

Deepparse et l'utilisation des sous-mots pour l'analyse syntaxique d'adresses multinationales

David Beauchemin, M. Sc.

david.beauchemin@ift.ulaval.ca

30 mai 2023





DAVID BEAUCHEMIN

Directeur général et candidat
au doctorat en informatique

- B. Sc. en actuariat
- M. Sc. en informatique
- Candidat au doctorat en informatique
- Membre fondateur de la coopérative de solidarité Baseline en intelligence artificielle ↗*
- Fondateur et créateur d'OpenLayer ↗*, un podcast bilingue sur l'IA

Objectifs de la conférence

- Comprendre Deepparse.
- Connaitre les cas d'application de Deepparse.
- Connaitre comment Deepparse fonctionne (technique).

Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

4 Plongements de sous-mots

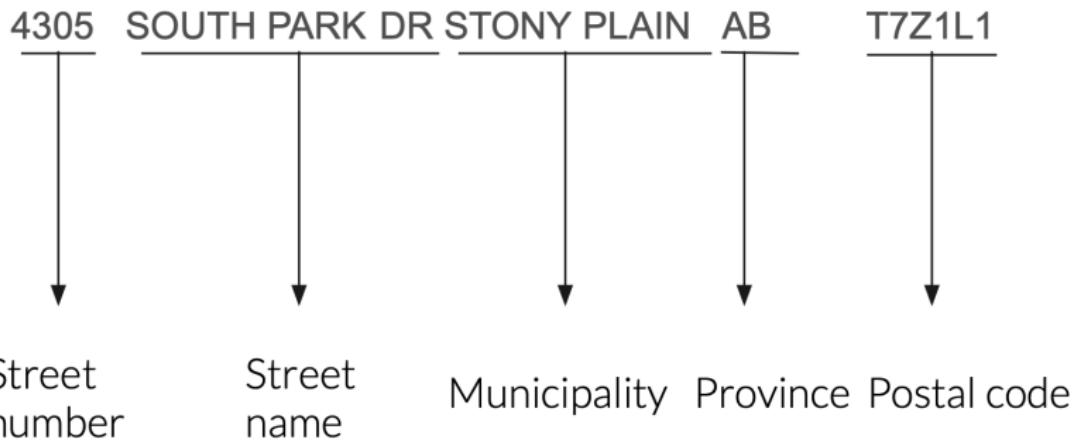
5 Architecture

6 Entraînement

7 Conclusion

Introduction

Qu'est-ce que l'analyse d'adresses ?

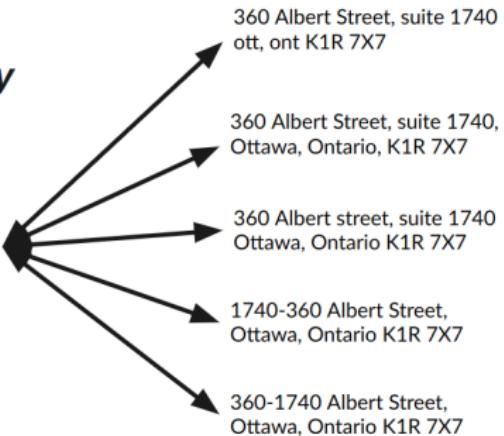


Utile pour des tâches telles que *Record Linkage* et *Geocoding*.

Introduction

Données = Messy

360 Albert Street, suite 1740
Ottawa, Ontario K1R 7X7



Agenda

1 Introduction

2 Depparse

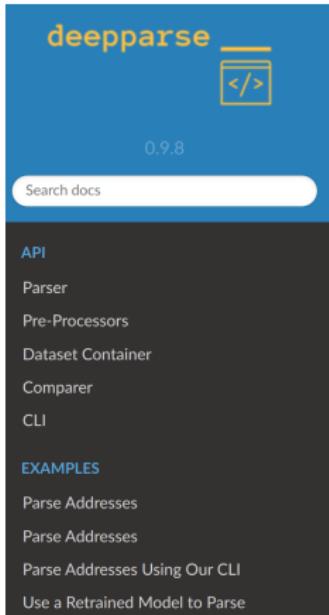
3 Cas d'utilisation

4 Plongements de sous-mots

5 Architecture

6 Entraînement

7 Conclusion



The screenshot shows the Deepparse documentation page. At the top left is the logo "deepparse" with a small icon of a document with code tags. Below it is the version "0.9.8". A search bar says "Search docs". On the left sidebar under "API", there are links for "Parser", "Pre-Processors", "Dataset Container", "Comparer", and "CLI". Under "EXAMPLES", there are links for "Parse Addresses", "Parse Addresses Using Our CLI", and "Use a Retrained Model to Parse".

