A Mixed Trigrams Approach for Context Sensitive Spell Checking

Davide Fossati Barbara Di Eugenio

Department of Computer Science
University of Illinois at Chicago
Chicago, IL, USA

Lecture in Natural Language Processing, 2010

Outline

Introduction

Abstract

The paper addresses the problem of real-word spell checking, i.e., the detection and correction of typos that result in real words of the target language. The paper proposes a methodology based on a mixed trigrams language model.

Introduction

- *Spell checking* is the process of finding misspelled words in a written text, and possibly correcting them.
- We can classify spelling errors in two main groups:
  - *Non-word errors*, which are spelling errors that result in words that do not exist in the language. E.g. “The *bok* was on the table.”
  - *Real-word errors* are errors that by chance end up as actual words. E.g. “I saw *tree* tress in the park.”
- Detecting and correcting real-word errors is the main focus of the paper.

Approaches

- Different approaches to real-word spell checking are present in the literature, e.g.
  - Symbolic approaches check for grammatical anomalies.
  - Statistical methods using *n*-gram models, PoS tagging, e.t.c.

Problems with statistical methods

- The problem with statistical methods using only *n*-grams is the data sparseness problem.
- PoS methods suffer less from sparseness problems, but are unable to detect misspelled that are of the same part of speech.

Outline

Definition

Definition

Given a sentence, a *mixed trigram* is a sequence of three elements (e_i, e_{i+1}, e_{i+2}), where e_k is either the k-th word of the sentence or its part of speech. Furthermore, at most one of the elements can be a word.

Example

Consider the sentence

The /DET kids /NOUN eat /VERB fresh /ADJ apples /NOUN

A complete set of mixed trigrams derived from it are:

(The, NOUN, VERB)	(NOUN, VERB, fresh)
(DET, kids, NOUN)	(NOUN, VERB, ADJ)
(DET, NOUN, eat)	(eat, ADJ, NOUN)
(DET, NOUN, VERB)	(VERB, fresh, NOUN)
(kids, VERB, ADJ)	(VERB, ADJ, apples)
(NOUN, eat, ADJ)	(VERB, ADJ, NOUN)

Outline

Definition

Definition

Given a dictionary W, a distance function d defining a metric on W, and a word $w \in W$, a *confusion set* $C(w) \subseteq W$ is a set of words such that $w_c \in C(w)$ if and only if $d(w, w_c) \leq k$, where k is a constant.

In practice the confusion set of a word contains all the words "similar enough" to that word.

Outline

Definition

Definition

Levensthein Distance is the minimum number of edition operations necessary to transform one word into another. An edition operation is either a character insertion, deletion or substitution.

- Reflects well on common typing mistakes.
- A word misspelled by pressing a key twice, typing two keys instead of one, skipping a key or typing a wrong key gives a Levensthein distance of 1 from the intended word.
- Switching two characters gives a Levensthein distance of 2.

Outline

Method

Given a sentence $S = w_1 \ldots w_k \ldots w_n$, find the most likely sequence of elements $E = t_1 \ldots w_k^c \ldots t_n$, where:

- w_i, $1 \leq i \leq n$ are words;
- w_k is the *central word* (i.e. the word to be checked);
- t_i, $1 \leq i \leq n$, $i \neq k$ are part of speech tags;
- w_k^c is a word belonging to the confusion set of the central word w_k.

Method (cont'd)

The observed word w_k is likely to be a spelling mistake of w_k^c if:

1. $w_k \neq w_k^c$ and
2. the probability of the sequence E is less than the probability of another sequence $\overline{E} = \overline{t_1} \ldots w_k \ldots \overline{t_n}$. I.e. the sequence $\overline{E}$ is the most likely sequence of elements where the central word w_k is assumed to be correct.

The criterion above means that a word will be detected as a spelling mistake if another word in the confusion set has higher likelihood of fitting into the same context. In particular, the word in the confusion set belonging to the sequence with the highest probability will be selected by the correction algorithm.

Method (cont'd)

This maximization problem can be solved using the Markov Model approach traditionally applied to PoS tagging, adopting the same simplifying assumptions, and using mixed trigrams instead of word trigrams. The resulting formula is:

$$\underset{E}{\operatorname{argmax}} \prod_{i=1}^{n} P(w_i|e_i)P(e_i|e_{i-1}, e_{i-2})$$

where w_i are words and e_i are either words or PoS tags.

