Middle East Technical University Department of Computer Engineering

# CENG 463

## Introduction to Natural Language Processing

Fall ’2012-2013

## Programming Assignment 1

Due date: 4 November 2012, Sunday, 23:55

## 1 Objectives

In this assignment you are expected to implement spell checker for English language. You will try to correct misspelled words in English.
You are also expected to write a report containing your implementation idea and possible problems and behavior of your algorithm under these conditions. State possible improvements and solutions to these problems.

**Keywords:** *Spelling Correction, Spell Checker, Noisy Channel Model, Edit Distance*

## 2 Noisy Channel Model

Noisy Channel Model (Shannon 1948) has been used in variety of areas in information theory. The idea is simple; you want to transfer data(a message) from your source (yourself) to your target (your friend) through a channel which is ’noisy’. You have no control over transfer channel and no idea how your data may altered. On the other side of the channel, your friend will try to understand your message which may have changed (reconstruct your original data).

An example would be ’Telephone’ game which you whisper a word or a sentence to the person next to you and the word is transferred ’from ear to ear’. In the end your team wins if the last person correctly calls the initial word.

Bayesian Decision is a good way of reconstruction. Suppose word ’w’ is altered into word ’m’. You are given ’m’ and and trying to find ’w’. Best decision is to choose the word ’w’ that maximizes the probability of ’w’ given ’m’.

$$argmax_w P(w|m) \tag{1}$$

where we can use Bayes Theorem

$$P(w|m) = \frac{P(m|w)P(w)}{P(m)} \tag{2}$$

- P(w) is the probability of a word to be transmitted and comes from your dictionary or corpus (language model).

- p(m|w) is the probability of an error generating the word m out of word w and comes from your error model.

- p(m) is the probability that you see m at the end of any transmission.
  (Hint: This will be same for each comparison and can be dropped from equation.)

# 3 Edit Distance

You will need a metric to compare word similarities. Edit distance is the distance or similarity measure between words and may be defined differently depending on which operations are allowed.

## 3.1 Levenshtein distance

In information theory and computer science, the Levenshtein distance is a string metric for measuring the difference between two sequences and equal to the number of single-character edits required to change one word into the other. Those single character edits are insertion, deletion, or substitution of a single character.

## 3.2 Damerau–Levenshtein distance

Damerau–Levenshtein distance allows the transposition of two adjacent characters besides remaining three operations. Damerau not only distinguished these four edit operations but also stated that they correspond to more than 80% of all human misspellings.

# 4 Single Character Edits in Damerau–Levenshtein distance

1. **Add a single letter**
   A new letter may be inserted at any position of the word
   e.g. something → somethying

2. **Delete a single letter**
   A letter may be deleted from any position of the word
   e.g. something → somethin

3. **Replace one letter with another**
   A letter at any position of the word may be replaced with another
   e.g. something → samething

4. **Transpose two adjacent letters**
   Two adjacent letters may be replaced with each other
   e.g. something → something

```
distance('ab','b') = 1
distance('ab','ba') = 1
distance('ab','ac') = 1
distance('ab','abc') = 1
distance('ab','acb') = 1 ( 'ab'->'abc'->'acb' or 'ab'->'abb'->'acb' ...)
```

# 5 Specifications

1. **N**ame your module spelling_corrector

2. **Y**ou will implement a function that will have single misspelled word and return at most five probable suggestions and their probabilities from your system.

```
correct_word(word)
```

3. **Y**ou are expected to train your system with Brown Corpus which has over one million tokens. Brown Corpus will be available on the test system

```
from nltk.corpus import brown
```

4. **Y**our module will be tested with an additional test.py file which imports your module and calls implemented functions.

```
from spelling_corrector import correct_word
```

# 6 Samples

```
>>> correct_word('said')
[('said', 0.9902), ('aids', 0.0049), ('maids', 0.0049)]
```

| | original | misspelled | suggestions |
|---|---|---|---|
| 1 | summit | sumit | [('summit', 0.56), ('suit', 0.44)] |
| 2 | we | le | [('le', 1.0)] |
| 3 | cold | ocld | [('old', 0.77), ('cold', 0.23)] |
| 4 | something | somerthing | [('something', 1.0)] |
| 5 | Schenk | Schnk | [('Schenk', 1.0)] |
| 6 | average | vverage | [('average', 1.0)] |
| 7 | said | saids | [('said', 0.99), ('aids', 0.00), ('maids', 0.00)] |
| 8 | The | me | [('me', 1.0)] |
| 9 | Company | ompany | [('company', 1.0)] |
| 10 | appeals | appeales | [('appeals', 0.58), ('appealed', 0.42)] |

Table 1: Several examples

# 7 Regulations

1. **Programming Language:** You will use Natural Language Toolkit with Python language.

2. **Late Submission:** Penalty for each day is calculated by (number of days) *10.

3. **R**emember that students of this course are bounded to code of honor and its violation is subject to severe punishment.

4. **Cheating:** We have zero tolerance policy for cheating. In case of cheating, all parts involved (source(s) and receiver(s)) get zero. People involved in cheating will be punished according to the university regulations.

5. **Newsgroup:** You must follow the newsgroup (news.ceng.metu.edu.tr) for discussions and possible updates on a daily basis.

6. **Evaluation:** Your program will be evaluated automatically using ”black-box” technique so make sure to obey the specifications. Your module will be tested with an additional test.py file which imports your module and calls implemented functions.

   - **Y**our ’correct_word(word)’ function will be run on 3 sets of 100 misspelled words from Brown Corpus.
   - **S**ets will be generated randomly with spelling errors described above.
   - **Y**ou will have 1.0 point if your first suggestion is the original word, 0.8 for second and so on.
   - **Y**our best score from these three tests will be your final score.

```
# example test
from spelling_corrector import correct_word
misspelled_words = generate_misspelling(100)
for word in misspelled_words
   # word is a pair -> [original, corrupted]
   evaluate(word[0],correct_word(word[1]))
```

# 8 Submission

- Submission will be done via COW.
  Create a tar.gz file named `hw1.tar.gz` that contains all your source code files and a copy of your report in pdf format.
- Submit a hardy copy of your report to A203.

# 9 References

- Eric Brill, Robert C. Moore: An Improved Error Model for Noisy Channel Spelling Correction. ACL 2000
- http://norvig.com/spell-correct.html
- http://nlp.stanford.edu/IR-book/html/htmledition/spelling-correction-1.html