Home / Here is Deepparse

[Edit on GitHub](#)

[DOWNLOAD DATASET](#)

Here is Deepparse

Deepparse is a state-of-the-art library for parsing multinational street addresses using deep learning.

Use deepparse to

- parse multinational address using one of our pretrained models with or without attention mechanism,
- parse addresses directly from the command line without code to write,
- retrain our pretrained models on new data to improve parsing on specific country address patterns,
- retrain our pretrained models with new prediction tags easily,
- retrain our pretrained models with or without freezing some layers,
- train a new Seq2Seq addresses parsing models easily using a new model configuration.

Deepparse is compatible with the **latest version of PyTorch and Python >= 3.8**.

Figure – Documentation



Particularités

```
● ● ●  
1 address_parser(model_type="fasttext", attention_mechanism=True)  
2  
3 address_parser("350 rue des Lilas Ouest Québec Québec G1L 1B6")
```

```
● ● ●  
1 retrain fasttext ./train_dataset_path.csv  
2 test fasttext ./test_dataset_path.csv
```

Particularités



```
1 address_parser([
2     "305 rue des Lilas, QC, G1L 1B6",
3     "305 Lilac Street, Qc, G1L 1B6",
4     "610065 中国 四川省 成都市 一环路南一段 24号"
5 ])
```

Particularités



```
1 address_parser.retrain(  
2     training_container,  
3     train_ratio=0.8,  
4     epochs=5,  
5     batch_size=8,  
6     num_workers=2,  
7 )
```

Particularités

```
● ● ●  
1 address_parser.retrain(  
2     training_container,  
3     train_ratio=0.8,  
4     epochs=5,  
5     batch_size=8,  
6     num_workers=2,  
7     prediction_tags=tag_dictionary,  
8 )
```

Particularités

```
● ● ●  
1 address_parser.retrain(  
2     training_container,  
3     train_ratio=0.8,  
4     epochs=5,  
5     batch_size=8,  
6     num_workers=2,  
7     seq2seq_params=seq2seq_params,  
8 )
```

Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

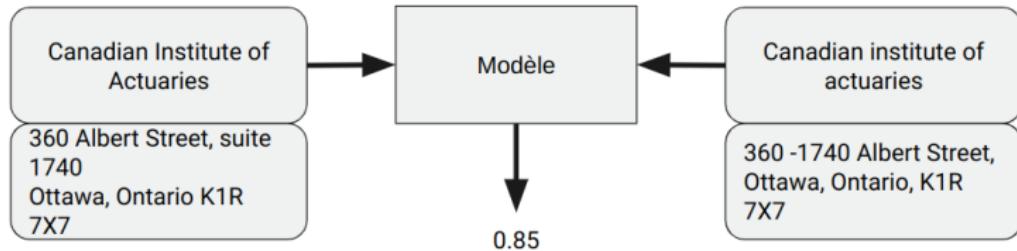
4 Plongements de sous-mots

5 Architecture

6 Entraînement

7 Conclusion

Similarité d'entité

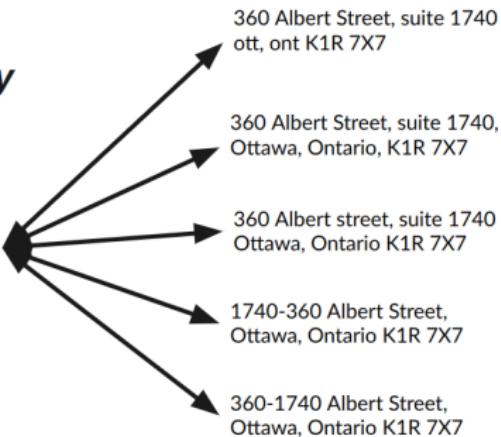


Détection de doublons parmi des informations non structurées provenant de sources de données différentes ↗*

Similarité d'entité

Données = Messy

360 Albert Street, suite 1740
Ottawa, Ontario K1R 7X7



Similarité d'entité

360 Albert Street, suite 1740 | 360 Albert Street, Bureau 1740

Adresse #1

Adresse #2

Numéro civique	360	Numéro civique	360
Unité	suite 1740	Unité	bureau 1740
Nom de la rue	Albert Street	Nom de la rue	Albert Street
Orientation	∅	Orientation	∅
Code postal	∅	Code postal	∅

