A comprehensive assignment on Spell Checking and Correction

The purpose of this assignment is to give you a good insight into Statistical NLP on both generative and discriminative fronts:

1. Take the parallel list of incorrect and correct words in the URL <http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings/For_machines>
2. Let *C* be the correct word corresponding to *W* the wrong word. *W*- input, *C*- Output
3. Solve the spell checking problem with the following DIFFERENT approaches and COMPARE them:
   1. Edit Distance based
      1. Error due to insertion *(I: distance/penalty=1), deletion (D: distance/penalty=1), substitution (S: distance/penalty=2)* and *Transposition (X: distance/penalty=2)*
      2. Compute the distance *D(C,W)*
      3. Rank the candidate *Cs* according to *D*; in case of ‘tie’, apply additional knowledge (need to think!)
   2. Generative

*C\*=argmaxC[P(C|W)]=argmaxC[P(C).P(W|C)]*

*P(C)=product of bigrams P(Ci|Ci-1)*

*P(W|C)* will be modeled in three ways:

* + 1. *P(W|C)=P(W|C)I+ P(W|C)D+ P(W|C)S+ P(W|C)X*
    2. *P(W|C)* modeled by alignment as in statistical machine translation (IBM models)
    3. *P(W|C)* modeled through its sound pattern in the CMU pronunciation dictionary

<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

* 1. Discriminative

P(C|W) will be modeled DIRECTLY, through features using an exponential model. If *f1, f2, f3*… are features of the words, then *P(C|W)=e(k1f1+k2f2+k3f3+…)* , where *ki*s are feature weights to be learnt from the training corpus.