Conditional probability estimation for the central word

In the formula on the previous slide, if $i = k$, the term $P(w_i|e_i) = P(w_k|w_k^c)$ means the probability of getting the word w_k from a misspelling of the word w_k^c. The estimation of this probability is

$$P(w_k|w_k^c) = \frac{\alpha(1-\alpha)^{d(w_k^c, w_k)}}{\left|\{\lambda \in W : d(\lambda, w_k) = d(w_k^c, w_k)\}\right|},$$

where α is a parameter that can be tuned with empirical investigation.

Example

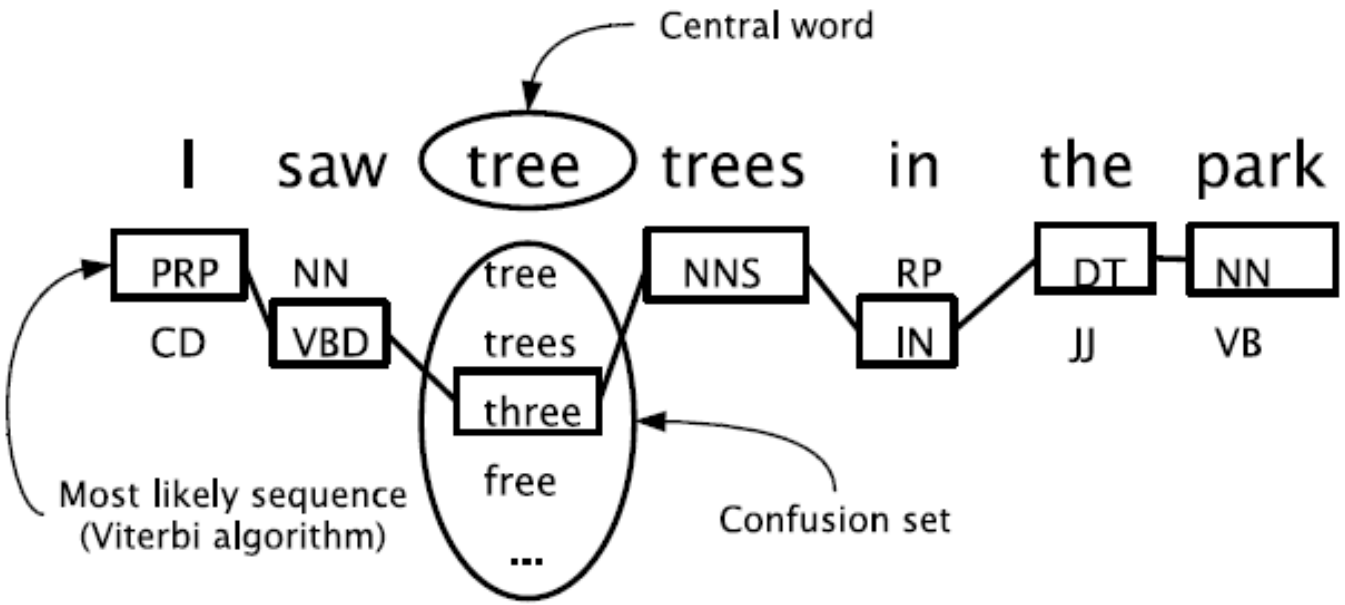

Figure: An example of the detection process.

Algorithm training

Training the algorithm requires

- creating a vocabulary,
- estimating probabilities of the mixed trigrams and conditional probabilities of each word given a PoS,
- calculating the distance between words in the vocabulary,
- calculating the conditional probability of each word given another word in its confusion set.

Algorithm training (cont'd)

- Vocabulary created from the Penn Treebank corpus.
- Least frequent words were not inserted.
- Three training data sets of increasing size were used.
- Probabilities were computed using MLE.
- Three different values used for α $(0.25, 0.50, 0.75)$.
- For each sentence in the test set, one spelling mistake was artificially inserted.
- Words shorter than three characters were skipped.