Similarité d'entité

Similarité de Jaccard

$$\text{Jaccard}(A, B) = \begin{cases} 0 & \text{si } |A \cap B| = 0 \text{ ou } |A \cup B| = 0 \\ \frac{|A \cap B|}{|A \cup B|} & \text{autrement} \end{cases}$$

Exemple

Jaccard(Laval University, Montreal University)



$$\frac{|\{\text{University}\}|}{|\{\text{University, Montreal, Laval}\}|} = \frac{1}{3} = \mathbf{0,333}$$

Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

4 Plongements de sous-mots

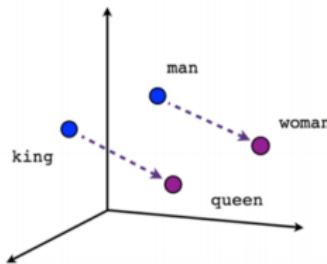
5 Architecture

6 Entraînement

7 Conclusion

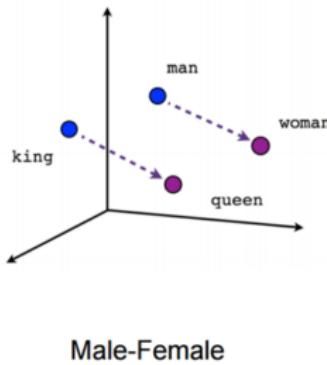
Plongements de sous-mots

Plongement de mot : représentation vectorielle d'un mot



Plongements de sous-mots

Plongement de mot : représentation vectorielle d'un mot



Subword embedding : représentation d'une plus petite unité

- Sous-chaîne de n -grams (e.g : Le bi-gram de "H1A 1B1" est {H1, 1A, A1, 1B, B1})

Pourquoi ?

Support multilingue (MultiBPEmb) et support pour mot hors vocabulaire (FastText).

Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

4 Plongements de sous-mots

5 Architecture

6 Entraînement

7 Conclusion

Modèles de plongements de mots

- FastText
- MultiBPEmb

Modèle de segmentation

Nous utilisons un modèle Seq2Seq composé d'

Modèle de segmentation

Nous utilisons un modèle Seq2Seq composé d'

- un encodeur LSTM unidirectionnel à une couche,

Modèle de segmentation

Nous utilisons un modèle Seq2Seq composé d'

- un encodeur LSTM unidirectionnel à une couche,
- un décodeur LSTM unidirectionnel à une couche,

Modèle de segmentation

Nous utilisons un modèle Seq2Seq composé d'

- un encodeur LSTM unidirectionnel à une couche,
- un décodeur LSTM unidirectionnel à une couche,
- une couche linéaire entièrement connectée pour mapper la représentation dans la dimensionnalité de l'espace des étiquettes, et

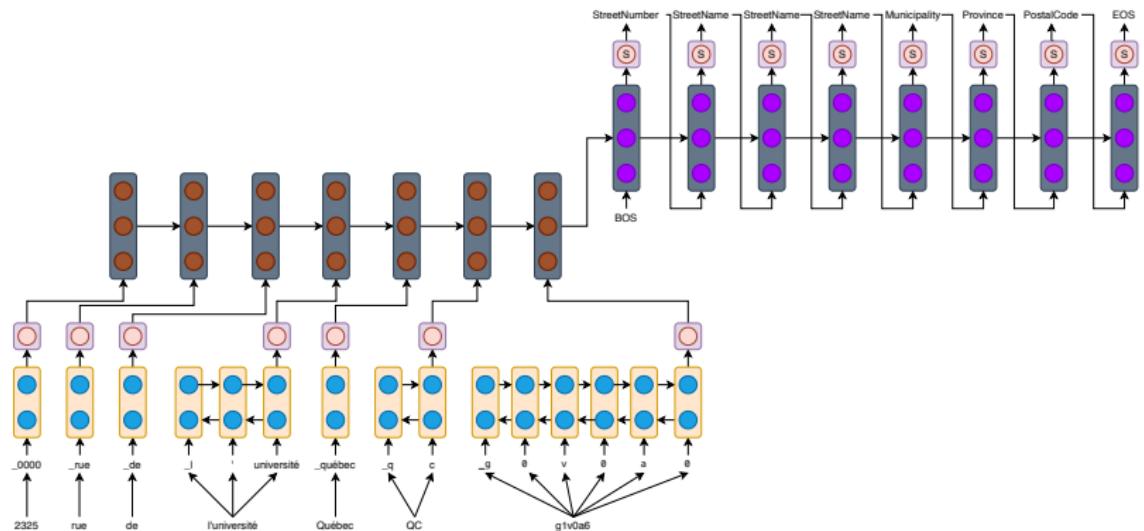
Modèle de segmentation

Nous utilisons un modèle Seq2Seq composé d'

- un encodeur LSTM unidirectionnel à une couche,
- un décodeur LSTM unidirectionnel à une couche,
- une couche linéaire entièrement connectée pour mapper la représentation dans la dimensionnalité de l'espace des étiquettes, et
- une fonction d'activation.

