Soutenance :
Correction de texte chat par approche de traduction automatique

ALOUI Cindy
DARY Franck

Introduction

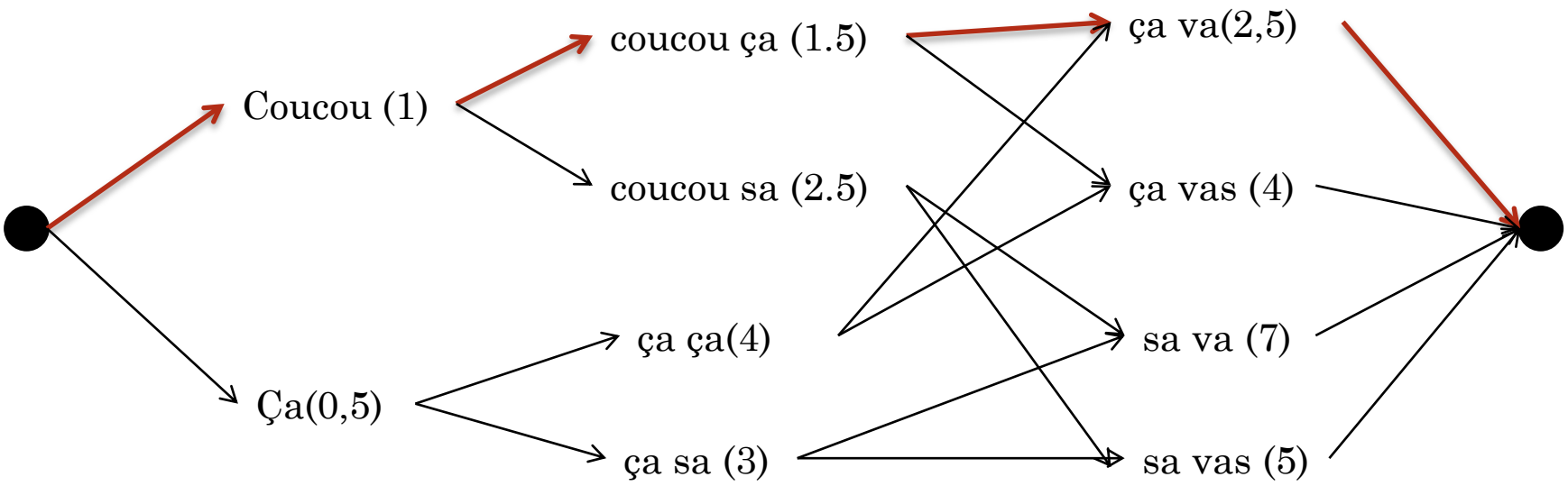
#Begin tchat1
[00:11:09] _TC_: Bjr, je suis _TC_, que puis-je pour vous ?
[00:13:24] _CLIENT_: bonjour _TC_ volia depuis samedi j ai changer toute mon instalation livebox plus decodeur mais quand on regarde la tv sa s arrete souvent et apres sa repart
[00:13:29] _TC_: Merci de me cofirmer votre numéro ligne fixe afin que je soiss sûr d'avoir le bon dossier.
[00:13:58] _CLIENT_: _NUMTEL_

#Begin tchat1
[00:11:09] _TC_: Bonjour, je suis _TC_, que puis-je pour vous ?
[00:13:24] _CLIENT_: bonjour _TC_ voilà depuis samedi j'ai changé toute mon installation livebox plus décodeur mais quand on regarde la tv ça s'arrête souvent et après ça repart
[00:13:29] _TC_: Merci de me confirmer votre numéro ligne fixe afin que je sois sûr d'avoir le bon dossier.
[00:13:58] _CLIENT_: _NUMTEL_

Méthode de base

• Construction d'un treillis et résolution avec Viterbi

Phrase de base : « cc, ca va? »



Phrase corrigée : « Coucou, ça va? »

Méthode d'évaluation

• On dispose d'un texte de chat (999 lignes, 8683 mots) .

• On dispose de sa correction par un être humain.

• On utilise l'outil Sclite qui calcul le taux d'erreur entre un texte et sa correction.

• Le texte avant correction a un taux d'erreur de 11,5%.

• On cherche à améliorer ce taux.

Distance de Levenshtein

T É L E C O M A N D E

↓ Substitution e en é

T É L É C O M A N D E

↓ Insertion d'un m

T É L É C O M M A N D E

⇒Télecomande est à distance 2 de télécommande

Nouveau taux d'erreur : **9,00%**

Modèle 3-grams

« J ’ai un problème » est plus fréquent que « j’est un problème ».

- Apprendre un modèle statistique d’un grand texte.
- Utiliser les probabilités des suites de mots dans l’algorithme de Viterbi
- Modèle appris sur des chats, Germinal, correspondance de Bizet.

Nouveau taux d’erreur : **9,00%**

Table Traduction

- A partir d'un texte et de sa version corrigé par un humain
- Apprendre automatiquement certaines corrections (Algorithme EM)
- Indispensable pour les abréviations

Nouveau taux d'erreur : **7,9%**

Intérêt du modèle 3-grams

Taux d'erreur en augmentant l'effet du modèle 3-grams :

- Valeur *2 : **8,1%**
- Valeur ^2 : **8,3%**

J'ai bas compris ⟶ J'ai bien compris

Taux d'erreur en retirant le modèle 3-grams : **7,6%**

Doit-on abandonner le modèle 3-grams?

Correction de mots corrects ?

On souhaiterait pouvoir obtenir :

Je suit satisfait ⟶ Je suis satisfait

On a donc décider de corriger les mots ‘corrects’ à l’aide des 3-grams.

Cela rend l’exécution très lente (explosion combinatoire)

Nouveau taux d’erreur : **8,00%**

Cela introduit trop de problème :

Mon écran est neuf ⟶ Mon écran est noir

Fusion des mots coupés

On souhaiterait pouvoir obtenir :

Je n ai pas de passe tem ps ⟶ Je n'ai pas de passe-temps

On tente de fusionner toutes les paires de mots adjacents :

- Sans rien qui les sépare : tem ps -> temps
- Avec ` ' ` : n ai -> n'ai
- Avec ` - ` : passe temps -> passe-temps

On utilise le modèle 3-grams pour déterminer si la fusion est pertinente.

Nouveau taux d'erreur : **7,00%**

Viterbi sans le modèle 3-grams

• On a déterminé que le modèle 3-grams empirait la correction lorsqu'on l'utilisait dans Viterbi.

• Cependant il est très utile pour la partie « Fusion des mots coupés »

• Donc on a décidé de n'utilisé le modèle 3-grams seulement pour la fusion des mots coupés

Nouveau taux d'erreur : **6,6%**

Conclusion

- Ingrédient d'un bon correcteur :
  - Levenshtein
  - Pas de modèles 3-grams pour Viterbi
  - Table de traduction (abréviations)
  - Fusion des mots coupés
- Fautes oubliées par le programme : grammaire, oubli d'espaces, phonétique.
- Pistes pour de possibles améliorations :
  - Séparer les mots faux pour en créer deux corrects : jesuis -> je suis
  - Trouver des correspondances phonétiques aux mots faux : cé -> c'est
  - Implémenter un correcteur grammatical qui corrigerait même les mots corrects : tu est -> tu es