Hit rates

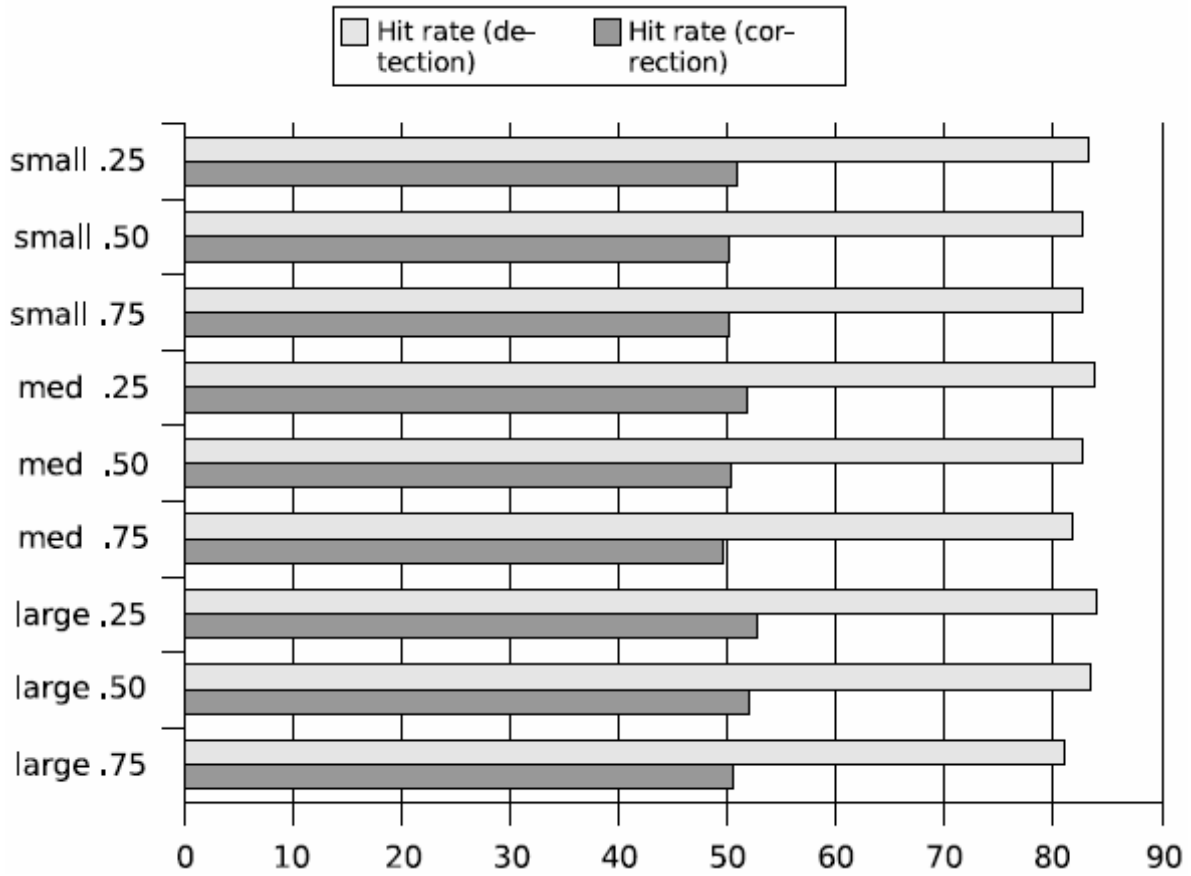

Figure: Hit rates.