Les états cachés du codeur et du décodeur sont tous deux de dimension 1024.

Architecture



Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

4 Plongements de sous-mots

5 Architecture

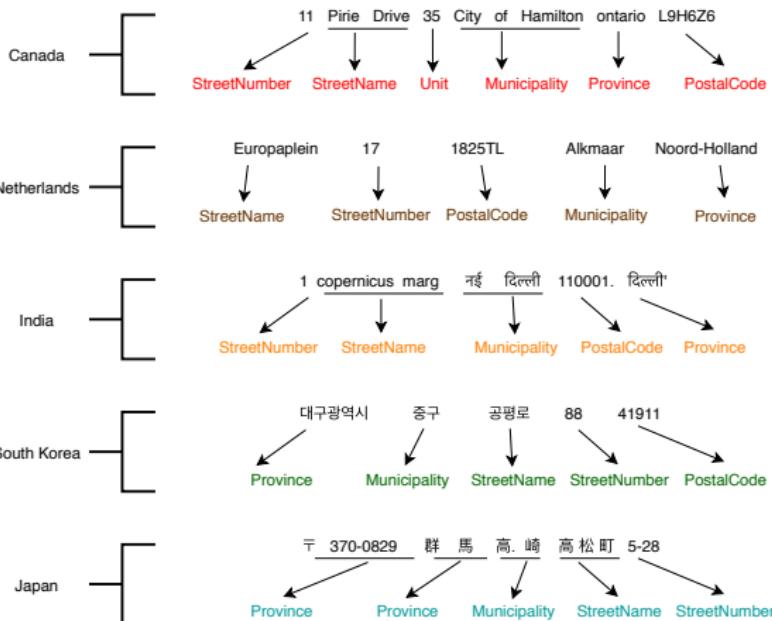
6 Entraînement

7 Conclusion

À propos de l'entraînement

- Jeu de données public en ligne qui contient des millions d'adresses.
- 61 pays sélectionner (avec un minimum de 100 adresses).
- 8 étiquettes : StreetNumber, StreetName, Unit, Municipality, Province, PostalCode, Orientation, et GeneralDelivery.

Exemples d'adresses d'entraînement



Performance

Pays	FastText	BPEmb	Pays	FastText	BPEmb
United States	99.61 ± 0.09	98.55 ± 2.19	Poland	99.69 ± 0.07	99.19 ± 1.39
Brazil	99.40 ± 0.10	98.54 ± 1.68	Norway	99.46 ± 0.06	97.98 ± 1.31
South Korea	99.96 ± 0.01	99.99 ± 0.02	Austria	99.28 ± 0.03	98.28 ± 1.56
Australia	99.68 ± 0.05	99.21 ± 1.17	Finland	99.77 ± 0.03	99.72 ± 0.30
Mexico	99.60 ± 0.06	98.55 ± 2.22	Denmark	99.71 ± 0.07	99.20 ± 1.38
Germany	99.77 ± 0.04	99.23 ± 1.30	Czechia	99.57 ± 0.09	98.77 ± 2.22
Spain	99.75 ± 0.05	98.65 ± 2.36	Italy	99.73 ± 0.05	98.91 ± 1.76
Netherlands	99.61 ± 0.07	99.26 ± 1.23	France	99.66 ± 0.08	98.65 ± 2.00
Canada	99.79 ± 0.05	99.19 ± 1.33	United Kingdom	99.61 ± 0.10	98.66 ± 2.11
Switzerland	99.53 ± 0.09	99.49 ± 0.53	Russia	99.03 ± 0.24	97.52 ± 4.23

- Évaluation en zero-shot.
- Modèle avec attention.
- Modèle avec entraînement par adaptation de domaine (ADAN).
- Tests de significativité.

Agenda

1 Introduction

2 Deeparse

3 Cas d'utilisation

4 Plongements de sous-mots

5 Architecture

6 Entraînement

7 Conclusion

Conclusion

Deepparse permet

- de séparer les composantes d'une adresse en plusieurs langues, notamment en français et en anglais,
- d'ajuster nos modèles sur son jeu de données, et
- de changer les étiquettes.

Pour en savoir plus, lire nos deux articles

<https://arxiv.org/abs/2006.16152> et

<https://arxiv.org/abs/2112.04008>

Questions

