Données Séquentielles Symboliques: Translittération automatique

Alexandre Bérard, Mathias Millet, Charles Robin

January 10th, 2014

Sommaire

Introduction : Translittération

Definition

Conversion de texte d'un système d'écriture à un autre, en substitutant des *graphèmes* ou des *phonèmes*.

Exemple : anglais - hindi

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

Pourquoi ?

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

Introduction : 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.

Méthodologie

Échantillonage ?

Les données fournies sont déjà séparées entre données d'apprentissage et données de test
$\Longrightarrow$ L'échantillonage n'est pas nécessaire

Métriques

Deux métriques :

- Précision
- Distance de Levenshtein

Espagnol - Portugais

Premières considérations

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

```
is#->e#
ción#->ção#
ido#->ídeo#
```

nous obtenons 57% de précision

$\Longrightarrow$ De bons résultat peuvent être obtenus en déterminant des règles automatiquement

Apprentissage

Objectif

Trouver un compromis entre :

- Fiabilité : la règle ne conduit pas à de fausses translittérations
- Généralisation : la règle peut s'appliquer à beaucoup de mots

$\Longrightarrow$ Nous allons essayer de générer de règles en prenant ces deux paramètres en compte

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
```

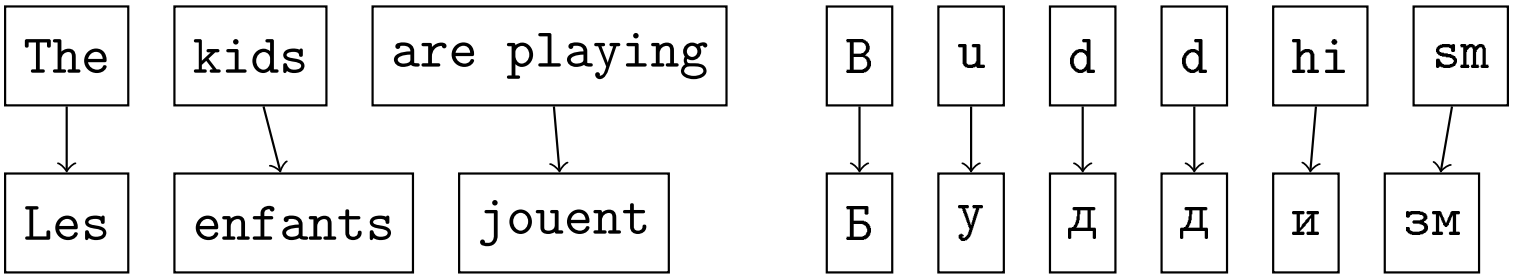
Partie droite :

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

Traduction statistique



Présentation

Équation principale des CRF

$$p(\mathbf{s}|\mathbf{o}) = \frac{1}{Z_0} exp(\sum_{t=1}^{T} \sum_{k} \lambda_k f_k(...)) \quad (1)$$

CRF

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

Métriques

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**

Résultats

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

Précision	Distance d'édition
49.8%	1.61

Bibliography


Kevin Knight and Jonathan Graehl.
Machine transliteration.
Computational Linguistics, 24(4) :599–612, 1998.


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.


Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, and Christine Moran.
Moses : Open source toolkit for statistical machine translation.
In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, 2007.

Conclusion

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