False positive rates

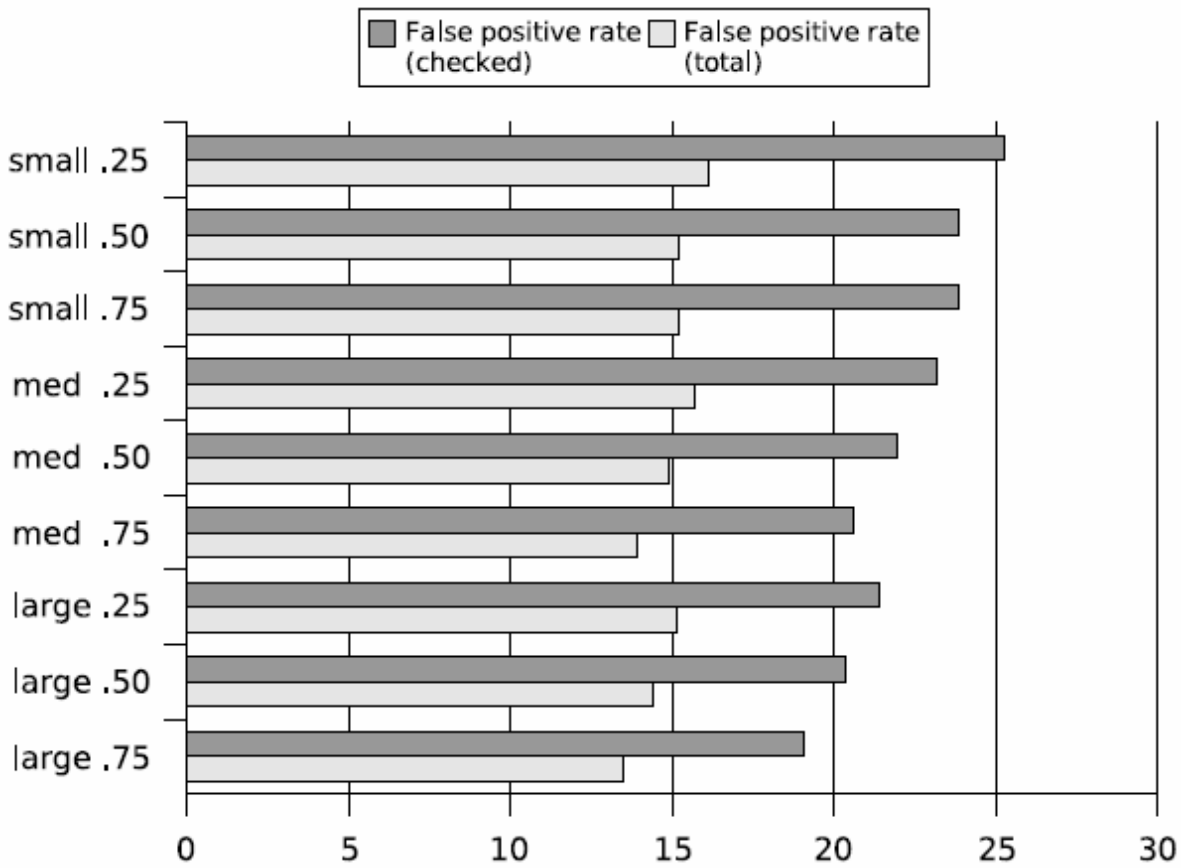

Figure: False positive rates.

