Données Séquentielles et Symboliques : Translittération automatique

Alexandre Bérard, Mathias Millet, Charles Robin

11 janvier 2014

Résumé

1 Introduction

1.1 Jeux de données

Les données nous ont été fournies par Vincent Claveau. Nous disposons de deux jeux de données, concernant respectivement la translittération de l'*espagnol* au *portugais* et de l'*anglais* vers le *russe*. Chacun de ces jeux de données est divisé en deux fichiers, un fichier pour l'apprentissage et un fichier pour l'évaluation. Le fichier d'apprentissage contient plusieurs milliers d'entrées (3057 pour le portugais, et 7262 pour le russe), chaque entrée correspondant à un mot dans la langue originale et sa transcription dans le langage cible. Les entrées dans le fichier d'évaluation, elles, peuvent cependant contenir plusieurs transcriptions pour un même mot.

FIGURE 1 – Entrées dans le fichier d'apprentissage Espagnol-Portugais

```
2996: #fotocopiado# #fotocópia#
2997: #hexanoles# #hexanóis#
2998: #catalasa# #catalase#
```

FIGURE 2 – Entrées dans le fichier d'évaluation Espagnol-Portugais

```
182: #centrifugación# #centrifugação# other unknown
183: #centriolo# #centríolo# OR #centríolos# other unknown
184: #ceramida# #ceramida# OR #ceramidas# other unknown
```

2 Règles de substitution

Nous avons observé que dans le cas de l'espagnol et le portugais, les mots sont très similaires dans les deux langages. Dans les données d'apprentissage, la distance d'édition moyenne entre un mot espagnol et sa transcription en portugais est de *2.0*.

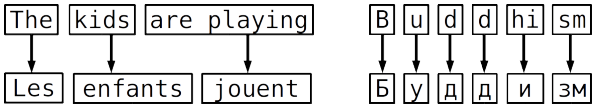
Nous avons ensuite constaté qu'un système mettant en jeu les trois règles de substitution suivantes, obtenait une précision de *58.6%*, et une distance d'édition moyenne de *0.76*.

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

3 Translittération statistique

Le problème de la translittération est relativement proche de la traduction automatique. En traduction automatique, on traduit des séquences de mots d'un langage à un autre. Cela se caractérise par l'association à chaque mot (ou groupe de mots) de la phrase original, d'un mot (ou groupe de mots) dans le langage cible, avec un réordonnancement possible.

FIGURE 3 – Alignement de mots en traduction, et alignment de lettres en translittération



4 Conclusion et perspectives

Références

[1] Yaser Al-Onaizan and Kevin Knight. Machine transliteration of names in arabic text. In *Proceedings of the ACL-02 Workshop on Computational Approaches to Semitic Languages*, SEMITIC '02, pages 1–13. Association for Computational Linguistics, 2002.

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

[3] Philipp Koehn. *Moses, Statistical Machine Translation System, User Manual and Code Guide*. University of Edinburgh, 2014.

[4] Jong-Hoon Oh, Key-Sun Choi, and Hitoshi Isahara. A comparison of different machine transliteration models. *J. Artif. Int. Res.*, 27(1) :119–151, 2006.

[5] Anil Kumar Singh, Sethuramalingam Subramaniam, and Taraka Rama. Transliteration as alignment vs. transliteration as generation for crosslingual information retrieval. *TAL*, 51(2) :95–117, 2010.