Spelling Correction in Context BORATOIRE D'INFORMATIC DE NANTES ATLANTIC



Guillaume PINOT and Chantal ENGUEHARD

Laboratoire d'Informatique de Nantes Atlantique (LINA) — Université de Nantes — France

Introduction

The goal is to correct real-word errors. They occur when one or more modifications of a word transform it into another word which is present in the dictionary.

example : This chocolate cake is a famous desert.

The omission of an s in *dessert* reveals the word *desert*.

Intuitive idea

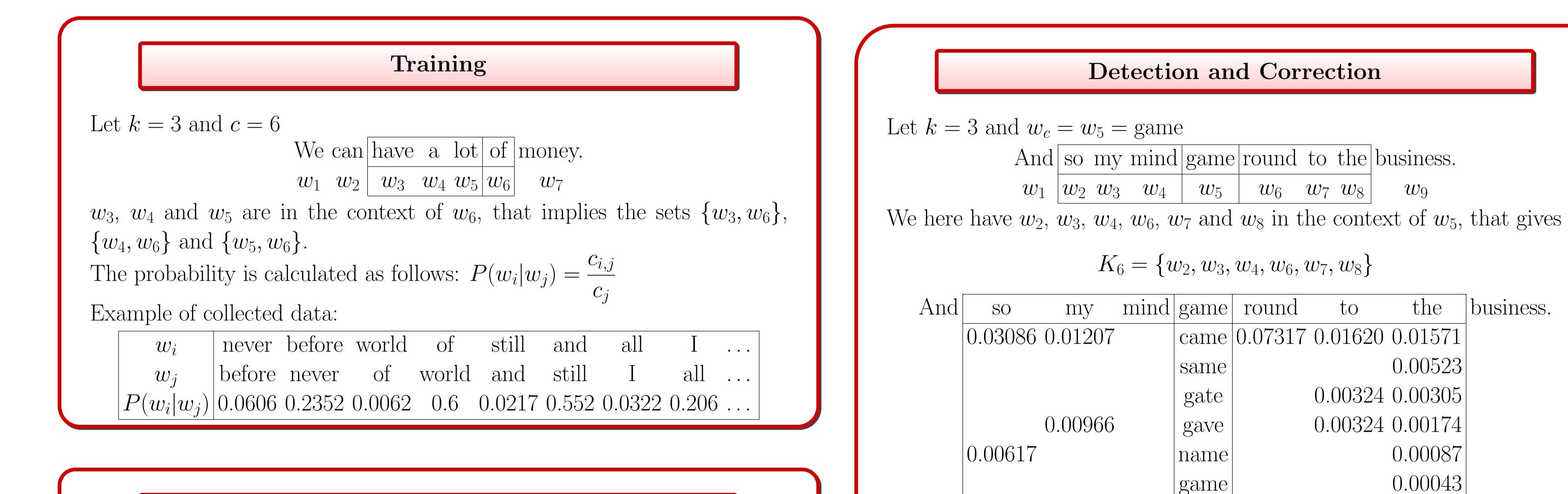
Traning: Learn contextual probabilities.

Correction :

1. Get all words that can have the current context. 2. Keep the words similar to the word to correct.

3. Score and sort these propositions.

4. Determine if the word is erroneous or not



Similarity Between Two Words

let $edist(w_i, w_j)$ the minimal edition distance. Let ϵ be the empty string, we can define $sim(w_i, w_j)$:

$$\sin(w_i, w_j) = \begin{cases} \text{true if } \text{edist}(w_i, w_j) \leq \frac{\text{edist}(w_i, \epsilon) + \text{edist}(w_j, \epsilon)}{\gamma} + c \\ \text{false } \text{else} \end{cases}$$

In practice, we will take $\gamma = 8$ and c = 1. These values have been determined after several experimentations.

We now need a heuristic to give a score to each proposition in order to have them in a pertinent order.

 $H_{\text{came}} = 5 + 0.030 \times 0.012 \times 0.073 \times 0.016 \times 0.0157 = 5.00000006601$ $H_{\rm game} = 1 + 0.0004 = 1.0004$

Then, the detection:

 $H1: H_{\text{came}} > H_{\text{game}} \Rightarrow \text{error}$ $H2: [H_{\text{came}}] > [H_{\text{game}}] \Rightarrow \text{error}$

Experimentations

Corpus: *les Misérables* by Victor HUGO divided into two parts:

- The training part (480,588 words);
- The part to be corrected (53,405 words) in which errors have been added.

Adding Errors:

pire

Results

Err.	Det.	Precision	Recall	Prop. Cor.	PCA	NPA
10%	H1	0.1081926	0.9363030	0.9622054	2.57	14.49
10%	H2	0.1561469	0.8800168	0.9635141	2.63	14.63
1%	H1	0.0206164	0.9615384	0.9500000	2.29	14.16
1%	H2	0.0301319	0.9120603	0.9696969	2.56	14.56

pie pirée paire poire pitre pires spire

PCA: Position of the Correction in Average

NPA: Number of Propositions in Average

 w_9

the

business.

Conclusion					
The advantages of this algorithm are:• simplicity;	 The disadvantages are: the significant size of the data generated by the training. the low precision on detection: the algorithm proposes corrections for a lot of correct words. 	Futur works:To search better heuristics for the detection.			
 independence from any linguistic information; use of a raw corpus for the training; few parameters have to be regulated. 		 To evaluate the influence of the size of the training co To define ordered contexts to use the syntax in additi semantics. 			