Detected over false positive ratio

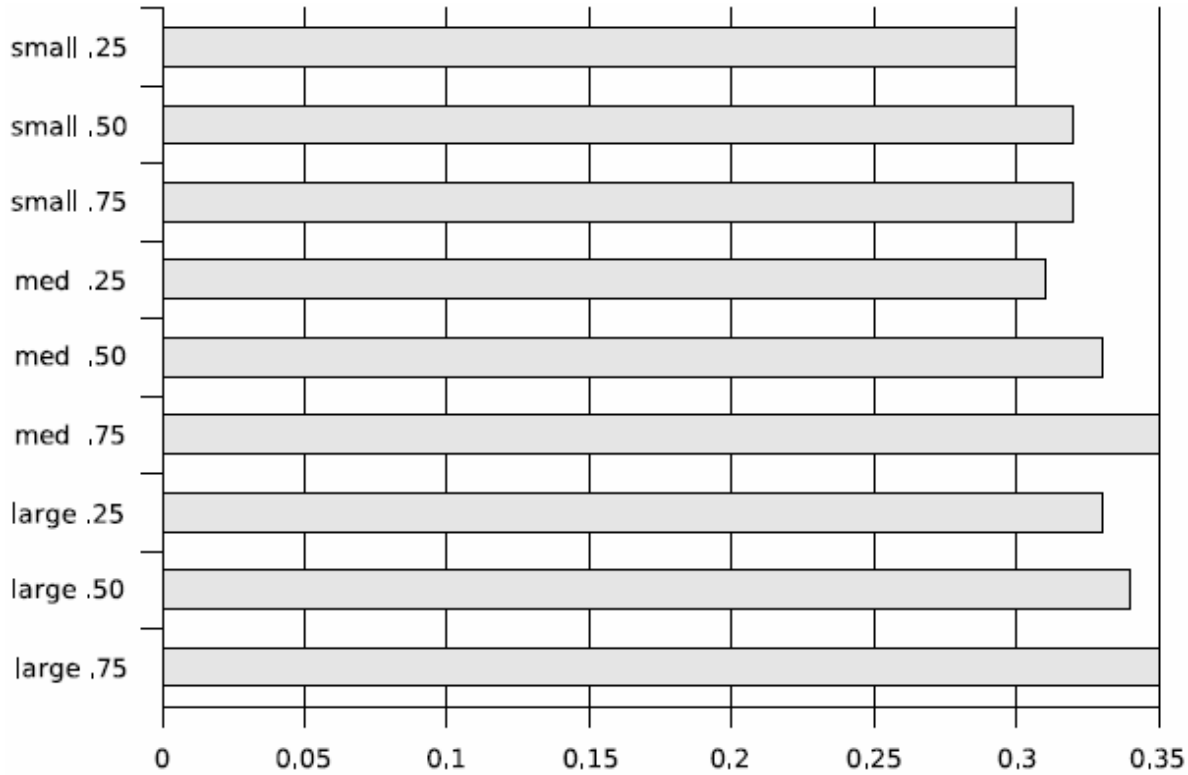

Figure: Detected over false positive ratio.

Coverage

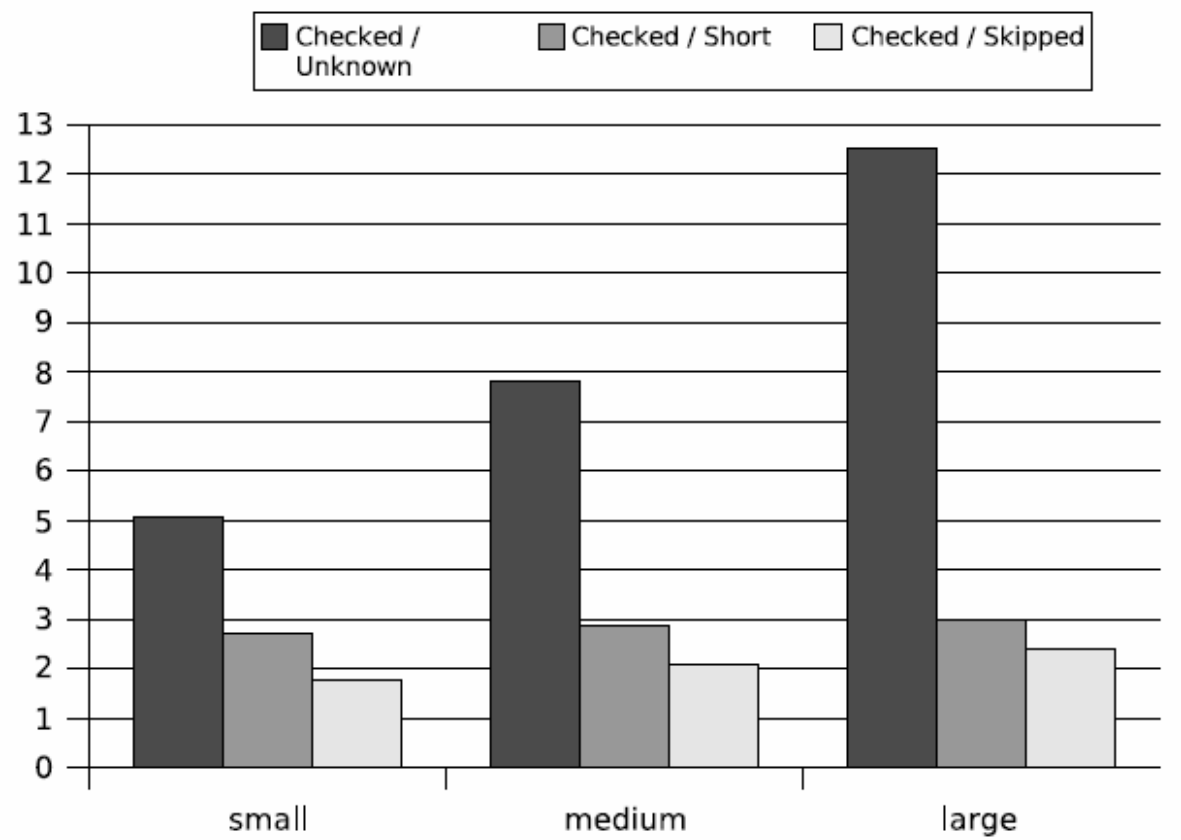

Figure: Coverage of the spell checker.

Conclusion

The results of these experiments are promising, and represent a good starting point for future research. Among the others, there are several points that would be worthy of further investigation:

- Improve estimation of the probability tables.
- Deal with problems of short words.
- Run experiments on ecologically valid data.
- Explore the usage of alternative word distance measures and conditional probabilities estimations for words with respect to their confusion set.