.

- the wrongly spelt word is identified.
- confusion set for this word is constructed - this set consists of all those words which are candidate replacements to the wrongly spelt word. This set consists of words which are at `edit distances` of 1 and 2, and words which are `homophones` of the wrongly spelt word.
- Two methods are used for suggesting the correction - context words and the syntax words (collocations) for each of the words in the confusion set.

**Using Context Words**

$p\left(w_i|c_{-k},...,c_{-1},c_1,...,c_k\right) = \frac{p(c_{-k},...,c_{-1},c_1,...,c_k|w_i)p(w_i)}{p(c_{-k},...,c_{-1},c_1,...,c_k)}$

The likelihood term $p\left(c_{-k},...,c_{-1},c_1,c_k\right)$ is difficult to estimate from the training data, due to the sparse data problem (we would need to count the number of times entire context occured along with the given word $w_i$). Hence, `conditional independence` is assumed to hold in this situation - the presence of one word in the context is independent of any other word. Hence, the above equation reduces to the following.

$p\left(w_i|c_{-k},...,c_{-1},c_1,...,c_k\right) \propto \prod_{j\in -k,...,-1,1,...,k} p\left(c_j|w_i\right) p\left(w_i\right)$

**Training Phase**

- The entire dataset is parsed and confusion sets for each of the words in the dataset are constructed.

  $Candidate\ words \ \leftarrow \ \{Words\ at\ edit\ distance\ 1\ and\ 2\ sorted\ based\ on\ their\ scores\}$

  $Confusion\ set \ \leftarrow \ Candidate\ words \ \cup \ \{Homophones\ to\ the\ given\ word\}$

- Along with the confusion sets, the counts for each word in the corpus is also maintained - this is used for `prior` calculation.
- Also for each word in the corpus, its context feature set is also constructed along with their counts. Context words include all those words which occur within a window of $k$ around the given word in the corpus. The counts for each context word are also maintained - this is used for `likelihood` calculation.
- If the number of context words is large, the spell correction could become ecomputationally expensive. Hence there is a need to prune the set of context words. Pruning is done based on the fact - some of the context words rarely occur (their count very low) and some of the context words occur in context of every word in the confusion set - in this case, they are not so useful in differentiating between the various confused words. Hence, such words are removed from the context set. The threshold used for this pruning is set to $T_{min}$ after experimenting with various values.
- All the information about each word in the training corpus is stored - this information includes the following.
  - count of the word in the corpus
  - its pruned context feature set consisting of context words and their counts
  - its pruned syntax feature set consisting of syntax words (collocations, described in the later part) and their counts

**Testing**

- Given a phrase (or a sentence), the wrongly spelt word is identified based on the dictionary we maintain.
- The confusion set for the wrongly spelt word is extracted, and the `posterior` for each of the words is calculated, based on the above formula.
- The word (or top three words) in the confusion set with the highest probability is (are) suggested as the correction(s).

The value of the context window $k$ is set to 2 for phrase level spell correction and to 3 for sentence level spell correction. This was based on the observations with other higher and lower values of $k$.

**Using Syntax Words (Collocations)**

This procedure is similar to the above procedure of using context words. In the training phase, the syntactic features, instead of context features are extracted and stored using the `POS tagging`. These features are pruned according to the two measures mentioned above. We tried this for a few cases. But the results are reported only based on context words.

# 6 Advantages and disadvantages

## 6.1 Strengths

## 6.2 Limitations

- How much ever rigorous the probability computations are and how much ever deep the pruning is done, the final result is highly data-driven. If the data set we chose does not contain the actual word, then it is very likely that it is not suggested as the correction (smoothing only increases its prior probability to a small extent).

# 7 Results of test cases provided

## 7.1 Word-level Spell Correction

| Misspelt words | Suggestions | Mean Reciprocal Rank |
|---|---|---|
| *belive* | $q$ | modens ponens |
| *bouyant* | $\neg p$ | modus tollens |
| *comitte* | $p \rightarrow r$ | hypothetical syllogism |
| *distarct* | $q$ | disjunctive syllogism |
| *extacy* | $p \vee q$ | addition |
| *pfailr* | $p$ | simplification |
| *hellpp* | $p \wedge q$ | conjunction |
| *gracefull* | $q \vee r$ | resolution |
| *liason* | $q \vee r$ | resolution |
| *ocassion* | $q \vee r$ | resolution |
| *possable* | $q \vee r$ | resolution |
| *thruout* | $q \vee r$ | resolution |
| *volly* | $q \vee r$ | resolution |
| *tatoos* | $q \vee r$ | resolution |
| *respe* | $q \vee r$ | resolution |

## 7.2 Phrase-level Spell Correction

| Misspelt phrases | Suggestions | Mean Reciprocal Rank |
|---|---|---|
| from the *eath* to the moon | $q$ | modens ponens |
| *wate* fountain at paris | $\neg p$ | modus tollens |
| *roff* of the house | $p \rightarrow r$ | hypothetical syllogism |
| a *geant* leap for mankind | $q$ | disjunctive syllogism |
| interested in your *opinin* | $p \vee q$ | addition |
| hope for the *futre* | $p$ | simplification |
| taking an *extreem* step | $p \wedge q$ | conjunction |
| walking down the *aisel* | $q \vee r$ | resolution |
| cops and *robers* | $q \vee r$ | resolution |
| *presient* of united states | $q \vee r$ | resolution |
| a few *goode* men | $q \vee r$ | resolution |
| a *briht* sunny day | $q \vee r$ | resolution |
| *rainig* cats and *doggs* | $q \vee r$ | resolution |
| chill down the *spene* | $q \vee r$ | resolution |
| broken *hart* | $q \vee r$ | resolution |
| coyote fox | $q \vee r$ | resolution |

## 7.3 Sentence-level Spell Correction

| **Misspelt sentences** | **Suggestions** | **Mean Reciprocal Rank** |
|---|---|---|
| The parliament passsed the *resoltion* to discuss the *bil*. | $q$ | modens ponens |
| Private *hopitals* to provide *frea* treatment to the poor. | $\neg p$ | modus tollens |
| The *fotball* match was very interesting. | $p \rightarrow r$ | hypothetical syllogism |
| The food served in the *restarant* was very *godd*. | $q$ | disjunctive syllogism |
| The departments for the institutes offer *corses* conducted by highly qualified staff. | $p \vee q$ | addition |
| In great *powrr* lies great responsibility. | $p$ | simplification |
| All divisions of the *ramed* forces participated in the parade. | $p \wedge q$ | conjunction |
| To be or to *bea* is not the question. | $q \vee r$ | resolution |
| The crime *raet* seems to be under control. | $q \vee r$ | resolution |
| A great *victry* has come but at a great *cort*. | $q \vee r$ | resolution |
| You cannot *handel* the truth. | $q \vee r$ | resolution |
| The *powre* has now shifted to the east. | $q \vee r$ | resolution |
| Keep your *frinds* close and your *enemis* closer. | $q \vee r$ | resolution |
| The best part of the *stiry* is yet to come. | $q \vee r$ | resolution |
| Who said it is difficult it is *impossile* | $q \vee r$ | resolution |