

Données Séquentielles Symboliques: Translittération automatique

Alexandre Bérard, Mathias Millet, Charles Robin

January 10th, 2014

Sommaire

- 1 Introduction
 - Translittération
 - Le projet
 - Méthodologie
- 2 Translittération automatique
 - Translittération par règles de substitution
 - Traduction statistique
 - CRF
- 3 Conclusion

Introduction : Translittération

Definition

Conversion d'un mot d'un système d'écriture à un autre, en substituant des *graphèmes* ou des *phonèmes*.

Exemple : anglais - hindi

Type	Word	Acceptable transliterations
English word	Azure	अज्योर, अज्यॉर, अज़्योर, अज़्यॉर, एज्योर, एज्यॉर, एज़्योर, एज़्यॉर, अज्यौर, अज्यौर, अज्यौर, अज्यौर
Foreign name	Norfolk	नोरफोल्क, नोरफ़ोल्क, नोरफॉल्क, नोरफ़ॉल्क,

Pourquoi ?

- Traduction automatique de termes techniques, noms propres, ou requêtes Web.
- Plus besoin de maintenir un dictionnaire !

Le projet

L'objectif

Implémenter des techniques de translittération automatique

- De l'espagnol vers le portugais (*SPA-POR*)
- De l'anglais vers le russe (*ENG-RUS*)

Les données

Pour chaque paire de langages :

- Un corpus d'apprentissage : 3057 entrées pour SPA-POR, et 7262 entrées pour ENG-RUS.
- Un corpus de test : 1000 entrées.

Le projet : exemples

ENG-RUS

```
#vitellin# #вителлин# OR #вителлины# other unknown  
#telephone# #телефон# other unknown
```

SPA-POR

```
#teléfono# #telefone# other unknown  
#thermosphaera# #thermosphaera# other unknown
```

Méthodologie

Échantillonnage ?

Les données fournies sont déjà séparées entre données d'apprentissage et données de test

⇒ L'échantillonnage n'est pas nécessaire

Métriques

Deux métriques :

- Précision
- Moyenne de la distance d'édition (Levenshtein)

⇒ Ces métriques seront utilisées tout au long du projet.

1 Introduction

2 Translittération automatique

- Translittération par règles de substitution
- Traduction statistique
- CRF

3 Conclusion

Espagnol - Portugais

Premières considérations

- Les deux langues sont très proches (51% de précision, distance de 1.06 sans rien faire)
- En appliquant les trois règles suivantes :

is#->e#
ción#->ção#
ido#->ídeo#

Nous obtenons 58.6% de précision (distance de 0.76)

⇒ De bons résultats peuvent être obtenus en déterminant des règles automatiquement

Génération des règles

Partie gauche :

- Aligner les mots :
[('catars', 'catars'), ('is', 'e'), ('', '')]
- Enumérer la liste des candidats de gauche, de longueur $\geq l$:
['is', 'sis', 'is#', 'sis#', ...]
- Garder les candidats ayant un support suffisant :
`len([w for w in words if 'is' in w]) >= s`

Partie droite :

Prendre la partie droite de règle offrant le meilleur *taux de confiance* ($\geq c$).

Résultats

Support s	Longueur l	Confiance c	Précision	Distance	Règles
2	6	0.80	56.8%	0.82	104
2	5	0.80	69.0%	0.58	315
2	4	0.80	68.1%	0.59	584
3	5	0.80	67.3%	0.61	195
5	5	0.80	67.2%	0.61	128
1	5	0.80	64.5%	0.68	1490
2	5	0.75	68.5%	0.59	340
2	5	0.85	68.4%	0.59	287

1 Introduction

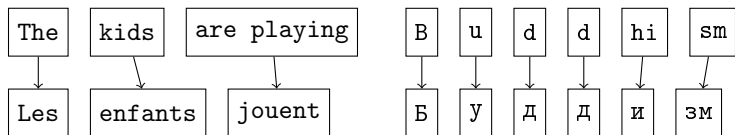
2 Translittération automatique

- Translittération par règles de substitution
- Traduction statistique
- CRF

3 Conclusion

Traduction statistique

Figure : Traduction et translittération



Loi de Bayes

$$e = \arg \max_{e \in e^*} p(e|f) = \arg \max_{e \in e^*} p(f|e)p(e)$$

- e^* est l'ensemble des traductions possibles
- $p(e)$ est le modèle de langue
- $p(f|e)$ est le modèle de traduction

Framework Moses

Framework de traduction statistique “syntagmatique”. Traduction de groupes de mots (par ex. “are playing”/“jouent”). Utilisation de *Giza++* pour l’alignement de mots, d’*IRSTLM* pour le modèle de langue.

Pré-traitement des données

Insertion d’espaces entre les lettres (c a t a r s i s)

Paramètres

- Poids des modèles (langage, traduction, pénalité de longueur, et réordonnancement)
- Longueur des syntagmes
- Ordre des n-grammes

Résultats

Table : Résultats de la baseline (paramètres par défaut)

Données	Précision	Distance moyenne	Variance
ENG-RUS	59.1%	2.10	15.41
SPA-POR	56.9%	1.21	4.30

Table : Résultats avec paramètres optimaux

Données	Précision	Distance moyenne	Variance
ENG-RUS	67.8%	1.75	14.47
SPA-POR	71.5%	0.73	3.46

1 Introduction

2 Translittération automatique

- Translittération par règles de substitution
- Traduction statistique
- CRF

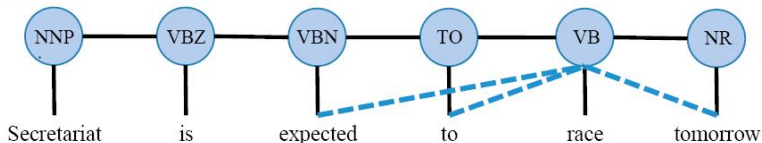
3 Conclusion

Présentation

Équation principale des CRF

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_0} \exp\left(\sum_{t=1}^T \sum_k \lambda_k f_k(\dots)\right) \quad (1)$$

CRF



Champs conditionnels aléatoires

Méthode

- Alignement grâce à *dpalign*
- Utilisation de *CRF++* :

```
# Unigram
U0:%x[0,0]
U1:%x[-1,0]
U2:%x[1,0]
U3:%x[-2,0]
U4:%x[2,0]
U5:%x[-3,0]
U6:%x[3,0]
# Bigram
B
```

Résultats

Table : Résultats pour le jeu de données *Espagnol-Portugais*

Règles	Précision	Distance d'édition
Baseline	51.0%	1.06
3 subst.	58.6%	0.76
CRF	63.9%	0.69

Table : Résultats pour le jeu de données *Anglais-Russe*

Précision	Distance d'édition
49.8%	1.61

- 1 Introduction
- 2 Translittération automatique
- 3 Conclusion

Bibliography



Philipp Koehn.

Slides of the statistical machine translation book.

<http://www.statmt.org/book/>, 2010.



Philipp Koehn.

Moses, Statistical Machine Translation System, User Manual and Code Guide.

University of Edinburgh, 2014.



John Lafferty, Andrew McCallum, and Fernando Pereira.

Conditional random fields : Probabilistic models for segmenting and labeling sequence data.

In Proceedings of the International Conference on Machine Learning, ICML '01, 2001.



David Matthews.

Machine transliteration of proper names.

Master's thesis, University of Edinburgh, 2007.

Conclusion

Améliorations et perspectives

- Possibilité d'optimiser le réglage des paramètres
- Bruit dans les données
- Approches futures : Étudier la translittération inverse

<https://code.google.com/p/transliteration>