Correction orthographique par apprentissage Bayésien

Edouard Leurent
Bastien Debras

31 mai 2013

Cours d'apprentissage artificiel

Table des matières

1 Introduction

1.1 Démarche

Comment réagir face à une faute d'orthographe ? Une première idée pourrait être de regarder dans un **dictionnaire**. Si l'utilisateur tape un mot qui ne fait pas partie du dictionnaire, on le remplace simplement par le mot du dictionnaire le plus proche (en un sens à préciser).

Cependant, on se rend vite compte que cette définition du mot "le plus proche" est ambigüe : que faire si un grand nombre de mots sont tous également candidats à être le mot le plus proche ?

Dans cette situation, notre démarche pour lever l'ambiguïté a été de choisir mot le plus **probable**, c'est à dire le plus fréquent, celui qui apparait le plus dans l'usage. Pour apprendre ces fréquences, il faut donc disposer non pas d'un simple dictionnaire, mais plutôt d'un **corpus** de textes.

2 Formalisation

Nous allons ici préciser nos problèmes, et expliquer notre choix d'algorithmes pour les traiter

2.1 Les problèmes

Nous allons voir que le problème de correction orthographique peut se ramener à un problème de machine learning.

- On considère un mot tapé par l'utilisateur comme une **observation** du mot réel qu'il a voulu écrire. Les mots d'une langue sont alors interprétées comme des **classes**. Corriger une faute d'orthographe revient donc à trouver à quelle classe appartient une observation donnée. Nous avons un premier **problème de classification**, très fortement multiclasse (par exemple, la langue française contient environ 100 000 mots).
- Pour rester général, nous avons décider de ne pas considérer uniquement le français mais de traiter le cas d'un **langage quelconque**. Une fois plusieurs langages appris, corriger une phrase donnée nécessite de savoir dans quel langage il faut raisonner. On doit donc être capable d'identifier dans quelle langue est écrite une phrase donnée, ce qui constitue notre deuxième **problème de classification**.

2.2 Méthode Bayésienne

On a choisit de considérer les mots les plus probables, et on a donc adopté une approche probabiliste. Une méthode adéquate est donc d'utiliser un classifieur naïf de Bayes.

Il est formalisé de la manière suivante :

Étant donné l'ensemble C des corrections possibles d'un mot donné, on s'intéresse à la probabilité qu'une correction particulière soit la bonne.

On note m le mot tapé par l'utilisateur
c une correction possible de ce mot
$P(c|m)$ la probabilité que le mot lu m soit une observation du mot correct c

On a alors, d'après la **formule de Bayes**

$$P(c|m) = \frac{P(m|c) \bullet P(c)}{P(m)} \tag{1}$$

On essaie de maximiser cette quantité. Le dénominateur $P(m)$ n'est donc qu'un facteur normalisateur qui est constant sur l'ensemble C des corrections, et on peut donc choisir la convention $P(m) = 1$

On cherche donc

$$\max_{c \in C} P(m|c) \bullet P(c) \tag{2}$$

On constate alors que notre fonction à maximiser est constituée de deux termes

- $P(m|c)$ la probabilité d'avoir fait la faute m en voulant taper le mot c. C'est le **modèle d'erreur**.
- $P(c)$ la probabilité que l'utilisateur ait voulu taper le mot c. C'est le **modèle de langage**.

Cette séparation en deux termes est très intuitive. Comment corriger la faute de frappe *histre* par exemple ? On peut proposer les deux corrections *bistre* et *histoire* par exemple. Laquelle est la meilleure ?

- D'un coté la mot *bistre* est plus proche de *histre* que la mot *histoire*, car une seule modification est nécessaire contre deux pour *histoire*. Il est plus probable que l'utilisateur ait faite une seule faute plutôt que deux fautes dans un même mot.
- D'un autre coté, le mot *histoire* est bien plus courant que le mot *bistre*, et il est donc plus probable que l'utilisateur ait voulu écrire ce mot.

Il faut donc bien considérer ces deux critères du modèle de langage et du modèle d'erreur pour pouvoir lever l'ambiguïté et choisir la meilleure correction.

2.3 Le modèle de langage

2.4 Le modèle d'erreur

On doit mettre en place un modèle permettant de calculer $P(m|c)$, la probabilité qu'en voulant écrire c l'utilisateur fasse une ou plusieurs fautes et aboutisse au mot m.

Nous avons choisit un modèle simple, basé sur la notion d'**opération élémentaire**.

Une opération élémentaire sur un mot est définie comme étant au choix :

- La suppression d'une lettre du mot
- Le remplacement d'une lettre du mot
- L'insertion du lettre dans le mot
- La transposition de deux lettres du mot

On interprète alors une faute de frappe comme une opération élémentaire, à laquelle on attribue une certaine probabilité d'erreur P_e. Cette probabilité est un paramètre de l'algorithme, choisi ici à $1/20^{\text{ème}}$.

On calcule ensuite la distance d'édition d entre deux mots, c'est à dire le nombre minimal d'opérations élémentaires pour passer d'un mot à l'autre.

On pose finalement le modèle d'erreur suivant :

$$P(m|c) = Pe^{d(m,c)} \tag{3}$$

3 Implémentation

3.1 Modèle de langage

3.1.1 La structure de dictionnaire

3.1.2 Phase d'apprentissage

3.2 Modèle d'erreur

3.2.1 Génération des corrections

3.2.2 Sélection des meilleurs corrections

3.3 L'interface IHM

4 Conclusion